UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

Teoria del Senyal i de Comunicacions

On the Majorization-Minimization framework and g-convex optimization: Exploiting diversity using sparse-aware and information theoretic criteria

DOCTORAL THESIS

Submitted for the degree of Doctor of Philosophy by:

Carlos Alejandro Lopez Molina

Under the supervision of:
Dr. Jaume Riba Sagarra

2025

A mis padres. Querían que fuese doctor y me convertí en uno de verdad.

Acknowledgements

Qui em coneix sap que em costa expressar segons quines coses. És per aquest motiu, i abans de començar a parlar sobre entropies, problemes d'optimització i la varietat de Grassmann, que aprofito aquestes línies per agrair a les persones que m'han acompanyat durant aquests anys de doctorat.

Primerament, vull agrair a en Jaume per tots aquests anys. Portem des del TFG treballant i, sent honest, part del procés de maduració que he fet durant tot aquest temps és gràcies a ell. Estic molt orgullós de tota la feina que hem fet. Sempre he gaudit les nostres converses. A més, agraeixo a tots els professors que formen part del SPCOM amb qui he coincidit: Xavi, Gregori, Xell, Josep i Francesc. Tenint en compte que soc una persona amb molta introversió (el COVID no ha ajudat gens), m'he sentit molt acollit.

En segon lloc, vull recordar-me dels meus companys de doctorat que han estat més propers a mi. Per una banda, tinc molt presents a en Ferran i a en Sergi. M'han ajudat moltíssim quan els he atabalat de preguntes per la docència i per tràmits del doctorat. Això també va per en Marc i l'Aniol. Encara que la meva presència al D5 era molt fugaç, guardo bon record de les converses i el viatge que vam compartir a Madrid. Els desitjo un bon final de doctorat.

Per acabar, vull mencionar també a aquelles persones més properes per fer-me companyia de manera incondicional. Per descomptat, agraeixo als meus pares i a l'Alejandra, a qui espero acompanyar per un camí similar en un futur. Ara que ja estem aquest punt, les bromes que sempre he fet s'han convertit en una realitat. Pel que fa als meus amics que m'han seguit de prop durant la tesi (Lluís, Júlia, Tere, Josep, María, Cristian, Dan, Pando...), gràcies per haver estat allà.

This dissertation has been supported by:

- Fellowship FI 2021 by the Secretary for University and Research of the Generalitat de Catalunya and the European Social Fund.
- Spanish Ministry of Science and Innovation through project RODIN (PID2019-105717RB-C22/AEI/10.13039/501100011033) and project MAYTE (PID2022-136512OB-C21 financed by MICIU/AEI/10.13039/501100011033 and by ERDF/EU).
- Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, under the grant 2021 SGR 01033.

Abstract

Diversity is a well-established concept in wireless communications whose purpose is to quantify the potential robustness of a receiver when multiple independent copies of the informative signal are received. Indeed, there exists a formal definition of this concept within the context of wireless communications that takes into account its practical usage, i.e. it is defined with respect to the symbol error probability averaged over the channel statistical fluctuation. However, there is no consensus on the generalization of the previous definition to other forms of signal processing applications. For this reason and being inspired by an intuitive definition of diversity extracted from the multimodal data fusion framework, the purpose of this dissertation is to explore the concept of diversity through the lenses of Information theory, a numerical optimization framework based on the Majorization-Minimization principle and the Grassmann manifold. The motivation behind the Majorization-Minimization algorithms is that they fit perfectly to the optimization problems arising from information theoretic cost functions, while the Grassmann manifold emerges naturally in the context of sparse-aware signal processing problems that exhibit some sort of diversity. All these ideas are surveyed through three different scenarios: the multisensor fusion, the Covariance Conversion from wireless MIMO communications and the detection of correlation. All of the scenarios share the fact that the intrinsic dimension of the data is much smaller than the ambient space dimension.

In the multisensor fusion problem, we analyze the intuitive definition of diversity in a straightforward manner for three fusion policies. Firstly, the Covariance Intersection principle is reviewed to highlight its connection to the minimum error entropy criterion and the waterfilling algorithm for optimal power allocation in communications. Secondly, we derive a bounded descriptor based on the Rényi entropy of a sensor network contamination worst-case scenario (unbounded variance). Thanks to the aforementioned descriptor, it is possible to provide an operational interpretation to the commonly used L0 norm regularization particularized for this problem. Finally, we consider a fusion scheme that incorporates a subspace-based regression technique into the fusion operation. This proposal, which is inspired by a duality with the problem of unstructured interference mitigation in navigation receivers, is motivated by the fact that it is possible to obtain a measure of the fusion integrity when the temporal redundancy of the measurements and the intersensor covariance matrix are estimated in a joint manner.

Besides, a different kind of diversity is unveiled in the Covariance Conversion problem for Frequency Division Duplexing schemes from wireless communications. In essence, this problem consists in the estimation of the Downlink channel covariance matrix using a prior estimation of the Uplink channel covariance matrix. Particularly, we are interested in those cases where sparsity can be defined on the second-order statistics, which are found in the mmWave and ultra-wide band channels. Through a detailed analysis of this problem, we show a promising conversion algorithm founded on the Alternating Direction Method of Multipliers.

Lastly, the detection of correlation between two Gaussian vectors problem serves as a way to explore an information theoretic approach for the quantification of diversity. In fact, we transform this setting into a Mutual Information estimation problem of M parallel Gaussian channels to yield the aforementioned information theoretic measure. However, the Maximum Likelihood estimation of the Mutual Information suffers from bias when a subset of these channels provide no information. In light of this, we propose the adoption of model-order selection rules, well-known in other areas, as a means for estimating information under a bias-variance trade-off.

Resumen

La diversidad es un concepto bien conocido en comunicaciones inalámbricas cuyo propósito es cuantificar la robustez potencial de un receptor cuando se reciben múltiples copias independientes de una señal. De hecho, existe una definición formal de este concepto que tiene en cuenta su uso práctico dentro del contexto de comunicaciones inalámbricas, es decir, se define con respecto a la probabilidad de error de símbolo promediada sobre la estadística del canal. Sin embargo, no existe consenso sobre la generalización de la definición para otro tipo de aplicaciones. Inspirándonos en una definición intuitiva de diversidad extraída de fusión de datos multimodal, el objetivo de esta tesis es explorar el concepto de diversidad bajo la perspectiva de la teoría de la información, la optimización numérica basada en el principio de MM y la variedad de Grassmann. La motivación detrás de los algoritmos de MM es que se ajustan perfectamente a los problemas de optimización que surgen de la teoría de la información, mientras que la variedad de Grassmann aparece naturalmente en el contexto de problemas con esparsidad que muestran diversidad. Todas estas ideas se analizan a través de tres escenarios diferentes: la fusión multisensor, la conversión de covarianza en comunicaciones inalámbricas MIMO y la detección de correlación. Estos escenarios comparten el hecho de que la dimensión intrínseca de los datos es mucho más pequeña que la dimensión del espacio global.

En el problema de fusión multisensor, analizamos la definición intuitiva de diversidad para tres esquemas de fusión. En primer lugar, revisamos el principio de Intersección de Covarianza para resaltar su conexión con el criterio de mínima entropía y el algoritmo de "waterfilling" en comunicaciones. En segundo lugar, derivamos un descriptor basado en la entropía de Rényi del peor caso de contaminación de una red de sensores. Gracias a este descriptor, es posible interpretar la regularización de norma L0 para el problema de fusión de sensores con argumentos extraídos de la teoría de la información. Finalmente, consideramos un esquema de fusión en la que se procesa conjuntamente la propia fusión de los sensores y una técnica de regresión basada en subespacios lineales. Este planteamiento, que se inspira en una dualidad con el problema de la mitigación de interferencias no estructuradas en receptores de navegación, está motivada por el hecho de que es posible obtener una medida de la integridad de la fusión cuando la redundancia temporal de las mediciones y la matriz de covarianza intersensores se estiman conjuntamente.

Por otra parte, se revela un tipo diferente de diversidad en el problema de conversión de covarianza para esquemas de duplexación por división de frecuencia en comunicaciones. Este problema consiste en la estimación de la matriz de covarianza de canal del enlace de bajada utilizando una estimación previa de la matriz de covarianza de canal del enlace de subida. Consideramos especialmente aquellos casos donde se puede definir esparsidad sobre las estadísticas de segundo orden, que ocurre en los canales "mmWave" y de banda ultra-ancha. A través de un análisis detallado de este problema, mostramos un algoritmo de conversión prometedor basado en el ADMM.

Por último, el problema de detección de correlación entre dos vectores Gaussianos sirve como una forma de explorar un enfoque basado en teoría de la información para cuantificar la diversidad. De hecho, transformamos este problema en uno de estimación de la información mutua de M canales Gaussianos paralelos para conseguir dicha medida de la diversidad. Sin embargo, la estimación de máxima verosimilitud de la información mutua sufre de sesgo cuando un subconjunto de estos canales no proporcionan información. Por este motivo, proponemos la adopción de reglas de selección de orden de modelo, bien conocidas en otras áreas, como un medio para estimar información bajo un compromiso de sesgo-varianza.

Resum

La diversitat és un concepte ben conegut en les comunicacions, el qual quantifica la robustesa d'un receptor quan es reben múltiples còpies independents d'un senyal informatiu. De fet, existeix una definició formal d'aquest concepte en el context de comunicacions sense fil que té en compte el seu ús pràctic, és a dir, es defineix en funció de la probabilitat d'error de símbol promitjada sobre la fluctuació estadística del canal. Tanmateix, no hi ha consens sobre la generalització de la definició anterior en altres aplicacions de processament de senyal. Per aquest motiu i inspirant-nos en una definició intuïtiva de la diversitat extreta de l'entorn de fusió de dades multimodal, l'objectiu d'aquesta tesi és explorar el concepte de diversitat des de la perspectiva de la teoria de la informació, l'optimització numèrica basada en el principi de MM i la varietat de Grassmann. La motivació dels algorismes de MM és que s'ajusten perfectament als problemes d'optimització derivats de les funcions de cost basades en mesures d'informació, mentre que la varietat de Grassmann sorgeix de manera natural en el context de problemes de processament de senyals amb esparsitat que presenten diversitat. Totes aquestes idees s'estudien a través de tres escenaris diferents: la fusió multisensor, la conversió de covariància de comunicacions MIMO sense fil i la detecció de correlació. Aquests escenaris comparteixen el fet que la dimensió intrínseca de les dades és molt més petita que la dimensió de l'espai global.

En el problema de la fusió multisensor, analitzem la definició intuïtiva de diversitat per a tres esquemes de fusió. En primer lloc, revisem el principi d'intersecció de covariància per destacar la seva connexió amb el criteri de mínima entropia i l'algorisme de "waterfilling" en comunicacions. A continuació, derivem un descriptor basat en l'entropia de Rényi del cas pitjor de contaminació en una xarxa de sensors. Gràcies al descriptor esmentat, és possible donar una interpretació a la regularització de la norma L0 amb arguments extrets de la teoria de la informació. Finalment, considerem un esquema de fusió que processa conjuntament la fusió i una tècnica de regressió basada en subespais lineals. La proposta anterior, que s'inspira en una dualitat amb el problema de la mitigació d'interferències no estructurades en receptors de navegació, està motivada pel fet que és possible obtenir una mesura de la integritat de la fusió quan la redundància temporal de les mesures i la matriu de covariància intersensor s'estimen de manera conjunta.

En segon lloc, es presenta un tipus diferent de diversitat en el problema de conversió de covariància per als esquemes de duplexació per divisió de freqüència de comunicacions. Aquest problema consisteix en l'estimació de la matriu de covariància de canal de l'enllaç de baixada mitjançant una estimació prèvia de la matriu de covariància de canal de l'enllaç de pujada. En particular, estem interessats en aquells casos en què l'esparsitat es pot definir sobre les estadístiques de segon ordre, els quals es troben als canals de "mmWave" i de banda ultra-ample. Mitjançant una anàlisi detallada d'aquest problema, mostrem un algorisme de conversió prometedor basat en el ADMM.

En darrer lloc, el problema de detecció de correlació entre dos vectors Gaussians serveix com a forma d'explorar una manera alternativa de quantificar la diversitat basada en teoria de la informació. De fet, transformem aquesta configuració en un problema d'estimació d'informació mútua de M canals Gaussians paral · lels per obtenir la mesura teòrica de la informació esmentada anteriorment. No obstant això, l'estimació de màxima versemblança de la informació mútua pateix de biaix quan un subconjunt d'aquests canals no proporcionen informació. En vista d'això, proposem l'adopció de regles de selecció de l'ordre de model, conegudes en altres àrees, com a una manera d'estimar la informació mútua sota un compromís de biaix-variància.

Table of Contents

List of Figures

List of Tables

Acronyms

ADMM Alternating Direction Method of Multipliers.
AIC Akaike Information Criterion.
BCD Block Coordinate Descent.
BIC Bayesian Information Criterion.
BPD Basis Pursuit Denoising.
CC Covariance Conversion.
CCA Canonical Correlation Analysis.
CCP Concave-Convex Procedure.
CI Covariance Intersection.
CML Conditional Maximum Likelihood.
CRLB Crámer-Rao Lower Bound.
CS Compressed Sensing.
CSD Cosine-Sine Decomposition.
CSI Channel State Information.
E-BLUE Entropic Best Linear Unbiased Estimator.
EM Expectation-Maximization.
FBSA Forward-Backward Splitting Algorithms.
FDD Frequency Division Duplexing.
GIC Generalized Information Criterion.
GMM Gaussian Mixture Model.
GMSE Geometric Mean Squared Error.
ISTA Iterative Soft-Thresholding Algorithm.
ITL Information Theoretic Learning.
IW Inverse Wishart.
KKT Karush-Kuhn-Tucker.
KL Kullback-Leibler.
LASSO Least Absolute Shrinkage and Selection Operator.
LS Least Squares.
MAP Maximum A Posteriori.
MCD Minimum Covariance Determinant.
MEE Minimum Error Entropy.
MI Mutual Information.
ML Maximum Likelihood.
MM Majorization-Minimization.
mmWave Millimeter Wave.
MP Matching Pursuit.
MSE Mean Squared Error.

NFQM Normalized Fusion Quality Measure.
ONB Orthonormal basis.
PABS Principal Angles Between Subspaces.
PCA Principal Component Analysis.
PDF Probability Density Function.
PHD Probability Hypothesis Density.
PMEE Parametric Minimum Error Entropy.
PMF Probability Mass Function.
RIP Restrictive Isometry Property.
SCP Sequential Convex Programming.
SNR Signal-to-Noise Ratio.
SVD Singular Value Decomposition.
UDCC Uplink-Downlink Covariance Conversion.

Nomenclature

General notation

$\exp(\cdot)$	Exponential with base e
$\log(\cdot)$	Natural logarithm
s. t.	subject to

Manifolds

$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
$\mathrm{Gr}(N, D)$	Grassmann manifold. Subspaces of dimension D with ambient space with N dimensions
$\mathrm{St}(N, D)$	Stiefel manifold. Orthogonal matrices of dimension D with ambient space with N dimensions
$\mathrm{O}(N)$	Orthogonal group in dimension N
$\mathrm{GL}(N)$	General linear group in dimension N
$\mathcal{S}_{++}^{M}$	Manifold of $M \times M$ positive definite matrices
$\mathcal{S}_{+}^{M}$	Manifold of $M \times M$ positive semidefinite matrices

Operators

$\|\|\cdot\|\|_p$	ℓ_p norm
$\|\|\cdot\|\|_F$	Frobenius norm
$\mathrm{E}[\cdot]$	Expected value
$\nabla_{\mathbf{x}}$	Gradient with respect to $\mathbf{x}$
$\nabla_{\mathbf{x}}^2$	Hessian with respect to $\mathbf{x}$
$\odot$	Hadamard (element-wise) product

$\frac{\mathrm{d}}{\mathrm{d}x}$ Derivative with respect to $\mathbf{x}$

$\frac{\partial}{\partial x}$ Partial derivative with respect to $\mathbf{x}$

$|\cdot|$ Absolute value

$\lfloor \cdot \rfloor$ Floor function

Vectors and matrices

$\mathbf{x}$ Column vector

$[\mathbf{x}]_n$ n-th component of $\mathbf{x}$

$\mathbf{X}$ Matrix

$\mathbf{X} \in \mathbb{R}^{N \times M}$ Real matrix of dimension $N \times M$

$[\mathbf{X}]_{m,n}$ (m, n)-th entry of $\mathbf{X}$

$\mathbf{X}_\perp$ Vector basis of the orthogonal complement of $\mathbf{X}$

$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ Singular Value Decomposition of $\mathbf{X} \in \mathbb{R}^{N \times M}$, where $\mathbf{\Sigma}$ is a $R \times R$ square matrix and R is the rank of $\mathbf{X}$. The dimensions of $\mathbf{U}$ and $\mathbf{V}$ are fixed accordingly (Compact SVD)

$\mathbf{1}_N$ N-dimensional column vector whose components are all 1

$\mathbf{0}_N$ N-dimensional column vector whose components are all 0

$\mathbf{0}_{N,D}$ $N \times D$ matrix whose components are all 0

$\mathbf{I}_N$ $N \times N$ identity matrix

$\mathrm{tr}(\mathbf{X})$ Trace of the square matrix $\mathbf{X}$

$\det(\mathbf{X})$ Determinant of the square matrix $\mathbf{X}$

$\mathrm{rank}(\mathbf{X})$ Rank of matrix $\mathbf{X}$

$\mathrm{diag}(\mathbf{x})$ Diagonal matrix whose diagonal elements are given by the vector $\mathbf{x}$

$\mathrm{diag}(\mathbf{X})$ Vector composed by the diagonal elements of the square matrix $\mathbf{X}$

$\cos(\mathbf{X})$ Element-wise cosine function on the diagonal elements of $\mathbf{X}$

$\sin(\mathbf{X})$ Element-wise sine function on the diagonal elements of $\mathbf{X}$

Chapter 1

Introduction

1.1 Motivation and goals of this thesis

This dissertation is the result of the intersection of three concepts in signal processing: multimodal data fusion [110], the Grassmann manifold [20] and the Information Theoretic Learning (ITL) paradigm [151]. Although it may seem that, at first glance, these ideas are unrelated, the focus of this thesis is to explore the potential connections between these frameworks with a strong emphasis on numerical optimization theory. Firstly, we reflect on the ideas that ignited our interest in these three topics.

Diversity and degrees of freedom are two key concepts in wireless communications that quantify the potential performance gains of the processing of multiple channels [72], which complement the information provided by the Signal to Noise Ratio (SNR). Unfortunately, there is no formal definition of these quantities (similar to the definitions given in [72]) that can be applied to other signal processing applications up to the authors knowledge. In the pursuit of a more general definition of these concepts, our attention was drawn towards multimodal data fusion for one reason: there is an intuitive definition of diversity in this problem [110]. Basically, two (or more) datasets have some sort of diversity if their joint processing results in the extraction of complementary information that could not be recovered otherwise, which is essentially what happens when multiple signals are processed jointly with the aim of mitigating the effects of the fading channel when multiple independent copies of the informative signal are received. [186]. In this regard, the focus of this thesis is to explore problems and cost functions that exploit the implicit diversity of multiple datasets. Particularly, provided that the multimodal data fusion framework encompasses a wide range of casuistics that are out of the scope of this dissertation, e.g. the processing of heterogeneous data such as images or audio signals, we focus our gaze on the multisensor fusion [103], the Covariance Conversion in wireless MIMO communications [125], [136] and the detection of correlation [120], [159] problems, to explore other alternative expressions of diversity in signal processing applications. While in the first two settings there is a necessity of exploiting diversity, we are interested in the quantification of the diversity on the last application.

Besides, our interest in the Grassmann manifold was triggered by the cycle-slips detection problem from the Precise Point Positioning (PPP) mode of the Global Navigation Satellite System (GNSS) [167], where the identification of the noise subspace is a crucial step to compute the test statistic. Just to briefly introduce the Grassmann manifold, it is intuitively defined as the set that contains the subspaces of a given dimension in which some notion of distances and curves connecting any two points of this set can be defined [20]. In light of this intuitive definition, it is clear that the Grassmann manifold plays a fundamental role in any signal processing application with orthogonality constraints (in addition to a particular homogeneity condition [57]), such as in the previously mentioned detector of cycle-slips that require the determination of the noise subspace [167], Principal Component Analysis (PCA) [96], [206], Subspace Learning [16], [54], [191], Matrix Factorization [7], [45], Non-Coherent [208] and Opportunistic [211] Communications, to name a few examples. Still, it is worth noting that the geometry of the previous problems can also be studied in terms of the Stiefel manifold since any function satisfying the homogeneity condition of the Grassmann manifold is parameterized by either manifold [57]. In

fact, the Grassmann manifold is equivalent to the Stiefel manifold up to an invariance with respect to a rotation, i.e. only the spanned subspace is relevant for the Grassmann manifold whereas the Stiefel manifold is the set of matrices satisfying the orthogonality constraints. Nevertheless, although there are known tools for signal processing that utilize the Stiefel manifold [31], [44], we prefer the Grassmann manifold to study any problem with orthogonality constraints because, as it will be argued, its geometric properties are much more versatile for the applications considered in this dissertation (see [7] for an example). The main challenge that results from the utilization of the Grassmann manifold in a numerical optimization problem is that the orthogonality constraints define a non-convex set [29], implying that already known convex optimization tools cannot be used straightforwardly.

Lastly, in the context of estimation of general parameters and filtering problems, the core of the ITL paradigm is to derive alternative cost functions to the Mean Squared Error (MSE) such that they include more information of the training data than the first and second-order moments [149]. A consequence of the previous rationale is that the resulting criteria are robust, even when the data deviates from the Gaussian assumption. Evidently, Information Theory is the mathematical foundation in which the ITL paradigm is grounded [151]. In this sense, information theoretic measures, such as entropies, divergences, and Mutual Information, are the main subject of study in this framework. While divergences play an important role as measures of discrepancy between distributions [108], [116], we are particularly interested in entropies and Mutual Information. On the one hand, the concept of entropy captured our attention to exploit diversity in signal processing not only for its inherent robustness, but also due to its close link to the sparse-aware signal processing [83], [161]. Indeed, sparsity and entropy can be considered as different expressions of the same underlying phenomenon, as explained in a future chapter. For this reason, in addition to finding practical interpretations of entropic quantities [78], another objective of this dissertation is to investigate the hidden relationship between sparsity and entropy. Similarly to what occurs with the orthogonality constraints, entropic measures tend to define non-convex optimization problems. On the other hand, provided that quantifying the amount of diversity between two datasets can be re-casted as a dependence measurement problem, Mutual Information is an ideal candidate for this task due to its information theoretic nature (robustness, mainly) and to its interpretability. The use of information theoretic quantities in signal processing is a continuation of [50].

As a result of the fact that most of the cost functions encountered throughout this dissertation are non-convex, we had to resort to a robust numerical optimization framework capable of engaging with both convex and non-convex optimization problems. Owing to a link identified between the Direction of Arrival in the presence of independent interferences [172] and multisensor fusion problems, we found that the Majorization-Minimization (MM) framework [179] was suitable for the cost functions and constraint sets considered in this dissertation. In this regard, the general MM framework allows us to avoid descent-like algorithms [117], which we discourage because they are strongly tied to a user-defined parameter (e.g. the gradient step), and to incorporate the Grassmann manifold in a natural manner by means of the geodesically convex optimization paradigm [192]. Considering our growing interest in the algorithmic perspective of signal processing, the three main topics are tainted by numerical optimization theory.

To summarize, the goal of this thesis is to study the concept of diversity in signal processing using information theoretic criteria. Throughout this research, it is shown that the Grassmann manifold appears in a natural manner since sparsity can also be understood in terms of a low-rank linear model.

1.2 Thesis outline and contributions

This thesis is comprised of seven chapters. Excluding the introductory and concluding chapters, two supporting chapters specify *what* kind of cost functions are encountered in this dissertation and *how* we solve their resulting optimization problems. The remaining three chapters are devoted to the study of diversity, exploiting the tools that were introduced in the supporting chapters. Below, we briefly outline each chapter and the research contributions that resulted from them.

- **Chapter 2** details the sparse and other related information theoretic criteria that will be used in subsequent chapters. Particularly, we evaluate three different manners of studying sparsity:

the classical sense of sparsity, low-rank subspace models and model-order selection rules. Within the first idea, we relate the classical definition of sparsity with entropy, and introduce other information theoretic measures. Secondly, this chapter provides an intuitive introduction to the Grassmann manifold and analyzes it from a numerical optimization perspective. Lastly, we review in detail the information theoretic model-order selection rules, as they are shown to be an alternative form of obtaining sparse solutions in Chapter 6, emphasizing on the Bayesian Information Criterion and the Generalized Information Criterion.

Utilizing ideas that emanated from this chapter, the author played a supportive role in the following technical work:

- M. Vilà, C. A. López and J. Riba, "Affine Projection Subspace Tracking," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3705-3709, DOI: 10.1109/ICASSP39728.2021.9415032.

- **Chapter 3** builds the algorithmic framework of this dissertation. This algorithmic framework consists in the intersection of the Majorization-Minimization paradigm with the geodesically convex optimization. Although the ideas are presented in a didactic manner, we concentrate on the key concepts that are relevant for this thesis. Notably, we revisit known results in convex optimization and proximal algorithm theory since they are needed throughout this dissertation. In addition to the previous ideas, this chapter generalizes the MM framework and the Principal Component Analysis problem with geometric ideas extracted from the Grassmann manifold.

  The technical work that resulted from this chapter is:

  - C. A. Lopez and J. Riba, "On the Convergence of Block Majorization-Minimization Algorithms on the Grassmann Manifold," in IEEE Signal Processing Letters, vol. 31, pp. 1314-1318, 2024, DOI: 10.1109/LSP.2024.3396660.

- **Chapter 4** studies thoroughly the multisensor fusion problem. Specifically, we explore three different fusion policies that are tightly related with the Minimum Entropy Criterion. It is in this chapter that we derived the majority of the results of this thesis. Firstly, we found connections between the Covariance Intersection fusion policy and the waterfilling algorithm from wireless communications. Secondly, we derived an information theoretic justification of the ℓ_0 norm regularization for the multisensor fusion problem under a worst-case scenario of contamination. Thirdly, we proved that a parameterized Minimum Error Criterion yields from the Conditional Maximum Likelihood principle in the blind fusion and regression problem, resulting in a cost function that admitted an implementation of a generalized MM algorithm over the Grassmann manifold.

  The technical works that resulted from this chapter are:

  - C. A. Lopez and J. Riba, "Data Driven Joint Sensor Fusion and Regression Based on Geometric Mean Squared Error," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, DOI: 10.1109/ICASSP49357.2023.10095018.
  - C. A. Lopez, F. de Cabrera and J. Riba, "Minimum Error Entropy Estimation Under Contaminated Gaussian Noise," in IEEE Signal Processing Letters, vol. 30, pp. 1457-1461, 2023, DOI: 10.1109/LSP.2023.3324295.
  - C. A. Lopez and J. Riba, "Parametric Minimum Error Entropy Criterion: A Case Study in Blind Sensor Fusion and Regression Problems," in IEEE Transactions on Signal Processing, vol. 72, pp. 5091-5106, 2024, DOI: 10.1109/TSP.2024.3488554.

- **Chapter 5** unveils a hidden expression of diversity in the Covariance Conversion problem from Frequency Division Duplex schemes in wireless communications. In particular, we consider a setting where a certain notion of sparsity of the second-order statistics of the channel can be defined. As a result, we reformulate this problem to an equivalent sparse regression problem. Our proposed solution is capable of surpassing known state of the art approaches to the Covariance Conversion problem when the sparse assumption holds.

  The technical work that resulted from this chapter is:

  - C. A. Lopez and J. Riba, "Sparse-Aware Approach for Covariance Conversion in FDD

Systems," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 1726-1730, DOI: 10.23919/EUSIPCO55093.2022.9909956.

- **Chapter 6** focuses on the quantification of diversity between two Gaussian random vectors by means of the information theoretic coherence, which is a measure built on the MI between two random variables. For this reason, the main contributions of this chapter consist in a regularized estimation of the Mutual Information via model-order selection rules.

  The technical work that resulted from this chapter is:

  – C. A. López, F. de Cabrera and J. Riba, "Estimation of Information in Parallel Gaussian Channels via Model Order Selection," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 5675-5679, DOI: 10.1109/ICASSP40776.2020.9053506.

- **Chapter 7** concludes this dissertation by outlining the achieved objectives and reflecting on the future lines of research.

A summary of this thesis can be found in Figure 1.1, where we relate the previously described chapters (light green are the supporting chapters and light blue are the chapters that explicitly explore diversity) and their inspiration from general frameworks (light red).

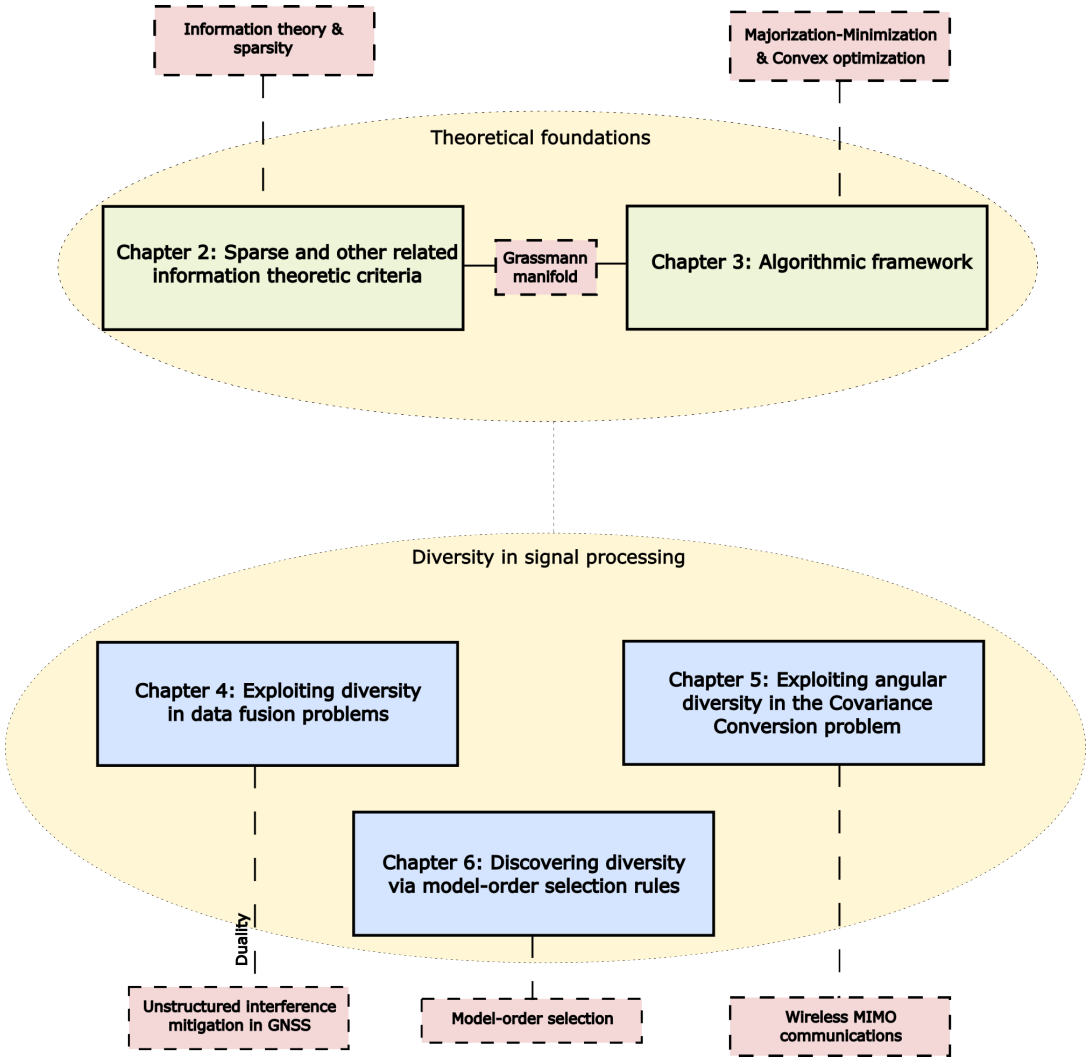


Figure 1.1: Thesis outline.

Chapter 2

Sparse and other related information theoretic criteria

The purpose of this chapter is to introduce the cost functions that are studied in this dissertation. Particularly, we review the concepts of sparsity, information theoretic measures, subspace learning and model-order selection with a mathematical optimization perspective. Although there are mentions to *convexity* and *non-convexity* in this chapter, we refer to Chapter 3 for a review on those concepts.

There has been a increased interest in sparse-aware ideas in the Signal Processing community since the publication of the seminal paper in Compressed Sensing (CS) [37]. From that point, sparse-aware methods have been widely explored in signal processing applications, deepening the exploration of the CS methodology [161], or even extending those ideas to statistical learning [79] and sparse filtering [205]. The classical idea of a sparse solutions comes from the problem of finding the solution with the most amount of components set to zero in an underdetermined system of equations [81], which is referred to as a *sparse solution*. Yet, in the context of this dissertation, not only we use sparse-aware statistical learning ideas to induce sparse solutions, but we also use them to introduce structural priors on a given model. In this regard, we show that there are several kinds of problems emanating from the classical sparse definition that are equivalent from a mathematical optimization perspective (and also application-wise). Thus, we aim at encompassing these kinds of problems in a generalized sparse signal processing framework. In order to specify the aforementioned framework, we consider a toy example that illustrates the fact that sparsity can be understood from different perspectives. Let an arbitrary signal $\mathbf{y}$ be modeled as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{2.1}$$

being $\mathbf{x} \in \mathbb{R}^D$ the encoding of $\mathbf{y}$ using $\mathbf{H} \in \mathbb{R}^{N \times D}$, which we term the *matrix of regressors*, and $\mathbf{w} \in \mathbb{R}^N$ is an additive noise vector. The corresponding *inverse problem* [6] of (2.1) consists on the estimation of $\mathbf{x}$ given $\mathbf{y}$. Two different sparse-aware frameworks for solving the presented inverse problem are identified depending on the interplay between D and N.

1. If $D \geq N$, then $\mathbf{H}$ is considered a *dictionary* of the input signal. In this setting, the sparse assumption considers that the optimal value of $\mathbf{x}$ has a limited amount of non-zero entries. Denoting as $S \ll D$ the presumed non-zero entries, $\mathbf{x}$ is said to be an S-sparse vector. In this case, it is said that $\mathbf{y}$ admits a sparse representation on $\mathbf{H}$ or, equivalently, (2.1) is a sparse model of $\mathbf{y}$ [33], [58]. The previous properties on $\mathbf{x}$ and $\mathbf{y}$ is what describes the classical definition of sparsity [37]. There are two possible kinds optimization problems that result from the previous interpretation of sparsity.

   When $\mathbf{H}$ is known, the recovery of the sparse representation, i.e. estimating $\mathbf{x}$, is known as *sparse coding* [5]. As shown in the sequel, the sparse solution is induced from what we term a *sparse inducing* penalty that is added to the cost function for solving the original inverse problem. Note that this interpretation can also be related to the use of frames in signal processing [106], where $\mathbf{H}$ would be an overcomplete basis of vectors.

On the other hand, when both $\mathbf{H}$ and $\mathbf{x}$ are unknown, the resulting optimization problem that finds those two parameters is known as *dictionary learning*. Problems of this kind have been found useful in signal processing applications [202] due to its relationship with the known matrix factorization (see [95] and [181] for a comparison between the two problems). However, in this dissertation we only deal with the problems of the first kind because we avoid the dictionary learning problem in favor of subspace-based approaches since the subspace-based approaches imply the manipulation of smaller matrices.

2. For $D < N$, the signal component in (2.1) lies in a subspace with smaller dimension than $\mathbb{R}^N$, enabling the incorporation of the Grassmann manifold [20] into the sparse-aware framework. Thanks to the geometric properties of the Grassmann manifold [20], which often arises in signal processing due to a homogeneity property of the cost functions [57], the estimation of $\mathbf{H}$ and $\mathbf{x}$ can be performed efficiently by means of the geodesically convex optimization framework [192]. As a consequence, the subject of study of this scenario is encompassed in the *subspace learning* problem [62], in which the geometry of the Grassmann manifold plays an important role.

   The subspace learning problem is related to other well-known signal processing techniques, such as the matrix factorization problem [7], subspace tracking [54], [191] and the Principal Component Analysis (PCA) framework [96], [206], to name a few. In this regard, it can also be shown that the dictionary learning problem is equivalent to the subspace learning one. This is seen from the fact that learning a low-rank subspace representation of a signal is equivalent to firstly learn an overcomplete representation of $\mathbf{y}$ by estimating $\mathbf{H}$ and $\mathbf{x}$ (for the case where $D > N$), and then to drop the components of $\mathbf{x}$ that are zero and the respective columns of $\mathbf{H}$. The main difference between the dictionary learning and subspace learning approaches is found on the techniques that are used to solve them. While the dictionary learning problem is based on convex optimization (often resorting to sparse regularizers), the subspace problem uses orthogonality constraints [57] that enable geodesically convex optimization framework. The main advantage of the subspace learning methods as compared to the dictionary learning is that they resort to the manipulation of a smaller matrix of regressors. Hence, the subspace learning has the possibility of being computationally faster and requires less memory than the dictionary learning counterpart. However, it is shown in the sequel that in subspace learning methods global optimality must be sacrificed, so one can only ensure a local optimality of the resulting optimization algorithms.

   As for the connection to the classical definition of sparsity, note that the total number of singular values of the signal component in (2.1) is equal to D, the *intrinsic dimension* of the problem [36], while the dimension of the ambient space is N. In this sense, the vector containing the singular values is D-sparse. As an additional remark, this scenario can become agnostic of prior information by the estimation of the intrinsic dimension (see [135] for an example), but the added layer of estimating D is out of the scope of this dissertation.

The intuition behind preferring sparse solutions in practical applications comes from the Occam's razor bias which states that *the simplest explanation of a phenomenon of interest is the most likely to be true* [139]. In this dissertation, the use of sparse-aware techniques is motivated by the fact that a model with the least amount of free parameters is less prone to modeling errors. Not only that, but they also provide a natural robustness to the resulting parameter estimation. This is also the reason why we are also interested in model-order selection rules [177] and their role at discovering the amount of sparsity in a given problem.

As it is expanded in the sequel, sparsity can also be understood in terms of entropic measures [165], which can be applied in many inverse problems that share the same structure as the one described in (2.1) for $D \geq N$. Intuitively, given that sparse solutions have few non-zero components, they are also related to a minimum entropy criterion [63]. This is inferred from the fact that entropy [150] can be interpreted as the amount of dispersion in the measured statistical distribution. Hence, with some abuse of notation, searching for the minimum value of a cost function using an entropic measure regularizer on $\mathbf{x}$ must be equivalent to use a classical sparse inducing measure. Moreover, even though this dissertation is motivated by the sparse-aware framework, we have found that *anti-sparse* measures [59] are also useful in the general data fusion problem. This kind of measures are useful for those applications that require having the most amount of non-zero elements, but its respective cost function

promotes sparsity, e.g. the Geometric Mean Squared Error (GMSE) [130]. Therefore, anti-sparse measures promote maximum entropy solutions.

The structure of the remainder of this chapter is straightforward, where in each section we explore a particular aspect of the toy inverse problem presented in (2.1). In Section 2.1, we relate and review the concepts of sparsity and anti-sparsity with minimum and maximum entropy measures, respectively, while studying their role (and measures that are derived from them) as cost functions. In Section 2.2, we review the geometry of the Grassmann manifold as a way to study the sparsity in the subspace sense under a numerical optimization perspective. Lastly, we complement the previous two sections with an extensive review on model-order selection rules in Section 2.3.

2.1 Sparsity and information theoretic cost functions

Sparsity and entropy are two different alternatives of quantifying the same phenomenon. In order to highlight the similarities between those two concepts, we consider an intuitive definition of sparsity and entropy that encompasses the properties that are used for solving the inverse problems that appear in this dissertation [146].

Definition 2.1 (Sparsity measures)**.** *Let* $\mathbf{x} \in \mathbb{R}^D$*. Then, any sparsity measure accounts for the concentration of the energy in few coefficients in* $\mathbf{x}$*.*

Definition 2.2 (Entropic measures)**.** *Let* $\mathbf{x} \in \mathbb{R}^D$ *be a random variable whose associate Probability Mass Function (PMF) is* $\mathbf{p} \in \mathbb{R}^D$ *(or any function that shares the same properties as the PMF of* $\mathbf{x}$*). Then, any entropic measure quantifies the dispersion in* $\mathbf{p}$ *and, as a result, in* $\mathbf{x}$*.*

Remark 2.1. The dispersion of the energy in most of the coefficients in $\mathbf{x}$, i.e. the antagonistic concept to sparsity, is also referred to as *diversity* [185] (not to be confused to the diversity in data fusion, see Definition 4.1). Therefore, an entropic measure grows as the diversity of a signal increases.

The above definitions describe two antagonistic phenomena on $\mathbf{x}$. In fact, entropy and sparsity have had a joint trajectory in several prior works. Particularly, generalized entropies [56] and an empiric measure called the *entropy function* of a signal [83] are used in the literature as sparsity-promoting regularizers for sparse signal recovery. The following example, which is extracted from [146], offers an additional intuition on the relationship between sparsity and entropic measures.

Example 2.1. Let $X \in \{x_1, x_2\}$ be a random variable whose associated PMF is $\mathbf{p} = [p_1, p_2]$. Also, let $s(\cdot)$ be any sparse measure and $h(\cdot)$ be any entropic measure. Then, these measures must satisfy:

1. Assuming initially that $p_1 > p_2$, concentrating even more probability mass on p_1 results in x_1 being more certain to appear in any realization. As a consequence, $\mathbf{p}$ is more compressible. Under the previous procedure, the measures of sparsity and entropy evolve in the following way:

$$\uparrow s(\mathbf{p}) \equiv \downarrow h(X), \tag{2.2}$$

2. Consider that $\mathbf{p}$ is $= [1, 0]^T$ initially. Then, any increment in p_2 implies that both x_1 and x_2 are now possible outcomes. Consequently, $\mathbf{p}$ becomes less compressible, so the complexity of $\mathbf{x}$ increases, yielding:

$$\downarrow s(\mathbf{p}) \equiv \uparrow h(X), \tag{2.3}$$

where we denote with $\uparrow$ and $\downarrow$ when a measure is increasing and decreasing, respectively.

In light of the previous ideas, we are interested in the study of sparsity and information theoretic measures, which are a generalization of entropic measures, in two families of optimization problems: the ones that actively seek sparse solutions via a sparse regularizer and the ones whose cost function is a parameterized information theoretic measure. Regarding the first kind of optimization problems, we consider regularized optimization problems of the following form:

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda r(\mathbf{x}), \tag{2.4}$$

where $f(\mathbf{x})$ is a fitting function constructed in such a way that its minimization retrieves the optimal solution of an inverse problem, e.g. the quadratic adjustment between the estimated model and the measurements, $r(\mathbf{x})$ is a regularization function that induces sparsity on $\mathbf{x}$ and $\lambda \geq 0$ is a regularization

parameter that specifies the sparsity level of the optimal solution. An alternative formulation to the one in (2.4) would be:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s. t. } r(\mathbf{x}) \leq R_s, \tag{2.5}$$

where $R_s \geq 0$. Note that the reasoning behind the sign of the inequality constraint in (2.5) is that $r(\mathbf{x})$ is often assumed to be a convex function, and hence the overall constraint set would be convex. From now on, (2.4) and (2.5) are termed the unconstrained and constrained formulations, respectively. In the following proposition we show that both formulations (constrained and unconstrained) are equivalent as long as the constrained formulation is feasible and $\lambda \geq 0$ in (2.4).

Proposition 2.1 (Equivalence between unconstrained and constrained regularized optimization problems)**.** *Let any two optimization problems be as the ones depicted in (2.4) and (2.5). Provided that* (2.5) *is feasible and that* $\lambda \geq 0$ *in* (2.4)*, there exists two pairs of values* (R_s, λ) *such that (2.4) and (2.5) are equivalent.*

Proof. In order to prove this proposition, we firstly show the conditions of the optimal solution of (2.5) and relate them to the ones of (2.4). In this sense, the optimal solution of (2.5) must satisfy the Karush-Kuhn-Tucker (KKT) conditions of the problem [29]. For the purpose of stating the equations that depict the KKT conditions of (2.5), let us consider its respective Lagrangian:

$$\mathcal{L} = f(\mathbf{x}) + \mu \left(r(\mathbf{x}) - R_s \right), \tag{2.6}$$

where μ is the inequality constraint Lagrange multiplier. Then, the KKT conditions of the constrained formulation are given by the following set of equations [29]:

$$\nabla_{\mathbf{x}} \mathcal{L} = \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \mu^* \nabla_{\mathbf{x}} r(\mathbf{x}^*) = \mathbf{0}, \tag{2.7a}$$

$$r(\mathbf{x}^*) - R_s \leq 0, \tag{2.7b}$$

$$\mu^* \geq 0, \tag{2.7c}$$

$$\mu^* \left(r(\mathbf{x}^*) - R_s \right) = 0, \tag{2.7d}$$

where $\mathbf{x}^*$ and μ^* are the optimal values of the primal and dual variables, respectively. Assuming that there exists $\mathbf{x}^*$ and μ^* such that all equations in (2.7) are fulfilled, i.e. the constrained formulation is feasible, substituting $\lambda = \mu^*$ in (2.4) would yield the same optimal value of $\mathbf{x}$ since (2.7a) is also the only optimal condition of the unconstrained formulation in (2.4). ■

Corollary 2.1.1. *There exists a pair of values* (C_s, R_s) *such that the following two optimization problems:*

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s. t. } r(\mathbf{x}) \leq R_s, \tag{2.8a}$$

$$\min_{\mathbf{x}} r(\mathbf{x}) \quad \text{s. t. } f(\mathbf{x}) \leq C_s, \tag{2.8b}$$

have the same optimal solution.

Remark 2.2. The proof of the previous corollary follows straightforwardly from the proof of the previous proposition by linking (2.8a) and (2.8b) to a common unconstrained formulation.

The previous proposition offers a way to reformulate any given optimization problem. Particularly, we are interested in the use of Corollary 2.1.1 to build optimization problems in such a way that the user-defined parameters can be set in a more intuitive manner. For instance, it is difficult to set the regularization parameter in (2.4) without resorting to cross-validation techniques since it is challenging to translate the pragmatic interpretation of λ (directly related to the sparsity of the solution) to the actual application. As seen in a future chapter, the formulations that are similar to the one in (2.8b) are much more useful from a practical point of view. Taking as a reference the inverse problem depicted by (2.1), we argue that the following optimization problem (inspired by (2.8b)):

$$\min_{\mathbf{x}} r(\mathbf{x}) \quad \text{s. t. } ||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 \leq \varepsilon, \tag{2.9}$$

is much more suited for a practical implementation than the following one:

$$\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 \quad \text{s. t. } r(\mathbf{x}) \leq R_s. \tag{2.10}$$

The reasoning behind the previous observation is that the constraint set in (2.9) has a much clearer interpretation. Here, ε accounts for the expected modeling errors due to the uncertainty of the measurements. In contrast, there is no clear reasoning behind R_s in (2.9), apart from controlling the sparsity of the solution (whose optimal level is not known a priori). For the previous reasons, we favor optimization problems that have the same structure as the one in (2.9) in front of the formulation given in (2.10), or even an unconstrained formulation.

As for the remaining family of optimization problems considered in this dissertation, we are interested in optimization problems of the following form:

$$\min_{\mathbf{x}} I(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X}, \tag{2.11}$$

where $I(\mathbf{x})$ is any parameterized information theoretic function, which can be based on the Rényi entropy or the Mutual Information (MI), and $\mathcal{X}$ is any constraint set that can be non-convex. Provided that the properties of $I(\mathbf{x})$ are highly dependent on the considered information theoretic measure, we contextualize the benefits of the previous formulation for each measure in the sequel. As a general insight, information theoretic measures are known for providing robustness in a natural manner [52], which comes in contrast to the ℓ_p norms as a sparsity measure. The aim of this approach within the context of this thesis is to find a middle ground between the generality of non-parametric methods and the optimality of model-based approaches. As a result, the proposed information theoretic formulations are aligned with the robust signal processing framework [212], whose goal is to construct algorithms that are close to optimal under nominal conditions, even if those conditions are only approximately valid.

In the following subsections, we expand on the foundations of the cost functions on which future chapters of this dissertation are substantiated. While we review ℓ_p norms as sparse inducing regularizers in the first subsection, we study the properties of a parameterized Rényi entropy [155], [165] and of the Shannon MI with a numerical optimization perspective in the second subsection.

2.1.1 ℓ_p norms

Definition 2.3 (ℓ_p norm)**.** *Let $\mathbf{x} \in \mathbb{R}^D$. Then, its ℓ_p norm is defined as:*

$$||\mathbf{x}||_p = \left(\sum_{d=1}^{D} |x_d|^p \right)^{\frac{1}{p}}, \tag{2.12}$$

where x_d denotes the d-th component of $\mathbf{x}$ and $p \geq 0$.

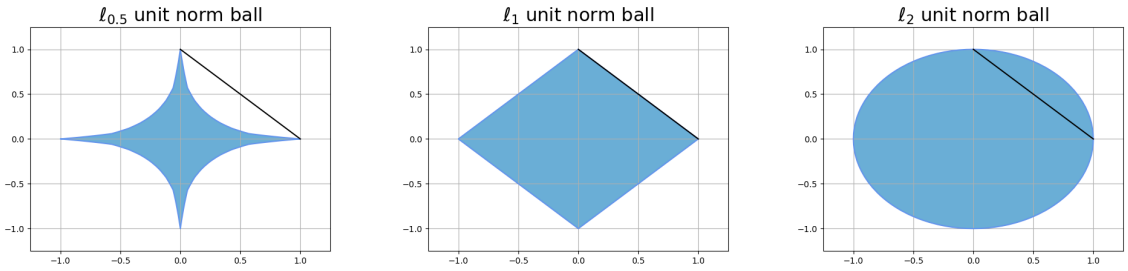
The properties of an ℓ_p norm depend on the value of p. For $p \geq 1$, the ℓ_p norms satisfy the distance from origin axioms in normed vector spaces and thus they are *norms* in the usual sense. From a numerical optimization perspective, it means that they are convex functions of its input arguments [29, Section 3.1.5]. In contrast, some of the axioms of normed spaces are not fulfilled for $p < 1$, and hence the ℓ_p norms are referred to as *pseudonorms* in this case. Pseudonorms are known to be non-convex functions. In order to illustrate the convexity of norms and non-convexity of pseudonorms, we show graphically the unit ℓ_p norm balls for p equal to 0.5, 1 and 2 in Figure 2.1. The non-convexity of each set is determined by the relative position of the (black) segment connecting $[1, 0]$ and $[0, 1]$ with respect to the given unit norm ball in this figure. For instance, the $\ell_{0.5}$ unit norm ball defines a non-convex set due to the fact that the aforementioned segment lies outside the unit norm ball. In contrast, the ℓ_1 and ℓ_2 unit norm balls contain the black segment, where the ℓ_1 unit norm is the limiting case for a unit norm ball to be convex.

A useful property of the ℓ_p norms that complements the previous observations from Figure 2.1 is given in the following lemma.

Lemma 2.2 (Inequality of the ℓ_p norm)**.** *Let $1 < p < q < \infty$. Then, the ℓ_p norms satisfy:*

$$||\mathbf{x}||_q \leq ||\mathbf{x}||_p. \tag{2.13}$$

Remark 2.3. Note that (2.13) implies that the unit norm ball has a greater *volume* for increasing p. This is clearly observed in Figure 2.1.



(a) $\ell_{0.5}$ unit norm ball. **(b)** ℓ_1 unit norm ball. **(c)** ℓ_2 unit norm ball.

Figure 2.1: Insights on the convexity of ℓ_p norms.

Regarding the applicability of ℓ_p norms in the sparse coding problem, there are three cases of interest: the ℓ_0, ℓ_1 and ℓ_∞ norms. One can obtain the ℓ_0 and ℓ_∞ norms by computing the respective limits. Particularly, the ℓ_0 norm yields:

$$||\mathbf{x}||_0 = \lim_{p \longrightarrow 0} \left(\sum_{d=1}^{D} |x_d|^p \right)^{\frac{1}{p}} = \text{card}(\mathbf{x}), \tag{2.14}$$

where $\text{card}(\cdot)$ denotes the cardinality of a vector, meaning that it is the function that counts the non-zero elements of its argument. The previous result suggests that the ℓ_0 norm is an ideal sparsity inducing regularizer. However, the main issue with the ℓ_0 norm as a regularizer is its non-convexity and its non-continuity. Consequently, its use often results in combinatorial NP-hard problems [48]. As an alternative, the ℓ_1 norm is used as a surrogate since it is the tightest convex relaxation of the ℓ_0 norm. Not only the ℓ_1 norm is convex, but it also promotes sparsity in practical applications [5], [37], [205]. Given that there exist efficient algorithms to solve regularized ℓ_1 problems (see [145] and [19] for examples), we favour the ℓ_1 convex relaxation to induce sparse solutions.

Similarly to the ℓ_0 norm, the ℓ_∞ norm is derived from the following limit:

$$||\mathbf{x}||_\infty = \lim_{p \longrightarrow \infty} \left(\sum_{d=1}^{D} |x_d|^p \right)^{\frac{1}{p}} = \max(\mathbf{x}), \tag{2.15}$$

which, in contrast to the ℓ_0 norm, is a convex function of its input argument. Given that ℓ_0 and ℓ_1 norms are well-known regularizers that induce sparse solutions, it is intuitive to think that as $p \to \infty$ the sparsity promoting property of ℓ_p norms fades away in favour of anti-sparsity. In fact, the ℓ_∞ norm have been proven useful for the task of promoting dispersed (anti-sparse) solutions [59], although its anti-sparse inducing property acts in a tricky way. In the sequel, we show how this mechanism works.

2.1.1.1 Using ℓ_p norms in inverse problems

The particular cases of ℓ_0 and ℓ_1 norms have had a long trajectory in sparse inverse problems since the earlier works on CS [37], where the Matching Pursuit (MP) [24] and Basis Pursuit Denoising (BPD) algorithms [17] have been the staple approaches. While the MP formulation can be seen as the classic formulation of CS, the BPD is much more suited for the signal processing applications that are considered in this dissertation. In the following paragraphs, we review both formulations and comment on their applicability.

Using as a reference the inverse problem depicted in (2.1), earlier approaches of the MP could be used to obtain $\mathbf{x}$, which are based on the following optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{x}||_0 \quad \text{s.t. } \mathbf{H}\mathbf{x} = \mathbf{y}. \tag{2.16}$$

Note that since $\mathbf{y} = \mathbf{H}\mathbf{x}$ is an underdetermined system of equations ($D > N$), it has an infinite amount of solutions. Thus, the above optimization problem is feasible in general. Nevertheless, the

optimization problem in (2.16) is known to be NP-hard [24] due to the non-convexity of the ℓ_0 norm. As a matter of fact, the above optimization requires a combinatorial search that grows exponentially with D for the purpose of ensuring the global optimality of the final solution. This is the reason why the ℓ_0 is often relaxed to the ℓ_1 norm [37], [156], yielding the following convex program:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \quad \text{s. t. } \mathbf{H}\mathbf{x} = \mathbf{y}. \tag{2.17}$$

A matrix is said to fulfill the RIP for some S if every $N \times S$ submatrix of the given matrix is approximately orthogonal [37]. If $\mathbf{H}$ fulfills the so-called Restrictive Isometry Property (RIP) [37] for some integer S, then the optimization problem given in (2.17) is able to retrieve the sparsest solution of the aforementioned system of equations, which contains S non-zero entries. Provided that ensuring that a matrix satisfies this property is NP-hard in general and that the formulations similar to the one in (2.17) are avoided, the RIP is out of the scope of this dissertation.

The main issue with the previous two formulations is that they often yield poor performance in the presence of noise due to the fact that the equality constraint in (2.16) and (2.17) does not allow room for errors in the signal model or in the measurements. This is the reason why, in practical signal processing applications, the BPD approach [9], [17], [21] is preferred over the MP. The BPD is based on the following level-set formulation:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \quad \text{s. t. } ||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 \leq \epsilon, \tag{2.18}$$

where ϵ accounts for expected fitting errors, which can consist on the additive noise or modeling errors. Notice that we have already introduced the convex relaxation of the ℓ_0 norm. Clearly, the BPD collapses to the MP formulation for $\epsilon \to 0$. Invoking Corollary 2.1.1, we can rewrite (2.18) into the following expression:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 \quad \text{s. t. } ||\mathbf{x}||_1 \leq \sigma, \tag{2.19}$$

where σ regulates the sparsity of $\mathbf{x}$ and the multiplicative constant of the cost function is considered for convenience. This latter formulation is termed the LASSO [184] in the literature. The main advantage of (2.19) with respect to (2.18) is that the minimization of (2.19) is much easier than the minimization of (2.18) because the constraint set is simpler to handle [9]. In fact, the solution of (2.19) can be found by first minimizing the objective and then projecting onto the scaled ℓ_1 norm ball. Still, (2.18) can be solved by means of the Alternating Direction Method of Multipliers (ADMM), which is an iterative optimization scheme that is extensively reviewed in Chapter 3. As a summary, (2.19) admits a faster minimization, but we prefer optimization problems that have the same structure as the one in (2.18) to have an easier interpretation of the user-defined parameters, e.g. it seems easier to fix ϵ than σ in an actual application.

Regarding the ℓ_∞ norm, we show how its anti-sparse mechanism works in a toy optimization setting, which is depicted by:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} g(\mathbf{x}) \quad \text{s. t. } ||\mathbf{x}||_\infty \leq c_0, \mathbf{x}^T\mathbf{1} = 1, \mathbf{x} \succeq \mathbf{0}_D, \tag{2.20}$$

where $g(\mathbf{x})$ is any cost function that naturally induces sparse solutions, e.g. GMSE criterion [130], and $\frac{1}{D} \leq c_0 \leq 1$ is the regularization constraint. The last two constraints are introduced to facilitate the exposition of the anti-sparse mechanism. Assume that the optimal solution of (2.20) without the ℓ_∞ norm ball constraint is a sparse vector with only one non-zero component (whose value is 1 due to the second constraint). Then, if we activate the norm ball constraint and c_0 satisfies $c_0 < 1$, the optimal solution of (2.20) is forced to allocate more energy in other components of $\mathbf{x}$, achieving the maximum dispersion for $c_0 = \frac{1}{D}$, i.e. $\mathbf{x} = \frac{1}{D}\mathbf{1}_D$. Note that c_0 must be bounded in $[\frac{1}{D}, 1]$. The reasoning behind the previous bound on c_0 is that (2.20) is unfeasible for $c_0 < \frac{1}{D}$ since the remaining two constraints cannot be satisfied. Likewise, the ℓ_∞ norm constraint becomes nuisance for $c_0 > 1$ since the remaining two constraints forces the values of each component in $\mathbf{x}$ to be bounded in $[0, 1]$.

2.1.1.2 Interpretability of the ℓ_p norms

Although the use of the ℓ_p norms has been driven by a pragmatic reasoning, it admits an interpretation from a Bayesian point of view. This interpretation is an example of how it is possible to relate a statistical prior with a structural prior (and viceversa). Thus, the Bayesian interpretation of the BPD formulations from (2.18) or (2.19) are a fundamental first step to understand the incorporation of structural priors in other cost functions widely used in signal processing. In this regard, the aforementioned BPD formulations can be shown to be equivalent to a Maximum a Posteriori (MAP) estimation of $\mathbf{x}$ by means of Proposition 2.1. This equivalence is achieved after the consideration of a Laplacian prior on $\mathbf{x}$ [13]. In order to showcase this connection, let the statistical model of $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ be:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \gamma_w \mathbf{I}_N), \tag{2.21a}$$

$$x_d \sim \text{Laplace}(0, b_l) \quad \forall d = 1, ..., D, \tag{2.21b}$$

where x_d is the d-th component of $\mathbf{x}$ and $b_l > 0$ is the parameter of the Laplacian distribution. In this case, $\mathbf{H}$ is a deterministic matrix. The joint distribution of $\mathbf{x}$, assuming independence between components, is obtained after the multiplication of the marginal priors:

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{d=1}^{D} \frac{1}{2b_l} \exp\left(-\frac{|x_d|}{b_l}\right) = \frac{1}{(2b_l)^D} \exp\left(-\frac{||\mathbf{x}||_1}{b_l}\right). \tag{2.22}$$

Also, let the conditional distribution of $\mathbf{y}$ be:

$$f_{\mathbf{w}}(\mathbf{y}|\mathbf{H}, \mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \gamma_w^N}} \exp\left(-\frac{1}{2\gamma_w}||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2\right). \tag{2.23}$$

Then, the MAP estimation of $\mathbf{x}$ consists on the maximization of the MAP function of the data given the prior obtained by the multiplication of (2.22) and (2.23). The previous procedure yields the following optimization problem:

$$\hat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x}} \log(f_{\mathbf{w}}(\mathbf{y}|\mathbf{H}, \mathbf{x})) + \log(f_{\mathbf{x}}(\mathbf{x})) = \arg\max_{\mathbf{x}} -\frac{1}{2\gamma_w}||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 - \frac{||\mathbf{x}||_1}{b_l} + C, \tag{2.24}$$

where C gathers additive constants that do not depend on $\mathbf{x}$. After rearranging terms and ignoring C, the resulting optimization problem is:

$$\hat{\mathbf{x}}_{MAP} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 + \frac{\gamma_w}{b_l}||\mathbf{x}||_1. \tag{2.25}$$

Notice that (2.25) is equivalent to the Lagrangian form of (2.18) or (2.19). Indeed, the MAP framework offers a Bayesian interpretation to the user-defined parameters of the BPD formulations. Still, choosing the parameter of the Laplacian distribution in a practical setting may still seem too arbitrary (and prone to modeling errors) as compared to ϵ in (2.18), whose value may be fixed using pragmatic arguments.

In a similar fashion to the ℓ_1 norm, the ℓ_∞ norm regularization also has a similar Bayesian interpretation. Indeed, the ℓ_∞ regularization is equivalent to assume a democratic prior [59] on $\mathbf{x}$, whose PDF is:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{D! 2^D b_d^D} \exp\left(-\frac{||\mathbf{x}||_\infty}{b_d}\right), \tag{2.26}$$

where $b_d > 0$ is the parameter of the democratic prior. With the same procedure as the one used to obtain (2.25), the resulting optimization problem that is built using the democratic prior yields:

$$\hat{\mathbf{x}}_{MAP} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 + \frac{\gamma_w}{b_d}||\mathbf{x}||_\infty, \tag{2.27}$$

which is the specific formulation that was used in [59] to enforce anti-sparsity on the toy inverse problem depicted by (2.1).

2.1.2 Information theoretic criteria

Even though there are several formal definitions of entropic measures, which are the foundations of information theoretic measures, such as the Shannon or Tsallis entropies [195], we have chosen the Rényi entropy [165] as the main building block for constructing information theoretic cost functions. As shown in the sequel, the motivation behind the utilization of the Rényi entropy is that, not only it is a generalization of the Shannon entropy [165], but its inherent structure provides of a natural robustness to the final solution and a practical implementation. The Rényi entropy is defined in the following way.

Definition 2.4 (Rényi entropy)**.** *Let X be a discrete random variable with Probability Mass Function (PMF) given by $P_X(X)$. Then, its generalized Rényi's entropy is:*

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{d=1}^{D} P_X^\alpha(X_d) \right), \tag{2.28}$$

where X_d for $d = 1, ..., D$ are the possible outcomes of X and α is the entropic index. Similarly, let $\mathbf{x} \sim p_\mathbf{x}(\mathbf{x})$ be a bounded Lebesgue integrable Probability Density Function (PDF). Then, its generalized (differential) Rényi's entropy is defined as:

$$h_\alpha(\mathbf{x}) = \frac{1}{1-\alpha} \log \left(\int_{-\infty}^{\infty} p_\mathbf{x}^\alpha(\mathbf{x})\, d\mathbf{x} \right). \tag{2.29}$$

The previous two expressions can be expressed more compactly using the notation from ℓ_p and L^p spaces respectively. In this way, (2.28) is rewritten as:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(||\mathbf{p}_X||_\alpha^\alpha \right), \tag{2.30}$$

where the d-th component of $\mathbf{p}_X$ is $P_X(X_d)$. In a similar manner, (2.29) is equivalent to the following expression:

$$h_\alpha(\mathbf{x}) = \frac{1}{1-\alpha} \log \left(||p_\mathbf{x}(\mathbf{x})||_\alpha^\alpha \right), \tag{2.31}$$

where:

$$||p_\mathbf{x}(\mathbf{x})||_\alpha^\alpha = \int_{-\infty}^{\infty} p_\mathbf{x}^\alpha(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{2.32}$$

which holds for the PDFs that were considered in Definition 2.4. What is more, the previous definition is known to collapse to the Shannon entropy and Shannon differential entropy (respectively) taking the limit for $\alpha \to 1$ and using L'Hôpital rule:

$$H_1(X) = -\sum_{d=1}^{D} P_X(X_d) \log \left(P_X(X_d) \right), \tag{2.33}$$

and:

$$h_1(\mathbf{x}) = -\int_{-\infty}^{\infty} p_\mathbf{x}(\mathbf{x}) \log(p_\mathbf{x}(\mathbf{x})) \mathrm{d}\mathbf{x}, \tag{2.34}$$

which verifies that the Rényi entropy is indeed a generalization of the Shannon entropy. As an intuitive comment, the exchange between the logarithm and the integral of the Rényi entropies (in contrast to the Shannon entropy) facilitates its use in optimization problems due to the fact that this structure is much more tractable in practice. As an example, the particular case of $\alpha = 2$, also known as *collision entropy*, promotes the use of second-order methods (see [35], [51] and references therein).

In the grand scheme, we motivate the study of information theoretic criteria that is based on Rényi entropies by two ideas. On the one hand, we are interested in the search for an operational meaning in the context of signal processing of the measures that are based on the Rényi entropy, given that its definition and utilization was mainly driven by a pragmatic reasoning. In this sense, the original purpose of Rényi entropies was to define an entropic measure that satisfied most of the Shannon entropy axioms, but tailored in a way that an information theoretic proof of the Central Limit Theorem could

be derived [78], [165]. Our objective is to find a natural way to answer the natural questions that emerge in signal processing applications using information theoretic descriptors, as it happens with the interpretation of the Shannon entropy as the compression rate of an information source [78] and also with the Rényi entropy when it is seen as a particular case of the *Rényi divergence* [61]. In the particular case of this dissertation, we look for interpretations of information theoretic criteria in signal processing applications, e.g. the properties of the solutions or the entropic parameters can be intuitively explained within the context of the particular application. This comes in contrast to the use of ℓ_1 norm regularization, which is often praised by practitioners, but its use may appear as an arbitrary additive penalty to the original cost function.

On the other hand, information theoretic measures have the advantage of providing a robustness that is inherent to their own nature. Indeed, this robustness is acquired by the fact that entropy depends on the probability of the events rather than in their magnitude [50]. Thanks to this feature, this kind of measures can also be used as a way to be not too restricted by the assumed model. Indeed, the known Minimum Error Entropy (MEE) criterion has been widely used in signal processing applications to deal with non-Gaussianity in a non-parametric manner, often resorting to kernelized measures of the Rényi entropy [60], [77], [200]. Nevertheless, we are not interested in information theoretic measures to face non-Gaussianity. Instead, we want to modify MEE-like criteria (and related approaches) to ensure the optimality under nominal conditions (Gaussianity) while still being robust to deviations of the assumed model. This is the reason why the information theoretic criteria that we consider in this dissertation lie, in part, in the robust signal processing framework [212], in contrast to earlier works that use the MEE criterion whose aim is to have a wide range of validity. An example of the previous information theoretic framework is found in [52], where the authors use a particularized expression of the Rényi entropy to robustify the estimation of the data covariance determinant, a widely used statistic in signal processing.

Yet, we remark that the use of entropic criteria to promote sparse and robust solutions is not new in signal processing applications. In [82], [83], entropy functions were used to solve inverse problems such as the one depicted in (2.1). Those alternatives result in NP-hard non-convex problems, but they have been proved to have higher probability of successful signal recovery than classical ℓ_1 norm approaches.

In the following subsections, we show the two main information theoretic cost functions that are considered in subsequent chapters, which are based on parameterized expressions of entropy and the MI. While we only study their properties in this chapter, these cost functions naturally appear in signal processing problems when the Conditional Maximum Likelihood (CML) function [134], [160], [166] is considered, as shown in Chapter 4.

2.1.2.1 Parametric Minimum Error Entropy criterion: Particularization to Gaussian random matrices

As a way to exploit the natural robustness of the generalized Rényi entropies, we define the Parametric Minimum Error Entropy (PMEE) criterion to estimate a general parameter. As stated in an earlier paragraph, the motivation behind this criterion is that the Rényi entropy serves a way to avoid being too constrained by the assumed model (either statistical or structural) while still keeping as much as possible the optimality of a model-driven method under the Gaussian assumption. Thus, this criterion lies in the previously explained robust signal processing framework [212]. We refer to this way of approaching a signal processing problem as a *semi-data-driven* approach.

With the purpose of defining the PMEE, let $\mathbf{X} \sim \mathcal{MN}_{N,M}(\mathbf{0}, \mathbf{K}, \mathbf{Q})$ be a zero-mean $N \times M$ random Gaussian matrix [4], whose associated PDF is:

$$p_{\mathbf{X}}(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{MN}{2}} \det(\mathbf{Q})^{\frac{N}{2}} \det(\mathbf{K})^{\frac{M}{2}}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{Q}^{-1}\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X})\right), \tag{2.35}$$

where $\mathbf{K} \in \mathbb{R}^{N\times N}$ and $\mathbf{Q} \in \mathbb{R}^{M\times M}$ model the covariances among rows and columns of $\mathbf{X}$, respectively. Then, the generalized Rényi entropy of $\mathbf{X}$ is given by:

$$h_\alpha(\mathbf{X}) = \frac{MN}{2}\log(2\pi) + \frac{N}{2}\log(\det(\mathbf{Q})) + \frac{M}{2}\log(\det(\mathbf{K})) + \frac{MN}{2}\frac{\log(\alpha)}{\alpha - 1}, \tag{2.36}$$

whose derivation is detailed in Appendix 8.1.1 using Definition 2.4 and (2.35). For the purpose of unveiling the inherent robustness of the entropic measure, we consider an alternative expression of (2.36):

$$h_\alpha(\mathbf{X}) = \frac{MN}{2}\left(\log(\det(\mathbf{Q}))^{\frac{1}{M}} + \log(\det(\mathbf{K}))^{\frac{1}{N}} + \log(2\pi) + \frac{\log(\alpha)}{\alpha - 1}\right). \tag{2.37}$$

It is clear from the previous expression that the Rényi entropy of a Gaussian random matrix is based on the geometric mean of the eigenvalues of $\mathbf{Q}$ and $\mathbf{K}$. In fact, the geometric mean by itself is known to provide robustness to estimation problems [130] when used as a cost function. Moreover, the geometric mean is linked to the Minimum Covariance Determinant (MCD) criterion, which is also known to be useful against outliers and contaminated sources [85]. In this regard, an intuitive explanation of the MCD estimators is that their goal is to construct a covariance matrix of the observed random variable such that it minimizes the influence of outliers. The aforementioned link to the MCD provides more insights on the performance of (2.37) as a cost function for other non-Gaussian distributions. It has been shown in [85] that the MCD criterion is optimal for elliptically symmetric unimodal distributions, e.g. a Gaussian distribution. Additionally, contaminated Gaussian random variables [188], which are well modeled using Gaussian Mixture Models (GMM), are distributions that are also suited for MCD-like estimators, as seen in [85]. Note that the value of α does not alter the previous observations since α only affects in the positive additive term, $\frac{\log(\alpha)}{\alpha-1}$, which decreases monotonically with α and converges to 1 for $\alpha = 1$ (the Shannon entropy). Inspired by (2.37), we define the PMEE estimator of a general parameter, $\boldsymbol{\theta}$, as follows.

Definition 2.5 (Parametric Minimum Error Entropy criterion)**.** *Let two estimators of the covariances parameterized by $\boldsymbol{\theta}$ be $\hat{\mathbf{K}}(\boldsymbol{\theta})$ and $\hat{\mathbf{Q}}(\boldsymbol{\theta})$. Then, the PMEE estimator of $\boldsymbol{\theta}$ is:*

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{M}\log(\det(\hat{\mathbf{Q}}(\boldsymbol{\theta}))) + \frac{1}{N}\log(\det(\hat{\mathbf{K}}(\boldsymbol{\theta}))). \tag{2.38}$$

Remark 2.4. The PMEE criterion minimizes the parametric error entropy for all α.
Remark 2.5. We have divided the expression in (2.37) by $\frac{1}{MN}$ and ignored the additive constants that do not depend on the estimated covariances to obtain (2.38).
Remark 2.6. Provided that the optimization problem in (2.38) consists on the minimization of a concave function (see Lemma 3.3), it is non-convex in general, being the main drawback of this criterion.

The robustness of the PMEE criterion is highlighted by the clear link to the MCD and the geometric mean of the eigenvalues of the respective estimators. Thus, it inherits the optimality for elliptically symmetric unimodal and contaminated distributions. As a final remark, not only the Gaussian assumption serves as a way to keep the optimality of the PMEE under nominal conditions, but it is also a natural way to specify the parameterized estimators of the covariances, $\hat{\mathbf{K}}(\boldsymbol{\theta})$ and $\hat{\mathbf{Q}}(\boldsymbol{\theta})$. In fact, the natural estimators of the aforementioned covariance matrices that appear in the applications within this dissertation are obtained using the CML principle [134], [160], [166] on a Gaussian likelihood function.

2.1.2.2 Mutual Information of two random variables

Up to this point, we have assumed that the diversity in the dataset of interest, consisting on the parameters that describe the sparsity of said dataset, has already been discovered. For instance, it is often assumed that the intrinsic dimension is known in the inverse problem depicted in (2.1). Yet, we are also interested in the problem of estimating the parameters that depict the diversity in a dataset. For the previous task, we resort to the MI between two random variables as a measure that quantifies the degree of dependence between two datasets. For simplicity, only the Shannon MI is considered, although the MI can also be defined using the Rényi entropy. The (continuous) Shannon MI is defined as follows [173], [183].

Definition 2.6 (Mutual information of two random variables)**.** *Let $\mathbf{x}$ and $\mathbf{y}$ be two random N-dimensional vectors. For simplicity, we do not consider vectors with different dimensions. Then, the MI of $\mathbf{x}$ and $\mathbf{y}$ is given by the following expression:*

$$I(\mathbf{x};\mathbf{y}) = \iint_{-\infty}^{\infty} p_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log\left(\frac{p_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{x}}(\mathbf{x})p_{\mathbf{y}}(\mathbf{y})}\right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \tag{2.39}$$

where $p_{xy}(\mathbf{x},\mathbf{y})$, $p_{\mathbf{x}}(\mathbf{x})$ and $p_{\mathbf{y}}(\mathbf{y})$ are the joint and marginal PDFs of $\mathbf{x}$ and $\mathbf{y}$, respectively. The MI can be alternatively rewritten as follows:

$$I(\mathbf{x};\mathbf{y}) = h_1(\mathbf{x}) + h_1(\mathbf{y}) - h_1(\mathbf{x},\mathbf{y}), \tag{2.40}$$

where $h_1(\mathbf{x},\mathbf{y})$ is the joint entropy of $\mathbf{x}$ and $\mathbf{y}$.

While the previous definition inherits the robustness of the entropy, there are some new properties that are exclusive to the MI. In fact, a useful property of the MI that is not satisfied by the Rényi entropy is detailed in the following lemma [107, Appendix], whose proof is remade since it is difficult to find a sufficient amount of detail in known references.

Lemma 2.3 (Invariance of MI with respect to homeomorphic transformations of the original variables)**.** *Let $\mathbf{x}$ and $\mathbf{y}$ be two Gaussian random vectors. Also, let $\mathbf{x}' = f(\mathbf{x})$ and $\mathbf{y}' = g(\mathbf{y})$, where $f : \mathbb{R}^N \to \mathbb{R}^N$ and $g : \mathbb{R}^N \to \mathbb{R}^N$ are homeomorphisms*[1]*. Then, it is verified that:*

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{x}';\mathbf{y}'). \tag{2.41}$$

Remark 2.7. Even though the MI is invariant to homeomorphisms, the PDFs of the associated random variables may change in general.

Proof. Since $\mathbf{x}'$ and $\mathbf{y}'$ are transformations of $\mathbf{x}$ and $\mathbf{y}$, respectively, their associated PDFs are given by [182]:

$$p_{\mathbf{x}'}(\mathbf{x}') = \frac{p_{\mathbf{x}}(\mathbf{x})}{|\det(\mathbf{J}_f(\mathbf{x}))|}\bigg|_{\mathbf{x}=f^{-1}(\mathbf{x}')}, \tag{2.42a}$$

$$p_{\mathbf{y}'}(\mathbf{y}') = \frac{p_{\mathbf{y}}(\mathbf{y})}{|\det(\mathbf{J}_g(\mathbf{y}))|}\bigg|_{\mathbf{y}=g^{-1}(\mathbf{y}')}, \tag{2.42b}$$

$$p_{\mathbf{x}'\mathbf{y}'}(\mathbf{x}',\mathbf{y}') = \frac{p_{\mathbf{x}\mathbf{y}}(\mathbf{x},\mathbf{y})}{|\det(\mathbf{J}_f(\mathbf{x}))||\det(\mathbf{J}_g(\mathbf{y}))|}\bigg|_{\mathbf{x}=f^{-1}(\mathbf{x}'),\mathbf{y}=g^{-1}(\mathbf{y}')}, \tag{2.42c}$$

where $\det(\mathbf{J}_f(\mathbf{x}))$ and $\det(\mathbf{J}_g(\mathbf{y}))$ denote the determinant of the Jacobian matrices of f and g. Note that the previous three equations are only valid when f and g are homeomorphisms [182]. Then, the MI of $\mathbf{x}'$ and $\mathbf{y}'$ is given by:

$$I(\mathbf{x}';\mathbf{y}') = \iint_{-\infty}^{\infty} p_{\mathbf{x}'\mathbf{y}'}(\mathbf{x}',\mathbf{y}') \log\left(\frac{p_{\mathbf{x}'\mathbf{y}'}(\mathbf{x}',\mathbf{y}')}{p_{\mathbf{x}'}(\mathbf{x}')p_{\mathbf{y}'}(\mathbf{y}')}\right) \mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}' = \tag{2.43a}$$

$$\iint_{-\infty}^{\infty} \frac{p_{\mathbf{x}\mathbf{y}}(f^{-1}(\mathbf{x}'),g^{-1}(\mathbf{y}))}{|\det(\mathbf{J}_f(f^{-1}(\mathbf{x}')))||\det(\mathbf{J}_g(g^{-1}(\mathbf{y}')))|} \log\left(\frac{p_{\mathbf{x}\mathbf{y}}(f^{-1}(\mathbf{x}'),g^{-1}(\mathbf{y}'))}{p_{\mathbf{x}}(f^{-1}(\mathbf{x}'))p_{\mathbf{y}}(g^{-1}(\mathbf{y}'))}\right) \mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}'. \tag{2.43b}$$

We introduce the change of variables $\mathbf{u} = f^{-1}(\mathbf{x}')$ and $\mathbf{v} = g^{-1}(\mathbf{y}')$ into the last expression, yielding:

$$I(\mathbf{x}';\mathbf{y}') = \iint_{-\infty}^{\infty} \frac{p_{\mathbf{x}\mathbf{y}}(\mathbf{u},\mathbf{v})}{|\det(\mathbf{J}_f(\mathbf{u}))||\det(\mathbf{J}_g(\mathbf{v}))|} \log\left(\frac{p_{\mathbf{x}\mathbf{y}}(\mathbf{u},\mathbf{v})}{p_{\mathbf{x}}(\mathbf{u})p_{\mathbf{y}}(\mathbf{v})}\right) \mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}' = \tag{2.44a}$$

$$\iint_{-\infty}^{\infty} p_{\mathbf{x}\mathbf{y}}(\mathbf{u},\mathbf{v}) \log\left(\frac{p_{\mathbf{x}\mathbf{y}}(\mathbf{u},\mathbf{v})}{p_{\mathbf{x}}(\mathbf{u})p_{\mathbf{y}}(\mathbf{v})}\right) \frac{\mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}'}{|\det(\mathbf{J}_f(\mathbf{u}))||\det(\mathbf{J}_g(\mathbf{v}))|} = \tag{2.44b}$$

$$\iint_{-\infty}^{\infty} p_{\mathbf{x}\mathbf{y}}(\mathbf{u},\mathbf{v}) \log\left(\frac{p_{\mathbf{x}\mathbf{y}}(\mathbf{u},\mathbf{v})}{p_{\mathbf{x}}(\mathbf{u})p_{\mathbf{y}}(\mathbf{v})}\right) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} = I(\mathbf{x};\mathbf{y}), \tag{2.44c}$$

where the last two steps are justified from the following fact:

$$\mathbf{u} = f^{-1}(\mathbf{x}'), \mathbf{v} = g^{-1}(\mathbf{y}') \implies \mathbf{x}' = f(\mathbf{u}), \mathbf{y}' = g(\mathbf{v}) \implies \tag{2.45a}$$

$$\mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}' = |\det(\mathbf{J}_f(\mathbf{u}))||\det(\mathbf{J}_g(\mathbf{v}))|\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \implies \tag{2.45b}$$

$$\frac{\mathrm{d}\mathbf{x}'\mathrm{d}\mathbf{y}'}{|\det(\mathbf{J}_f(\mathbf{u}))||\det(\mathbf{J}_g(\mathbf{v}))|} = \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}. \tag{2.45c}$$

■

[1] An homeomorphism is a bijective function such that said function and its inverse are both continuous and smooth [46, Definition 6.1.11].

The previous lemma allows us to find such a transformation of the measurements that the overall computation (and estimation) of the MI is much easier. Indeed, the previous observation is exploited in Chapter 6, where the MI is the considered information theoretic descriptor for the task of quantifying the amount of diversity between two datasets.

2.2 Preliminaries on Differential Geometry: the Grassmann manifold

This section introduces the Grassmann manifold as a mean to introduce the tools that exploit the sparsity in the subspace sense from the model in (2.1). Our aim is to provide some insights to facilitate the understanding of the practical use of the Grassmann manifold in geodesically convex optimization. For an extensive treatment on this manifold, we refer to [3], [20], [57].

The Grassmann manifold appears naturally in two kinds of optimization problems. On the one hand, this manifold appears when there is an invariance with respect to a full rank matrix, which we refer to as the *homogeneity* condition of the Grassmann manifold [57]. In order to illustrate this idea, let an arbitrary function be $f : \mathbb{R}^{N\times D} \to \mathbb{R}$ with $D < N$. Then, the aforementioned invariance is depicted as:

$$f(\mathbf{X}) = f(\mathbf{XM}), \tag{2.46}$$

where $\mathbf{M}$ is any $D \times D$ full rank matrix. Any function, $f(\mathbf{X})$, that is expressed in terms of a projection matrix of $\mathbf{X}$, i.e. $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, satisfies the homogeneity condition. The latter statement is true since, for any $\mathbf{X}_* = \mathbf{XM}$, we get that:

$$\mathbf{X}_*(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T = \mathbf{XM}(\mathbf{M}^T\mathbf{X}^T\mathbf{XM})^{-1}\mathbf{M}^T\mathbf{X}^T = \tag{2.47a}$$

$$\mathbf{XMM}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{M}^T)^{-1}\mathbf{M}^T\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P}_X. \tag{2.47b}$$

In addition to the previous kind of functions, the Grassmann manifold also appears in problems with orthogonality constraints, e.g. $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, in which we are interested only in the subspace that $\mathbf{X}$ generates. In fact, one could also add the orthogonality constraint to a problem whose cost function satisfies the homogeneity assumption. The added constraint is often advantageous since the constraints $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ define a compact set, which is often a necessary assumption for an iterative algorithm to be convergent.

An example of the previous rationale is given by the toy inverse problem depicted by the second case in (2.1), where $D < N$. A cost function that could be used in the solution of the aforementioned toy problem is:

$$f(\mathbf{H}) = \min_{\mathbf{x}} ||\mathbf{y} - \mathbf{Hx}||_2^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_H)\mathbf{y}, \tag{2.48}$$

where $\mathbf{P}_H = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$. Clearly, the previous cost function satisfies the homogeneity condition since it is a function of a projection matrix. Thus, restricting to an orthogonal $\mathbf{H}$ does not change $f(\mathbf{H})$. Notice that (2.48) is the cost function that is often considered in subspace learning problems [16]. Motivated by similar cost functions to the one in (2.48), in the following subsections we review the concepts behind the geometry of the Grassmann manifold.

2.2.1 Geometry of the Grassmann manifold

The Grassmann manifold, also known as the Grassmannian, is depicted as the set of D dimensional subspaces in $\mathbb{R}^N$, which is a set that has rich geometrical properties that can be exploited to derive computationally fast and memory efficient optimization algorithms. More formally, the Grassmann manifold is defined as follows [20]:

Definition 2.7 (The Grassmann manifold)**.** *Given an ambient space, $\mathbb{R}^N$, let the Grassmann manifold be:*

$$\mathrm{Gr}(N, D) = \{[\mathbf{X}] \subset \mathbb{R}^N : [\mathbf{X}] \text{ is a subspace}, dim([\mathbf{X}]) = D\}, \tag{2.49}$$

where $[\mathbf{X}]$ is any element of the Grassmannian and $D < N$.

In Figure 2.2, we show a representation of toy Grassmann manifold, $\mathrm{Gr}(3,1)$. Note that $\mathrm{Gr}(3,1)$ is illustrated by a semi-sphere, given that antipodal points depict the same subspace.

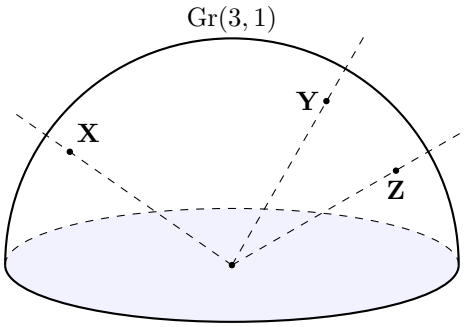


Figure 2.2: Example: Visualization of $\mathrm{Gr}(3,1)$

In general, there is not a unique way to determine a signal subspace. For this reason, there are several alternatives based on the so-called *equivalence classes*. Informally speaking, representing a point with an equivalence class is equivalent to assign an entire set with some notion of congruence to a particular element of the aforementioned set, e.g. an orthonormal matrix or a projection matrix. Particularizing on the Grassmannian, the fact that any subspace can be represented by any linear combination of a given basis is an example of congruence. The set of equivalence classes is referred to as a *quotient space*, which is a concept that is often used to describe the Grassmann manifold. With the above idea in mind, the different approaches to represent points of the Grassmannian are the following ones [20]:

- The *basis perspective* identifies a point $[\mathbf{X}] \in \mathrm{Gr}(N,D)$ with any set of D vectors such that they span the same subspace as $[\mathbf{X}]$. In other words, each point is represented by the equivalence class of all rank D matrices whose columns spans $[\mathbf{X}]$. Mathematically, a point in the Grassmann manifold using the basis perspective is (non-uniquely) represented by any matrix that belongs to the following set:
$$\mathrm{St}_{nc}(N,D) = \{\mathbf{X} \in \mathbb{R}^{N\times D} : \mathrm{rank}(\mathbf{X}) = D\}, \tag{2.50}$$
where $\mathrm{St}_{nc}(N,D)$ denotes the non-compact Stiefel manifold [3]. Since any representative $\mathbf{X} \in \mathrm{St}_{nc}(N,D)$ is non unique, the equivalence class that represents $[\mathbf{X}]$ is:
$$[\mathbf{X}] = \{\mathbf{XM} : \mathbf{X} \in \mathrm{St}_{nc}(N,D), \forall \mathbf{M} \in \mathrm{GL}(D)\}, \tag{2.51}$$
where $\mathrm{GL}(D)$ is the set of the $D \times D$ non-singular matrices, also termed the *General Linear* group. In this manner, the Grassmann manifold is represented by the quotient space $\mathrm{St}_{nc}(N,D)/\mathrm{GL}(D)$. Although the storage and manipulation of $N \times D$ matrices is computationally efficient, the non-compactness of $\mathrm{St}_{nc}(N,D)$ is undesired for iterative algorithms. This perspective is reviewed in [3].
- Analogously to the basis perspective, the *orthonormal basis (ONB) perspective* uses any orthonormal basis to represent $[\mathbf{X}] \in \mathrm{Gr}(N,D)$. As a consequence, any matrix that belongs to the Stiefel manifold [57] is used as a representative in this approach:
$$\mathrm{St}(N,D) = \{\mathbf{X} \in \mathbb{R}^{N\times D} : \mathbf{X}^T\mathbf{X} = \mathbf{I}\}, \tag{2.52}$$
where the previous constraint is referred to as an orthogonality constraint. In a similar way, any rotation of those D orthonormal vectors, i.e. the columns of $\mathbf{X}$, spans the same subspace. Hence, there is an infinite number of elements in $\mathrm{St}(N,D)$ to depict $[\mathbf{X}]$ and, for this reason, each subspace can be represented by the following equivalence class:
$$[\mathbf{X}] = \{\mathbf{XR} : \mathbf{X} \in \mathrm{St}(N,D), \forall \mathbf{R} \in \mathrm{O}(D)\}, \tag{2.53}$$
where $\mathrm{O}(D)$ is the set of $D \times D$ orthonormal matrices, i.e. $\mathbf{R}$ satisfies $\mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathbf{I}_D$, also known as the *Orthogonal* group. The equivalence class depicted in (2.53) defines the quotient

space $\mathrm{St}(N,D)/\mathrm{O}(D)$. Not only this approach is memory efficient due to the manipulation of $N \times D$ matrices, but it also has the advantage of using representatives that belong to a compact set. The compactness can be verified observing the constraint in (2.52), where the equality ensures the compactness of the representatives. For more insights on compact sets, we refer to Definition 3.4. This perspective is extensively studied in [57].

- In contrast to the previous two approaches, in the *projection perspective* each point in $\mathrm{Gr}(N,D)$ is uniquely identified by a projection matrix. In consequence, the Grassmann manifold can be represented by the following set:

$$\mathrm{Gr}(N,D) = \{\mathbf{P} \in \mathbb{R}^{N\times N} : \mathbf{P} = \mathbf{P}^T, \mathbf{P}\mathbf{P} = \mathbf{P}, \mathrm{tr}(\mathbf{P}) = D\}. \tag{2.54}$$

  Note that this approach is relatable to the previous ones. Let $[\mathbf{X}]$ be an arbritrary point in $\mathrm{Gr}(N,D)$ identified by $\mathbf{P}_X$ (a projection matrix), $\mathbf{X}_{nc} \in \mathrm{St}_{nc}(N,D)$ and $\mathbf{X} \in \mathrm{St}(N,D)$. Then, all those matrices can be related as follows:

$$\mathbf{P}_X = \mathbf{X}_{nc}(\mathbf{X}_{nc}^T\mathbf{X}_{nc})^{-1}\mathbf{X}_{nc}^T = \mathbf{X}\mathbf{X}^T, \tag{2.55}$$

  where it is remarked that the second and third expressions in (2.55) are equivalent for any representative that belongs to the equivalence classes (2.51) and (2.53), respectively.

  Even though this approach has the advantage of identifying each point in the Grassmannian with a unique element, it is not a computationally efficient approach since it requires to do computations with $N \times N$ matrices, a problem that is aggravated in high-dimensional datasets. This perspective is surveyed in [18].

- Finally, in addition to the previous alternatives, the *Lie group perspective* uses equivalence classes based on the Orthogonal group to represent a subspace. This approach is very useful to obtain the geometrical properties of the Grassmann manifold, as expanded in [57]. However, one must delve in the intricacies of Lie groups theory for the proper utilization of this perspective in practical scenarios, which is not as thoroughly studied as the previous alternatives. For a practical application of this approach, see [70].

As already stated, all these approaches are closely related and, as a result, deriving the Grassmann Riemannian geometry with each perspective relies on Lie group theory to some extend [20]. With the advantages and disadvantages of the previous approaches in mind, the ONB perspective is one of the better fits to represent subspaces for optimization purposes, which have had a long trajectory among optimization practitioners. Henceforth, and with a slight abuse of notation, our proposed perspective to describe points in the Grassmann manifold is the defined in the following way:

Definition 2.8 (Representative of a subspace (ONB perspective))**.** *We identify with* $\mathbf{X} \in \mathrm{St}(N,D)$*, or any rotation* $\mathbf{X}_r = \mathbf{X}\mathbf{R}$*, the entire equivalent class defined by:*

$$[\mathbf{X}] = \{\mathbf{X}\mathbf{R} : \mathbf{X} \in \mathrm{St}(N,D), \forall \mathbf{R} \in \mathrm{O}(D)\}, \tag{2.56}$$

which represents the quotient space $\mathrm{St}(N,D)/\mathrm{O}(D)$*.*

After defining how we represent each point in the Grassmannian, we next describe its Riemannian geometry. The first step is the definition of the *tangent space*. The tangent space at $\mathbf{X}$ in the Grassmann manifold can be defined informally as the set of possible directions (more specifically, a vector field) in which any line passes tangentially through $\mathbf{X}$. This concept is important in optimization since the Riemannian gradient (and any descent direction) of a function is a vector in the tangent space. From now on, we refer to a single element of the tangent space as a *tangent direction*.

The derivation of the Grassmann manifold tangent space relies on the exploitation of its quotient structure. Considering that the Grassmann manifold can be represented as $\mathrm{Gr}(N,D) \cong \mathrm{St}(N,D)/\mathrm{O}(D)$, it is intuitive to think that both tangent spaces of $\mathrm{Gr}(N,D)$ and $\mathrm{St}(N,D)$ have some sort of relationship. In fact, the tangent space of $\mathrm{Gr}(N,D)$ coincides with the *horizontal space* of $\mathrm{St}(N,D)$ [57], which is obtained as the orthogonal complement of the *vertical space*. Intuitively speaking, the vertical space is the set of tangent vectors in $\mathrm{St}(N,D)$ at $\mathbf{X}$ whose movements along those directions stay in the same equivalence class, i.e. $[\mathbf{X}]$ as in (2.53). Thus, the orthogonal complement of the vertical space are those directions that point to different equivalence classes in the Grassmannian.

With the previous informal definition in mind, consider the general expression of a tangent direction in $\mathrm{St}(N, D)$ at $\mathbf{X}$ [57, Eq. (2.5)]:

$$\boldsymbol{\xi} = \mathbf{X}\mathbf{A} + (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{B}, \tag{2.57}$$

where $\mathbf{A}$ is a $D \times D$ skew-symmetric matrix and $\mathbf{B}$ is an arbitrary $N \times D$ matrix. The vertical and horizontal spaces corresponds to the first and second terms of (2.57), respectively. Then, the tangent space in the Grassmann manifold is defined as follows.

Definition 2.9 (Tangent space of the Grassmann manifold)**.** *Let any $\mathbf{X} \in \mathrm{Gr}(N, D)$, then the tangent space is defined by the following set:*

$$\mathcal{T}_{\mathbf{X}} \mathrm{Gr}(N, D) = \{\boldsymbol{\Delta} \in \mathbb{R}^{N \times D} : \mathbf{X}^T \boldsymbol{\Delta} = \mathbf{0}_{D \times D}\}, \tag{2.58}$$

whose general form is given by:

$$\boldsymbol{\Delta} = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{D}, \tag{2.59}$$

where $\mathbf{D}$ is any arbitrary $N \times D$ matrix. The previous expression is interpreted as the projection of $\mathbf{D}$ into the tangent space at $\mathbf{X}$. Notice that this projection is invariant to any rotation of $\mathbf{X}$.

Remark 2.8. Notice that (2.59) correspond to the second term in (2.57).

In Figure 2.3, we illustrate an insightful example of the tangent space at the point $\mathbf{X} = [\cos(\frac{\pi}{4}), \sin(\frac{\pi}{4})]$ in $\mathrm{Gr}(2, 1)$. The tangent space is depicted by the points that belong to the highlighted tangent line to the semi-circle, which is given by the following expression:

$$\mathcal{T}_{\mathbf{X}} \mathrm{Gr}(2, 1) = \left\{ x, y \in \mathbb{R} : y = -\frac{\cos(\frac{\pi}{4})}{\sin(\frac{\pi}{4})} x \right\}. \tag{2.60}$$

Note that, in Figure 2.3, we placed the tangent space with an affine translation in such a way that it is clear that it depicts the tangent directions at $\mathbf{X}$, although it should be placed in the origin so it is a vector space.

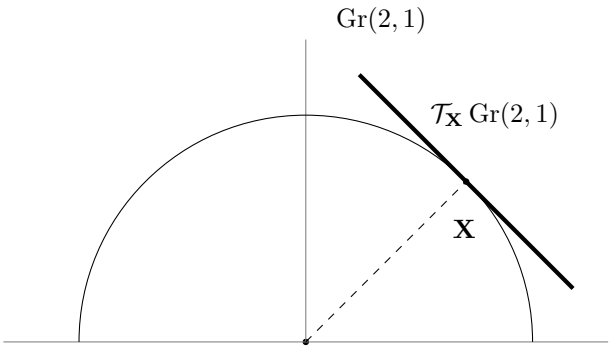


Figure 2.3: Example: Visualization of $\mathcal{T}_{\mathbf{X}} \mathrm{Gr}(2, 1)$.

The last step to fully describe the geometry of the Grassmann manifold is the definition of its canonical metric, also known as Riemannian metric [20]. From the canonical metric, one can define distances, geodesics and even relate tangent spaces from different points by means of the parallel translation. This metric is defined in the following way.

Definition 2.10 (Canonical metric of the Grassmann manifold)**.** *Let $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathbf{X}} \mathrm{Gr}(N, D)$ be two arbitrary directions at $\mathbf{X}$. Then, the canonical metric of the Grassmann manifold is defined as:*

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{X}} = \mathrm{tr}(\mathbf{A}^T \mathbf{B}). \tag{2.61}$$

Note that the dependence of the canonical metric at $\mathbf{X}$ appears in a tricky way, i.e. $\mathbf{X}$ does not appear in the previous expression, which can be made more explicit using (2.59) as follows:

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{X}} = \mathrm{tr}(\mathbf{A}^T (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{B}), \tag{2.62}$$

where now $\mathbf{A}$ and $\mathbf{B}$ can be any arbitrary matrices in $\mathbb{R}^{N \times D}$.

Now that we have defined the basic geometric concepts, in the following sections we focus on the principal angles, geodesics and distances in the Grassmannian. Those concepts are crucial in the generalization of convex optimization for variables constrained in the Grassmann manifold and thus they are fundamental in our algorithmic framework.

2.2.2 Principal Angles Between Subspaces

The use of Stiefel representatives to represent points on the Grassmannian suggests that the use of classical matrix norms, such as the Frobenius norm of two matrices, is not useful for $\mathrm{Gr}(N, D)$. This is due to the fact that the norm of a Stiefel representative is not a relevant figure to distinguish it from another representative. In fact, taking as an example the simpler case $\mathrm{Gr}(2, 1)$, using the angles between lines yields an intuitive way to define distances. Figure 2.4 depicts this intuition. Following this insight, the Principal Angles Between Subspaces (PABS) aim at generalizing the concept of angles between lines in the semi-circle to the general case, $\mathrm{Gr}(N, D)$.

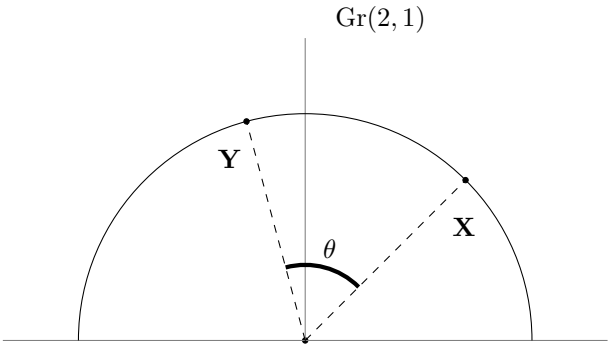


Figure 2.4: Example: Visualization of the principal angle θ between two subspaces in $\mathrm{Gr}(2, 1)$. Note that the principal angles are a generalization of this particular case.

Intuitively, the principal angles are the minimal angles between all possible basis of two subspaces, yielding the following definition [69], [207].

Definition 2.11 (PABS)**.** *The principal angles between any two points* $\mathbf{X}, \mathbf{Y} \in \mathrm{Gr}(N, D)$ *are defined recursively for* $d = 1, ..., D$ *as follows:*

$$\cos(\theta_d) = \max_{\mathbf{x}_d \in \mathbf{X}, \mathbf{y}_d \in \mathbf{Y}} \mathbf{x}_d^T \mathbf{y}_d \quad s.t. \; ||\mathbf{x}_d||_2 = ||\mathbf{y}_d||_2 = 1, \; \mathbf{x}_d^T \mathbf{x}_k = 0, \; \mathbf{y}_d^T \mathbf{y}_k = 0 \; \forall k < d, \tag{2.63}$$

where $\theta_d \in [0, \frac{\pi}{2}]$.

Although (2.63) is derived from the intuitive definition of the PABS, there is a more practical way to compute them using the SVD. After noting that (2.63) is an alternative way of defining the singular values of $\mathbf{X}^T\mathbf{Y}$ [104, Theorem 1.1], the PABS can be obtained from the following Singular Value Decomposition (SVD):

$$\mathbf{X}^T\mathbf{Y} = \mathbf{U} \cos(\mathbf{\Theta}) \mathbf{V}^T, \tag{2.64}$$

where $\cos(\cdot)$ is applied element-wise on its input main diagonal and $\mathbf{\Theta}$ is a diagonal matrix containing the D principal angles between $\mathbf{X}$ and $\mathbf{Y}$. Provided that the SVD orders the singular values (real and non-negative by definition) in a descending order, the diagonal entries of $\mathbf{\Theta}$, which are bounded in $[0, \frac{\pi}{2}]$, are ordered in an ascending way.

For the purpose of manipulating and finding clean expressions in terms of the PABS, it is often advisable to operate with *aligned representatives*. This kind of representatives are described in the following lemma [7, Proposition 1].

Lemma 2.4 (Principal alignment)**.** *For any two points* $\mathbf{X}, \mathbf{Y} \in \mathrm{Gr}(N, D)$*, one can find two aligned representatives* $\mathbf{X}_a$ *and* $\mathbf{Y}_a$ *such that* $\mathbf{X}_a^T \mathbf{Y}_a = \cos(\mathbf{\Theta})$.

Proof. Considering that $\mathbf{X}^T\mathbf{Y} = \mathbf{U} \cos(\mathbf{\Theta}) \mathbf{V}^T$, we can rotate the original representative using the singular vectors of the previous SVD to obtain the aligned representatives, i.e. $\mathbf{X}_a = \mathbf{X}\mathbf{U}$ and $\mathbf{Y}_a = \mathbf{Y}\mathbf{V}$.

Then:

$$\mathbf{X}_a^T\mathbf{Y}_a = \mathbf{U}^T\mathbf{X}^T\mathbf{Y}\mathbf{V} = \mathbf{U}^T\mathbf{U}\cos(\mathbf{\Theta})\mathbf{V}^T\mathbf{V} = \cos(\mathbf{\Theta}). \tag{2.65}$$

■

From now on, $\mathbf{X}_a$ denotes the aligned representative of $\mathbf{X}$. Now that we have shown how to compute the principal angles between $\mathbf{X}$ and $\mathbf{Y}$, it is proven useful in future chapters to relate the previous principal angles to the ones between $\mathbf{X}$ and $\mathbf{Y}_\perp$, where $[\mathbf{Y}, \mathbf{Y}_\perp] \in \mathrm{O}(N)$. In other words, we are interested in finding the angles between $\mathbf{X}$ and the orthogonal complement of $\mathbf{Y}$. In the following lemma we complement the principal alignment idea from Lemma 2.4 for the principal angles between $\mathbf{X}$ and $\mathbf{Y}_\perp$ using ideas presented in [76].

Lemma 2.5 (Complementarity of the principal angles)**.** *Let $\mathbf{\Theta}$ be the principal angles between $\mathbf{X}$ and $\mathbf{Y}$ and let $[\mathbf{Y}, \mathbf{Y}_\perp]$ be a $N \times N$ orthonormal matrix. Without loss of generality, assume that $D > N - D$ and let $\mathbf{Y}_{\perp'} = [\mathbf{0}_{N,2D-N}, \mathbf{Y}_\perp]$ (i.e. a zero-padded $\mathbf{Y}_\perp$). Then, the principal angles between $\mathbf{X}$ and $\mathbf{Y}_{\perp'}$ are contained in the following matrix:*

$$\mathbf{X}^T\mathbf{Y}_{\perp'} = \mathbf{U}\sin(\mathbf{\Theta})\mathbf{W}^T. \tag{2.66}$$

Remark 2.9. $\mathbf{Y}_{\perp'}$ is a $N \times D$ matrix.

Remark 2.10. The left singular vectors in (2.66) are equal to the left singular vectors in $\mathbf{X}^T\mathbf{Y}$ and, therefore, the aligned representative $\mathbf{X}_a$ is also aligned with $\mathbf{Y}_{\perp'}$.

Remark 2.11. Since $D > N - D$, there are $\lfloor D - (N - D) \rfloor$ non-trivial principal angles, i.e. $\theta_d \neq 0$ or $\theta_d \neq \frac{\pi}{2}$ for $d = 1, ..., D$. Hence, the zero entries in the diagonal of $\sin(\mathbf{\Theta})$ coincide with the zero-padded columns of $\mathbf{Y}_{\perp'}$.

Proof. Consider the following equation:

$$\mathbf{I}_D = \mathbf{X}^T\mathbf{X} = \mathbf{X}^T(\mathbf{Y}\mathbf{Y}^T + \mathbf{Y}_{\perp'}\mathbf{Y}_{\perp'}^T)\mathbf{X} = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X} + \mathbf{X}\mathbf{Y}_\perp\mathbf{Y}_\perp^T\mathbf{X}, \tag{2.67}$$

which can be rewritten using the fact that $\mathbf{X}^T\mathbf{Y} = \mathbf{U}\cos(\mathbf{\Theta})\mathbf{V}^T$ as:

$$\mathbf{U}\cos^2(\mathbf{\Theta})\mathbf{U}^T + \mathbf{X}^T\mathbf{Y}_\perp\mathbf{Y}_\perp^T\mathbf{X} = \mathbf{I}_D. \tag{2.68}$$

Now, let the SVD of $\mathbf{X}^T\mathbf{Y}_{\perp'}$ be $\mathbf{U}_*\mathbf{\Lambda}\mathbf{W}^T$. Then, (2.68) is further rewritten as:

$$\mathbf{U}\cos^2(\mathbf{\Theta})\mathbf{U}^T + \mathbf{U}_*\mathbf{\Lambda}^2\mathbf{U}_*^T = \mathbf{I}_D, \tag{2.69}$$

from where the only possible values of $\mathbf{U}_*$ and $\mathbf{\Lambda}$ are:

$$\mathbf{U}_* = \mathbf{U}, \tag{2.70a}$$

$$\mathbf{\Lambda} = \sin(\mathbf{\Theta}). \tag{2.70b}$$

Similarly to $\cos(\cdot)$, $\sin(\cdot)$ is applied element-wise to its input main diagonal. ■

The motivation behind the left zero padding in $\mathbf{Y}_{\perp'}$ and the constraint $D > N - D$ is to ensure that the matrix containing the principal angles, $\mathbf{\Theta}$, is a $D \times D$ matrix. Otherwise, we would need to express the SVD of $\mathbf{X}^T\mathbf{Y}$ and $\mathbf{X}^T\mathbf{Y}_\perp$ with matrices having different dimensions. As it is shown in Subsection 3.2.2.1 from Chapter 3, this particular case is more than enough to retrieve insightful results.

2.2.3 Geodesics and distances in the Grassmann manifold

A key issue in this thesis framework is the concept of *geodesics*. Intuitively, the geodesics are defined as the path with shortest length between any two points in the Grassmannian. Note that all the points along this path are also points in the Grassmann manifold. This intuitive definition leads to the variational problem that defines the geodesic path connecting $\mathbf{X}$ with $\mathbf{Y}$ [57]:

$$\mathbf{\Gamma}(t) = \arg\min_{\mathbf{\Gamma}(t)} \int_0^1 \left\langle \frac{\mathrm{d}\mathbf{\Gamma}(t)}{\mathrm{d}t}, \frac{\mathrm{d}\mathbf{\Gamma}(t)}{\mathrm{d}t} \right\rangle_{\mathbf{\Gamma}(t)}^{-\frac{1}{2}} \mathrm{d}t \quad \text{s.t. } \mathbf{\Gamma}(0) = \mathbf{X}, \mathbf{\Gamma}(1) = \mathbf{Y}, \tag{2.71}$$

where this integral accounts for the distance (also referred to as the *arclength*) between $\mathbf{X}$ and $\mathbf{Y}$. Given that (2.71) is quite tedious to compute, an alternative derivation of Grassmann geodesics is done from the geodesic equation (a differential equation that describes a geodesic path) in the Stiefel manifold [57, Eq. (2.7)]. On account of the fact that the scope of this section is to learn how to use the Grassmann manifold in an optimization framework, we do not detail the derivation of Grassmann geodesics. Instead, we limit ourselves to study the geometric properties of the Grassmann geodesics, which are defined as follows.

Definition 2.12 (Grassmann geodesics)**.** *Let the tangent vector $\mathbf{T} \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$ pointing to $\mathbf{Y}$ and its compact SVD be $\mathbf{T} = \mathbf{U}_y \mathbf{\Theta} \mathbf{V}_y^T$, where $\mathbf{\Theta} \prec \frac{\pi}{2}\mathbf{I}$. Then, the expression of the geodesic that connects $\mathbf{X}$ with $\mathbf{Y}$ is:*

$$\mathbf{\Gamma}(t) = \mathbf{X}\mathbf{V}_y \cos(\mathbf{\Theta} t)\mathbf{V}_y^T + \mathbf{U}_y \sin(\mathbf{\Theta} t)\mathbf{V}_y^T, \tag{2.72}$$

and the expression of the aligned geodesic, i.e. computing (2.72) using aligned representatives, is:

$$\mathbf{\Gamma}_a(t) = \mathbf{X}_a \cos(\mathbf{\Theta} t) + \mathbf{\Delta}_a \sin(\mathbf{\Theta} t), \tag{2.73}$$

where $\mathbf{\Delta}_a \in \mathbb{R}^{N \times D}$ is an orthogonal matrix such that $\mathbf{\Gamma}_a(1) = \mathbf{Y}_a$.

Remark 2.12. $\mathbf{\Theta}$ is the matrix containing the principal angles between $\mathbf{X}$ and $\mathbf{Y}$.

Remark 2.13. The rationale behind the constraint $\mathbf{\Theta} \prec \frac{\pi}{2}\mathbf{I}$ is to ensure the uniqueness of the geodesic expression. Indeed, when the principal angles between $\mathbf{X}$ and $\mathbf{Y}$ are all smaller than $\frac{\pi}{2}$, it is known that the geodesic that joins them is unique [199]. Nonetheless, when any of the principal angles equals $\frac{\pi}{2}$, $\mathbf{X}$ and $\mathbf{Y}$ are *conjugate points*, meaning that there are multiple geodesics joining them [57]. Actually, there is an additional non-unique geodesic for every principal angle that is equal to $\frac{\pi}{2}$.

Remark 2.14. $\mathbf{\Gamma}(0) = \mathbf{X}$, $\mathbf{\Gamma}(1) = \mathbf{Y}$ and $\frac{\mathrm{d}\mathbf{\Gamma}(t)}{\mathrm{d}t}\big|_{t=0} = \mathbf{T}$. We verify the latter expression to provide more insights on the manipulation of Grassmann expressions:

$$\left.\frac{\mathrm{d}\mathbf{\Gamma}(t)}{\mathrm{d}t}\right|_{t=0} = \left[-\mathbf{X}\mathbf{V}_y\mathbf{\Theta}\sin(\mathbf{\Theta} t)\mathbf{V}_y^T + \mathbf{U}_y\mathbf{\Theta}\cos(\mathbf{\Theta} t)\mathbf{V}_y^T\right]\big|_{t=0} = \mathbf{U}_y\mathbf{\Theta}\mathbf{V}_y^T = \mathbf{T}. \tag{2.74}$$

Remark 2.15. The intuition behind (2.73) can be grasped from the following expression:

$$\mathbf{X}_a^T\mathbf{Y}_a = \mathbf{X}_a^T\mathbf{\Gamma}_a(1) = \cos(\mathbf{\Theta}) + \mathbf{X}_a^T\mathbf{\Delta}_a\sin(\mathbf{\Theta}), \tag{2.75}$$

from where it can be deduced that $\mathbf{\Delta}_a$ must comply with the following constraint so that Lemma 2.4 holds:

$$\mathbf{X}_a^T\mathbf{\Delta}_a = \mathbf{0}_{D\times D}. \tag{2.76}$$

In other words, $\mathbf{\Delta}_a$ belongs to the tangent space at $\mathbf{X}$, $\mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$.

Considering the expression of the (unique) geodesic, the *canonical distance* in the Grassmann manifold can be obtained using the integral expression in (2.71). Luckily, there is an equivalent (and more convenient) integral to evaluate the arclength between two points:

$$d_{arc}(\mathbf{X}, \mathbf{Y}) = \int_0^1 \langle \mathbf{T}_x, \mathbf{T}_x \rangle_{\mathbf{X}}^{-\frac{1}{2}} \mathrm{d}t = \int_0^1 \langle \mathbf{T}_y, \mathbf{T}_y \rangle_{\mathbf{Y}}^{-\frac{1}{2}} \mathrm{d}t = ||\mathbf{\Theta}||_F, \tag{2.77}$$

where $\mathbf{T}_x \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$ and $\mathbf{T}_y \in \mathcal{T}_{\mathbf{Y}} \operatorname{Gr}(N, D)$ are the tangent vectors at $\mathbf{X}$ and $\mathbf{Y}$ that point to $\mathbf{Y}$ and $\mathbf{X}$, respectively, defined in a similar way to $\mathbf{T}$ in Definition 2.12. Note that the singular values of $\mathbf{T}_x$ and $\mathbf{T}_y$ are both equal to $\mathbf{\Theta}$ and hereby the arclength between two subspaces is symmetric.

Up to this point, it has not been yet discussed how to compute tangent vectors between two subspaces. For this purpose, it is useful to define the *exponential* and *logarithm* maps. Both of these mappings represent functions that relate directions and points in the Grassmannian and are based on Definition 2.12 [57], [207].

Definition 2.13 (Grassmann exponential map)**.** *For every $\mathbf{X} \in \operatorname{Gr}(N, D)$, the exponential map is a function defined as $\exp_{\mathbf{X}} : \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D) \to \operatorname{Gr}(N, D)$ such that:*

$$\exp_{\mathbf{X}}(\mathbf{T}) = \mathbf{\Gamma}(1), \tag{2.78}$$

where $\mathbf{\Gamma}(t)$ is a geodesic fulfilling $\mathbf{\Gamma}(0) = \mathbf{X}$ and $\frac{d\mathbf{\Gamma}(t)}{dt}\big|_{t=0} = \mathbf{T}$.

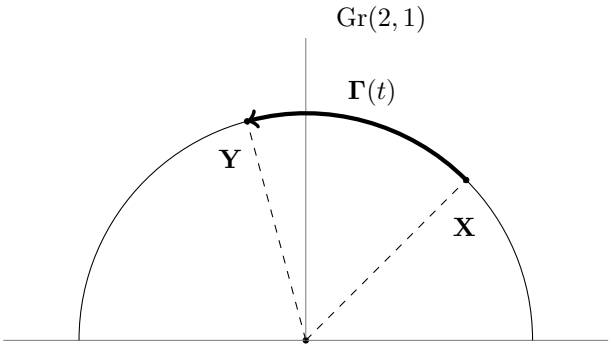


Figure 2.5: Example: Visualization of the geodesic, $\mathbf{\Gamma}(t)$, that connects $\mathbf{X}$ with $\mathbf{Y}$ in $\mathrm{Gr}(2,1)$. Note that the distance is the arclength in this semi-circle, coinciding with the only principal angle between $\mathbf{X}$ and $\mathbf{Y}$.

Definition 2.14 (Grassmann logarithm map)**.** *For every* $\mathbf{X}, \mathbf{Y} \in \mathrm{Gr}(N, D)$*, the logarithm map is a function defined as* $\log_{\mathbf{X}} : \mathrm{Gr}(N, D) \to \mathcal{T}_{\mathbf{X}} \mathrm{Gr}(N, D)$ *such that:*

$$exp_{\mathbf{X}}(\log_{\mathbf{X}}(\mathbf{Y})) = \mathbf{Y}. \tag{2.79}$$

In simpler words, the logarithm map at $\mathbf{X}$ *of* $\mathbf{Y}$ *computes the tangent direction at* $\mathbf{X}$ *that points to* $\mathbf{Y}$*.*

While the exponential map is computed using the known expression of the Grassmann geodesics, the logarithm map requires the Cosine-Sine Decomposition (CSD). As a matter of fact, the Grassmann logarithm map is defined by the following system of equations [57], [207]:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{Y} \\ (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_y \cos(\mathbf{\Theta})\mathbf{V}_y^T \\ \mathbf{U}_y \sin(\mathbf{\Theta})\mathbf{V}_y^T \end{pmatrix}, \tag{2.80}$$

where $\log_X(\mathbf{Y}) = \mathbf{T} = \mathbf{U}_y\mathbf{\Theta}\mathbf{V}_y^T$. The practical (and general) way to solve the above system of equations is by means of the generalized SVD [190]. However, we propose an alternative way of computing the logarithm map using aligned representatives. Invoking Lemma 2.4, equation (2.80) becomes:

$$\begin{pmatrix} \mathbf{X}_a^T\mathbf{Y}_a \\ (\mathbf{I}_N - \mathbf{X}_a\mathbf{X}_a^T)\mathbf{Y}_a \end{pmatrix} = \begin{pmatrix} \cos(\mathbf{\Theta}) \\ \mathbf{\Delta}_a \sin(\mathbf{\Theta}) \end{pmatrix}, \tag{2.81}$$

where the desired result of the *aligned* logarithm map is $\mathbf{T}_a = \mathbf{\Delta}_a\mathbf{\Theta}$. Notice that (2.81) only requires to compute the SVD of $\mathbf{X}^T\mathbf{Y}$ in contrast to the generalized SVD in (2.80).

2.3 Information theoretic Model-Order Selection

In many signal processing applications, there is an inherent need to determine the degrees of freedom of a given parameter. For instance, in the setting described in the previous section and in the sparse signal model in (2.1), where the signal component lies in a low-rank subspace, the estimation of intrinsic dimension D is a pragmatic procedure in a practical application [135]. The previous idea is formalized in the *model-order selection* framework [127] and it is considered in this dissertation as an auxiliary tool that is used to discover the diversity of a dataset. Particularly, we are interested in model-order selection criteria that are based on information theoretic arguments [127], [177]. The motivation behind the consideration of this kind of approaches for the model-order selection problem is twofold. On the one hand, we show and prove that the resulting criterion from the information theoretic model-order selection is based on a simple (but powerful) additive penalty term on the log-likelihood function. On the other hand, the information theoretic roots of the considered approach are clearly aligned with the general tone of this thesis.

For the purpose of introducing the information theoretic model-order selection, let us consider a random variable $\mathbf{y}$ with an associated PDF given by $p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}_D)$, where $\boldsymbol{\theta}_D \in \mathbb{R}^D$ is the vector that

contains the parameters of the model. In this setting, D is the true value of the model-order. The objective of the model-order selection problem is to detect the value of the model-order, D, using the available data. In other words, the solution of the model-order selection consists on choosing the best PDF among the set of possible models denoted as $\{p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}_L)\}_{L\in\mathcal{M}}$, where each item is characterized by the model-order L. We consider that the set of possible model-orders is given by:

$$\mathcal{M} = \{L \in \mathbb{N} : 1 \leq L \leq N\}, \tag{2.82}$$

where each item denotes the dimension of the tested parameter, $\boldsymbol{\theta}_L$. The identification of the best model-order, D, using information theoretic arguments is based on the maximization of the following cost [127], [177]:

$$\hat{D} = \arg\max_{L} \sum_{k=1}^{K} \log(p_{\mathbf{y}}(\mathbf{y}_k|\hat{\boldsymbol{\theta}}_L)) - \frac{L}{2}\eta(K), \tag{2.83}$$

where the first term is the summation of the likelihoods of the K (independent) realizations of $\mathbf{y}$ for the model-order L particularized on the ML estimation of the parameter, $\hat{\boldsymbol{\theta}}_L$, and $\eta(L, K)$ is the penalty term. In the previous equation, we differentiate the subscript L, the tested model-order, from D and $\hat{D}$, which are the true value and the final model-order decision obtained with a particular procedure, respectively. The focal point of the information theoretic model-order selection is the penalty term, whose goal is to induce a trade-off between an overfitted model (high L) and an underfitted model (low L) as a function of the data sample size. Although there are several particularizations of the penalty term [127], we consider the ones summarized in Table 2.1 for concreteness.

Criterion	**Penalty**, $\eta(K)$
Bayesian Information Criterion (BIC) [168], [171]	$\ln(K)$
Akaike Information Criterion (AIC) [8], [89]	2
Generalized Information Criterion (GIC) [177]	$\lambda + 1$, for $\lambda > 1$

Table 2.1: Known expressions of the penalty in model-order selection rules based on information theoretic criteria.

Yet, there are some implicit assumptions that one must check before the utilization of (2.83) in a practical problem. Depending on the chosen criterion from Table 2.1, there exists two conditions that the Fisher information matrix of $\mathbf{y}$ with respect to $\boldsymbol{\theta}_L$ must fulfill so that (2.83) is an adequate optimization problem for the model-order selection task. The Fisher information matrix is defined as follows.

Definition 2.15 (Fisher information matrix)**.** *Let* $\mathbf{y}$ *be a random variable with an associated PDF given by* $p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta})$*. Then, its Fisher information matrix is given by:*

$$\mathbf{F}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \log(p_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]. \tag{2.84}$$

In the following subsections, we detail the derivations of the criteria stated in Table 2.1 with the aim of specifying the reasons behind the constraints on the Fisher information. We remark that these derivations can also be found in [22], [41], [177]. Just to anticipate the conclusions of the following paragraphs, the non-singularity of the Fisher information is required for all the alternatives, while the BIC relies on a specific asymptotic convergence of this matrix.

2.3.1 Bayesian Information Criterion (BIC)

From all of the alternatives shown in Table 2.1, the BIC stands out for its interpretability. For large sample sizes, not only the BIC is known to be a consistent estimator of D, but it is also related to the Minimum Description Length (MDL) principle [171] with a well-known operational meaning, i.e. an MDL model is the one that best describes the data with the least amount of parameters [168]. In simpler terms, the BIC is an MAP estimation of the model-order. Thus, the introduction of the

Bayesian framework to the model-order selection problem means that $\hat{D}$ is obtained from the following optimization problem:

$$\hat{D} = \arg\max_{L} p_L(L|\mathbf{y}, \boldsymbol{\theta}_L) = \arg\max_{L} \int_{-\infty}^{\infty} p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L) \prod_{k=1}^{K} p_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L) \mathrm{d}\boldsymbol{\theta}_L, \tag{2.85}$$

where $p_L(L|\mathbf{y}, \boldsymbol{\theta}_L)$ denotes the MAP function of L (with respect to the measurements and the parameter), and $p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L)$ is the prior density of $\boldsymbol{\theta}_L$. In (2.85), we have already considered the K possible realizations in the form of the product of the K marginal likelihoods. The main issue with the integral in (2.85) is that it is difficult to compute in the general case. As a result, the BIC is derived from an approximation of the aforementioned integral based on the Laplace approximation [22], which is summarized and proved in the following lemma.

Lemma 2.6 (Laplace approximation of an integral)**.** *Let $f : \mathbb{R}^D \to \mathbb{R}$ be a function with a single global maximum (rapidly decaying from that point). Then, we have the following approximation:*

$$\int_{-\infty}^{\infty} \exp(f(\mathbf{x})) d\mathbf{x} \approx \sqrt{\frac{(2\pi)^D}{\det(-\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0))}} \exp(f(\mathbf{x}_0)), \tag{2.86}$$

where $\mathbf{x}_0$ is the maximizer of $f(\mathbf{x})$ and $\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)$ is the (non-singular) Hessian matrix of $f(\mathbf{x})$ at $\mathbf{x}_0$.

Proof. Given the assumptions on $f(\mathbf{x})$, it can be approximated around $\mathbf{x}_0$ by its second-order Taylor expansion:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0), \tag{2.87}$$

where the first-order term of the Taylor expansion does not appear since $\mathbf{x}_0$ is the global maximizer, i.e. the gradient vanishes at this point. Plugging the previous approximation into $\int_{-\infty}^{\infty} \exp(f(\mathbf{x}))\mathrm{d}\mathbf{x}$, we obtain:

$$\int_{-\infty}^{\infty} \exp(f(\mathbf{x}))\mathrm{d}\mathbf{x} \approx \int_{-\infty}^{\infty} \exp\left(f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\right) \mathrm{d}\mathbf{x} = \tag{2.88a}$$

$$\exp(f(\mathbf{x}_0)) \int_{-\infty}^{\infty} \exp\left(\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\right) \mathrm{d}\mathbf{x}. \tag{2.88b}$$

We want to rewrite the integrand in (2.88b) so that the whole integral is equal to 1. With the previous goal in mind, we get:

$$\exp(f(\mathbf{x}_0)) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \left(-\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)\right)(\mathbf{x} - \mathbf{x}_0)\right) \mathrm{d}\mathbf{x} = \tag{2.89a}$$

$$\exp(f(\mathbf{x}_0)) C \int_{-\infty}^{\infty} \frac{1}{C} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \left(-\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)\right)(\mathbf{x} - \mathbf{x}_0)\right) \mathrm{d}\mathbf{x}, \tag{2.89b}$$

where:

$$C = \sqrt{\frac{(2\pi)^D}{\det(-\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0))}}, \tag{2.90}$$

is a constant such that the integral in (2.89b) is equal to 1. The previous statement is true since the integrand in (2.88a) has the same structure as the PDF of a Gaussian random variable of the following form:

$$\mathbf{x}' \sim \mathcal{N}\left(\mathbf{x}_0, \left(-\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0)\right)^{-1}\right). \tag{2.91}$$

In this manner, the original integral is approximated by:

$$\int_{-\infty}^{\infty} \exp(f(\mathbf{x}))\mathrm{d}\mathbf{x} \approx C \exp(f(\mathbf{x}_0)) = \sqrt{\frac{(2\pi)^D}{\det(-\nabla_{\mathbf{x}}^2 f(\mathbf{x}_0))}} \exp(f(\mathbf{x}_0)). \tag{2.92}$$

■

For the purpose of approximating the integral in (2.85) using the previous lemma, we rewrite it as follows:

$$\hat{D} = \arg\max_{L} \int_{-\infty}^{\infty} p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L) \prod_{k=1}^{K} p_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L) \mathrm{d}\boldsymbol{\theta}_L = \tag{2.93a}$$

$$\arg\max_{D} \int_{-\infty}^{\infty} \exp\left(\log\left(p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L) \prod_{k=1}^{K} p_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L)\right)\right) \mathrm{d}\boldsymbol{\theta}_L = \arg\max_{D} \int_{-\infty}^{\infty} \exp\left(g(\mathbf{y}, \boldsymbol{\theta}_L)\right) \mathrm{d}\boldsymbol{\theta}_L, \tag{2.93b}$$

where $g(\mathbf{y}, \boldsymbol{\theta}_L) = \log(p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L) \prod_{k=1}^{K} p_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L))$. The previous expression allows us to invoke Lemma 2.6, yielding:

$$\int_{-\infty}^{\infty} \exp\left(g(\mathbf{y}, \boldsymbol{\theta}_L)\right) \approx \sqrt{\frac{(2\pi)^L}{\det(-\nabla^2_{\boldsymbol{\theta}_L} g(\mathbf{y}, \hat{\boldsymbol{\theta}}_L))}} \exp(g(\mathbf{y}, \hat{\boldsymbol{\theta}}_L)), \tag{2.94}$$

where $\hat{\boldsymbol{\theta}}_L$ is the ML estimation of $\boldsymbol{\theta}_L$ (MAP if $p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L)$ is also considered). After plugging the previous approximation into (2.85), we obtain the optimization problem in which the BIC is based:

$$\hat{D} = \arg\max_{L} \sqrt{\frac{(2\pi)^L}{\det\left(-\nabla^2_{\boldsymbol{\theta}_L} g(\mathbf{y}, \hat{\boldsymbol{\theta}}_L)\right)}} \exp(g(\mathbf{y}, \hat{\boldsymbol{\theta}}_L)), \tag{2.95}$$

which can be further parsed by taking the logarithm of the previous expression:

$$\hat{D} = \arg\max_{L} \sum_{k=1}^{K} \log(p_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L)) + \log(p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L)) + \frac{L}{2}\log(2\pi) - \frac{1}{2}\log\left(\det(-\nabla^2_{\boldsymbol{\theta}_L} g(\mathbf{y}, \hat{\boldsymbol{\theta}}_L))\right). \tag{2.96}$$

The second key aspect of the BIC is the consideration of a non-informative prior. An example of a non-informative prior is:

$$p_{\boldsymbol{\theta}_L}(\boldsymbol{\theta}_L) = \lim_{v \to \infty} \frac{1}{\sqrt{(2\pi v)^D}} \exp\left(-\frac{1}{2v}||\boldsymbol{\theta}_L||_2^2\right). \tag{2.97}$$

While the consideration of a non-informative prior is useful from a practical point of view (it does not require any knowledge of the particular problem), it has been object of criticism due to the fact that this idea is contradictory to the Bayesian framework [193]. Nevertheless, the practicality and interpretability of the BIC outweighs the aforementioned critique. Notice that any non-informative prior, e.g. (2.97), implies that the Hessian of $g(\mathbf{y}, \boldsymbol{\theta}_L)$ yields:

$$\nabla^2_{\boldsymbol{\theta}_L} g(\mathbf{y}, \hat{\boldsymbol{\theta}}_L) = \nabla^2_{\boldsymbol{\theta}_L} \log(p_{\boldsymbol{\theta}_L}(\hat{\boldsymbol{\theta}}_L) \prod_{k=1}^{K} p_{\mathbf{y}}(\mathbf{y}_k|\hat{\boldsymbol{\theta}}_L)) = \sum_{k=1}^{K} \nabla^2_{\boldsymbol{\theta}_L} \log(p_{\mathbf{y}}(\mathbf{y}_k|\hat{\boldsymbol{\theta}}_L), \tag{2.98}$$

which is an expression that becomes:

$$\sum_{k=1}^{K} \nabla^2_{\boldsymbol{\theta}_L} \log(p_{\mathbf{y}}(\mathbf{y}_k|\hat{\boldsymbol{\theta}}_L) \to -K\mathbf{F}(\hat{\boldsymbol{\theta}}_L), \tag{2.99}$$

for $K \to \infty$ due to the Law of Large Numbers (see Definition 2.15). Motivated by the previous result, we plug (2.99) into (2.96) and let $K \to \infty$. Taking the limit for $K \to \infty$ implies that we drop the terms that do not increase with K in (2.96), resulting in the BIC optimization problem.

$$\hat{D} = \arg\max_{L} \sum_{k=1}^{K} \log(p_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L)) - \frac{1}{2}\log\left(\det(K\mathbf{F}(\hat{\boldsymbol{\theta}}_L))\right). \tag{2.100}$$

Although the previous optimization problem is effective to determine the model-order, the BIC methodology further simplifies the second term of the previous expression by dropping the terms that

do not grow with K in (2.100). In this regard, note that the determinant in second term in (2.100) can be written as follows:

$$\log\Big(\det(K\mathbf{F}(\hat{\boldsymbol{\theta}}_L))\Big) = \log(K^L \det(\mathbf{F}(\hat{\boldsymbol{\theta}}_L)) = L\log(K) + \log(\det(\mathbf{F}(\hat{\boldsymbol{\theta}}_L)). \tag{2.101}$$

The second term in (2.101) can be dropped as long as:

$$\frac{1}{K}\mathbf{F}(\hat{\boldsymbol{\theta}}_L) \to \mathbf{K}, \tag{2.102}$$

for $K \to \infty$, where $\mathbf{K}$ is a constant matrix with respect to $\boldsymbol{\theta}_L$. As a result of (2.102), the optimization problem of the BIC is the following one:

$$\hat{D} = \arg\max_L \sum_{k=1}^{K} \log(f_{\mathbf{y}}(\mathbf{y}_k|\boldsymbol{\theta}_L)) - \frac{L\log(K)}{2}. \tag{2.103}$$

In summary, the conditions that the Fisher information matrix of $\mathbf{y}$ must fulfill to invoke the BIC expression given in (2.103) are the non-singularity of $\mathbf{F}(\boldsymbol{\theta}_L)$ for all $\boldsymbol{\theta}_K$ (so that Lemma 2.6 is valid) and the asymptotic convergence of $\mathbf{F}(\hat{\boldsymbol{\theta}}_L)$ described in (2.102) (to be able to drop the Fisher information matrix term). As a last remark, it is possible to drop (or to modify) any of the previous two constraints on the Fisher information matrix and still obtain an BIC-like model-order selection rule (see [135] for an example). Yet, other BIC-like estimations of the model-order are out of the scope of this dissertation.

2.3.2 Akaike Information Criterion (AIC) and Generalized Information Criterion (GIC)

An alternative approach to the Bayesian framework for the detection of the model-order is the minimization of the Kullback-Leibler (KL) divergence between the true and the tested model PDFs with respect to all possible models. This measure is defined as follows:

$$D_{KL}(p_D||p_L) = \mathrm{E}\left[\log\left(\frac{p_D(\mathbf{y})}{p_L(\mathbf{y})}\right)\right] = \tag{2.104a}$$

$$\mathrm{E}\left[\log(p_D(\mathbf{y}))\right] - \mathrm{E}\left[\log(p_L(\mathbf{y}))\right], \tag{2.104b}$$

where the expected value is taken with respect to $p_D(\mathbf{y})$. In the previous expression, $p_D(\mathbf{y})$ and $p_L(\mathbf{y})$ are the PDF of the data (evidence function in the Bayesian framework) using the true and tested models, respectively. Note that only the second term in (2.104b) is relevant for the model-order selection framework since it is the only term between the previous two KL divergences that depends on L. Also, in contrast to the BIC derivation, we do not need to implicitly consider the PDFs of the K different realizations in (2.104b) thanks to the expected value. As a consequence of the previous observation, the AIC and the GIC are based on the maximization (for a minimum KL divergence) of the following function:

$$d_{KL}(L) = \mathrm{E}[\log(p_L(\mathbf{y}|\boldsymbol{\theta}_L))]. \tag{2.105}$$

The previous expression is referred to as the relative KL information [177] or the KL discrepancy [41]. The main issue with $d_{KL}(L)$ as a cost function is that the expected value cannot be evaluated since the true distribution of the data is unknown a priori [177]. This is the reason why in the AIC and GIC literature [8], [41], [177] the authors often resort to a surrogate of the KL discrepancy of (2.105). For the purpose of obtaining a suitable surrogate function, let us define a new random variable $\mathbf{y}'$ that is identically distributed as $\mathbf{y}$ and independent from $\mathbf{y}$. Also, let K' be the number of fictitious samples of $\mathbf{y}'$ and let $\hat{\boldsymbol{\theta}}'_L$ be the ML estimation of $\boldsymbol{\theta}_L$ constructed from these samples. The goal of introducing $\mathbf{y}'$ is to obtain an approximation of the true distribution by means of a cross-validatory procedure [177]. Then, a surrogate of the KL discrepancy is:

$$\hat{d}_{KL}(L) = \mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{y}'}\left[\log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}'_L))\right]\right], \tag{2.106}$$

where:

$$\log(p_L(\mathbf{y}|\boldsymbol{\theta}_L)) \approx \mathrm{E}_{\mathbf{y}'}\left[\log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}'_L))\right], \tag{2.107}$$

is a negatively biased estimation of $\log(p_L(\mathbf{y}|\boldsymbol{\theta}_L))$ [40], [41]. Equation (2.107) is the reason why the AIC and GIC are labeled cross-validatory model-order approaches in the literature [177]. The subscript in the expected value operators in the previous two expressions specifies with respect to which random variable each expected value is being computed. Notice that the expected value operator in (2.107) eliminates the dependency with the fictitious random variable, $\mathbf{y}'$. Yet, in order to avoid the potentially difficult integrals, we approximate the argument of the expected value in (2.107) by means of its second-order Taylor expansion around $\hat{\boldsymbol{\theta}}_L$ (the ML estimation of $\boldsymbol{\theta}_L$ obtained from samples of the original random variable, $\mathbf{y}$):

$$\log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}'_L)) \approx \log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}_L)) + \frac{1}{2}(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L)^T \left(\nabla^2_{\hat{\boldsymbol{\theta}}'_L} \log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}'_L))\right)\Big|_{\hat{\boldsymbol{\theta}}'_L = \hat{\boldsymbol{\theta}}_L} (\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L) \approx \tag{2.108a}$$

$$\log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}_L)) - \frac{1}{2}(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L)^T \mathbf{F}(\hat{\boldsymbol{\theta}}_L)(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L). \tag{2.108b}$$

After plugging the previous asymptotic approximation into (2.106), we get:

$$\hat{d}_{KL}(L) \approx \mathrm{E}_{\mathbf{y}}\left[\log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}_L))\right] - \frac{1}{2}\mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{y}'}\left[(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L)^T \mathbf{F}(\hat{\boldsymbol{\theta}}_L)(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L)\right]\right], \tag{2.109}$$

where the first term gets out of the expected value with respect to $\mathbf{y}'$ since it is independent of this random variable (recall that $\hat{\boldsymbol{\theta}}_L$ is constructed using realizations from $\mathbf{y}$). The last step to obtain the AIC and GIC is to compute the double expected value from the second term in (2.109), which yields:

$$\mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{y}'}\left[(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L)^T \mathbf{F}(\hat{\boldsymbol{\theta}}_L)(\hat{\boldsymbol{\theta}}'_L - \hat{\boldsymbol{\theta}}_L)\right]\right] = \tag{2.110a}$$

$$\mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{y}'}\left[\left((\hat{\boldsymbol{\theta}}'_L - \boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta})\right)^T \mathbf{F}(\hat{\boldsymbol{\theta}}_L)\left((\hat{\boldsymbol{\theta}}'_L - \boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta})\right)\right]\right] = \tag{2.110b}$$

$$\mathrm{tr}\left(\mathbf{F}(\hat{\boldsymbol{\theta}}_L)\,\mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{y}'}\left[\left((\hat{\boldsymbol{\theta}}'_L - \boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta})\right)\left((\hat{\boldsymbol{\theta}}'_L - \boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta})\right)^T\right]\right]\right) = \tag{2.110c}$$

$$\mathrm{tr}\left(\mathbf{F}(\hat{\boldsymbol{\theta}}_L)\left(\mathrm{E}_{\mathbf{y}'}\left[(\hat{\boldsymbol{\theta}}'_L - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}'_L - \boldsymbol{\theta})^T\right] + \mathrm{E}_{\mathbf{y}}\left[(\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_L - \boldsymbol{\theta})^T\right]\right)\right) = \tag{2.110d}$$

$$\mathrm{tr}\left(\mathbf{F}(\hat{\boldsymbol{\theta}}_L)\left(\mathbf{F}^{-1}(\hat{\boldsymbol{\theta}}'_L) + \mathbf{F}^{-1}(\hat{\boldsymbol{\theta}}_L)\right)\right) = \mathrm{tr}\left(\mathbf{F}(\hat{\boldsymbol{\theta}}_L)\left(\lambda\mathbf{F}^{-1}(\hat{\boldsymbol{\theta}}_L) + \mathbf{F}^{-1}(\hat{\boldsymbol{\theta}}_L)\right)\right) = L(1+\lambda), \tag{2.110e}$$

where:

$$\lambda = \frac{K}{K'}. \tag{2.111}$$

The meaning behind λ is that, intuitively, the risk of overfitting (chossing a $\hat{D}$ larger than D) might be reduced if the validation set (samples from $\mathbf{y}$) is larger than the estimation sample (samples from $\mathbf{y}'$) [177]. The previous observation implies that λ must be greater or equal to 1. In fact, the GIC is obtained for $\lambda > 1$, whereas the AIC is the particular case resulting from $\lambda = 1$. In (2.110), equation (2.110d) is obtained using the fact that $\mathbf{y}$ and $\mathbf{y}'$ are independent (and so are their respective ML estimators of $\boldsymbol{\theta}_L$), while (2.110e) holds from the asymptotic second-order moment of the ML estimators [101, Theorem 7.3] and from the following fact [177]:

$$\mathbf{F}(\hat{\boldsymbol{\theta}}'_L) = \frac{1}{\lambda}\mathbf{F}(\hat{\boldsymbol{\theta}}_L). \tag{2.112}$$

Finally, we obtain the desired approximation of the KL discrepancy by plugging (2.110e) into (2.109):

$$\hat{d}_{KL}(L) \approx \mathrm{E}_{\mathbf{y}}\left[\log(p_L(\mathbf{y}|\hat{\boldsymbol{\theta}}_L)) - L(1+\lambda)\right], \tag{2.113}$$

from where model-order selection cost function of the AIC and GIC are obtained taking the argument of the expected value in (2.113). The reason behind choosing the argument of the previous expected value as the final criterion is it is a biased estimate of (2.113). Thus, the final criterion is the one shown in (2.83) and particularized for the AIC and GIC penalty terms given in Table 2.1.

As a summary, the AIC and the GIC only require the non-singularity of the Fisher information matrix so that (2.110e) is a valid step. In a similar fashion to the BIC, these two criteria are expected to have a better performance for large sample sizes. The reason behind the previous observation is that all the previously described approximations (see (2.99) and (2.108)) improve as $K \to \infty$. With some abuse of notation, we can say that these approaches are asymptotically consistent.

2.4 Concluding remarks

In this chapter, we introduced the kind of cost functions that are considered in this dissertation. As it has been noted by the link between sparsity and entropy, all these cost functions are encompassed in the information theoretic framework. Not only that, but we have also shown that subspace learning techniques can also be related to a sparse signal processing methodology. In this sense, our main motivation behind the consideration of the Grassmann manifold is the introduction of structural priors, which are thought to be more versatile than classical statistical priors. Thus, the previous kind of priors are much more suitable for practical applications.

The main issue with the previous framework is that it leads to challenging non-convex optimization problems (see Section 2.2 or Subsection 2.1.2). In this regard, there is no effective implementation of a globally optimal algorithm for the proposed globally optimal non-convex criteria (up to the authors knowledge). This is the reason why we renounce to globally optimal solutions in favour of locally optimal solutions, which are much faster to compute. As it is expanded in the sequel, the previously mentioned locally optimal solutions have a reasonable performance. The previous observation, in addition to the natural robustness and interpretability of the cost functions that are detailed in this chapter, suggests that the framework detailed in this chapter can have a practical implementation in signal processing applications.

Chapter 3

Algorithmic framework

The purpose of this chapter is to introduce the framework that is used to solve the optimization problems shown in Chapter 2. As it has already been mentioned, both convex and non-convex optimization problems are considered in this dissertation. For instance, while the cost functions that are built using an ℓ_1 norm regularization are convex, the problems that are based on information theoretic measures often yield non-convex optimization problems (see (2.38) for an example). Likewise, the problems whose constraint set is the Grassmann manifold are also non-convex due to the orthogonality constraints.

As it is often said in convex optimization theory, once a problem is formulated as a convex optimization, then it is already solved. The motivation behind the previous statement is summarized by the following two properties of convex formulations [29]: feasible convex optimization problems are always solvable and there exist fast known algorithms that solve them. This is the reason why convex optimization problems are preferred over non-convex formulations. In this regard, algorithms that solve non-convex optimization problems, such as the ones mentioned in Chapter 2, must renounce to one of the previous two properties due to the challenging nature of non-convex problems [48]. As a result, non-convex methods can be classified into two categories [55]:

1. The *global optimization methods* search and verify the global optimum of a non-convex optimization problem, but they tend to be really slow for this purpose. An example of this methodology is found in the branch and bound methods [144]. See [48, Section 8] for more instances of global optimization methods.
2. The *local optimization methods* thrive in their convergence speed but they do not necessarily find a global optimum. Even if they found an optimal solution of the problem, they cannot distinguish it from a stationary point.

In this dissertation, we focus on the latter family of algorithms to solve non-convex optimization problems. In this sense, we are interested in the local optimization algorithms that are constructed using the Sequential Convex Programming (SCP) framework. Informally speaking, this methodology solves an optimization problem by means of iterative schemes based on the optimization of a convex surrogate, whose optimal value is easily found, of the original cost. For non-convex problems, the SCP framework splits the cost functions or the constraint sets into two independent components: the convex and non-convex blocks of the original cost. While the convex components are unmodified, the non-convex terms are approximated by convex surrogates. In this manner, the overall problem is solved iteratively using efficient algorithms. In fact, the previous rationale is encompassed in the known *Majorization-Minimization* (MM) framework, which will be thoroughly surveyed in this chapter. Albeit this framework may fail to retrieve (or certify) the global optimum, a stationary point is often a sufficiently well-behaved solution in practice. Besides, one could always repeat the given procedure for different initialization points to achieve a better estimate. Regarding the problems that consider the subspace-based sparsity from Chapter 2 (see Section 2.2), we show that the *geodesically convex optimization* (g-convex for short) framework can complement the classical MM methodology. Indeed, g-convex optimization can be seen as an alternative efficient way of deal with some non-convex sets (Riemannian manifolds). Although this framework is general to any Riemannian manifold, we only

focus on its particularization for the Grassmann manifold with the main goal of generalizing some of the known results of the MM framework to the Grassmann manifold constraint set.

Not only the rationale behind the MM framework is useful for non-convex optimization, but it can also be applied to convex optimization problems as well. Whilst local optimization methods ensure the converge to a stationary point of non-convex problems, they converge to a global optimum in convex problems. Actually, the MM framework is the foundation of known algorithms, such as the Gradient Descent [169], and the Iterative Soft-Thresholding Algorithm [19], among others, that are widely used in the convex optimization literature.

The structure of this chapter consists of three sections. In Section 3.1, we introduce several background concepts that will be needed to analyze the MM framework. Despite the fact that some of those background concepts may seem familiar to the reader, this first section also serves as a way to set the notation of the presented ideas. Secondly, we review some of the key ideas from the convex optimization theory and compare them to their Riemannian counterparts (particularized on the Grassmann manifold) in Section 3.2. Finally, the core of the MM framework that will be used in the remaining chapters of this dissertation is described in Section 3.3. It is in Section 3.3 where the main results regarding the MM framework are detailed. Particularly, our proposed generalization of the MM framework to the Grassmann manifold can be found in Subsection 3.3.4.

3.1 Preliminaries on optimization theory

For clarity in the exposition, the concepts that are presented in this chapter are targeted towards minimization problems, unless stated otherwise. Yet, we remark that these concepts can be easily translated for maximization problems.

3.1.1 Level sets

Level sets are fundamental tools to assess the existence of optimal values of a given function. Thanks to the extreme value theorem [152], the compactness of the sublevel and supralevel sets is a sufficient condition for a minimum and a maximum (respectively) of a function to exist. In fact, the previous statement is fundamental in optimization theory since it is one of the conditions for the feasibility of an optimization problem. In the particular case of our algorithmic framework, we define the *sublevel sets*, which are a particular case of the level sets, as follows.

Definition 3.1 (Sublevel set)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be any arbitrary function. Then, its sublevel set is defined as:*

$$\mathcal{S}(f, C) = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \leq C\}, \tag{3.1}$$

where C is a constant.

Remark 3.1. While sublevel sets are important for minimization problems, the *supralevel sets* are their counterparts for maximization problems. The supralevel sets are defined as follows:

$$\mathcal{S}'(f, C) = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \geq C\}. \tag{3.2}$$

We are interested in the assessment of the compactness of a function level set. For this purpose, we review the properties that are needed for this analysis, which are the closedness and boundedness of a set [46].

Definition 3.2 (Open and closed sets)**.** *A set $\mathcal{U}$ is said to be open if for every $\mathbf{u} \in \mathcal{U}$ there exists a positive number r such that the ball of radius r is contained in $\mathcal{U}$. On the contrary, a set is said to be closed if its complementary set is open.*

The previous definition can be simplified in an informal manner for the cases that are encountered in this dissertation. Particularly, we are interested on sets that are defined using functions of the optimization variables. In this regard, the following sets:

$$g(\mathbf{x}) \leq 0, \tag{3.3a}$$

$$g(\mathbf{x}) \geq 0, \tag{3.3b}$$

$$g(\mathbf{x}) = 0, \tag{3.3c}$$

where $g : \mathbb{R}^N \to \mathbb{R}$ is a continuous function, are closed. Contrarily, the following sets:

$$g(\mathbf{x}) < 0, \tag{3.4a}$$

$$g(\mathbf{x}) > 0, \tag{3.4b}$$

are open. Additionally, $\mathbb{R}^N$ is open and closed by definition [46]. The previous examples are useful to quickly assess the closedness of most of the sets that are encountered in this dissertation. The remaining concept that is needed to asses the compactness of a set is whether a set is bounded or not. The boundedness of a set is defined as follows.

Definition 3.3 (Bounded set)**.** *A set $\mathcal{B}$ is said to be bounded if there exists a positive value C such that:*

$$||\mathbf{b}||_2^2 < C, \tag{3.5}$$

for any $\mathbf{b} \in \mathcal{B}$.

In contrast to the closedness of a set, there is no general rule of thumb for the assessment of the boundedness of a set. Thus, we verify this property in a case by case basis henceforth. Finally, the definition of compact sets emanates from the previous two definitions.

Definition 3.4 (Compact set)**.** *A set is said to be compact if it is closed and bounded.*

There are two cases in which the sublevel sets of a function are compact: $\mathcal{X}$ is itself compact or $f(\mathbf{X})$ is a *coercive* function [92]. For completeness, we review the definition of coercive functions [15].

Definition 3.5 (Coercive function)**.** *Let any arbitrary function be $f : \mathbb{R}^N \to \mathbb{R}$. Then, it is said to be coercive, or also radially unbounded, if for every $c > 0$ there exists an $r > 0$ such that:*

$$\forall \mathbf{x} \in \mathbb{R}^N : ||\mathbf{x}||_2^2 > c \implies f(\mathbf{x}) > r. \tag{3.6}$$

Remark 3.2. The previous definition implies that if f is coercive, then it has a global minimum value. However, (3.6) is not useful to verify whether f has a global maximum value. For the purpose of assessing the existence of a global maximum value of f, (3.6) can be adapted as follows:

$$\forall \mathbf{x} \in \mathbb{R}^N : ||\mathbf{x}||_2^2 > c \implies f(\mathbf{x}) < r, \tag{3.7}$$

for every $c > 0$. A function that satisfies (3.7) is said to be *negatively* coercive.

One simple way to test coerciveness of a function is to study the behaviour of $f(\mathbf{x})$ in terms of $||\mathbf{x}||_2^2$. If $f(\mathbf{x}) \to +\infty$ when $||\mathbf{x}||_2^2 \to \infty$, then $f(\mathbf{x})$ is coercive. By the comparison of Definition 3.1 with Definition 3.5, it is clear that the coerciveness of a function implies that its corresponding sublevel set is compact.

3.1.2 Derivatives and stationary points of a function

Derivatives are used to determine descent directions of a function and to study the stationarity of a solution in the numerical optimization theory, among other applications. For this reason, they are a fundamental tool in any optimization framework. Given that variables constrained in the Grassmann manifold are also considered in our algorithmic framework, there is a necessity for the generalization of the directional derivatives. For this reason, we particularize the concept of Gateaux differentials, which are the generalization of directional derivative for Banach spaces (a vector space with some notion of metric) [2], to the context considered in this dissertation. The Gateaux differentials are defined as follows.

Definition 3.6 (Gateaux differentials)**.** *Let any arbitrary function be $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ can be $\mathrm{Gr}(N, D)$ or $\mathbb{R}^D$. Then, the Gateaux differential of f at $\mathbf{x} \in \mathcal{X}$ is defined as:*

$$Df(\mathbf{x})[\boldsymbol{\delta}] = \lim_{t \to 0^+} \frac{f(\mathbf{x} + t\boldsymbol{\delta}) - f(\mathbf{x})}{t}, \tag{3.8}$$

where $\boldsymbol{\delta} \in \mathcal{X}$ is the direction of the Gateaux differential.

Remark 3.3. The rationale behind the positive constraint in t is to avoid the sign ambiguity in the direction of the derivative. In this way, we emphasize the direction of the differential, $\boldsymbol{\delta}$, which is critical for the determination of descent directions. Not only that, but the aforementioned constraint also causes that the set of Gateaux differentiable functions is larger than the set of differentiable functions in the classical sense.

Remark 3.4. For those functions that are differentiable in the classical sense, the Gateaux differential can be alternatively computed as follows:

$$\left.\frac{\mathrm{d}f(\mathbf{x}+t\boldsymbol{\delta})}{\mathrm{d}t}\right|_{t=0} = \nabla_{\mathbf{x}} f^T(\mathbf{x})\boldsymbol{\delta}, \tag{3.9}$$

where now $\boldsymbol{\delta}$ coincides with the classical directional derivative. Likewise, the equivalent Gateaux differential expression for the (differentiable) functions whose input variables are matrices is:

$$Df(\mathbf{X})[\boldsymbol{\Delta}] = \lim_{t\to 0^+} \frac{f(\mathbf{X}+t\boldsymbol{\Delta}) - f(\mathbf{X})}{t} = \left.\frac{\mathrm{d}f(\mathbf{X}+t\boldsymbol{\Delta})}{\mathrm{d}t}\right|_{t=0} = \mathrm{tr}(\nabla_{\mathbf{X}} f^T(\mathbf{X})\boldsymbol{\Delta}), \tag{3.10}$$

where $\boldsymbol{\Delta}, \mathbf{X} \in \mathcal{X}$.

Gateaux directional derivatives are useful for several tasks in our algorithmic framework. Firstly, they can be used to approximate functions in a similar fashion to the first-order Taylor expansion. For a given function $f(\mathbf{x})$, this approximation yields:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + Df(\mathbf{x}_0)[\boldsymbol{\delta}], \tag{3.11}$$

where $\mathbf{x}_0$ is the expansion point and $\boldsymbol{\delta}$ is the direction in which the approximation is computed. Note that (3.11) is the generalization of the first-order Taylor expansion, which could be recovered from (3.11) if f were a differentiable function in the classical sense (see (3.9)) and if $\boldsymbol{\delta} = (\mathbf{x} - \mathbf{x}_0)$. As it is formalized in the sequel, for convex (and g-convex) functions, (3.11) is also a lower bound of the original function. Similarly, it becomes an upper bound for concave (and g-concave) functions.

Secondly, there are non-differentiable functions in the classical sense whose Gateaux differential does exists. In the following example, we show that the absolute value, which is the building block of the ℓ_1 norm, is differentiable under Definition 3.6.

Example 3.1 (Gateaux differentiability of the absolute value)**.** Let $f(x) = |x|$. Then, its Gateaux differential at $x = 0$ yields:

$$Df(0)[d] = \lim_{t\to 0^+} \frac{|td|}{t} = |d| \lim_{t\to 0^+} \frac{|t|}{t} = |d|. \tag{3.12}$$

Besides, its Gateaux differential at $x \neq 0$ is given by:

$$Df(x)[d] = \lim_{t\to 0^+} \frac{|x+td| - |x|}{t} = \left.\frac{\mathrm{d}f(x+td)}{\mathrm{d}t}\right|_{t=0} = d\,\mathrm{sign}(x), \tag{3.13}$$

which follows from the fact that f is differentiable for every $x \neq 0$. As a summary, the Gateaux differential of $f(x) = |x|$ is:

$$Df(x)[d] = \begin{cases} |d| & x = 0 \\ d\,\mathrm{sign}(x) & x \neq 0 \end{cases}. \tag{3.14}$$

Note that the Gateaux differential is linear on d at the values of x where the function is differentiable, while it is non-linear on d for the values of x such that the derivative does not exist.

Lastly, the use of directional derivatives is necessary to assess whether a solution of an iterative optimization scheme is a *stationary point* of a function. In essence, finding stationary points is the main goal of iterative optimization schemes. The stationary points are built on Definition 3.6 and there are two kinds: the local minimum and local maximum points.

Definition 3.7 (Local minimum points)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be any arbitrary function. Then, any $\mathbf{x} \in \mathcal{X}$ is a local minimum if:*

$$Df(\mathbf{x})[\boldsymbol{\delta}] \geq 0 \quad \forall \mathbf{x} + \boldsymbol{\delta} \in \mathcal{X}. \tag{3.15}$$

Definition 3.8 (Local maximum points)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be any arbitrary function. Then, any $\mathbf{x} \in \mathcal{X}$ is a local maximum if:*

$$Df(\mathbf{x})[\boldsymbol{\delta}] \leq 0 \quad \forall \mathbf{x} + \boldsymbol{\delta} \in \mathcal{X}. \tag{3.16}$$

Remark 3.5. Stationary points are also defined using an alternative definition of directional derivatives, which are known as *Dini derivatives* [46, Appendix B]. The lower Dini derivative is defined as follows:

$$f_l'(\mathbf{x}; \boldsymbol{\delta}) = \lim_{t \to 0} \inf \frac{f(\mathbf{x} + t\boldsymbol{\delta}) - f(\mathbf{x})}{t}, \tag{3.17}$$

where lim inf denotes the limit inferior of a sequence, which is used to obtain the local minimum points of a function. Similarly, the upper Dini derivative is the equivalent counterpart for the determination of local maximum points. Its expression is given by:

$$f_s'(\mathbf{x}; \boldsymbol{\delta}) = \lim_{t \to 0} \sup \frac{f(\mathbf{x} + t\boldsymbol{\delta}) - f(\mathbf{x})}{t}, \tag{3.18}$$

where lim sup denotes the limit superior of a sequence. For simplicity, we do not consider the above kind of directional derivatives.

It can be shown that the previous definition of stationary points is intuitive. As an example, consider the absolute value function and its Gateaux differential (see Example 3.1). Clearly, $x = 0$ is the only stationary point of $f(x) = |x|$ since it is the only point in $\mathbb{R}$ such that $Df(x)[d] \geq 0$ for all $d \in \mathbb{R}$. The fact that $Df(0)[d] \geq 0$ means that evaluating $f(x)$ after taking a step in the direction of the differential, d, increases the value of the function. Thus, it must be a local minimum (global minimum in the previous example). In this regard, notice that the previous two definitions include the classical definition of stationary points, i.e. gradients equal to zero, but they are much more precise in the distinction of local minimum and maximum points.

Given that we are also interested in the block extension of optimization algorithms, such as the Block Coordinate Descent (BCD) approach [154] or the block extension of the MM framework [123], [162], it is useful to describe the cost functions whose coordinate-wise stationary points are equivalent to stationary points as in Definitions 3.7 and 3.8. The reasoning behind the previous statement is that the block extension of the iterative optimization schemes is only able to search for coordinate-wise stationary points. Coordinate-wise stationary points are described as follows.

Definition 3.9 (Coordinate-wise minimum points)**.** *Let $f : \mathcal{X} \subseteq \mathbb{R}^M \to \mathbb{R}$ be any arbitrary function. Also, let the partition of $\mathbf{x}$ into N blocks of variables be $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, being the stacking of $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$. The dimensions each block of variables are denoted by N_n for $n = 1, ..., N$ and satisfy:*

$$M = \sum_{n=1}^{N} N_n. \tag{3.19}$$

Furthermore, let $\boldsymbol{\delta}_n$ be an all zeroes vector except for the n-th block, which contains the direction $\mathbf{d}_n$ of dimensions N_n, i.e. $\boldsymbol{\delta}_n = (\mathbf{0}, ..., \mathbf{d}_n, ..., \mathbf{0})$.. Then, any $\mathbf{x} \in \mathcal{X}$ is a coordinate-wise minimum with respect to the n-th block if it satisfies:

$$Df(\mathbf{x})[\boldsymbol{\delta}_n] \geq 0 \quad \forall \mathbf{x} + \boldsymbol{\delta}_n \in \mathcal{X}. \tag{3.20}$$

Remark 3.6. Coordinate-wise maximum points are defined in the same manner by changing the sign of the inequality in (3.20).

With the previous definition in mind, we consider the concept of *regular functions*, which is a formal description of the functions whose coordinate-wise stationary points are equivalent to the global stationary points. Cost functions that satisfy this property are essential for block relaxations of iterative schemes [162], [187]. We consider the definition of regular functions presented in [187, Lemma 3.1].

Definition 3.10 (Regular function)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be any arbitrary function. Also, let the partition of $\mathbf{x}$ into N blocks of variables be $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$ where the dimensions of the n-th block of variables are denoted by N_n for $n = 1, ..., N$. In addition, let $\boldsymbol{\delta}_n$ be an all-zeroes vector except for the n-th block,*

which contains the direction $\mathbf{d}_n$ *of dimensions* N_n*, i.e.* $\boldsymbol{\delta}_n = (\mathbf{0}, ..., \mathbf{d}_n, ..., \mathbf{0})$*. Then, any* $\mathbf{x} \in \mathcal{X}$ *is said to be a regular point of* f *if it satisfies the following implication:*

$$\begin{aligned} Df(\mathbf{x})[\boldsymbol{\delta}_n] \geq 0 \ \forall \mathbf{x} \in \mathcal{X} \quad \text{s.t. } \mathbf{x} + \boldsymbol{\delta}_n \in \mathcal{X}, \forall \boldsymbol{\delta}_n \in \mathbb{R}^{N_n}, \ n = 1, ..., N \implies \\ Df(\mathbf{x})[\boldsymbol{\delta}] \geq 0 \quad \text{s.t. } \mathbf{x} + \boldsymbol{\delta} \in \mathcal{X}, \forall \boldsymbol{\delta} \in \mathcal{X}. \end{aligned} \tag{3.21}$$

Note that the converse statement is not true in general.
Remark 3.7. Regular functions can be thought of as a relaxation of smooth (differentiable) functions. An example of a non-regular function is found in [162, Figure 2.1].
Remark 3.8. The assumptions that ensure the regularity of a function are shown in [187, Section 3].

3.2 Riemannian optimization

Riemannian optimization, also referred to as geodesically convex optimization (g-convex optimization for short), is a generalization of convex optimization for Riemannian manifolds. This framework is a tool that can be used to bypass the challenges behind the optimization of some particular non-convex sets, e.g. the Grassmann manifold. In addition to this, g-convex optimization can be used to provide a geometric interpretation to some well-known signal processing problems, such as the Principal Component Analysis (PCA) and the Matrix Factorization problem [7].

The goal of this section is to review the particularization of g-convex optimization to the Grassmann manifold and convex optimization. This review is mainly based on [29] (convex optimization) and [57], [192] (g-convex optimization on the Grassmannian), which are very general, with an effort on extracting the relevant concepts for this thesis. In this regard, we also show the intuition behind the idea that the g-convex optimization is the generalization of convex optimization.

3.2.1 Convex optimization review

As it has been mentioned in Chapter 2, there is a solid literature on convex optimization problems [29], [74]. Indeed, it is widely known that convex programs (optimization problems) can be solved by means of relatively simple optimization algorithms that are fast and efficient. This is the main motivation behind looking for convex formulations in any kind of signal processing problem, often relying in well-known techniques such as the *convex relaxation* or the use of convex surrogates. A general convex optimization problem is depicted as follows:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{C}, \tag{3.22}$$

where $f(\mathbf{x})$ is a *convex function* and $\mathcal{C}$ is a *convex set*. We refer to $\mathcal{C}$ as the constraint or feasible set. Notice that the fundamental definitions in convex optimization are *convex sets* and *convex functions* [29].

Definition 3.11 (Convex set)**.** *Any set* $\mathcal{C}$ *is said to be convex if and only if the following statement is true for every* $\mathbf{x}, \mathbf{y} \in \mathcal{C}$*:*

$$t\mathbf{x} + (1-t)\mathbf{y} \in \mathcal{C} \quad \forall t \in [0, 1], \tag{3.23}$$

where the above linear combination is also known as convex combination*.*
Remark 3.9. A convex combination is a geodesic (or, equivalently, an exponential map) in the Euclidean space [192].

The above definition is essential in optimization theory because not only it describs the constraint sets that enable line search methods [117], e.g. Gradient Descent, but they also describe the geometric properties that a set must fulfill in order to be well-behaved in optimization. As it is shown in the sequel, the previous idea encourages its generalization to g-convex optimization, where the very restricted convex combination idea from (3.23) is substituted by a geodesic in a Riemannian manifold. Moreover, the definition of convex and concave functions is also dependant on Definition 3.11.

Definition 3.12 (Convex function)**.** *Any function* $f : \mathcal{C} \to \mathbb{R}$ *is said to be convex if it satisfies for every* $\mathbf{x}, \mathbf{y} \in \mathcal{C}$*:*

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \quad \forall t \in [0, 1]. \tag{3.24}$$

Definition 3.13 (Concave function)**.** *Any function $f : \mathcal{C} \to \mathbb{R}$ is said to be concave if it satisfies for every $\mathbf{x}, \mathbf{y} \in \mathcal{C}$:*

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \geq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \quad \forall t \in [0,1]. \tag{3.25}$$

Remark 3.10. f is convex if and only if the restriction to any line that remains on $\mathcal{C}$ is convex. To be more specific, $f(\mathbf{x})$ is convex with respect to $\mathbf{x}$ if the function $g(t) = f(\mathbf{x} + t\mathbf{y})$ subject to $\mathbf{x} + t\mathbf{y} \in \mathcal{C}$ is convex with respect to t. An equivalent argument also holds for concave functions.
Remark 3.11. A function g is said to be concave if $-g$ is convex. Also, when a function is both concave and convex, i.e. (3.24) and (3.25) are fulfilled with an equality, it is said to be affine.
Remark 3.12. A local minimum is equivalent to a global minimum of f if it is convex. Equivalently, a local maximum is a global maximum of f if it is concave.

There are two main implications emanating from the previous two definitions that are fundamental in convex optimization theory. Both of them are based on first and derivatives of a convex/concave function and they are summarized in the following two theorems.
Theorem 3.1 (First-order characterization of convex functions)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a convex differentiable function defined on the convex set $\mathcal{C}$. Then:*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{C}. \tag{3.26}$$

Remark 3.13. The vector $(\mathbf{y} - \mathbf{x})$ can be seen as a tangent direction from $\mathbf{x}$ to $\mathbf{y}$ in the Euclidean space.
Theorem 3.2 (Second-order characterization of convex functions)**.** *Let $f : \mathcal{C} \to \mathbb{R}$ be a convex (twice) differentiable function defined on the convex set $\mathcal{C}$. Then:*

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) \succeq 0 \quad \forall \mathbf{x} \in \mathcal{C}. \tag{3.27}$$

Remark 3.14. The above two theorems can be particularized for concave functions by reversing the sign of the previous inequalities.

In the sequel, we are using both of these theorems that characterize convex functions to derive surrogate functions in the MM framework and to assess the convexity of a function. Note that a function that is convex in the whole space becomes no longer convex when it is restricted in a non-convex set. This is caused by the fact that there would exist convex combinations between two points that lie outside the feasible set. Thus, it would not be possible to satisfy Definitions 3.12 or 3.13.

Provided that the log-determinant function plays an important role in this dissertation (see (2.38) for an example), we prove its concavity as an example [29], [75] in the following lemma.
Lemma 3.3 (Concavity of the log-determinant function)**.** *The function $f(\mathbf{X}) = \log(\det(\mathbf{X}))$ is concave if $\mathbf{X} \in \mathcal{S}^M_{++}$.*

Proof. In order to proof this, we need to invoke Remark 3.10. Let us define the following function:

$$g(t) = \log(\det(\mathbf{X} + t\mathbf{Y})) \quad \text{s.t. } \mathbf{X} + t\mathbf{Y} \in \mathcal{S}^M_{++}, \tag{3.28}$$

for $t \in [0,1]$. If (3.28) is proven to be concave, then the log-determinant function is also concave. Given that $\mathbf{X} \in \mathcal{S}^M_{++}$, the square root of $\mathbf{X}$ exists. As a result, $g(t)$ can be rewritten as follows:

$$g(t) = \log(\det(\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} + t\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}})) = \log(\det(\mathbf{X}^{\frac{1}{2}}(\mathbf{I} + t\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}})\mathbf{X}^{\frac{1}{2}})) = \tag{3.29a}$$

$$\log(\det(\mathbf{X})) + \log(\det((\mathbf{I} + t\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}))) = \log(\det(\mathbf{X})) + \sum_{m=1}^{M} \log(1 + t\lambda_m), \tag{3.29b}$$

where λ_m for $m = 1, ..., M$ are the eigenvalues of $\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}$. Then, the second derivative of $g(t)$ is:

$$\frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2} = -\sum_{m=1}^{M} \frac{\lambda_m^2}{(1 + t\lambda_m)^2}, \tag{3.30}$$

which satisfies $\frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2} \leq 0$ for every $t \in \mathbb{R}$ since λ_m is positive for $m = 1, ..., M$. Invoking Theorem 3.2, we get that $f(\mathbf{X})$ is concave since $g(t)$ is concave. ■

The previous proof provides some insights on the assessment of g-convex functions. Notice that, in the above rationale, the restriction to a line can be interpreted as a geodesic on the manifold defined by $\mathcal{S}_{++}^{M}$. G-convex functions are aimed at generalizing the idea presented in the previous example.

Even though convex functions are important for any kind of optimization scheme, there is a relaxation of this kind of functions that is equally important for iterative algorithms. This relaxation of convexity, which is known as *quasiconvexity* [29], is based on the convexity of the level sets of a function and it is defined as follows.

Definition 3.14 (Quasiconvex functions)**.** *Any function $f : \mathcal{C} \to \mathbb{R}$ is said to be quasiconvex in the convex set $\mathcal{C}$ if any sublevel set of this function, i.e.:*

$$\mathcal{S}(f, \alpha) = \{\mathbf{x} \in \mathcal{C} : f(\mathbf{x}) \leq \alpha\}, \tag{3.31}$$

is convex.

Definition 3.15 (Quasiconcave functions)**.** *Any function $f : \mathcal{C} \to \mathbb{R}$ is said to be quasiconcave in the convex set $\mathcal{C}$ if any supralevel set of this function, i.e.:*

$$\mathcal{S}'(f, \alpha) = \{\mathbf{x} \in \mathcal{C} : f(\mathbf{x}) \geq \alpha\}, \tag{3.32}$$

is convex.

Remark 3.15. Similarly to convex and concave functions, the restriction to any line of a quasiconvex or quasiconcave function also results in a quasiconvex or quasiconcave function, respectively.

Remark 3.16. The definition of quasiconvex functions implies:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq \max(f(\mathbf{x}), f(\mathbf{y})) \ \forall t \in [0, 1], \tag{3.33}$$

from where it can be seen that quasiconvexity is a relaxation of convexity since:

$$(1-t)f(\mathbf{x}) + tf(\mathbf{y}) \leq \max(f(\mathbf{x}), f(\mathbf{y})) \ \forall t \in [0, 1]. \tag{3.34}$$

Similarly, quasiconcave functions satisfy:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \geq \min(f(\mathbf{x}), f(\mathbf{y})) \ \forall t \in [0, 1], \tag{3.35}$$

whose relation with concavity can be seen from the following expression:

$$(1-t)f(\mathbf{x}) + tf(\mathbf{y}) \geq \min(f(\mathbf{x}), f(\mathbf{y})) \ \forall t \in [0, 1]. \tag{3.36}$$

The above remarks are highlighted for two reasons. While the first remark is useful to verify whether a function is quasiconvex, the second remark highlights the fact that verifying the convexity or concavity of a function is a sufficient condition to show its quasiconvexity or quasiconcavity, respectively, which is relevant in the MM framework.

Although it is out of the scope of this thesis, there is a standard way to deal with quasiconvex problems called *quasiconvex programming* [74]. It is especially useful for those functions whose convexity nor concavity cannot be guaranteed. As an example, we analyze the previous properties of a function that plays an important role in Section 4.3 from Chapter 4.

Example 3.2 (Quasiconvexity of the logarithm of a positive semidefinite quadratic form)**.** Let $f : \mathbb{R}^M \to \mathbb{R}$ be defined as:

$$f(\mathbf{x}) = \log(\mathbf{x}^T \mathbf{C} \mathbf{x}), \tag{3.37}$$

where $\mathbf{C} \in \mathcal{S}_{++}^{M}$. Firstly, it can be shown that the previous function is concave for $M = 1$. For $M = 1$, $f(\mathbf{x})$ becomes:

$$f(x) = \log(cx^2) = \log(c) + 2\log(x), \tag{3.38}$$

which is a concave function since:

$$\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} = -\frac{2}{x^2} < 0 \quad \forall x \in \mathbb{R}. \tag{3.39}$$

Next, we show that $f(\mathbf{x})$ is not convex nor concave when $M > 1$. For the purpose of proving the non-convexity of $f(\mathbf{x})$, let us consider the Hessian matrix of $f(\mathbf{x})$:

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \mathbf{H}(\mathbf{x}) = \frac{\mathbf{C}}{\mathbf{x}^T\mathbf{C}\mathbf{x}} - 2\frac{\mathbf{C}\mathbf{x}\mathbf{x}^T\mathbf{C}}{(\mathbf{x}^T\mathbf{C}\mathbf{x})^2}. \tag{3.40}$$

Clearly, in order to assess the convexity or concavity of $f(\mathbf{x})$, one needs to evaluate the definiteness of $\mathbf{H}(\mathbf{x})$. The previous evaluation can be performed by evaluating the positiveness or the negativeness of $h(\mathbf{x}, \mathbf{v}) = \mathbf{v}^T\mathbf{H}(\mathbf{x})\mathbf{v}$ for all $\mathbf{v}, \mathbf{x} \in \mathbb{R}^M$, where $\mathbf{v}$ is an arbitrary vector. To facilitate the assessment of the positiveness/negativeness of $h(\mathbf{x}, \mathbf{v})$, we rewrite this function as follows:

$$h(\mathbf{x}, \mathbf{v}) = \mathbf{v}^T\mathbf{H}(\mathbf{x})\mathbf{v} = \tag{3.41a}$$

$$\frac{\mathbf{v}^T\mathbf{C}\mathbf{v}}{\mathbf{x}^T\mathbf{C}\mathbf{x}} - 2\frac{\mathbf{v}^T\mathbf{C}\mathbf{x}\mathbf{x}^T\mathbf{C}\mathbf{v}}{(\mathbf{x}^T\mathbf{C}\mathbf{x})^2} = \tag{3.41b}$$

$$\frac{\mathbf{v}^T\mathbf{C}\mathbf{v}}{\mathbf{x}^T\mathbf{C}\mathbf{x}} - 2\frac{(\mathbf{x}^T\mathbf{C}\mathbf{v})^2}{(\mathbf{x}^T\mathbf{C}\mathbf{x})^2} = \tag{3.41c}$$

$$\frac{(\mathbf{v}^T\mathbf{C}\mathbf{v})(\mathbf{x}^T\mathbf{C}\mathbf{x}) - 2(\mathbf{x}^T\mathbf{C}\mathbf{v})^2}{(\mathbf{x}^T\mathbf{C}\mathbf{x})^2}, \tag{3.41d}$$

from where the quadratic forms of the previous expression can be interpreted as the norms and dot products of $\mathbf{v}$ and $\mathbf{x}$ since $\mathbf{C}$ is a positive definite matrix [138, Chapter 2]. Note that the sign of $h(\mathbf{x}, \mathbf{v})$ can be evaluated from the numerator of (3.41d) because its denominator is always positive. We show that there are two instances where $h(\mathbf{x}, \mathbf{v})$ is positive and negative. Taking a $\mathbf{v}$ that is orthogonal to $\mathbf{x}$ in the vector space whose inner product is constructed using $\mathbf{C}$, we get that the numerator in (3.41d) yields:

$$(\mathbf{v}^T\mathbf{C}\mathbf{v})(\mathbf{x}^T\mathbf{C}\mathbf{x}) > 0, \tag{3.42}$$

because $\mathbf{x}^T\mathbf{C}\mathbf{v} = 0$ in this case. Contrarily, by setting $\mathbf{v} = \alpha\mathbf{x}$, the numerator in (3.41d) becomes:

$$(\mathbf{v}^T\mathbf{C}\mathbf{v})(\mathbf{x}^T\mathbf{C}\mathbf{x}) - 2(\mathbf{x}^T\mathbf{C}\mathbf{v})^2 = \alpha^2(\mathbf{x}^T\mathbf{C}\mathbf{x})^2 - 2\alpha^2(\mathbf{x}^T\mathbf{C}\mathbf{x})^2 = -\alpha^2(\mathbf{x}^T\mathbf{C}\mathbf{x})^2 < 0 \quad \forall \mathbf{x} \in \mathbb{R}^M. \tag{3.43}$$

As a result, there are combinations of values of $\mathbf{x}$ and $\mathbf{v}$ such that $h(\mathbf{x}, \mathbf{v})$ is positive and negative. Hence, $\mathbf{H}(\mathbf{x})$ is not positive nor negative definite, so $f(\mathbf{x})$ is non-convex and non-concave for $M > 1$.

Interestingly, we show that $f(\mathbf{x})$ is quasiconvex for $M > 1$. With the previous goal in mind, consider the sublevel set of $f(\mathbf{x})$:

$$S(f, \alpha) = \{\mathbf{x} \in \mathbb{R}^M : \log(\mathbf{x}^T\mathbf{C}\mathbf{x}) \leq \alpha\}, \tag{3.44}$$

which can be rewritten as follows:

$$S(f, \alpha) = \{\mathbf{x} \in \mathbb{R}^M : \mathbf{x}^T\mathbf{C}\mathbf{x} \leq \exp(\alpha)\}. \tag{3.45}$$

Proving the convexity of $S(f, \alpha)$ is a sufficient condition to certify the quasiconvexity of f. With this purpose in mind, we verify the convexity of the set described by (3.45). First, let $\mathbf{x}, \mathbf{y} \in S(f, \alpha)$. Also, let $\mathbf{z} = t\mathbf{x} + (1-t)\mathbf{y}$. Then:

$$\mathbf{z}^T\mathbf{C}\mathbf{z} = \tag{3.46a}$$

$$(t\mathbf{x} + (1-t)\mathbf{y})^T\mathbf{C}(t\mathbf{x} + (1-t)\mathbf{y}) \leq t\mathbf{x}^T\mathbf{C}\mathbf{x} + (1-t)\mathbf{y}^T\mathbf{C}\mathbf{y} \leq \tag{3.46b}$$

$$t\exp(\alpha) + (1-t)\exp(\alpha) = \exp(\alpha), \tag{3.46c}$$

where (3.46b) holds since $\mathbf{x}^T\mathbf{C}\mathbf{x}$ is a convex function with respect to $\mathbf{x}$ ($\mathbf{C} \in \mathcal{S}^M_{++}$ ensures the convexity of the aforementioned quadratic form) and (3.46c) follows from the fact that $\mathbf{x}, \mathbf{y} \in S(f, \alpha)$. Therefore, $S(f, \alpha)$ is a convex set, so f is a quasiconvex function.

3.2.2 G-convex optimization on the Grassmann manifold

Building upon the earlier overview on convex optimization, in this subsection we review a generalization of convex optimization for variables constrained in the Grassmann manifold. Note that, although the surveyed concepts on g-convex optimization are particularized to the geometry of the Grasmannian shown in Section 2.2, these ideas are general to other Riemannian manifolds. Still, given that there are known results of g-convex optimization for the Grassmann manifold, we do not tackle the general case. Instead, this subsection serves as a way to show *how to use* g-convex optimization. In contrast to what happens in convex optimization with convex sets, there is a necessity to define two types of *Riemannian convex sets* since, in a Riemannian manifold, the paths that connect any two points within the manifold may not be unique in general. This leads to the distinction between *totally convex sets* and *geodesically convex sets* [192].

Definition 3.16 (Totally convex sets on the Grassmann manifold)**.** *Let $\mathcal{G} \subseteq \mathrm{Gr}(N, D)$. Then, it is said that $\mathcal{G}$ is a totally convex subset of the Grassmann manifold if, for any $\mathbf{X}, \mathbf{Y} \in \mathcal{G}$, any path within the Grassmann manifold connecting those points is also contained in $\mathcal{G}$.*

Remark 3.17. The above definition is often relaxed so that only geodesics (paths with minimum distance) are required to be contained in $\mathcal{G}$. In this manner, sets that satisfy this relaxation are defined as *geodesically convex sets* (or g-convex sets for short). In the Euclidean space, both definitions are equivalent since the only path connecting any two points is the straight line. In general, it is easier to deal with totally convex sets as compared to g-convex sets. In the case of the Grassmannian, the whole manifold is totally convex, but it does not exists any subset of the Grassmann manifold such that it is totally convex. One of the implications of the previous statement is seen in Lemma 3.4.

To exemplify the difference between g-convex sets and totally convex sets, in Figure 3.1 we show an example of a g-convex set that is not a totally convex set in the $\mathbb{R}^2$ unit sphere, i.e. the particular case of the Stiefel manifold given by $\mathrm{St}(2, 1)$ [192]. Note that there are two possible paths between two points in $\mathrm{St}(2, 1)$, i.e. the short and long arcs, and that the g-convex subset is depicted by the highlighted area in Figure 3.1. Clearly, the long arc does not belong to the highlighted area. Thus, it is not a totally convex set. On the contrary, given that every geodesic (the short arcs) between any two elements of the highlighted area is contained in the aforementioned set, it fulfills the definition of a g-convex set.

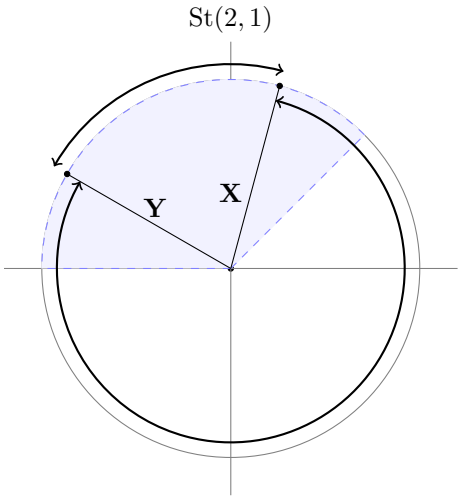


Figure 3.1: Example: Visualization of a g-convex set (highlighted area) in $\mathrm{St}(2, 1)$ which is not totally convex.

Luckily, totally convex sets are not mandatory for g-convex optimization. Instead, g-convex sets are well-behaved for the definition of g-convex functions, as seen in the sequel. Regarding the Grassmann manifold, note that the whole Grassmann manifold is a g-convex set (also, totally convex) since there exists a geodesic joining any two points. Yet, the assessment of g-convex subsets in the Grassmann

manifold must be done with care. For this purpose, we focus our analysis on *ball-like*[1] g-convex subsets on the Grassmannian. To this end, we firstly define balls in the Grassmann manifold.

Definition 3.17 (Ball on the Grassmann manifold)**.** *Let an arbitrary point in the Grassmann manifold be* $\mathbf{C} \in \text{Gr}(N, D)$*. Then, an open ball of radius* ϕ *in the Grassmann manifold is given by:*

$$B_\phi(\mathbf{C}) = \{\mathbf{X} \in \text{Gr}(N, D) : \mathbf{\Theta}_\mathbf{X} \prec \phi \mathbf{I}_D\} = \{\mathbf{X} \in \text{Gr}(N, D) : d_{arc}(\mathbf{X}, \mathbf{C}) < \phi\sqrt{D}\}, \tag{3.47}$$

where $\mathbf{\Theta}_\mathbf{X}$ *contains the principal angles between* $\mathbf{C}$ *and* $\mathbf{X}$*.*

Remark 3.18. Note that the intuition behind the upper bound on the canonical distance in (3.47) comes from the fact that the canonical distance is defined as the Frobenius norm of the matrices that contain the principal angles (see (2.77)). In this regard, by setting $\mathbf{\Theta}' = \phi \mathbf{I}_D$, we get that:

$$||\mathbf{\Theta}'||_F = ||\phi \mathbf{I}_D||_F = \phi\sqrt{D}, \tag{3.48}$$

which is the upper bound given in the last expression of (3.47).

Given the previous definition, the conditions that a ball in the Grassmann manifold must fulfill to be a g-convex subset on the Grassmann manifold are summarized in the following lemma [7, Lemma 2].

Lemma 3.4 (G-convexity of balls on the Grassmann manifold)**.** *Let a ball on the Grassmann manifold be* $B_\phi(\mathbf{C})$ *for any* $\mathbf{C} \in \text{Gr}(N, D)$*. Then,* $B_\phi(\mathbf{C})$ *is a g-convex subset of the Grassmannian if and only if:*

$$\phi \leq \frac{\pi}{4}. \tag{3.49}$$

The proof of this Lemma can be found in [7].

In Figure 3.2, we show an example that provides some insights on Lemma 3.4 in the simple case of $\text{Gr}(2, 1)$ (the semi-circle in $\mathbb{R}^2$). The intuition behind the maximum radius for a Grassmann ball to be g-convex is that the furthest two points in this subset of $\text{Gr}(2, 1)$, $\mathbf{Y}$ and $\mathbf{X}$ in Figure 3.2, can be connected by a geodesic with principal angle $\frac{\pi}{2}$. The previous geodesic is such that its expression is unique (see Remark 2.13). On the contrary, the geodesic that connects $\mathbf{X}$ and $\mathbf{Z}$ is outside the g-convex subset since $\mathbf{X}$ is close to $\mathbf{Z}'$, which belongs to the same equivalence class as $\mathbf{Z}$ ($\mathbf{Z} = -\mathbf{Z}'$).

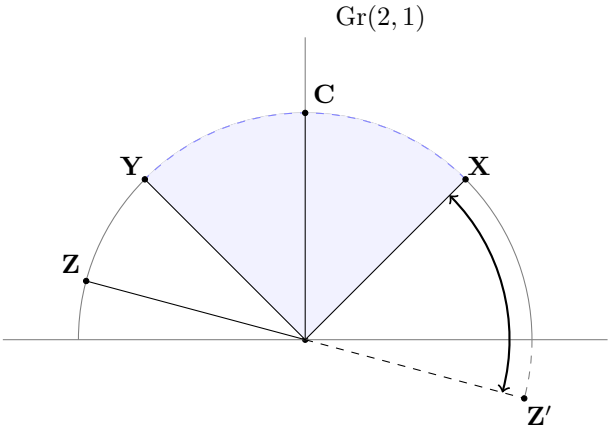


Figure 3.2: Example: Visualization of a g-convex set on $\text{Gr}(2, 1)$, $B_{\frac{\pi}{4}}(\mathbf{C})$.

Before diving into the details of g-convex functions, we firstly need to extend classical notions of calculus for the Grassmannian, such as gradients and Hessians [57], [192].

Definition 3.18 (Gradient of a function on the Grassmann manifold)**.** *Let* $f : \text{Gr}(N, D) \to \mathbb{R}$ *be a function such that its Gateaux differential exists. Then, the gradient on the Grassmann manifold (also referred to as Riemannian gradient) at a point* $\mathbf{X} \in \text{Gr}(N, D)$ *is defined as the unique tangent vector*

[1] By ball-like sets we mean those subsets that are defined as upper bounds of a distance with respect to a point, e.g. $\{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, \mathbf{y}) \leq r\}$, where $d(\cdot, \cdot)$ is any distance defined on $\mathcal{X}$ and $\mathbf{y}$ is the *center* of the ball.

grad $f(\mathbf{X}) \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$ *such that:*

$$\langle \text{grad } f(\mathbf{X}), \boldsymbol{\Delta} \rangle_{\mathbf{X}} = \left. \frac{\mathrm{d} f(\boldsymbol{\Gamma}(t))}{\mathrm{d}t} \right|_{t=0} = Df(\mathbf{X})[\boldsymbol{\Delta}] \quad \forall \boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D), \tag{3.50}$$

where $\boldsymbol{\Gamma}(t)$ *is a geodesic that points in the direction described by* $\boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$*. Invoking the expressions of the tangent space, the gradient on the Grassmannian is related to the classical gradient of* f *thanks to the following expression:*

$$\text{grad } f(\mathbf{X}) = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T) \nabla_{\mathbf{X}} f(\mathbf{X}). \tag{3.51}$$

Proof of (3.51). The expression of the Grassmann gradient is obtained differentiating $f(\boldsymbol{\Gamma}(t))$. Using the chain rule [147], this derivative yields:

$$\left. \frac{\mathrm{d} f(\boldsymbol{\Gamma}(t))}{\mathrm{d}t} \right|_{t=0} = \operatorname{tr} \left(\nabla_{\mathbf{X}} f^T(\boldsymbol{\Gamma}(t)) \frac{\mathrm{d}\boldsymbol{\Gamma}(t)}{\mathrm{d}t} \right) \Bigg|_{t=0}, \tag{3.52}$$

which, after plugging the derivative of the geodesics obtained in (2.74), yields:

$$\operatorname{tr} \left(\nabla_{\mathbf{X}} f^T(\boldsymbol{\Gamma}(t)) \frac{\mathrm{d}\boldsymbol{\Gamma}(t)}{\mathrm{d}t} \right) \Bigg|_{t=0} = \operatorname{tr} \left(\nabla_{\mathbf{X}} f^T(\mathbf{X}) \boldsymbol{\Delta} \right) = \operatorname{tr} \left(\nabla_{\mathbf{X}} f^T(\mathbf{X}) (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T) \boldsymbol{\Delta} \right), \tag{3.53}$$

where $\boldsymbol{\Gamma}(0) = \mathbf{X}$ and $\left. \frac{\mathrm{d}\boldsymbol{\Gamma}(t)}{\mathrm{d}t} \right|_{t=0} = \boldsymbol{\Delta}$ is any tangent direction at $\mathbf{X}$. The last expression is obtained thanks to the fact that $(\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\boldsymbol{\Delta} = \boldsymbol{\Delta}$ since $\boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$. Considering that:

$$\langle \text{grad } f(\mathbf{X}), \boldsymbol{\Delta} \rangle_{\mathbf{X}} = \operatorname{tr} \left((\text{grad } f(\mathbf{X}))^T \boldsymbol{\Delta} \right), \tag{3.54}$$

the only tangent vector, grad $f(\mathbf{X}) \in \mathcal{T}_{\mathbf{X}} \operatorname{Gr}(N, D)$, such that (3.53) and (3.54) are equivalent is grad $f(\mathbf{X}) = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T) \nabla f(\mathbf{X})$. ■

Remark 3.19. Although Definition 3.18 is particularized on the Grassmannian, the expression given in (3.50) is general for any differentiable manifold. In fact, equation (3.50) is the expression of the Gateaux differential for differentiable functions (in the classical sense) in the Euclidean space (see (3.9)).

Note that the properties of classical derivatives in optimization also apply for the Grassmann gradients, e.g. stationary points satisfy Definitions 3.7 and 3.8, and the negativeness or positiveness of (3.50) indicates that $\boldsymbol{\Delta}$ is a descent or crescent direction, respectively. In a similar manner, we also review the definition of the Hessian in the Grassmann manifold.

Definition 3.19 (Hessian of a function on the Grassmann manifold)**.** *The Hessian of a function* $f : \operatorname{Gr}(N, D) \to \mathbb{R}$ *at a point* $\mathbf{X}$ *on the Grassmann manifold (also referred to as the Riemannian Hessian) is defined with the following second derivative:*

$$\text{hess } f(\mathbf{X})[\boldsymbol{\Delta}, \boldsymbol{\Delta}] = \left. \frac{\mathrm{d}^2 f(\boldsymbol{\Gamma}(t))}{\mathrm{d}t^2} \right|_{t=0}, \tag{3.55}$$

where $\boldsymbol{\Gamma}(t)$ *is a geodesic departing from* $\mathbf{X}$ *with direction* $\boldsymbol{\Delta}$*.*

Notice that, in contrast to the classical Hessian, the Riemannian Hessian is a quadratic form (scalar) of the tangent direction. If one were interested in computing the Hessian with respect to two different directions, hess $f(\mathbf{X})[\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2]$, the standard process of metric polarization should be used (see [174, Eq. (21)]). This aforementioned process is useful to compute the so-called *Intrinsic Cramér-Rao Bound* [175].

In light of the previous overview on g-convex sets and Riemannian calculus, linking the concepts of geodesics and convex functions results in the definition of g-convex functions.

Definition 3.20 (G-convex functions)**.** *Let* $f : \mathcal{G} \to \mathbb{R}$ *be an arbitrary function, where* $\mathcal{G} \subseteq \operatorname{Gr}(N, D)$ *is a g-convex subset of the Grassmannian. Then, it is said to be g-convex in* $\mathcal{G}$ *if the following condition is satisfied:*

$$f(\boldsymbol{\Gamma}(t)) \leq (1-t) f(\mathbf{X}) + t f(\mathbf{Y}) \quad \forall t \in [0, 1], \forall \mathbf{X}, \mathbf{Y} \in \mathcal{G}, \tag{3.56}$$

where $\boldsymbol{\Gamma}(t)$ *is the geodesic that connects* $\mathbf{X}$ *and* $\mathbf{Y}$*. The particularization to the Grassmann manifold comes from the expression of the geodesic given in Definition 2.12.*

Remark 3.20. A convex function is a particular case of g-convexity with respect to the geodesic described by the convex combination between two points (see (3.23)).
Remark 3.21. The g-convexity of a function can be seen as a generalization of Remark 3.10. In other words, if $f(\mathbf{\Gamma}(t))$ is convex, then f is g-convex.
Remark 3.22. Unfortunately, it is known that a smooth function that is g-convex on a compact Riemannian manifold is a constant function (see [44, Remark 5.6] and references therein). Nevertheless, a function that is locally g-convex in a given neighbourhood, e.g. a g-convex ball on the Grassmann manifold (see Lemma 3.4), is well-behaved for optimization, as it is shown in the sequel. For simplicity and unless otherwise stated, we continue the analysis of g-convex optimization assuming that functions are globally g-convex.

In a similar manner to convex functions, there are first and second-order characterizations of g-convex functions. They are specially useful for g-convex functions since they are powerful tools to analyze (locally) g-convex functions. These characterizations are surveyed in the following theorems.
Theorem 3.5 (First-order characterization of g-convex functions)**.** *Let $\mathcal{G}$ be a g-convex subset of the Grassmannian. A g-convex differentiable function $f : \mathcal{G} \to \mathbb{R}$ satisfies:*

$$f(\mathbf{Y}) \geq f(\mathbf{X}) + \langle \text{grad } f(\mathbf{X}), (\mathbf{I}_N - \mathbf{X}\mathbf{X})^T\mathbf{Y}\rangle_{\mathbf{X}} \quad \forall \mathbf{X}, \mathbf{Y} \in \mathcal{G}. \tag{3.57}$$

Remark 3.23. Theorem 3.5 is particularized for g-concave functions by reversing the sign of the previous inequality.
Remark 3.24. Note that the result in (3.57) is a particularization of the following expression for g-convex functions [7], [203]:

$$f(\mathbf{Y}) \geq f(\mathbf{X}) + \langle \text{grad } f(\mathbf{X}), \mathbf{\Delta}\rangle_{\mathbf{X}}, \tag{3.58}$$

for any $\mathbf{Y}, \mathbf{X} \in \mathcal{G}$ and $\mathbf{\Delta} \in \mathcal{T}_{\mathbf{X}} \text{Gr}(N, D)$.

Proof. Let a geodesic on $\mathcal{G}$ be such that $\mathbf{\Gamma}(0) = \mathbf{X}$ and $\mathbf{\Gamma}(1) = \mathbf{Y}$. After rearranging terms in (3.56) and dividing by t both sides, we obtain:

$$f(\mathbf{Y}) \geq f(\mathbf{X}) + \frac{f(\mathbf{\Gamma}(t)) - f(\mathbf{X})}{t}, \tag{3.59}$$

from where, after taking the limit for $t \to 0$ and invoking equation (3.50), we get:

$$\lim_{t \to 0} \frac{f(\mathbf{\Gamma}(t)) - f(\mathbf{X})}{t} = \frac{\mathrm{d}f(\mathbf{\Gamma}(t))}{\mathrm{d}t}\bigg|_{t=0} = \langle \text{grad } f(\mathbf{X}), \mathbf{\Delta}\rangle_{\mathbf{X}}, \tag{3.60}$$

for any tangent direction $\mathbf{\Delta} \in \mathcal{T}_{\mathbf{X}} \text{Gr}(N, D)$ pointing to $\mathbf{Y}$. Intuitively, we can choose $\mathbf{\Delta} = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{Y}$. Plugging (3.60) into (3.59) yields the desired result. ■

Notice that, in (3.57), $(\mathbf{I}_N - \mathbf{X}\mathbf{X})^T\mathbf{Y}$ computes a tangent direction from $\mathbf{X}$ to $\mathbf{Y}$ with no regard for the *step length*, as it is not needed to obtain the lower bound in (3.57). A more formal lower bound would require the logarithmic map (see (2.80) or (2.81)), which would yield a similar expression to the one in (3.57). In the following theorem we generalize Theorem 3.2 using Definition 3.19 [192].
Theorem 3.6 (Second-order characterization of g-convex functions)**.** *Let $\mathcal{G}$ be a g-convex subset of the Grassmannian. A g-convex differentiable function $f : \mathcal{G} \to \mathbb{R}$ satisfies:*

$$\text{hess } f(\mathbf{X})[\mathbf{\Delta}, \mathbf{\Delta}] \geq 0 \quad \forall \mathbf{X} \in \mathcal{G}, \forall \mathbf{\Delta} \in \mathcal{T}_{\mathbf{X}} \text{Gr}(N, D). \tag{3.61}$$

Remark 3.25. G-concave functions would yield a negative Hessian.

Proof. [192] Let a geodesic on $\mathcal{G}$ be such that $\mathbf{\Gamma}(0) = \mathbf{X}$ and $\mathbf{\Gamma}(1) = \mathbf{Y}$. Also, let $\gamma_f(t) = f(\mathbf{\Gamma}(t))$. From Definition 3.20, we get:

$$f(\mathbf{\Gamma}(t)) \leq (1-t)f(\mathbf{X}) + tf(\mathbf{Y}), \tag{3.62a}$$

$$\gamma_f(t) \leq (1-t)\gamma_f(0) + t\gamma_f(1), \tag{3.62b}$$

since $f(\mathbf{X})$ and $f(\mathbf{Y})$ are equivalent to $\gamma_f(0)$ and $\gamma_f(1)$, respectively. In other words, $\gamma_f(t)$ is a convex function with respect to t. Hence, from Theorem 3.2 we get that:

$$\frac{\mathrm{d}^2\gamma_f(t)}{\mathrm{d}t^2} \geq 0, \tag{3.63}$$

from which, after particularizing the previous expression for $t = 0$, we get (3.61). ■

Similarly to the convex case and with the same motivations in mind, g-convex functions are relaxed to g-quasiconvex functions [203] in the following definition.

Definition 3.21 (G-quasiconvex functions)**.** *Any function $f : \mathcal{G} \to \mathbb{R}$ is said to be g-quasiconvex in the g-convex subset $\mathcal{G} \subseteq \mathrm{Gr}(N, D)$ with respect to the geodesic $\mathbf{\Gamma}(t)$ if it satisfies:*

$$f(\mathbf{\Gamma}(t)) \leq \max(f(\mathbf{X}), f(\mathbf{Y})) \ \forall \mathbf{X}, \mathbf{Y} \in \mathcal{G}. \tag{3.64}$$

Remark 3.26. If, for every possible geodesic in $\mathcal{G}$, the sublevel set defined as:

$$\mathcal{S}(f, \alpha) = \{t \in [0, 1] : f(\mathbf{\Gamma}(t)) \leq \alpha\}, \tag{3.65}$$

is convex, then f is g-quasiconvex.

Definition 3.22 (G-quasiconcave functions)**.** *Any function $f : \mathcal{G} \to \mathbb{R}$ is said to be g-quasiconcave in the g-convex subset $\mathcal{G} \subseteq \mathrm{Gr}(N, D)$ with respect to the geodesic $\mathbf{\Gamma}(t)$ if it satisfies:*

$$f(\mathbf{\Gamma}(t)) \geq \min(f(\mathbf{X}), f(\mathbf{Y})) \ \forall \mathbf{X}, \mathbf{Y} \in \mathcal{G}. \tag{3.66}$$

Remark 3.27. If, for every possible geodesic in $\mathcal{G}$, the supralevel set defined as:

$$\mathcal{S}'(f, \alpha) = \{t \in [0, 1] : f(\mathbf{\Gamma}(t)) \geq \alpha\}, \tag{3.67}$$

is convex, then f is g-quasiconcave.

Remark 3.28. G-quasiconvexity and g-quasiconcavity are particular cases of path-connected functions (see [92, Section III-A]).

Even though the classical definitions of quasiconvexity or quasiconcavity are based on the convexity of level set (and the Grassmann manifold admits a similar idea), we prefer the above definition to characterize g-quasiconvex functions because the intuition behind them is much easier to grasp in practice.

3.2.2.1 Toy problem: Riemannian perspective on Principal Component Analysis

In order to provide a use case of g-convex optimization, we show a geometric interpretation of the well-known PCA [97], which is based on a trace maximization problem with orthogonality constraints, in this subsection. Although the solution to this problem is well known in the literature, the assessment of its geometric structure is not known. In this manner, we exemplify the additional geometric insights that g-convex optimization provides to a classical optimization problem. Given that the Matrix Factorization problem is a particular case of the PCA problem, the following rationale is our proposed generalization to the ideas presented in [7]. Let us consider the following trace minimization problem:

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} \frac{1}{2} \mathrm{tr}(\mathbf{X}^T \mathbf{S} \mathbf{X}) \quad \text{s. t. } \mathbf{X} \in \mathrm{Gr}(N, D), \tag{3.68}$$

where $\mathbf{S} \in \mathcal{S}_+^N$ is usually the covariance matrix (or some estimation of it) of some signal and the $\frac{1}{2}$ scaling is to derive cleaner expressions of gradients and Hessians. Note that (3.68) is a non-convex problem (in the classical sense) due to two reasons: the constraint set, $\mathrm{Gr}(N, D)$, is non-convex and the optimization problem is a maximization of a convex function. That being said, we are interested in the study of the Riemannian gradient and Hessian of the cost function in (3.68) to provide a geometric interpretation of its stationary points. Additionally, we show the conditions in which (3.68) behaves (locally) as a g-convex optimization problem. Henceforward, we denote as $f_{PCA}(\mathbf{X}) = \frac{1}{2} \mathrm{tr}(\mathbf{X}^T \mathbf{S} \mathbf{X})$ the cost function in (3.68).

For the purpose of obtaining the stationary points of (3.68), let the gradient on the Grassmannian of $f_{PCA}(\mathbf{X})$ be:

$$\text{grad } f_{PCA}(\mathbf{X}) = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\nabla f_{PCA}(\mathbf{X}) = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{S}\mathbf{X}, \tag{3.69}$$

where $\nabla_{\mathbf{X}} f_{PCA}(\mathbf{X}) = \mathbf{S}\mathbf{X}$ is the classical gradient of $f_{PCA}(\mathbf{X})$. In order to unveil its geometric properties, we reformulate the previous Riemannian gradient in terms of the principal angles between $\mathbf{X}$ and the signal subspace of $\mathbf{S}$, i.e. the eigenvectors corresponding to the largest D eigenvalues of $\mathbf{S}$.

Without any loss of generality and with the objective of deriving clean expressions, we assume that $D > N - D$. In this regard, optimizing with a variable constrained in $\text{Gr}(N, D)$ or $\text{Gr}(N, N - D)$ is equivalent (assuming that the dimensions of the matrix multiplications match in both cases) due to the fact that $\text{Gr}(N, D)$ is diffeomorfic[2] to $\text{Gr}(N, N - D)$. As a result, one could always optimize with respect to $\text{Gr}(N, D)$ or $\text{Gr}(N, N - D)$ interchangeably. An alternative argument is based on the fact that estimating a subspace with D (signal subspace) or $N - D$ (noise subspace) dimensions yields an equivalent estimation error [175]. From a practical point of view, estimating the orthogonal space to the signal subspace in (3.68) would require a minimization problem instead of a maximization one, which is not restrictive for the rationale behind the geometric interpretation of the PCA problem shown in the next paragraphs. With the previous rationale in mind, let an alternative expression of the SVD of $\mathbf{S}$ be:

$$\mathbf{S} = \mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T + \mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T, \tag{3.70}$$

where the columns of $\mathbf{U}_s \in \mathbb{R}^{N\times D}$ contain the eigenvectors corresponding to the largest D eigenvalues of $\mathbf{S}$, which are contained on the diagonal matrix $\mathbf{D}_s \in \mathbb{R}^{D\times D}$. The remaining term in (3.70) consists of zero-padded matrices that contain the remaining eigenvectors and eigenvalues of $\mathbf{S}$. The zero-padding in $\mathbf{U}_n$ and $\mathbf{D}_n$ is such that their dimensions are equal to the ones of $\mathbf{U}_s$ and $\mathbf{D}_s$, respectively. In this regard, we denote as:

$$\mathbf{U}_n = \begin{bmatrix}\mathbf{0}_{N,2D-N} & \mathbf{U}_n'\end{bmatrix}, \tag{3.71}$$

where the columns of $\mathbf{U}_n' \in \mathbb{R}^{N\times N-D}$ contain the eigenvectors corresponding to the smallest $N - D$ eigenvalues of $\mathbf{S}$. Equivalently:

$$\mathbf{D}_n = \begin{bmatrix}\mathbf{0}_{2D-N,2D-N} & \mathbf{0}_{2D-N,N-D} \\ \mathbf{0}_{N-D,2D-N} & \mathbf{D}_n'\end{bmatrix}, \tag{3.72}$$

where $\mathbf{D}_n' \in \mathbb{R}^{N-D\times N-D}$ is the diagonal matrix containing the smallest $N - D$ eigenvalues of $\mathbf{S}$. The behaviour of (3.69) in terms of the principal angles between $\mathbf{X}$ and $\mathbf{U}_s$ is summarized in the following theorem.

Theorem 3.7 (First-order characterization of $f_{PCA}(\mathbf{X})$)**.** *Let the Riemannian gradient of $f_{PCA}(\mathbf{X})$ be given by (3.69) and let a decomposition of $\mathbf{S}$ be as in (3.70). Then, using aligned representatives of $\mathbf{X}$, $\mathbf{U}_s$ and $\mathbf{U}_n$, the Riemannian gradient of $f_{PCA}(\mathbf{X})$ in the Grassmann manifold is given by:*

$$\begin{aligned}\text{grad } f_{PCA}(\mathbf{X}) = \boldsymbol{\Delta}_s \sin(\boldsymbol{\Theta})\mathbf{V}_s^T\mathbf{D}_s\mathbf{V}_s\cos(\boldsymbol{\Theta}) + \boldsymbol{\Delta}_n\cos(\boldsymbol{\Theta})\mathbf{V}_n^T\mathbf{D}_n\mathbf{V}_n\sin(\boldsymbol{\Theta}) = \\ \boldsymbol{\Delta}_s\sin(\boldsymbol{\Theta})\boldsymbol{\Sigma}_s\cos(\boldsymbol{\Theta}) + \boldsymbol{\Delta}_n\cos(\boldsymbol{\Theta})\boldsymbol{\Sigma}_n\sin(\boldsymbol{\Theta}),\end{aligned} \tag{3.73}$$

where $\mathbf{V}_s, \mathbf{V}_n \in \mathbb{R}^{D\times D}$ are orthonormal matrices, $\boldsymbol{\Delta}_s \in \mathcal{T}_{\mathbf{X}}\,\text{Gr}(N, D)$, $\boldsymbol{\Delta}_n = \begin{bmatrix}\mathbf{0}_{N,2D-N} & \boldsymbol{\Delta}_n'\end{bmatrix}$, $\boldsymbol{\Delta}_n' \in \mathbb{R}^{N\times N-D}$ is an orthonormal matrix such that $\boldsymbol{\Delta}_n \in \mathcal{T}_{\mathbf{X}}\,\text{Gr}(N, D)$, $\boldsymbol{\Sigma}_s = \mathbf{V}_s^T\mathbf{D}_s\mathbf{V}_s$ and $\boldsymbol{\Sigma}_n = \mathbf{V}_n^T\mathbf{D}_n\mathbf{V}_n$.

Remark 3.29. The values of $\boldsymbol{\Theta}$ such that grad $f_{PCA}(\mathbf{X}) = \mathbf{0}_{N,D}$ are:

$$[\boldsymbol{\Theta}]_{d,d} = 0 \;\text{ or }\; [\boldsymbol{\Theta}]_{d,d} = \frac{\pi}{2} \;\; \forall d = 1, ..., D. \tag{3.74}$$

Therefore, any $\mathbf{X}$ whose columns consist of a subset of the N eigenvectors of $\mathbf{S}$ is a stationary point of (3.68).

[2]One could always find a one to one mapping between a subspace and its orthogonal complement. Thus, they are *equivalent* manifolds.

Proof. Firstly, we show that (3.73) is derived from:

$$\text{grad } f_{PCA}(\mathbf{X}) = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\mathbf{S}\mathbf{X} = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)(\mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T + \mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T)\mathbf{X}. \tag{3.75}$$

Invoking the principal alignment principle from Lemma 2.4 between $\mathbf{X}$ and $\mathbf{U}_s$ and the complementarity of the principal angles from Lemma 2.5, we get the following expressions:

$$\mathbf{U}_s^T\mathbf{X}_a = \mathbf{V}_s^T\cos(\mathbf{\Theta}), \tag{3.76a}$$

$$\mathbf{U}_n^T\mathbf{X}_a = \mathbf{V}_n^T\sin(\mathbf{\Theta}), \tag{3.76b}$$

$$(\mathbf{I} - \mathbf{X}_a\mathbf{X}_a^T)\mathbf{U}_s = \mathbf{\Delta}_s\sin(\mathbf{\Theta})\mathbf{V}_s, \tag{3.76c}$$

$$(\mathbf{I} - \mathbf{X}_a\mathbf{X}_a^T)\mathbf{U}_n = \mathbf{\Delta}_n\cos(\mathbf{\Theta})\mathbf{V}_n, \tag{3.76d}$$

where $\mathbf{V}_s, \mathbf{V}_n \in \mathrm{O}(D)$ are rotation matrices such that $\mathbf{U}_s$ and $\mathbf{U}_n$ are aligned with $\mathbf{X}_a$. The aligned logarithm map defined in (2.81) is used to derive (3.76c) and (3.76d). With (3.76) in mind, the first term in (3.75) yields:

$$(\mathbf{I} - \mathbf{X}\mathbf{X}^T)\mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T\mathbf{X} = \mathbf{\Delta}_s\sin(\mathbf{\Theta})\mathbf{V}_s^T\mathbf{D}_s\mathbf{V}_s\cos(\mathbf{\Theta}), \tag{3.77}$$

whereas the second term is given by:

$$(\mathbf{I} - \mathbf{X}\mathbf{X}^T)\mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T\mathbf{X} = \mathbf{\Delta}_n\cos(\mathbf{\Theta})\mathbf{V}_n^T\mathbf{D}_n\mathbf{V}_n\sin(\mathbf{\Theta}). \tag{3.78}$$

Lastly, the final expression of grad $f_{PCA}(\mathbf{X})$ is obtained by the summation of (3.77) with (3.78). ■

Notice that in the previous proof we can see the motivation behind the particular case of $D > N - D$ and the zero-padding in $\mathbf{U}_n$ and $\mathbf{D}_n$. In this regard, it is thanks to the aforementioned zero-padding and to the ascending ordering of $\mathbf{\Theta}$ (see Definition 2.11) that we are able to write (3.77) and (3.78) in terms of $\mathbf{\Theta}$ since the zero-padded columns coincide with the $\lfloor D - \frac{N}{2} \rfloor$ trivial principal angles entries in $\cos(\mathbf{\Theta})$, i.e. $[\mathbf{\Theta}]_{d,d} = 0$. In simpler words, the trivial principal angles are those that result from the intersection between two subspaces, which is always non-empty when $D > N - D$.

The above theorem shows that the stationary points of $f_{PCA}(\mathbf{X})$ consist of any subset of D eigenvectors of $\mathbf{S}$. In other words, any subset of D eigenvectors of $\mathbf{S}$ can be either a local maximum, a local minimum or a saddle point. However, Theorem 3.7 does not certify which subset of eigenvectors is the optimal solution of the optimization problem given in (3.68). In light of the previous observation, we want to verify the (at least local) g-concavity of the cost function of the problem in (3.68) since it is one of the possible ways of assessing whether a given point is a (local) maximum of the original cost. For the previous reason, the following theorem formalizes the local g-concavity of $f_{PCA}(\mathbf{X})$ around $\mathbf{U}_s$.

Theorem 3.8 (Second-order characterization of $f_{PCA}(\mathbf{X})$)**.** *Let an optimization problem be defined as in (3.68). Then, the Riemannian Hessian of $f_{PCA}(\mathbf{X})$ yields:*

$$\begin{aligned}\text{hess } f_{PCA}(\mathbf{X})[\mathbf{\Delta}, \mathbf{\Delta}] = &- \text{tr}(\cos(\mathbf{\Theta})\mathbf{\Sigma}_s\cos(\mathbf{\Theta})\mathbf{\Delta}^T\mathbf{\Delta}) + \text{tr}(\mathbf{\Delta}^T\mathbf{\Delta}_s\sin(\mathbf{\Theta})\mathbf{\Sigma}_s\sin(\mathbf{\Theta})\mathbf{\Delta}_s^T\mathbf{\Delta}) \\ &- \text{tr}(\sin(\mathbf{\Theta})\mathbf{\Sigma}_n\sin(\mathbf{\Theta})\mathbf{\Delta}^T\mathbf{\Delta}) + \text{tr}(\mathbf{\Delta}^T\mathbf{\Delta}_n\cos(\mathbf{\Theta})\mathbf{\Sigma}_n\cos(\mathbf{\Theta})\mathbf{\Delta}_n^T\mathbf{\Delta}),\end{aligned} \tag{3.79}$$

where $\mathbf{\Delta}$ is an arbitrary direction in $\mathcal{T}_{\mathbf{X}}\,\mathrm{Gr}(N, D)$ and the remaining parameters are defined in Proposition 3.7. Moreover, the Riemannian Hessian satisfies:

$$\text{hess } f_{PCA}(\mathbf{X})[\mathbf{\Delta}, \mathbf{\Delta}] \leq 0, \tag{3.80}$$

for any $\mathbf{X} \in B_{\frac{\pi}{4}}(\mathbf{U}_s)$. Hence, $f_{PCA}(\mathbf{X})$ is locally g-concave on $\mathbf{X} \in B_{\frac{\pi}{4}}(\mathbf{U}_s)$.

Remark 3.30. The result given in (3.80), in addition to Remark 3.29, implies that $\mathbf{U}_s$ is a maximum point (at least a local one) of $f_{PCA}(\mathbf{X})$.

Proof. Firstly we derive the Riemannian Hessian in terms of the principal angles between $\mathbf{X}$ and $\mathbf{U}_s$, and secondly we show which values of the principal angles satisfy (3.80). The Hessian of $f_{PCA}(\mathbf{X})$ on the Grassmann manifold is obtained from the following expression:

$$\text{hess } f_{PCA}(\mathbf{X})[\boldsymbol{\Delta},\boldsymbol{\Delta}] = \left.\frac{\mathrm{d}^2 f_{PCA}(\boldsymbol{\Gamma}(t))}{\mathrm{d}t^2}\right|_{t=0}, \tag{3.81}$$

where $\boldsymbol{\Gamma}(t)$ is a Grassmann geodesic as in (2.72). Notice that $f_{PCA}(\boldsymbol{\Gamma}(t))$ can be rewritten as $f_{PCA}(\boldsymbol{\Gamma}(t)) = \text{tr}\left(\mathbf{S}\boldsymbol{\Gamma}(t)\boldsymbol{\Gamma}^T(t)\right)$, implying that the Riemannian Hessian can be obtained by, firstly, differentiating $\boldsymbol{\Gamma}(t)\boldsymbol{\Gamma}^T(t)$ two times and then plugging the resulting expression into $f_{PCA}(\boldsymbol{\Gamma}(t))$. Differentiating $\boldsymbol{\Gamma}(t)\boldsymbol{\Gamma}^T(t)$ two times yields:

$$\left.\frac{\mathrm{d}^2\boldsymbol{\Gamma}(t)\boldsymbol{\Gamma}^T(t)}{\mathrm{d}t^2}\right|_{t=0} = -2\mathbf{X}\boldsymbol{\Delta}^T\boldsymbol{\Delta}\mathbf{X}^T + 2\boldsymbol{\Delta}\boldsymbol{\Delta}^T, \tag{3.82}$$

from where, given the shown decomposition of $\mathbf{S}$ (see (3.70)) and after plugging (3.82) into $f_{PCA}(\boldsymbol{\Gamma}(t)) = \text{tr}\left(\mathbf{S}\boldsymbol{\Gamma}(t)\boldsymbol{\Gamma}^T(t)\right)$, we get the initial expression of the Riemannian Hessian:

$$\begin{aligned}\text{hess } f_{PCA}(\mathbf{X})[\boldsymbol{\Delta},\boldsymbol{\Delta}] = &-\text{tr}(\mathbf{X}^T\mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T\mathbf{X}\boldsymbol{\Delta}^T\boldsymbol{\Delta}) + \text{tr}(\boldsymbol{\Delta}^T\mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T\boldsymbol{\Delta})\\ &-\text{tr}(\mathbf{X}^T\mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T\mathbf{X}\boldsymbol{\Delta}^T\boldsymbol{\Delta}) + \text{tr}(\boldsymbol{\Delta}^T\mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T\boldsymbol{\Delta}).\end{aligned} \tag{3.83}$$

With a view of assessing the geodesic concavity of $f_{PCA}(\mathbf{X})$ from the previous Hessian, we want to rewrite (3.83) as a function of the principal angles between $\mathbf{X}$ and $\mathbf{U}_s$, denoted as $\boldsymbol{\Theta}$. Importing (3.76) from Theorem 3.7 into (3.83) results in the following expression:

$$\begin{aligned}\text{hess } f_{PCA}(\mathbf{X})[\boldsymbol{\Delta},\boldsymbol{\Delta}] = &-\text{tr}(\cos(\boldsymbol{\Theta})\boldsymbol{\Sigma}_s\cos(\boldsymbol{\Theta})\boldsymbol{\Delta}^T\boldsymbol{\Delta}) + \text{tr}(\boldsymbol{\Delta}^T\boldsymbol{\Delta}_s\sin(\boldsymbol{\Theta})\boldsymbol{\Sigma}_s\sin(\boldsymbol{\Theta})\boldsymbol{\Delta}_s^T\boldsymbol{\Delta})\\ &-\text{tr}(\sin(\boldsymbol{\Theta})\boldsymbol{\Sigma}_n\sin(\boldsymbol{\Theta})\boldsymbol{\Delta}^T\boldsymbol{\Delta}) + \text{tr}(\boldsymbol{\Delta}^T\boldsymbol{\Delta}_n\cos(\boldsymbol{\Theta})\boldsymbol{\Sigma}_n\cos(\boldsymbol{\Theta})\boldsymbol{\Delta}_n^T\boldsymbol{\Delta}),\end{aligned} \tag{3.84}$$

where we have used the fact that $(\mathbf{I}_N - \mathbf{X}\mathbf{X}^T)\boldsymbol{\Delta} = \boldsymbol{\Delta}$ (since it is a tangent direction at $\mathbf{X}$) to obtain the positive terms.

In order to obtain the second conclusion of this theorem in (3.80), we need to further parse (3.84) and obtain for which values of $\boldsymbol{\Theta}$ the Hessian matrix of $f_{PCA}(\mathbf{X})$ is negative. Following a similar rationale as in [7], the above expression can be further simplified rewriting the direction $\boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{X}}\,\text{Gr}(N,D)$ as $\tilde{\boldsymbol{\Delta}}\sin(\boldsymbol{\Phi})$, where $\tilde{\boldsymbol{\Delta}}$ is orthonormal and $\boldsymbol{\Phi}$ is a diagonal matrix whose entries are bounded in $[0, \frac{\pi}{2}]$. This reformulation is inspired by the second equation from (2.81), which is a possible approach to depict any arbitrary direction to any point of the Grassmann manifold. Then, (3.84) is further rewritten as:

$$\begin{aligned}\text{hess } f_{PCA}(\mathbf{X})[\tilde{\boldsymbol{\Delta}}\sin(\boldsymbol{\Phi}),\tilde{\boldsymbol{\Delta}}\sin(\boldsymbol{\Phi})] = &-\text{tr}(\sin^2(\boldsymbol{\Phi})\cos^2(\boldsymbol{\Theta})\boldsymbol{\Sigma}_s)\\ &+\text{tr}(\sin(\boldsymbol{\Theta})\boldsymbol{\Delta}_s^T\tilde{\boldsymbol{\Delta}}\sin^2(\boldsymbol{\Phi})\tilde{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}_s\sin(\boldsymbol{\Theta})\boldsymbol{\Sigma}_s)\\ &-\text{tr}(\sin^2(\boldsymbol{\Phi})\sin^2(\boldsymbol{\Theta})\boldsymbol{\Sigma}_n)\\ &+\text{tr}(\cos(\boldsymbol{\Theta})\boldsymbol{\Delta}_n^T\tilde{\boldsymbol{\Delta}}\sin^2(\boldsymbol{\Phi})\tilde{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}_n\cos(\boldsymbol{\Theta})\boldsymbol{\Sigma}_n),\end{aligned} \tag{3.85}$$

where we used the product commutativity of diagonal matrices and the circularity of the trace to obtain the previous expression of the negative terms. To show the negativity of the Hessian, we resort to an upper bound of (3.85). This upper bound is derived by noting that the matrices inside the traces of the positive terms in (3.85) are majorized as follows:

$$\sin(\boldsymbol{\Theta})\boldsymbol{\Delta}_s^T\tilde{\boldsymbol{\Delta}}\sin^2(\boldsymbol{\Phi})\tilde{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}_s\sin(\boldsymbol{\Theta}) \preceq \frac{1}{2}\sin^2(\boldsymbol{\Phi}), \tag{3.86a}$$

$$\cos(\boldsymbol{\Theta})\boldsymbol{\Delta}_n^T\tilde{\boldsymbol{\Delta}}\sin^2(\boldsymbol{\Phi})\tilde{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}_n\cos(\boldsymbol{\Theta}) \preceq \frac{1}{2}\sin^2(\boldsymbol{\Phi}), \tag{3.86b}$$

which follows from:

$$\sin(\boldsymbol{\Theta}) \preceq \frac{1}{\sqrt{2}}\mathbf{I}_D, \tag{3.87a}$$

$$\cos(\boldsymbol{\Theta}) \preceq \frac{1}{\sqrt{2}}\mathbf{I}_D, \tag{3.87b}$$

for $\boldsymbol{\Theta} \preceq \frac{\pi}{4}\mathbf{I}_D$, and:

$$\tilde{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}_s \preceq \mathbf{I}_D, \tag{3.88a}$$

$$\tilde{\boldsymbol{\Delta}}^T\boldsymbol{\Delta}_n \preceq \mathbf{I}_D, \tag{3.88b}$$

since $\tilde{\boldsymbol{\Delta}}$, $\boldsymbol{\Delta}_s$ and $\boldsymbol{\Delta}_n$ are orthonormal. The justification of the last observation is a result of the fact that the eigenvalues of the identity matrix (which are all equal to 1) are greater or equal to the eigenvalues of the product of two $N \times D$ orthogonal matrices. As a result, (3.85) is upper bounded by:

$$\begin{aligned}\text{hess } f_{PCA}(\mathbf{X})[\boldsymbol{\Delta},\boldsymbol{\Delta}] \leq &-\operatorname{tr}(\sin^2(\boldsymbol{\Phi})\cos^2(\boldsymbol{\Theta})\boldsymbol{\Sigma}_s) + \operatorname{tr}\left(\frac{1}{2}\sin^2(\boldsymbol{\Phi})\boldsymbol{\Sigma}_s\right)\\ &- \operatorname{tr}(\sin^2(\boldsymbol{\Phi})\sin^2(\boldsymbol{\Theta})\boldsymbol{\Sigma}_n) + \operatorname{tr}\left(\frac{1}{2}\sin^2(\boldsymbol{\Phi})\boldsymbol{\Sigma}_n\right) = \end{aligned} \tag{3.89a}$$

$$\operatorname{tr}\left(\sin^2(\boldsymbol{\Phi})\left(\frac{1}{2}\mathbf{I}_D - \cos^2(\boldsymbol{\Theta})\right)\boldsymbol{\Sigma}_s\right) + \operatorname{tr}\left(\sin^2(\boldsymbol{\Phi})\left(\frac{1}{2}\mathbf{I}_D - \sin^2(\boldsymbol{\Theta})\right)\boldsymbol{\Sigma}_n\right) = \tag{3.89b}$$

$$\operatorname{tr}\left(\sin^2(\boldsymbol{\Phi})\left(\frac{1}{2}\mathbf{I}_D - \sin^2(\boldsymbol{\Theta})\right)(\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_s)\right) \leq 0, \tag{3.89c}$$

where the left hand side of (3.89c) is obtained thanks to the following expression:

$$\frac{1}{2}\mathbf{I}_D - \cos^2(\boldsymbol{\Theta}) = \frac{1}{2}\mathbf{I}_D - (\mathbf{I}_D - \sin^2(\boldsymbol{\Theta})) = -\frac{1}{2}\mathbf{I}_D + \sin^2(\boldsymbol{\Theta}) = -\left(\frac{1}{2}\mathbf{I}_D - \sin^2(\boldsymbol{\Theta})\right). \tag{3.90}$$

The upper bound in (3.89c) follows from $\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_s \preceq \mathbf{0}_{D,D}$ (see the definition of $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_n$ in Theorem 3.7) and from the fact that $\left(\frac{1}{2}\mathbf{I}_D - \sin^2(\boldsymbol{\Theta})\right) \succeq \mathbf{0}_{D,D}$ as long as the principal angles are all bounded in $[0, \frac{\pi}{4}]$. As a summary, the final expression found in (3.89) implies that the Hessian of $f_{PCA}(\mathbf{X})$ is non-positive for $\boldsymbol{\Theta} \preceq \frac{\pi}{4}\mathbf{I}_D$, which is the definition of $B_{\frac{\pi}{4}}(\mathbf{U}_s)$. ■

Theorem 3.8 only assesses that $\mathbf{U}_s$ is a local (possibly global) maximum point of the original problem in (3.68). In fact, it is the optimal value of (3.68) since:

$$f_{PCA}(\mathbf{U}_s) = \operatorname{tr}(\mathbf{U}_s^T\mathbf{S}\mathbf{U}_s) = \operatorname{tr}\left(\mathbf{U}_s^T(\mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T + \mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T)\mathbf{U}_s\right) = \operatorname{tr}(\mathbf{D}_s), \tag{3.91}$$

which is the sum of the D largest singular values, i.e. the known PCA solution [27], [97]. In this regard, Theorem 3.8 complements the previous result since it characterizes to which extent the PCA cost function behaves in a g-concave manner around $\mathbf{U}_s$. Actually, the previous observation proves to be useful in the iterative optimization schemes involving a Grassmann constrained variable that are shown in the sequel. Unfortunately and up to the authors knowledge, the remaining stationary points of the PCA cost function are too difficult to characterize due to the complex structure of the Hessian given in (3.79), contrary to the results shown in [7, Corollary 6]. This is the reason why we had to show the optimality of $\mathbf{U}_s$ with the intuitive derivation given in (3.91), in addition to theorems 3.7 and 3.8.

3.3 Majorization-minimization framework

The Majorization-Minimization (MM) framework [179], also known as the Successive Upper-bound Minimization (SUM) [162] framework, is an approach to construct optimization algorithms that are capable of providing a solution to both convex and non-convex problems. In essence, the goal of an MM algorithm is to obtain a stationary point of the original problem by means of successive optimizations of a surrogate function which majorizes (or minorizes in the case of maximization problems) the

original cost. Particularly, this surrogate is a tight upper bound (lower bound for maximization problems), which is constructed using the current iterate, of the original cost. The main advantage of this rationale is that the resulting sequential steps for optimization are easier than the original problem since the surrogates are designed in such a way that its optimal value is retrieved with ease (not necessarily with a closed-form solution). In this regard, we are also especially interested in the block extension of the MM framework, which consists in the combination of the Block Coordinate Descent (BCD) [154], [187] approach with the MM methodology. In essence, the block extension (or also, block relaxation) of the MM partitions the optimization variables into several blocks of variables and applies the MM methodology to a single block while keeping the values of the remaining blocks fixed [92], [179, Subsection II-B]. The previous procedure provides the MM framework of an additional desired properties that are highlighted throughout this section. In addition to the review of the previous ideas, the main goal of this section is to extend the MM framework and its block extension to variables constrained in the Grassmann manifold. As an additional remark and for clarity in the exposition, the MM framework detailed in this section is centered around minimization problems. Yet, it can be extended for maximization problems by reversing the sing of the inequalities and by the substitution of convex functions by concave functions (and viceversa). The previous steps result in a *Minorization-Maximization* framework [179].

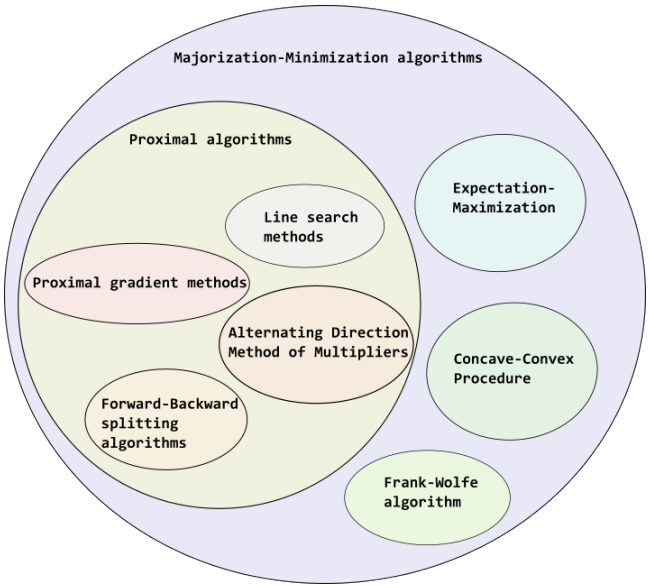


Figure 3.3: Relationship between several algorithmic frameworks that lie within the MM framework.

We show examples of several known algorithmic frameworks that can be encompassed within the MM methodology in Figure 3.3, where it is observed that the majority of those frameworks are gathered in the family of Proximal algorithms. While we address most of the algorithms that are depicted in Figure 3.3 (especially the proximal algorithms family), there are some of them that are not properly addressed in this dissertation since they are a combination of other algorithmic frameworks. Particularly, the Forward-Backward Splitting Algorithms (FBSA) [73] integrates the Gradient Descent framework [117], [169] with a *proximal iteration* [145], which are both extensively treated in subsections 3.3.3.2 and 3.3.5, respectively. Thus, delving into the ideas behind the FBSA is redundant.

Before the introduction of the MM framework, let us define the *global convergence* (not to be confused with the global optimum of a function) of an iterative optimization algorithm for a proper analysis of this general iterative optimization scheme.

Definition 3.23 (Global convergence of an iterative algorithm)**.** *Let an optimization problem be defined as:*

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X}, \tag{3.92}$$

and a sequence of iterates $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$ *be generated by an iterative optimization algorithm. Then, it is said*

that the given optimization algorithm is globally convergent if the previous sequence satisfies:

$$\lim_{i\to\infty} ||\mathbf{x}_{i+1} - \mathbf{x}_i||_2^2 = 0, \tag{3.93}$$

and:

$$\lim_{i\to\infty} f(\mathbf{x}_i) = f(\mathbf{z}), \tag{3.94}$$

where $\mathbf{z}$ *is a stationary point of (3.92) as in Definitions 3.7 or 3.8, for every initialization point* $\mathbf{x}_0$*. In other words,* $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$ *is a sequence convergent to a stationary point of (3.92).*

Remark 3.31. In simpler words, an optimization scheme is globally convergent if it converges to a stationary point for any initialization point. Since it may be confusing at first glance, we remark that the global convergence idea is not equivalent to a *global optimization method*, which require that the point of convergence is a global optimum (minimum or maximum) of the original cost.

Remark 3.32. For variables constrained in the Grassmann manifold, (3.93) becomes:

$$\lim_{i\to\infty} d_{arc}(\mathbf{X}_{i+1}, \mathbf{X}_i) = 0, \tag{3.95}$$

where $d_{arc}(\cdot,\cdot)$ is defined in (2.77).

3.3.1 The MM algorithm

Consider the following optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X}, \tag{3.96}$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is a continuous function and $\mathcal{X}$ is a closed convex set. Applying an MM algorithm to (3.96) would generate a sequence of feasible points, $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$, obtained from the optimization of a surrogate function of f (also referred to as the majorant function [92]). The majorant function, denoted as $g(\cdot|\mathbf{x}_i) : \mathcal{X} \to \mathbb{R}$, is built using the i-th iterate, $\mathbf{x}_i$. Mathematically, the previous procedure is summarized using the following sequential optimization problem:

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} g(\mathbf{x}|\mathbf{x}_i) \quad \text{s.t. } \mathbf{x} \in \mathcal{X}. \tag{3.97}$$

The conditions in which the sequence generated by (3.97) is a globally convergent one are summarized as follows [87], [162], [179].

(A1) The surrogate must be a tight approximation of the original cost:

$$g(\mathbf{x}|\mathbf{x}) = f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X}. \tag{3.98}$$

(A2) The surrogate must majorize the original cost:

$$g(\mathbf{x}|\mathbf{y}) \geq f(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}. \tag{3.99}$$

(A3) The first directional derivatives of the surrogate and of the original cost must be equivalent at the current iterate. Using Gateaux differentials, the previous condition is depicted by the following equation:

$$Dg(\mathbf{x}|\mathbf{x}')[\boldsymbol{\delta}] = Df(\mathbf{x})[\boldsymbol{\delta}] \quad \forall \boldsymbol{\delta} \in \mathbb{R}^N \quad \text{s.t. } \mathbf{x} + \boldsymbol{\delta} \in \mathcal{X}, \tag{3.100}$$

for $\mathbf{x}' = \mathbf{x}$.

(A4) $g(\mathbf{x}|\mathbf{y})$ must be continuous on $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}$.

For illustration purposes, Figure 3.4 depicts the MM principle emanating from (A1)-(A4). The previous conditions only ensure that, if the sequence generated by (3.97) reaches a limit point, then it is a stationary point. In this sense, an additional assumption on the original problem in (3.96) that ensures the convergence to limit points of $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$ (and their existence) is necessary. In other words, the sequence $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$ must be compact. The coerciveness of $f(\mathbf{x})$ in $\mathcal{X}$ [162] (see Definition 3.5) and the compactness of $\mathcal{X}$ [87], [179] are two of the most common assumptions on the original problem

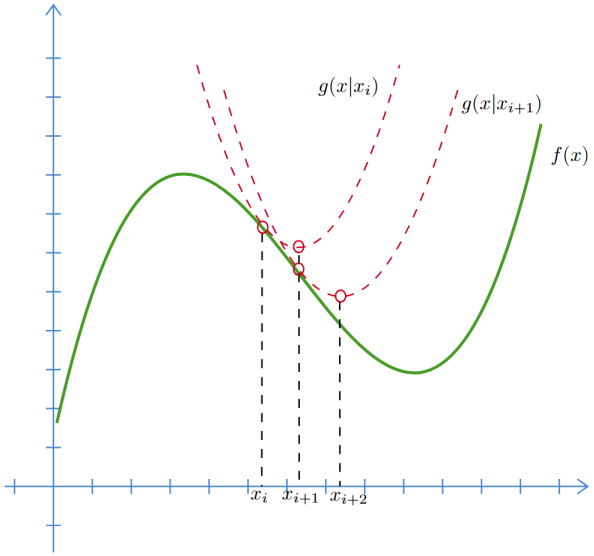


Figure 3.4: Example: Majorants of a function (red dashed lines). x_{i+1} and x_{i+2} are obtained by the minimization of the majorant functions.

that results in a compact sequence of iterates, $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$. Yet, we show in Section 4.4 from Chapter 4 that certain majorants force the compactness of the aforementioned sequence without the need of the previous two assumptions on $f(\mathbf{x})$ or $\mathcal{X}$. In light of the previous observations, we did not incorporate the previous ideas as an additional assumption on the MM framework since they are contingent on the specific scenario.

Regarding the convergence point, there are two possible outcomes of a convergent MM algorithm, which depends on the convexity of the optimization problem in (3.96). While an MM algorithm is capable of obtaining a global minimum of (3.96) if the original cost is convex, in the contrary case (non-convex f), the algorithm depicted by (3.97) is only guaranteed to converge to a stationary point without the verification of whether it is the global optimum or not. Still, this latter behaviour is enough in practical applications, where a fast convergence speed is much more valuable, especially in scalable applications, than the global optimality. Although a formal convergence proof of the algorithm described by (3.97) is found in [162], we remark that it can also be envisioned from the proof of Theorem 3.10 (by translating the Grassmann arguments to the Euclidean manifold).

3.3.2 MM block relaxation

Given that there are cases in which the feasible set can be expressed as the cartesian product of structured subsets, the MM framework is often combined with the BCD procedure to exploit the inherent structure of the feasible set. Indeed, this alternative provides even more flexibility on the design of the majorant functions, which often results in a faster convergence rate [179]. This improvement can be seen from the fact that constructing an specific majorant function for each block yields a potentially much better upper bound of the original cost than with the single block approach. In order to review the block relaxation of the MM algorithm, consider the following optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X}_1 \times ... \times \mathcal{X}_M, \tag{3.101}$$

where, again, $f : \mathbb{R}^N \to \mathbb{R}$ is a continuous and regular function. There are M different blocks of variables that are such that $\mathcal{X}_m \subseteq \mathbb{R}^{N_m}$ with $\sum_{m=1}^{M} N_m = N$. Consequently, the optimization variable in (3.101) can be decomposed as $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_M)$ where $\mathbf{x}_m \in \mathcal{X}_m$ for $m = 1, ..., M$. The previous decomposition is what enables the block relaxation in a similar way to the BCD [154], [187]. For clarity in the exposition, we write $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_M)$ and $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_M$ interchangeably when needed.

The MM block relaxation applied to (3.101) requires a majorant function for each block of variables. Thus, the successive update equations yield:

$$\mathbf{x}_{m,i+1} = \arg\min_{\mathbf{x}_m} g_m(\mathbf{x}_m|\mathbf{x}_1, ..., \mathbf{x}_M) \quad \text{s. t. } \mathbf{x}_m \in \mathcal{X}_m \ \forall m = 1, ..., M, \tag{3.102}$$

where $\mathbf{x}_{m,i}$ denotes the i-th iteration of the m-th block. Similarly to the non-block MM algorithm, there are a set of assumptions on the M majorants that ensure the convergence of the block MM algorithm, which are founded on the ones in (A1)-(A4) from Subsection 3.3.1. The majorants' conditions for the convergence of the block MM are reviewed as follows.

(B1) The majorants must be tight approximations of the original cost:

$$g_m(\mathbf{x}_m|\mathbf{x}) = f(\mathbf{x}) \ \ \forall \mathbf{x} \in \mathcal{X}, \mathbf{x}_m \in \mathcal{X}_m. \tag{3.103}$$

(B2) The majorants must majorize the original cost:

$$g_m(\mathbf{x}_m|\mathbf{y}) \geq f(\mathbf{y}) \ \ \forall \mathbf{y} \in \mathcal{X}, \forall \mathbf{x}_m \in \mathcal{X}_m. \tag{3.104}$$

(B3) The first directional derivatives of the surrogate and of the original cost must be equivalent at the current iterate. In terms of the Gateaux differentials of the majorants and of the original cost, the previous idea is depicted mathematically as follows:

$$Dg_m(\mathbf{x}_m|\mathbf{x})[\boldsymbol{\delta}_m] = Df(\mathbf{x})[\boldsymbol{\delta}] \quad \text{s. t. } \mathbf{x} + \boldsymbol{\delta} \in \mathcal{X}, \tag{3.105}$$

for all $\boldsymbol{\delta} = (\mathbf{0}, ..., \boldsymbol{\delta}_m, ..., \mathbf{0})$ and $m = 1, ..., M$.

(B4) The majorants, $g(\mathbf{x}_m|\mathbf{y})$, must be continuous on $(\mathbf{x}_m, \mathbf{y}) \in \mathcal{X}_m \times \mathcal{X}$ for all $m = 1, ..., M$.

(B5) All the majorants must be quasiconvex (quasiconcave if they define a maximization problem).

(B6) Each majorant must have a unique minimizer.

Notice that (B1)-(B4) are imported from the non-block MM since they are the assumptions that define the MM framework, whereas (B5)-(B6) are imported from the BCD procedure [187]. In this case, (B3) only ensures that the limit points are coordinate-wise stationary points (see Definition 3.9). This is the reason why the block extension of the MM framework requires f to be a regular function, so the coordinate-wise stationary points are equivalent to the global stationary points. The remaining conditions for the global convergence of the block MM methodology, i.e. the compactness of the generated sequence, are equivalent to the ones presented for the non-block MM case from Subsection 3.3.1. The convergence proof of the block MM can be found in [92], [162] and can also be grasped from the proof of the generalized case to the Grassmann manifold detailed in Subsection 3.3.4.2.

3.3.3 Construction of the majorant function

The most important aspect of the MM algorithm is the construction of the majorant function. Indeed, not only the choice of the majorant defines the performance and convergence rate of the resulting sequence, but also the selection of a suitable majorant is one of the most difficult tasks in this framework. In the broader picture, the properties of the majorant function that are cherished are the following ones [87], [179].

- Separability of the variables, which enables parallel computing.
- Avoiding large matrix inversions or any other expensive computations.
- Convexity (or quasiconvexity) and smoothness.
- The existence of a closed-form minimizer. The uniqueness of this minimizer is also preferred.

A surrogate that satisfies the above properties simplifies the resulting algorithm and, thus, it provides a huge improvement with respect to the original optimization problem. The iterative procedure is the price to pay for the simplification of the original cost. Even though the above items are well known to practitioners, there is no general ruleset to build the majorant functions. Instead, there are some known rules of thumb that are useful to derive the majorant functions [179], which are reviewed in the following subsections. For clarity in the exposition, we do not consider strategies that may apply specifically to the block relaxation. Nevertheless, the following guidelines can also be applied to each block of variables independently.

3.3.3.1 First-order majorants

Assume that the original cost, $f(\mathbf{x})$, can be decomposed in the following way:

$$f(\mathbf{x}) = f_0(\mathbf{x}) + h(\mathbf{x}), \tag{3.106}$$

where $f_0(\mathbf{x})$ is a convex function and $h(\mathbf{x})$ is a differentiable concave function. Invoking Theorem 3.1 for concave functions, it is verified that $h(\mathbf{x})$ is upper bounded as follows:

$$h(\mathbf{x}) \leq h(\mathbf{x}_i) + \nabla_{\mathbf{x}} h(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i), \tag{3.107}$$

where $\mathbf{x}_i$ is the i-th iterate. In this way, an upper bound of $f(\mathbf{x})$ is derived by plugging (3.107) into (3.106):

$$f(\mathbf{x}) \leq f_0(\mathbf{x}) + h(\mathbf{x}_i) + \nabla_{\mathbf{x}} h(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i). \tag{3.108}$$

Hence, after ignoring additive constants that do not depend on $\mathbf{x}$, the resulting majorant yields:

$$g(\mathbf{x}|\mathbf{x}_i) = f_0(\mathbf{x}) + \nabla_{\mathbf{x}} h^T(\mathbf{x}_i)\mathbf{x}. \tag{3.109}$$

The previous majorant has the advantage that it results in a simple and separable function as long as $f_0(\mathbf{x})$ also meets those requirements. The above derivation is the one used in the Concave-Convex Procedure (CCP) [204] and the Frank-Wolfe algorithm for non-convex optimization [94], [109]. In order to show the utility of the first-order majorants, we review the Frank-Wolfe algorithm for non-convex optimization in the following example.

Example 3.3 (The Frank-Wolfe algorithm [94])**.** Let an optimization problem be described as follows:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s. t. } \mathbf{x} \in \mathcal{C}, \tag{3.110}$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is a concave differentiable function. Note that (3.110) is a non-convex problem since it consists of the minimization of a concave function [80]. An intuitive and straightforward application of the MM framework using a first-order majorant in (3.110) would result in the following sequential optimization problem:

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)\mathbf{x} \quad \text{s. t. } \mathbf{x} \in \mathcal{C}. \tag{3.111}$$

Still, the main issue with the previous sequential convex program is that it does not admit any kind of line search acceleration scheme [117], i.e. improving the convergence speed by changing the implicit step-size. This is the reason why the Frank-Wolfe algorithm incorporates an upper bound of the original cost using differentials (or descent directions) such as the one given in (3.11). Following the previous idea, the upper bound that the Frank-Wolfe algorithm utilizes is the following one:

$$f(\mathbf{x}) \leq f(\mathbf{x}_i) + \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)\mathbf{d}, \tag{3.112}$$

where $\mathbf{d} \in \mathcal{C}$ is such that $\mathbf{x} + \mathbf{d} \in \mathcal{C}$. As a result, the Frank-Wolfe algorithm consists in the following two update equations:

$$\mathbf{d}_{i+1} = \arg\min_{\mathbf{d}} \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)\mathbf{d} \quad \text{s. t. } \mathbf{d} \in \mathcal{C}, \mathbf{x} + \mathbf{d} \in \mathcal{C}, \tag{3.113a}$$

$$\mathbf{x}_{i+1} = u(\mathbf{x}_i, \mathbf{d}_{i+1}), \tag{3.113b}$$

where $u(\mathbf{x}_i, \mathbf{d}_{i+1})$ is any function that performs a *step* in the descent direction, $\mathbf{d}_{i+1}$. Clearly, a $u(\mathbf{x}_i, \mathbf{d}_{i+1})$ such that:

$$u(\mathbf{x}_i, \mathbf{d}_{i+1}) = \mathbf{x}_i + \mathbf{d}_{i+1}, \tag{3.114}$$

is not only the simplest alternative but it is also equivalent to the solution provided by the resulting algorithm from (3.111) since the two problems are related via a change of variables. Instead, the classical Frank-Wolfe algorithm considers the following expression of $u(\mathbf{x}_i, \mathbf{d}_{i+1})$ [68], [94]:

$$u(\mathbf{x}_i, \mathbf{d}_{i+1}) = (1 - \gamma)\mathbf{x}_i + \gamma\mathbf{d}_{i+1}, \tag{3.115}$$

for $\gamma = \frac{2}{i+1}$. Notice that the previous expression can be interpreted as a low-pass infinite impulse response filter of the descent directions, i.e. averaging the sequence of descent directions. In this sense, γ is a forgetting factor. As a final remark, although we have particularized the Frank-Wolfe algorithm for a concave cost function, so the MM framework can be applied, practical implementations of this algorithm are able to retrieve the optimal solution (global minima) of a convex optimization problem [94].

3.3.3.2 Second-order majorants

A refinement of the first-order majorants that are capable of yielding a better approximation of the original cost are the second-order majorant functions. Not only they are simple to derive, but they often result in strongly convex functions [29], which are functions that have better numerical-properties in optimization (see Section 3.3.5 for more insights). In order to introduce the second-order majorants, assume that $f : \mathbb{R}^N \to \mathbb{R}$ is a continuous non-convex differentiable function. Then, the second-order Taylor expansion majorants are constructed using the following inequality:

$$f(\mathbf{x}) \leq f(\mathbf{x}_i) + \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T \mathbf{M}_i (\mathbf{x} - \mathbf{x}_i), \tag{3.116}$$

where $\mathbf{M}_i \in \mathcal{S}_+^N$ is such that $\mathbf{M} \succeq \nabla^2 f(\mathbf{x}_i)$. The resulting majorant from the previous expression is:

$$g_T(\mathbf{x}|\mathbf{x}_i) = \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)\mathbf{x} + \frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T \mathbf{M}_i (\mathbf{x} - \mathbf{x}_i). \tag{3.117}$$

The degree of approximation of this kind of majorants is dependant on the selection of $\mathbf{M}_i$. Indeed, the closer $\mathbf{M}_i$ is from $\nabla^2 f(\mathbf{x}_i)$, the better the second-order approximation of the original cost is. A rigurous approach to the previous problem is built on the following optimization problem:

$$\mathbf{M}_i = \arg\min_{\mathbf{X}} ||\mathbf{X} - \nabla_{\mathbf{x}}^2 f(\mathbf{x}_i)||_F^2 \quad \text{s. t. } \mathbf{X} \succeq \mathbf{0}_{N,N}. \tag{3.118}$$

Note that the previous convex program guarantees that $\mathbf{M}_i \succeq \nabla_{\mathbf{x}}^2 f(\mathbf{x}_i)$. The optimal solution of (3.118) is given by [29, p. 399]:

$$\mathbf{M}_i = \sum_{n=1}^{N} \max(0, \lambda_{n,i}) \mathbf{u}_{n,i} \mathbf{u}_{n,i}^T, \tag{3.119}$$

where $\lambda_{n,i}$ and $\mathbf{u}_{n,i}$ are the n-th eigenvalue and eigenvector of $\nabla^2 f(\mathbf{x}_i)$, respectively. In other words, the optimal solution of (3.118) is the eigendecomposition of $\nabla_{\mathbf{x}}^2 f(\mathbf{x}_i)$ after ignoring the terms with negative eigenvalues. The previous procedure is what ensures that $\mathbf{M}_i$ majorizes the Hessian at $\mathbf{x}_i$. Also, the resulting $\mathbf{M}_i$ is always semi-definite positive, meaning that the resulting majorant is always convex.

An alternative, and more general, second-order majorant is based on the following lemma [179, Lemma 12].

Lemma 3.9 (Descent Lemma)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a continuous differentiable function with Lipschitz constant L. Then, the following inequality is satisfied:*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla_{\mathbf{x}} f^T(\mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{C}{2}||\mathbf{x} - \mathbf{y}||_2^2, \tag{3.120}$$

for any $C \geq L$.

Remark 3.33. The Lipschitz constant of a function is defined as the minimum positive constant such that [19]:

$$\frac{||\nabla_{\mathbf{x}} f(\mathbf{x}) - \nabla_{\mathbf{x}} f(\mathbf{y})||_2}{||\mathbf{x} - \mathbf{y}||_2} \leq L. \tag{3.121}$$

Invoking Lemma 3.9, we can obtain a much simpler (as compared to (3.117)) second-order majorant as follows:

$$g_L(\mathbf{x}|\mathbf{x}_i) = \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)\mathbf{x} + \frac{C}{2}||\mathbf{x} - \mathbf{x}_i||_2^2, \tag{3.122}$$

with a C such that it is greater than the Lipschitz constant of the orginal function. Both of the previous majorants have complementary properties that make them suitable for different scenarios. While $g_T(\mathbf{x}|\mathbf{x}_i)$ in (3.117) is the majorant that best approximates the original cost, it requires more computational resources for its construction than the one in (3.122). What is more, $g_L(\mathbf{x}|\mathbf{x}_i)$ results in an easier optimization problem since the variables are separable. The separability of the variables can be seen from the fact that (3.122) can be rewritten as:

$$g_L(\mathbf{x}|\mathbf{x}_i) = \sum_{n=1}^{N} x_n[\nabla_{\mathbf{x}} f^T(\mathbf{x}_i)]_n + \frac{L}{2}(x_n - x_{n,i})^2, \tag{3.123}$$

where x_n, $x_{n,i}$ and $[\nabla f^T(\mathbf{x}_i)]_n$ are the n-th components of $\mathbf{x}$, $\mathbf{x}_i$ and $\nabla f^T(\mathbf{x}_i)$, respectively. The most known iterative scheme, the Gradient Descent [169], is based on the second-order majorant given in (3.122). For the purpose of contextualizing all the previous ideas, we review the Gradient Descent algorithm in the following example.

Example 3.4 (MM perspective on the Gradient descent algorithm)**.** Let an optimization problem be described as follows:

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{3.124}$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is a differentiable function with Lipschitz constant L. Again, invoking Lemma 3.9 and its resulting majorant from (3.122), we get that the use of a second-order majorant for the optimization problem given in (3.124) results in the following iterative scheme:

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} \nabla_{\mathbf{x}} f^T(\mathbf{x}_i)\mathbf{x} + \frac{C}{2}||\mathbf{x} - \mathbf{x}_i||_2^2, \tag{3.125}$$

with $C \geq L$. The solution of the previous sequential convex program is obtained from the following equations (obtained after taking the gradient of its cost function and substituting $\mathbf{x}$ by $\mathbf{x}_{i+1}$):

$$\nabla_{\mathbf{x}} f(\mathbf{x}_i) + C(\mathbf{x}_{i+1} - \mathbf{x}_i) = \mathbf{0}_N, \tag{3.126a}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \frac{1}{C}\nabla_{\mathbf{x}} f(\mathbf{x}_i), \tag{3.126b}$$

which is the known expression of the Gradient Descent [117], [169] with stepsize equal to $\frac{1}{C}$. Notice that the rationale behind the maximum value of the Gradient Descent stepsize is founded on Lemma 3.9.

3.3.3.3 Jensen's inequality

There is a way of constructing majorants that exploits the definition of convex functions (see Definition 3.12). Let a convex function be denoted as $f(\mathbf{x})$. Then, we have the following inequality:

$$f\left(\sum_{l=1}^{L} \mathbf{x}_l w_l\right) \leq \sum_{l=1}^{L} w_l f(\mathbf{x}_l), \tag{3.127}$$

where w_l for $l = 1, ..., L$ are such that $\sum_{l=1}^{L} w_l = 1$, i.e. it is a convex combination. The previous inequality is also known as the Jensen's inequality and it is widely used in Information Theory [150]. The motivation behind the convex inequality is to use the right hand side of (3.127) as a surrogate function when $f(\mathbf{x})$ is simple enough. In fact, by letting $L \to \infty$ we get:

$$f\left(\int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}\right) \leq \int_{-\infty}^{\infty} f(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{3.128}$$

where $p_{\mathbf{x}}(\mathbf{x})$ is a PDF of $\mathbf{x}$. Note that if $f(\mathbf{x})$ were concave, the previous inequalities ((3.127) and (3.128)) would be inverted. The inequality in (3.128) is the tool needed to obtain an MM perspective on the Expectation-Maximization (EM) algorithm [179], [209], which is reviewed in the following example.

Example 3.5 (Derivation of the EM from the MM framework)**.** The EM algorithm is generally used to retrieve the ML estimation of a model with incomplete observations [53, Chapter 9]. In order to show the connections between the EM and MM frameworks, let us denote the observed variable $\mathbf{x}$ whose corresponding PDF is $p_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. Then, the ML estimate of $\boldsymbol{\theta}$ is obtained by the following optimization problem:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log(p_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta})), \tag{3.129}$$

where the equality holds since the logarithm is a non-decreasing function [29]. There are some cases in which the above optimization problem is intractable. One of those cases arises when the available observations present missing values, being the case where the EM algorithm thrives. For the purpose of introducing the EM methodology, let us assume that there exists an unobserved random variable related to $\mathbf{x}$ such that it contains $\mathbf{x}$ and all its missing values, denoted as $\mathbf{z}$. The associated PDF to $\mathbf{z}$ is denoted as $p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$. In the EM terminology, $\mathbf{x}$ and $\mathbf{z}$ are the incomplete and complete random variables, respectively.

Ideally, one would want to obtain $\hat{\boldsymbol{\theta}}$ from the maximization of $p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$ if $\mathbf{z}$ were available and if the maximization of $p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$ were simple enough. Instead, the EM procedure looks to sequentially optimize a minorant function (since the ML estimation is a maximization problem) that approximates the complete random variable PDF, $p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$. With the previous idea in mind, note that the ML estimation of $\boldsymbol{\theta}$ can also be obtained as follows:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log(p_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta})) = \log\left(\mathrm{E}_{\mathbf{z}|\boldsymbol{\theta}}[p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})]\right), \tag{3.130}$$

where $\mathrm{E}_{\mathbf{z}|\boldsymbol{\theta}}[\cdot]$ denotes the expected value with respect to $p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$. In order to obtain the minorant function, note that the expected value term in (3.130) can be rewritten as:

$$\mathrm{E}_{\mathbf{z}|\boldsymbol{\theta}}\left[p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})\right] = \int_{-\infty}^{\infty} p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z} = \tag{3.131a}$$

$$\int_{-\infty}^{\infty} p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta}) \frac{p_{\mathbf{z}}(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i)}{p_{\mathbf{z}}(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i)} \mathrm{d}\mathbf{z} = \mathrm{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i}\left[\frac{p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})}{p_{\mathbf{z}}(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i)}\right], \tag{3.131b}$$

for any given iterate $\boldsymbol{\theta}_i$. After applying the logarithm to the last expression in (3.131) and invoking the Jensen's inequality, we get:

$$\log\left(\mathrm{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i}\left[\frac{p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})}{p_{\mathbf{z}}(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i)}\right]\right) \geq \mathrm{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i}\left[\log\left(\frac{p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})}{p_{\mathbf{z}}(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i)}\right)\right] = \tag{3.132a}$$

$$\mathrm{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i}\left[\log(p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta}))\right] - \mathrm{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i}\left[\log(p_{\mathbf{z}}(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i))\right], \tag{3.132b}$$

from where, after ignoring additive constants that do not depend on $\boldsymbol{\theta}$ in (3.132b), i.e. the second term, the minorant in which the EM is based is:

$$g_{EM}(\boldsymbol{\theta},\boldsymbol{\theta}_i) = \mathrm{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_i}[\log(p_{\mathbf{x},\mathbf{z}}(\mathbf{x},\mathbf{z}|\boldsymbol{\theta}))], \tag{3.133}$$

where $p_{\mathbf{x},\mathbf{z}}(\mathbf{x},\mathbf{z}|\boldsymbol{\theta}) = p_{\mathbf{x}}(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p_{\mathbf{z}}(\mathbf{z}|\boldsymbol{\theta})$ is the joint distribution of the complete and incomplete random variables. As a result, the update equation of the EM algorithm is:

$$\boldsymbol{\theta}_{i+1} = \arg\max_{\boldsymbol{\theta}} g_{EM}(\boldsymbol{\theta},\boldsymbol{\theta}_i). \tag{3.134}$$

In this manner, we have shown that the *Expectation* step consists of the construction of the minorant function using the previous estimate of $\boldsymbol{\theta}$ and that the *Maximization* is the optimization of the aforementioned minorant.

3.3.4 MM framework on the Grassmann manifold

As stated in [179], whenever the function is non-continuous or the constraint set of the optimization problems is non-convex, the convergence of the MM framework must be surveyed in a case by case basis. Particularly, given that the Grassmann manifold is a non-convex set, classical convergence proofs do not apply [92], [162]. Within this context, the study of orthogonally constrained variables in the MM framework is not new. In [31], the authors study the convergence and performance of the resulting MM algorithm for a particular type of majorants over the Stiefel manifold. Still, no formal proof was given for the block MM case, although they mention a possible approach to prove its convergence. The previously mentioned work can be a starting point for this analysis given the strong connections between the Stiefel and Grassmann manifold. In the following subsections, we analyze the MM algorithm and its block extensions for variables constrained in the Grassmann manifold, extending the works of [31] and [162].

3.3.4.1 MM algorithm on the Grassmann manifold

Let an optimization problem be:

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} f(\mathbf{G}) \quad \text{s.t.} \quad \mathbf{G} \in \mathcal{G}, \tag{3.135}$$

where $f : \mathcal{G} \to \mathbb{R}$ is any continuous function (additional constraints will be imposed later) bounded from below in $\mathcal{G}$ and $\mathcal{G} \subseteq \mathrm{Gr}(N, D)$ is any g-convex subset of the Grassmannian. Following the same rationale as in the classical MM algorithm, the MM procedure generates a sequence of iterates, $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$, from the sequential optimization of the following majorant:

$$\mathbf{G}_{i+1} = \arg\min_{\mathbf{G}} g(\mathbf{G}|\mathbf{G}_i) \quad \text{s.t.} \quad \mathbf{G} \in \mathcal{G}. \tag{3.136}$$

The conditions in which the sequence generated by the previous sequential optimization problem is globally convergent of the resulting algorithm are straightforward generalizations from the ones in Subsection 3.3.1. They are summarized as follows [123]:

(A1) The surrogate must be a tight approximation of the original cost:

$$g(\mathbf{G}|\mathbf{G}) = f(\mathbf{G}) \quad \forall \mathbf{G} \in \mathcal{G}. \tag{3.137}$$

(A2) The surrogate must majorize the original cost:

$$g(\mathbf{G}|\mathbf{H}) \geq f(\mathbf{G}) \quad \forall \mathbf{G}, \mathbf{H} \in \mathcal{G}. \tag{3.138}$$

(A3) The first Gateaux differential of the surrogate and of the original cost must be equivalent at the current iterate:

$$Dg(\mathbf{G}|\mathbf{G}')[\boldsymbol{\Delta}] = Df(\mathbf{G})[\boldsymbol{\Delta}] \quad \text{s.t.} \quad \boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{G}} \mathrm{Gr}(N, D), \tag{3.139}$$

for $\mathbf{G}' = \mathbf{G}$, where $\boldsymbol{\Delta}$ is restricted to tangent directions such that their corresponding geodesic stays in $\mathcal{G}$.

(A4) $g(\mathbf{G}|\mathbf{H})$ must be continuous on $(\mathbf{G}, \mathbf{H}) \in \mathcal{G} \times \mathcal{G}$.

Notice that, in the case of the MM algorithm without any block relaxation, (A3) is the only condition generalized to an equivalent expression for the Grassmann manifold. We generalize Theorem 1 in [162] to the algorithm depicted by (3.136).

Theorem 3.10 (Convergence of an MM algorithm on the Grassmann manifold)**.** *Let a sequence be generated by (3.136) and that the majorant function satisfies (A1)-(A4). Then, every limit point of this sequence is a stationary point of (3.135) for every initialization,* $\mathbf{G}_0$.

Proof. We firstly prove the stationarity of the limit points of this sequence and, secondly, the convergence of this sequence. Since $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$ is a compact sequence, it has at least one limit point, denoted as

$\bar{\mathbf{G}}$. Let a subsequence convergent to the aforementioned limit point be $\{\mathbf{G}_{i_k}\}_{k\in\mathbb{N}}$. From assumptions (A1)-(A4) and restricting to the previous subsequence, we have the following set of inequalities:

$$g(\mathbf{G}|\mathbf{G}_{i_k}) \underbrace{\geq}_{\text{Eq. (3.136)}} g(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_k}) \underbrace{\geq}_{(A2))} f(\mathbf{G}_{i_{k+1}}) \underbrace{=}_{(A1)} g(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_{k+1}}), \tag{3.140}$$

from where, after taking the limit for $k \to \infty$, we get:

$$g(\mathbf{G}|\bar{\mathbf{G}}) \geq g(\bar{\mathbf{G}}|\bar{\mathbf{G}}) \;\; \forall \mathbf{G} \in \mathcal{G}. \tag{3.141}$$

The previous inequality also implies that:

$$Dg(\mathbf{G}|\bar{\mathbf{G}})[\boldsymbol{\Delta}] \underbrace{=}_{(A3)} Df(\bar{\mathbf{G}})[\boldsymbol{\Delta}] \geq 0. \tag{3.142}$$

Thus, $\bar{\mathbf{G}}$ is a local minimum point (possibly global) of the original problem. ■

Remark 3.34. Notice that we did not use any argument based on the Grassmann manifold in the previous proof, so it can be assumed general for any variable constrained in any non-convex set, e.g. other Riemannian manifolds, such that its Gateaux differentials are well-defined. In this way, assumption (A3) can be rewritten in terms of the aforementioned non-convex set.

Similarly to [31, Proposition 1], the previous theorem does not ensure the convergence in terms of the variable $\mathbf{G}$. In fact, it only ensures that the sequence defined by $\{f(\mathbf{G}_i)\}_{i\in\mathbb{N}}$ converges to a stationary point, $f(\bar{\mathbf{G}})$, of the original problem. This fact is proven as follows. From assumptions (A2) and (A4), we get:

$$f(\mathbf{G}_{i_k}) = g(\mathbf{G}_{i_k}|\mathbf{G}_{i_k}) \underbrace{\geq}_{\text{Eq. (3.136)}} g(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_k}) \underbrace{\geq}_{(A2)} f(\mathbf{G}_{i_{k+1}}). \tag{3.143}$$

The previous inequalities, in addition to the continuity and the boundedness from below of $f(\mathbf{G})$, imply the following limit:

$$\lim_{k\to\infty} f(\mathbf{G}_{i_k}) = f(\bar{\mathbf{G}}), \tag{3.144}$$

where $\bar{\mathbf{G}}$ is any limit point of $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$. Note that the value of the previous limit, $f(\bar{\mathbf{G}})$, depends on the initialization point, $\mathbf{G}_0$. Additionally, equation (3.144) is independent of the particular subsequence of $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$, meaning that there could be two limit points, $\bar{\mathbf{G}}$ and $\bar{\mathbf{G}}'$, of this sequence such that $f(\bar{\mathbf{G}}) = f(\bar{\mathbf{G}}')$. Therefore, there is a need of additional constraints to the original problem that ensure the convergence to the limit points of $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$. Luckily, there are several known conditions in which the MM algorithm converges in terms of $\mathbf{G}$. The simplest case is found when there is a single stationary point of the original cost [93]. In some cases, the monotonic decrement of the objective is a sufficient condition (see [31] and references therein). In the sequel, we show an example where the uniqueness of the minimizer of the majorant, in addition to its g-quasiconvexity, ensures the convergence of the resulting algorithm.

3.3.4.2 Block MM algorithm on the Grassmann manifold

In a similar manner to the classical block MM algorithm, consider the following optimization problem [123]:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} f(\mathbf{X}) \;\; \text{s.t. } \mathbf{X} \in \mathcal{X}, \tag{3.145}$$

where $f(\mathbf{X})$ is any continuous function bounded from below in $\mathcal{X}$, $\mathcal{X} = \mathcal{G} \times \mathcal{C}$, $\mathcal{C}$ is any closed convex subset of $\mathbb{R}^M$ and $\mathcal{G} \subseteq \text{Gr}(N, D)$ is any g-convex subset of the Grassmann manifold. The reason of choosing only two blocks, one constrained in the Grassmannian and the other being a convex constrained set, is to emphasize the differences with respect to the non-Grassmann block MM algorithm from Subsection 3.3.2 without any loss of generality. This concreteness in the exposition is also motivated by a problem on the data fusion scenario that will be exposed in Section 4.4 from Chapter 4.

Given the structure of the feasible set, $\mathcal{X}$, the optimization variables can be split into two independent blocks, $\mathbf{X} = (\mathbf{G}, \mathbf{c})$ where $\mathbf{G} \in \mathcal{G}$ and $\mathbf{c} \in \mathcal{C}$. We rewrite $f(\mathbf{X})$ as $f(\mathbf{G}, \mathbf{c})$ and $\mathbf{X}$ as $(\mathbf{G}, \mathbf{c})$ when needed. The block MM rationale applied to (3.145) consists in the following update equations:

$$\mathbf{G}_{i+1} = \arg\min_{\mathbf{G}} g_G(\mathbf{G}|\mathbf{G}_i, \mathbf{c}_i) \quad \mathbf{G} \in \mathcal{G}, \tag{3.146a}$$

$$\mathbf{c}_{i+1} = \arg\min_{\mathbf{c}} g_c(\mathbf{c}|\mathbf{G}_{i+1}, \mathbf{c}_i) \quad \mathbf{c} \in \mathcal{C}, \tag{3.146b}$$

where $g_G(\mathbf{G}|\mathbf{G}_i, \mathbf{c}_i)$ and $g_c(\mathbf{c}|\mathbf{G}_i, \mathbf{c}_i)$ are the majorant functions of $f(\mathbf{X})$ constructed using the i-th iterates with respect to $\mathbf{G}$ and $\mathbf{c}$, respectively. As a result of the update equations in (3.146), a sequence of iterates is generated, denoted as $\{\mathbf{X}_i\}_{i\in\mathbb{N}} = \{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$. The following assumptions on $g_G(\mathbf{G}|\mathbf{G}_i, \mathbf{c}_i)$ are a generalization to the ones in (B1)-(B6) from the classical block MM [123].

(B1) The surrogate must be a tight approximation of the original cost:

$$g_G(\mathbf{G}|\mathbf{G}, \mathbf{c}) = f(\mathbf{G}, \mathbf{c}) \ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C}. \tag{3.147}$$

(B2) The surrogate must majorize the original cost function:

$$g_G(\mathbf{G}|\mathbf{H}, \mathbf{d}) \geq f(\mathbf{H}, \mathbf{d}) \ \forall \mathbf{G}, \mathbf{H} \in \mathcal{G}, \forall \mathbf{d} \in \mathcal{C}. \tag{3.148}$$

(B3) The Gateaux differentials of this majorant and of the original cost must coincide:

$$Dg_G(\mathbf{G}|\mathbf{G}', \mathbf{c})[\boldsymbol{\Delta}] = Df(\mathbf{G}, \mathbf{c})[\boldsymbol{\Delta}, \mathbf{0}_N] \quad \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C}, \tag{3.149}$$

for $\mathbf{G}' = \mathbf{G}$ and for all tangent directions $\boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{G}} \operatorname{Gr}(N, D)$ whose resulting geodesic remains on $\mathcal{G}$.

(B4) $g_G(\cdot|\cdot)$ must be continuous on its input arguments.

(B5) $g_G(\cdot|\cdot)$ must be g-quasiconvex on $\mathcal{G}$.

(B6) $g_G(\cdot|\cdot)$ must have a unique minimizer.

In addition to the previous assumptions on $g_G(\cdot|\cdot)$, the classical conditions of the block MM (see (B1)-(B6) from Subsection 3.3.2) are also imposed to $g_c(\cdot|\cdot)$. In the following theorem, we generalize Theorem 2 in [162] for blocks of variables constrained in the Grassmann manifold [123].

Theorem 3.11 (Convergence of the block MM on the Grassmann manifold)**.** *Supose that a sequence is generated by (3.146) and that the majorant functions satisfy their respective block MM conditions ((B1)-(B6) from this subsection and from Subsection 3.3.2). In addition, assume that the sequence* $\{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$ *is compact and that* $f(\mathbf{X})$ *is regular and continuous in* $\mathcal{G} \times \mathcal{C}$*. Then, this sequence converges to a stationary point of (3.145).*

Proof. The rationale of the proof consists of two steps. Firstly, we prove that the sequence $\{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$ converges to the limit point $\{\bar{\mathbf{G}}, \bar{\mathbf{c}}\}$ and, secondly, we prove that those limit points are stationary points of the problem depicted in (3.145).

From assumptions (B1)-(B2) and (3.146), we have the following series of inequalities:

$$f(\mathbf{G}_i, \mathbf{c}_i) \underbrace{=}_{(B1)} g_G(\mathbf{G}_i|\mathbf{G}_i, \mathbf{c}_i) \underbrace{\geq}_{\text{Eq. (3.146}a\text{)}} g_G(\mathbf{G}_{i+1}|\mathbf{G}_i, \mathbf{c}_i) \underbrace{\geq}_{(B2)} f(\mathbf{G}_{i+1}, \mathbf{c}_i) \underbrace{=}_{(B1)} \tag{3.150a}$$

$$g_c(\mathbf{c}_i|\mathbf{G}_{i+1}, \mathbf{c}_i) \underbrace{\geq}_{\text{Eq. (3.146}b\text{)}} g_c(\mathbf{c}_{i+1}|\mathbf{G}_{i+1}, \mathbf{c}_i) \underbrace{=}_{(B1)} f(\mathbf{G}_{i+1}, \mathbf{c}_{i+1}). \tag{3.150b}$$

The inequalities in (3.150) yield $f(\mathbf{G}_i, \mathbf{c}_i) \geq f(\mathbf{G}_{i+1}, \mathbf{c}_{i+1})$ for $i \in \mathbb{N}$ which, in addition to the continuity of $f(\mathbf{G}, \mathbf{c})$ and to the compactness of the sublevel set, implies that the sequence $\{f(\mathbf{G}_i, \mathbf{c}_i)\}_{i\in\mathbb{N}}$ is non-increasing and thus has at least one limit point, $f(\bar{\mathbf{G}}, \bar{\mathbf{c}})$. In fact, the previous set of inequalities also imply that $\{f(\mathbf{G}_i, \mathbf{c}_i)\}_{i\in\mathbb{N}}$ is convergent to $f(\bar{\mathbf{G}}, \bar{\mathbf{c}})$. Since the iterates belong to a compact set (the Grasmannian and boundedness from below of the cost function), the generated sequence also has limit points, denoted as $(\bar{\mathbf{G}}, \bar{\mathbf{c}})$.

Consider a subsequence $\{\mathbf{G}_{i_k}, \mathbf{c}_{i_k}\}_{k\in\mathbb{N}}$ that converges to $(\bar{\mathbf{G}}, \bar{\mathbf{c}})$. Henceforth, in order to be able to take the limits for $k \to \infty$, we need to highlight the fact that the Grassmann variable, $\mathbf{G}_{i_k}$, is updated infinitely often in $\{\mathbf{G}_{i_k}, \mathbf{c}_{i_k}\}_{k\in\mathbb{N}}$ due to equations (3.146). In this way, we can prove that $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$ converges to the aforementioned limit point by contradiction. Let us assume that the Grassmann variable does not converge. Thus, there exists a positive value $\bar{\gamma}$ such that:

$$d_{arc}(\mathbf{G}_{i_{k+1}}, \mathbf{G}_{i_k}) = \gamma_{i_{k+1}} \geq \bar{\gamma} > 0, \tag{3.151}$$

from where we show that it is contradictory with the previous assumptions. Let the aligned geodesic (see (2.73)) joining $\mathbf{G}_{i_{k+1}}$ and $\mathbf{G}_{i_k}$ with arclength $\gamma_{i_{k+1}}$ be:

$$\mathbf{\Gamma}_{i_{k+1}}(t) = \mathbf{G}_{a,i_k} \cos(\mathbf{\Theta}_{i_k} t) + \mathbf{\Delta}_{a,i_k} \sin(\mathbf{\Theta}_{i_k} t), \tag{3.152}$$

where $\mathbf{\Delta}_{a,i_k}$ and $\mathbf{\Theta}_{i_k}$ are such that $\mathbf{\Gamma}_{i_{k+1}}(1) = \mathbf{G}_{a,i_{k+1}}$. With this geodesic in mind, we obtain the following inequalities:

$$f(\mathbf{G}_{i_{k+1}}, \mathbf{c}_{i_{k+1}}) \underbrace{\leq}_{(B2)} g_G(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{=}_{\mathbf{G}_{i_{k+1}}=\mathbf{\Gamma}_{i_{k+1}}(1)} g_G(\mathbf{\Gamma}_{i_{k+1}}(1)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{\leq}_{\text{Eq. (3.146}a\text{)}} \tag{3.153a}$$

$$g_G(\mathbf{\Gamma}_{i_{k+1}}(t)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{=}_{\text{Eq. (3.152)}} g_G(\mathbf{G}_{a,i_k} \cos(\mathbf{\Theta}_{i_k} t) + \mathbf{\Delta}_{a,i_k} \sin(\mathbf{\Theta}_{i_k} t)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{\leq}_{(B5)} \tag{3.153b}$$

$$g_G(\mathbf{G}_{i_k}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{=}_{(B1)} f(\mathbf{G}_{i_k}, \mathbf{c}_{i_k}), \tag{3.153c}$$

where (B5) is used to obtain (3.153c) from (3.153b) since $\mathbf{G}_{i_{k+1}}$ the minimizer of $g_G(\mathbf{G}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k})$ (see (3.146a)) and:

$$\max(g_G(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}), g_G(\mathbf{G}_{i_k}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) = g_G(\mathbf{G}_{i_k}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}). \tag{3.154}$$

By summarizing (3.153), we get:

$$f(\mathbf{G}_{i_{k+1}}, \mathbf{c}_{i_{k+1}}) \leq g_G(\mathbf{\Gamma}_{i_{k+1}}(t)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \leq f(\mathbf{G}_{i_k}, \mathbf{c}_{i_k}). \tag{3.155}$$

Provided that $\mathbf{\Theta}_{i_k}$ (because the principal angles are all bounded in $[0, \frac{\pi}{2}]$) and $\mathbf{\Delta}_{a,i_k}$ (due to the orthonormality constraints, see (2.73)) belong to compact sets, the sequences that they generate have limit points $\bar{\mathbf{\Theta}}$ and $\bar{\mathbf{\Delta}}_a$, respectively. As a consequence, the geodesic defined by those limit points is denoted as $\bar{\mathbf{\Gamma}}(t)$, which, with some abuse of notation, can be referred to as the *limit point* geodesic. By further restricting to a subsequence that has limit points $\bar{\mathbf{\Theta}}$ and $\bar{\mathbf{\Delta}}_a$, invoking (B4) and letting $k \to \infty$, (3.155) yields:

$$f(\bar{\mathbf{G}}, \bar{\mathbf{c}}) \leq g_G(\bar{\mathbf{\Gamma}}(t)|\bar{\mathbf{G}}, \bar{\mathbf{c}}) \leq f(\bar{\mathbf{G}}, \bar{\mathbf{c}}), \tag{3.156}$$

which is equivalent to:

$$\begin{aligned} f(\bar{\mathbf{G}}, \bar{\mathbf{c}}) = g_G(\bar{\mathbf{\Gamma}}(t)|\bar{\mathbf{G}}, \bar{\mathbf{c}}) = \\ g_G(\bar{\mathbf{G}}_a \cos(\bar{\mathbf{\Theta}} t) + \bar{\mathbf{\Delta}}_a \sin(\bar{\mathbf{\Theta}} t)|\bar{\mathbf{G}}, \bar{\mathbf{c}}). \end{aligned} \tag{3.157}$$

However, (3.157) is contradictory with the unique minimizer assumption when the arclength of the limit point geodesic, $\bar{\gamma}$, is different from 0. From (B2), we know that:

$$g_G(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_{k+1}}, \mathbf{c}_{i_{k+1}}) \leq g_G(\mathbf{G}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \;\; \forall\, \mathbf{G} \in \mathcal{G}, \tag{3.158}$$

whose limit for $k \to \infty$ is:

$$g_G(\bar{\mathbf{G}}|\bar{\mathbf{G}}, \bar{\mathbf{c}}) \leq g_G(\mathbf{G}|\bar{\mathbf{G}}, \bar{\mathbf{c}}) \;\; \forall\, \mathbf{G} \in \mathcal{G}, \tag{3.159}$$

so $\bar{\mathbf{G}}$ is a minimizer of $g_G(\cdot|\bar{\mathbf{G}}, \bar{\mathbf{c}})$ and so is $\bar{\mathbf{G}}_a$. This means that the limit point geodesic in (3.157) must remain in the same point since we assumed that the minimizers of the majorants are unique. As a result, equation (3.157) is only true if $\bar{\mathbf{\Gamma}}(t) = \bar{\mathbf{G}}_a$ for $t \in [0, 1]$ and, given that $\mathcal{G}$ is a geodesically convex subset (there is only a unique path joining two points), it means that:

$$\lim_{k \longrightarrow \infty} d_c(\mathbf{G}_{i_{k+1}}, \mathbf{G}_{i_k}) = 0. \tag{3.160}$$

The above argument must be satisfied by every subsequence. Hence, the sequence $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$ converges to $\bar{\mathbf{G}}$. Note that this limit point depends on the initialization points, $\mathbf{G}_0$ and $\mathbf{c}_0$. Besides, the methodology that proves the convergence of the sequence $\{\mathbf{c}_i\}_{i\in\mathbb{N}}$, given the assumptions of this theorem, is shown in [162, Theorem 2]. In consequence, the joint sequence, $\{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$, also converges.

Finally, we prove that the limit point of the Grassmann variable iterates is a stationary point of the original problem. Notice that (3.159) implies, after taking the lower directional derivative of both sides at $\bar{\mathbf{G}}$, that:

$$Dg_G(\bar{\mathbf{G}}|\bar{\mathbf{G}}, \bar{\mathbf{c}})[\boldsymbol{\Delta}] \geq 0, \tag{3.161}$$

for all directions $\boldsymbol{\Delta} \in \mathcal{T}_{\bar{\mathbf{G}}}\mathrm{Gr}(N, D)$ whose respective geodesic emanating from $\bar{\mathbf{G}}$ stays in $\mathcal{G}$. Due to assumption (B3), we get that:

$$Df(\bar{\mathbf{G}}, \bar{\mathbf{c}})[\boldsymbol{\Delta}, \mathbf{0}_N] \geq 0. \tag{3.162}$$

In other words, $\bar{\mathbf{G}}$ is a coordinate-wise minimum of $f(\mathbf{G}, \mathbf{c})$. A similar argument is derived for the remaining convex block, $\mathbf{c}$, in [162, Theorem 2] with the assumptions specified in Subsection 3.3.2. Note that the previous arguments for each of the blocks of variables hold since $\mathbf{G}$ and $\mathbf{c}$ are both updated infinitely often in (3.146). Since $\bar{\mathbf{G}}$ and $\bar{\mathbf{c}}$ are coordinate-wise minimums and considering the regularity property of $f(\mathbf{G}, \mathbf{c})$, the aforementioned limit point is also a stationary point of $f(\mathbf{G}, \mathbf{c})$. ■

Remark 3.35. Other manifolds that share the same properties as the Grassmann manifold (particularly, if they define a compact set) may admit a similar convergence proof.
Remark 3.36. In contrast to Theorem 3.10, the convergence in terms of the variables $\mathbf{G}$ and $\mathbf{c}$ is needed to obtain the stationary points in the previous proof.

3.3.5 Proximal algorithms

As already shown in Figure 3.3, the proximal algorithms are one of the staple implementations of the MM framework [145]. Indeed, the proximal algorithms are often the go-to methodologies to deal with non-smooth convex problem, such as the ones based on the ℓ_1 norm (see Subsection 2.1.1). For the previous reason, the implementation of the proximal algorithms to ℓ_1 norm-based problems is of great interest in this dissertation, especially in the applications detailed in Chapter 5. Particularly, we are interested in the proximal algorithm perspective of the ADMM. The reason behind the fact that the proximal algorithms lie within the MM methodology [179] is that they are based on a successive optimization of a surrogate function termed the *proximal operator*. The proximal operator is defined as follows [19], [145].

Definition 3.24 (Proximal operator)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex continuous function. Then, the proximal operator, $\mathrm{prox}_{f,\lambda} : \mathbb{R}^N \to \mathbb{R}^N$, is defined as follows*

$$\mathrm{prox}_{f,\lambda}(\mathbf{v}) = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{2\lambda}||\mathbf{x} - \mathbf{v}||_2^2, \tag{3.163}$$

where $\lambda \geq 0$.
Remark 3.37. The cost function of the proximal operator is also a convex function. In fact, it is a *strongly* convex function. A strongly convex function, $h(x)$, is a function that satisfies the following inequality (see Theorem 3.2):

$$\frac{\mathrm{d}^2 h(x)}{\mathrm{d}x^2} \geq m \quad \forall x \in \mathbb{R}, \tag{3.164}$$

for some $m > 0$. The generalization of the previous condition to a multivariate function, $h(\mathbf{x})$, is given by:

$$\nabla_{\mathbf{x}}^2 h(\mathbf{x}) \succeq m\mathbf{I}_N \quad \forall \mathbf{x} \in \mathbb{R}^N. \tag{3.165}$$

Clearly, the objective function in (3.163) fulfills the previous conditions since f is a convex function and:

$$\nabla_{\mathbf{x}}^2 \left(f(\mathbf{x}) + \frac{1}{2\lambda}||\mathbf{x} - \mathbf{v}||_2^2 \right) = \nabla_{\mathbf{x}}^2 f(\mathbf{x}) + \frac{1}{\lambda}\mathbf{I}_N \succeq \frac{1}{\lambda}\mathbf{I}_N, \tag{3.166}$$

One of the possible interpretations of the proximal operators is that they are generalized projections depicted by the convex function f. In order to provide some insights on the generalized projection interpretation of the proximal operators, we need the formal definition of the indicator function.

Definition 3.25 (Indicator function of a set)**.** *Let S be any subset of $\mathbb{R}^N$. Then, $\mathcal{I}_S : \mathbb{R}^N \to \mathbb{R}$ denotes the indicator function of this set, which is defined as:*

$$\mathcal{I}_S(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in S \\ +\infty & \text{otherwise} \end{cases}. \tag{3.167}$$

Remark 3.38. If S were convex, its respective indicator function is also convex.

Using the previous definition, we highlight the fact that, when the convex function in Definition 3.24 is a convex indicator function for some convex set S, the resulting proximal operator can be rewritten as follows:

$$\text{prox}_{\mathcal{I}_S,\lambda}(\mathbf{v}) = \arg\min_{\mathbf{x}} \mathcal{I}_S(\mathbf{x}) + \frac{1}{2\lambda}||\mathbf{x}-\mathbf{v}||_2^2 = \arg\min_{\mathbf{x}\in S} ||\mathbf{x}-\mathbf{v}||_2^2, \tag{3.168}$$

which, in essence, is the orthogonal projection of $\mathbf{v}$ into S. The previous observation suggests that the proximal operators must share some of the properties of orthogonal projectors (see [145] for more details on this idea). Indeed, the proximal operators can be seen as a way to project $\mathbf{v}$ into the direction perpendicular to the level sets of f, which is an idea that is depicted in a clear manner in [145, Figure 1.1]. Regarding the connections of the proximal algorithms with the MM framework, notice that the cost function in the proximal operator, i.e. $g_P(\mathbf{x}|\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2\lambda}||\mathbf{x}-\mathbf{v}||_2^2$, satisfies the assumptions imposed on the majorants in the classical MM algorithm (see (A1)-(A4) in Subsection 3.3.1). We particularize the aforementioned assumptions on $g_P(\mathbf{x}|\mathbf{v})$ as follows.

(A1) The surrogate is a tight approximation of the original function:

$$g_P(\mathbf{x}|\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2\lambda}||\mathbf{x}-\mathbf{x}||_2^2 = f(\mathbf{x}). \tag{3.169}$$

(A2) The surrogate majorizes the original cost:

$$g_P(\mathbf{x}|\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2\lambda}||\mathbf{x}-\mathbf{v}||_2^2 \geq f(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{v} \in \mathbb{R}^N. \tag{3.170}$$

(A3) At the current iterate, the Gateaux differential of the surrogate coincides with the one of the original cost (the Gateaux differential is linear [2]):

$$Dg_P(\mathbf{x}|\mathbf{x})[\boldsymbol{\delta}] = Df(\mathbf{x})[\boldsymbol{\delta}] + D\left[\frac{1}{2\lambda}||\mathbf{x}-\mathbf{x}||_2^2\right][\boldsymbol{\delta}] = f'(\mathbf{x};\boldsymbol{\delta}). \tag{3.171}$$

(A4) The surrogate is continuous since it is the sum of two continuous functions.

Having defined the proximal operators, we show how to use them to obtain the minimum value of any convex function in the following theorem [145].

Theorem 3.12 (Fixed point of the proximal algorithm)**.** *A point $\mathbf{x}^*$ minimizes the convex function f if and only if:*

$$\mathbf{x}^* = \text{prox}_{f,\lambda}(\mathbf{x}^*), \tag{3.172}$$

for any $\lambda \geq 0$, as long as there exists a minimizer of (3.163)*. In other words, $\mathbf{x}^*$ is a fixed point of the proximal operator. The proof of this theorem can be found in [145, Subsection 2.3].*

The previous theorem, in addition to the MM interpretation of the proximal algorithms, suggests that the minimum value of a function $f : \mathbb{R}^N \to \mathbb{R}$ can be found using the following update equation [19], [145]:

$$\mathbf{x}_{i+1} = \text{prox}_{f,\lambda}(\mathbf{x}_i), \tag{3.173}$$

which is often referred to as *proximal minimization* or *proximal iteration*. For most non-smooth convex cost functions, e.g. the ℓ_1 norm, it is much easier to derive the proximal iteration due to the strong convexity of the proximal operator cost function (see Remark 3.37). Among other properties, the strong

convexity implies that its minimizer is unique [29]. Besides, the convergence of the update equation in (3.173) can be seen from two different perspectives. On the one hand, the previous update equation is a fixed point equation [77], [90], so its convergence can be analyzed from the fixed point algorithms view. In general fixed point algorithms, the existence of a fixed point of a function is ensured if its Lipschitz constant is less than 1 (see Remark 3.33), i.e. it is a *contraction*. In the case of a contraction, the fixed point can be found by repeatedly applying the contraction. Luckily, it has been proven that, as long as there exists a minimizer of (3.163), the fixed point iteration in (3.173) is sufficient to reach the desired fixed point [145, Subsection 4.1] (and references therein). On the other hand, provided that the proximal operator is, essentially, an MM algorithm, the mentioned convergence theorems from Subsection 3.3.2 ensure the convergence of the proximal algorithm described by (3.173). The latter perspective is our preferred one due to its link to the more general MM framework.

The main issue with the proximal minimization process is the computation of λ. As a matter of fact, cross-validation is the straightforward approach to compute the proximal parameter. Still, we discourage the use of cross-validation for the same reason that we avoid descent-like algorithms, i.e. setting in an arbitrary manner a user-defined parameter. An alternative to cross-validation that yields a faster convergence rate of the proximal iteration and that requires less computational resources consists in updating the proximal parameter, λ, in each iteration such that [145]:

$$\lambda_i > 0, \ \sum_{i=1}^{\infty} \lambda_i = \infty \ \text{ and } \ \lambda_i \leq \lambda_{i-1}, \tag{3.174}$$

where λ_i is the i-th iterate. The previous conditions on λ_i ensure the global convergence of the modified proximal minimization process while speeding the overall convergence. As an example, one may use $\lambda_i = \frac{C}{i}$, where C is any positive constant. For other alternatives, see [19].

Within the context of this dissertation, the main application of the proximal algorithms that we are interested in is found in the implementation of the ADMM to deal with ℓ_1 norm regularized problems (see Subsection 2.1.1). Before delving into the intricacies of the ADMM framework for the case of interest, we review the proximal operator of the absolute value function, which is the main building block of the ℓ_1 norm, in the following proposition [19], [145].

Proposition 3.13 (Proximal operator of the ℓ_1 norm)**.** *The proximal operator of the absolute value function is given by the following optimization problem:*

$$\text{prox}_{\ell_1,\lambda}(v) = \arg\min_x |x| + \frac{1}{2\lambda}||x - v||_2^2, \tag{3.175}$$

whose closed-form solution is:

$$\text{prox}_{\ell_1,\lambda}(v) = \max(0, |v| - \lambda)\,\text{sign}(v) = \begin{cases} v - \lambda & v > \lambda \\ 0 & |v| \leq \lambda \\ v + \lambda & v < -\lambda \end{cases}. \tag{3.176}$$

Remark 3.39. The previous proximal operator is also referred to as the soft-thresholding operator [19]. The proof of (3.176) can be found in [145, Subsection 6.5.2].

Remark 3.40. Since the ℓ_1 norm is separable in each of the components of its input vector, one can obtain the proximal operator of the ℓ_1 norm with respect to a vector $\mathbf{v}$ by applying entry-wise the expression given in (3.176). Henceforth, $\text{prox}_{\ell_1,\lambda}(\mathbf{v})$ denotes the entry-wise soft-threshold function of $\mathbf{v}$.

3.3.5.1 Alternating Direction Method of Multipliers (ADMM)

The purpose of this subsection is to review the Alternating Direction Method of Multipliers (ADMM) and provide a use case of this methodology that will be useful in a future chapter. In essence, the ADMM is a methodology, which is based on the Augmented Lagrangian and Dual Ascent techniques [30], [143], [178], that is useful for solving optimization algorithms of a certain structure. In this regard, our motivation behind the consideration of the ADMM is that, in Chapter 5, we find an optimization problem that can be mapped to the standard ADMM formulation. For the sake of exposing efficiently

the ideas behind the ADMM, we introduce the Augmented Lagrangian and Dual Ascent techniques as we construct the ADMM algorithm from a generalization of the aforementioned use case of interest (which is similar to the one in [26]), so this subsection provides an informal introduction of the ADMM. For a more rigorous and detailed explanation of the ADMM, we refer to [30]. Let a feasible optimization problem be as follows:

$$\min_{\mathbf{x}} f(\mathbf{x}) + h(\mathbf{x}) \quad \text{s.t. } \mathbf{H}\mathbf{x} = \mathbf{y}, \tag{3.177}$$

where $f(\mathbf{x})$ and $h(\mathbf{x})$ are convex, possibly non-smooth, and continuous functions, $\mathbf{x} \in \mathbb{R}^D$, $\mathbf{y} \in \mathbb{R}^M$, and $\mathbf{H} \in \mathbb{R}^{M \times D}$. For the purpose of introducing the ADMM methodology, we need to reformulate (3.177) into the standard ADMM formulation [30], [143], [178]. To this end, we incorporate a change of variables into (3.177) such that the overall optimization problem is equivalent but parameterized with two blocks of variables:

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + h(\mathbf{z}) \quad \text{s.t. } \mathbf{H}\mathbf{x} = \mathbf{y}, \mathbf{z} = \mathbf{x}, \tag{3.178}$$

where $\mathbf{z}$ is the newly introduced block of variables. Although the ADMM may also be derived from (3.178), the ADMM methodology joins the two constraints in (3.178) as follows:

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + h(\mathbf{z}) \quad \text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}, \tag{3.179}$$

where:

$$\mathbf{A} = \begin{bmatrix} \mathbf{H} \\ \mathbf{I}_D \end{bmatrix}, \tag{3.180a}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{0}_{N,D} \\ -\mathbf{I}_D \end{bmatrix}, \tag{3.180b}$$

$$\mathbf{c} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_D \end{bmatrix}. \tag{3.180c}$$

The previous expression in (3.179) is the standard ADMM formulation [30], [143], [178]. As we show in a future chapter, the rationale behind the previous change of variables is that, in practice, optimizing $f(\mathbf{x})$ and $h(\mathbf{x})$ independently via the block relaxation is much easier than their joint optimization. This is especially true if one of those functions is the ℓ_1 norm of $\mathbf{x}$. The next key idea of the ADMM is the Augmented Lagrangian technique [30], [178], which is the main reason why the ADMM can be viewed as a proximal algorithm (and lie in the MM framework). With the intention of introducing the Augmented Lagrangian technique, let us consider the Lagrangian of (3.179):

$$\mathcal{L}(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}) = f(\mathbf{x}) + h(\mathbf{z}) + \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}). \tag{3.181}$$

The Augmented Lagrangian technique consists in the addition of an extra term in the Lagrangian, whose goal is to enable the solution of the original problem using the proximal operators of f and h. Also, it is desirable that the added term do not modify the optimal value of the original problem in (3.179). A possible term that satisfies the previous two conditions is found in the following expression [30]:

$$\mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}) = f(\mathbf{x}) + h(\mathbf{z}) + \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\lambda}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2, \tag{3.182}$$

where λ is the penalty parameter of the Augmented Lagrangian [30]. Note that (3.182) can be interpreted as the Lagrangian of the following optimization problem:

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + h(\mathbf{z}) + \frac{\lambda}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2 \quad \text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}. \tag{3.183}$$

In fact, with some abuse of notation, the previous optimization problem can be seen as the *generalized equality constrained proximal operators*. Clearly, the added term in (3.182) (or (3.183)) does not modify the value of the original cost in the feasible set, i.e.:

$$S_f = \{\mathbf{x}, \mathbf{z} \in \mathbb{R}^N : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}\}, \tag{3.184}$$

since for any $\mathbf{x}, \mathbf{z} \in S_f$ we get that:

$$||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2 = ||\mathbf{c} - \mathbf{c}||_2^2 = 0. \tag{3.185}$$

Thus, the value of the Augmented Lagrangian is equivalent to the classical Lagrangian in S_f. For the purpose of deriving cleaner expressions, notice that the Augmented Lagrangian can be rewritten as follows:

$$\mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}) = f(\mathbf{x}) + h(\mathbf{z}) + \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\lambda}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2 = \tag{3.186a}$$

$$f(\mathbf{x}) + h(\mathbf{z}) + \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\lambda}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2 + \frac{1}{2\lambda}||\boldsymbol{\mu}||_2^2 - \frac{1}{2\lambda}||\boldsymbol{\mu}||_2^2 = \tag{3.186b}$$

$$f(\mathbf{x}) + h(\mathbf{z}) + \left|\left|\frac{1}{\sqrt{2\lambda}}\boldsymbol{\mu} + \sqrt{\frac{\lambda}{2}}\left(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\right)\right|\right|_2^2 - \frac{1}{2\lambda}||\boldsymbol{\mu}||_2^2 = \tag{3.186c}$$

$$f(\mathbf{x}) + h(\mathbf{z}) + \frac{\lambda}{2}\left|\left|\frac{1}{\lambda}\boldsymbol{\mu} + \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\right|\right|_2^2 - \frac{1}{2\lambda}||\boldsymbol{\mu}||_2^2, \tag{3.186d}$$

which is a much more compact expression of $\mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu})$. Actually, the *scaled* Augmented Lagrangian [30], whose purpose is to derive shorter expressions of the ADMM algorithm, is obtained by introducing a change of variables in (3.186d). Indeed, by denoting $\mathbf{u} = \frac{1}{\lambda}\boldsymbol{\mu}$, we get that the scaled Augmented Lagrangian is:

$$\mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}, \mathbf{u}) = f(\mathbf{x}) + h(\mathbf{z}) + \frac{\lambda}{2}||\mathbf{u} + \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2 - \frac{\lambda}{2}||\mathbf{u}||_2^2. \tag{3.187}$$

Henceforth, we write all the subsequent equations using the scaled Lagrange multipliers, $\mathbf{u}$, instead of the original multipliers, $\boldsymbol{\mu}$.

The ADMM solves the original optimization problem by using an iterative block optimization scheme in the (scaled) Augmented Lagrangian from (3.187) with respect to $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{u}$. While $\mathbf{x}$ and $\mathbf{z}$ are obtained by a straightforward minimization of $\mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}, \mathbf{u})$, i.e.:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}_k, \mathbf{u}_k), \tag{3.188a}$$

$$\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}} \mathcal{L}_\lambda(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{u}_k), \tag{3.188b}$$

for some k (denoting the current iteration number), the derivation of the optimal $\boldsymbol{\mu}$ follows in a trickier manner. In this regard, the ADMM determines the optimal value of $\boldsymbol{\mu}$ using the Dual Ascent algorithm [30]. Before diving into the Dual Ascent step of the ADMM, we need to review the concept of *duality* in optimization. In general optimization theory, the *dual function* of (3.179) is defined as follows:

$$g(\mathbf{u}) = \min_{\mathbf{x},\mathbf{z}} \mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}, \mathbf{u}), \tag{3.189}$$

which is known to be a lower bound of the optimization problem in (3.179) [29]. Although the dual function is initially defined using the classic Lagrangian, we already introduced the Augmented Lagrangian in $g(\mathbf{u})$ to obtain the expression of the dual function in which the ADMM is based. Intuitively, maximizing $g(\mathbf{u})$ with respect to $\mathbf{u}$ yields the best possible lower bound of the original problem, which is referred to as the *dual problem* of (3.179). Fortunately, there are several optimization problems in which the lower bound of the dual problem is tight, meaning that the maximum value of $g(\mathbf{u})$ and the minimum value of the original cost function coincide. An instance of those optimization problems is the one found in (3.179) since it satisfies the Slatter's condition [29, Subsection 5.2.3], i.e. the optimization problem is convex and feasible (for the particular case of equality constrained optimization problems). As a result, the dual problem is the optimization problem that is used in the ADMM to retrieve the optimal value of $\mathbf{u}$. The Dual Ascent solves the dual problem by means of the Gradient Ascent (the counterpart of the Gradient Descent) algorithm. Even though the Gradient Ascent algorithm would imply an update equation of the following form:

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \eta\nabla_{\mathbf{u}} g(\mathbf{u}), \tag{3.190}$$

for some step size, $\eta > 0$, the ADMM algorithm is based on (following from (3.188)):

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{c}, \tag{3.191}$$

which is a *subgradient* of $g(\mathbf{u})$ [29, Subsection 6.5.5]. The justification of the update equation in (3.191) is founded on the KKT conditions of (3.179). In the case given by (3.179), the particularized KKT conditions consist in the following two equations [29], [30]:

$$\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}, \tag{3.192a}$$

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*, \mathbf{z}^*, \mathbf{u}^*) = \nabla_{\mathbf{x}} f(\mathbf{x}^*) + \lambda \mathbf{A}^T \mathbf{u}^* = \mathbf{0}_N, \tag{3.192b}$$

$$\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{x}^*, \mathbf{z}^*, \mathbf{u}^*) = \nabla_{\mathbf{z}} h(\mathbf{z}^*) + \lambda \mathbf{B}^T \mathbf{u}^* = \mathbf{0}_N, \tag{3.192c}$$

where $\mathbf{x}^*$, $\mathbf{z}^*$ and $\mathbf{u}^*$ are the optimal values of their respective variables. We remark that the KKT conditions are derived using the classic Lagrangian with $\boldsymbol{\mu} = \lambda \mathbf{u}$ instead of the Augmented Lagrangian. Taking into account (3.188), $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ also minimizes $\mathcal{L}(\mathbf{x}, \mathbf{z}_k, \mathbf{u}_k)$ and $\mathcal{L}(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{u}_k)$, respectively, because of the fact that the Augmented Lagrangian does not change the optimal value of the classic Lagrangian. Considering the fact that $\mathbf{x}_k$, $\mathbf{z}_k$ and $\mathbf{u}_k$ must fulfill the KKT conditions for all $k \in \mathbb{N}$, we get from the gradient of the Augmented Lagrangian that:

$$\mathbf{0}_N = \nabla_{\mathbf{x}}\mathcal{L}_\lambda(\mathbf{x}_{k+1}, \mathbf{z}_{k+1}, \mathbf{u}_k) = \tag{3.193a}$$

$$\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) + \lambda \mathbf{A}^T(\mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{c}) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \lambda \mathbf{A}^T \mathbf{u}_{k+1}. \tag{3.193b}$$

An equivalent argument is derived after differentiating with respect to $\mathbf{z}$. In other words, the update equation in (3.191) is the one ensuring that the Augmented Lagrangian satisfies the KKT conditions in (3.192). In fact, it is thanks to the Augmented Lagrangian technique that the Dual Ascent update in (3.191) is well-behaved (globally convergent) [30], [143], [178] since the added term of the Augmented Lagrangian (see (3.183) or (3.187)) ensures that the overall cost function is strongly convex with respect to $\mathbf{x}$ and $\mathbf{z}$ for convex f and h. The verification of the latter observation can be found in Remark 3.37 applied to the cost function in (3.183).

With the previous ideas in mind, the update equations of the ADMM applied to the original problem given in (3.177) are depicted as follows:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \mathcal{L}_\lambda(\mathbf{x}, \mathbf{z}_k, \mathbf{u}_k) = f(\mathbf{x}) + \frac{\lambda}{2}\left|\left|\mathbf{u}_k + \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{c}\right|\right|_2^2, \tag{3.194a}$$

$$\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}} \mathcal{L}_\lambda(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{u}_k) = h(\mathbf{z}) + \frac{\lambda}{2}\left|\left|\mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_k - \mathbf{c}\right|\right|_2^2, \tag{3.194b}$$

$$\mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{c}, \tag{3.194c}$$

where we evidenced the fact that the respective update equations of $\mathbf{x}$ and $\mathbf{z}$ consist in the proximal operators of f and h. The last details of the ADMM algorithm are the initialization of the variables and setting the user-defined parameter λ. As for the initialization, it is not required that $\mathbf{x}$ and $\mathbf{z}$ are initialized with a feasible value. Yet, it is advisable that $\mathbf{u}$ is initialized to $\mathbf{0}_N$ since, in this case, the update equation of $\mathbf{u}$ becomes the running sum of residuals, i.e.:

$$\mathbf{u}_k = \sum_{k=1}^{K} \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{z}_k - \mathbf{c}, \tag{3.195}$$

where K is the total amount of available iterations. Hence, it provides the dual variable of an intuitive meaning. Regarding the penalty parameter, λ, although the ADMM always converges for any value of λ, its convergence speed is strongly dependent on this parameter [132]. For this reason, selecting the optimal value of λ must be done in a case by case basis. Despite the fact that we choose a value of λ by heuristic procedures in a future chapter, we remark that there are several proposals whose goal is to find a systematic way of selecting λ. For instance, see the approach given in [132]. In Algorithm 1, we summarize all the previous ideas.

Algorithm 1 ADMM algorithm summary

Initialization: λ, K_{max}, $\mathbf{u}_0 = \mathbf{0}_N$ and $\mathbf{z}_0$
for $k = 1$ to K_{max} **do**
  $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\lambda}{2} ||\mathbf{u}_k + \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{c}||_2^2$.
  $\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}} h(\mathbf{z}) + \frac{\lambda}{2} ||\mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z} - \mathbf{c}||_2^2$.
  $\mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{c}$.
  if Stopping condition is met **then**
    Set $k^* = k$.
    Break.
  end if
end for
return $\mathbf{x}_{k^*}$ or $\mathbf{z}_{k^*}$

3.4 Concluding remarks

This chapter is meant to be an exhaustive treatment of the MM framework and its extension to the Grassmann manifold. To this end, this chapter is centered in the analysis of the optimization tools that face the challenges posed by the cost functions proposed in Chapter 2, which are the ones that model (or discover) the diversity presented in a given signal processing application of interest. Although the ideas presented in this chapter were, initially, auxiliary tools to solve the optimization problems that appeared in subsequent chapters, our growing interest in these ideas resulted in the transformation of this chapter into one of the fundamental pillars of this dissertation. In fact, this is one of the reasons why one of the main publications (see [123]) that emanated from this dissertation focuses on the theoretic analysis of the block MM for the Grassmann manifold. The essence of the remaining chapters consists in the link of the ideas presented in chapters 2 and 3.

Chapter 4

Exploiting diversity in data fusion problems

In this chapter, we focus on the multisensor fusion problem as a first approach to explore the meaning of *diversity* in signal processing applications, which is motivated by the multimodal data fusion framework [110]. As stated in [110], diversity is defined as the property that a multimodal dataset has that enhances its performance, its insights and its uses in a particular application in such a way that it could not be achieved by a single modality. In essence, the multimodal data fusion framework consists of the interplay between two different agents: the multimodal dataset containing the information diversity and the fusion algorithm, which exploits the inherent diversity. The multimodal dataset gathers the information of a given phenomenon of interest using several data acquisition frameworks, where each particular acquisition technique is termed as *modality*. An example of a multimodal dataset is the data obtained by a camera and a microphone during a video recording. Nevertheless, we are not interested in the study of multimodal datasets. Indeed, the use of *multimodality* in the literature is often reserved to datasets with heterogeneous data types, e.g. audio, images and text, which is out of the scope of this dissertation. Instead, we focus on the particular case of multisensor fusion, which is based on a network of sensors that share the same acquisition framework. While the multisensor framework is technically a multimodal dataset, we do not use the multimodal label from now on in order to avoid any possible confusion.

Instead, this chapter is devoted to the study of fusion algorithms that exploit the inherent diversity in the multisensor dataset. For the previous reason and being inspired by the definition of diversity given in [110], we consider a particular definition of diversity that is suited for the multisensor fusion problem depicted in this chapter.

Definition 4.1 (Diversity (in a multisensor network))**.** *Diversity is the complementary information that is discovered when a fusion scheme combines the information retrieved from several sensors. This complementary information can be used to improve the accuracy of the sensor network (the variance of the fusion), the uncertainty of the combined measure (the entropy of the fused random variable), or the integrity of the final result.*

Although the Definition 4.1 is rather intuitive, as compared to the definition of diversity in wireless communications [72], it provides an interpretation of the advantage of the joint processing of several sensors. The insights given in Definition 4.1 follow from the common feeling that an ensemble of related datasets *is more than the sum of its parts*. Yet, in order to achieve the capabilities of the joint processing of the sensors, the sensor network must be statistically rich so that there exists the mentioned complementary information. As it will be seen in the analysis shown in this chapter, the correlation between the sensors noise hinders the potential gains of any fusion scheme, being the main challenge of the multisensor fusion framework [103] that we aim at solving in this chapter.

In the multisensor fusion framework, we differentiate two different kinds of diversity, the *spatial* and *temporal* diversity. Whenever there is a shared information between all sensors, it is said that there is *spatial diversity*. The use of this term is motivated by the fact that sensors are typically distributed

throughout a field with the aim of measuring a given phenomenon of interest. Note that, in the previous case, some correlation among the sensors noise may be introduced due to the placement of sensors. Besides, the *temporal diversity* describes the redundancy that appears in the measured phenomena. This kind of diversity is especially interesting if one wants to perform a regression analysis on the phenomenon of interest in addition to the information fusion of the sensor network. Algorithmically speaking, the spatial and temporal diversity consists of the redundancy that appears in the columns and rows, respectively, of a data matrix.

With the previous rationale in mind, we analyze and derive three fusion schemes to solve the general multisensor problem. Following from the rationale shown in Chapter 2, we emphasize in the use of information theoretic criteria as an alternative to classical cost functions, such as the MSE of the fusion. Particularly, we explore the use of the Rényi entropy and its role in the multisensor fusion framework. Regarding the surveyed approaches, while we review and particularize the known Covariance Intersection (CI) principle [43], [88], [99] into the multisensor fusion problem, where we show links with the waterfilling problem in communications [186] and the PMEE criterion (see Definition 2.5), we also derive two new fusion schemes as alternative approaches. The first one of them is based on the MEE fusion of a limiting case of a contaminated Gaussian random variable [188], and the second one is a practical implementation of the PMEE criterion that performs the fusion and the regression of the measurements in a joint manner. The joint fusion and regression approach is derived using the Conditional Maximum Likelihood (CML) principle [134], [160], [166], showing that the CML is naturally linked with the PMEE criterion.

The structure of this chapter is straightforward. Firstly, Section 4.1 states the general fusion problem, where not only we describe the problem statement of the multisensor fusion problem, but we also review the building blocks of fusion schemes. Next, we survey the classical and the information theoretic derivations of the CI principle while also providing a particularized analysis of the optimization of the intersection weights in Section 4.2. Finally, in Sections 4.3 and 4.4 we show an in-depth derivation and analysis of our proposed fusion schemes. In Figure 4.1, we summarize the proposed approaches to the multisensor fusion problem and highlight the main ideas that are explored within each fusion scheme.

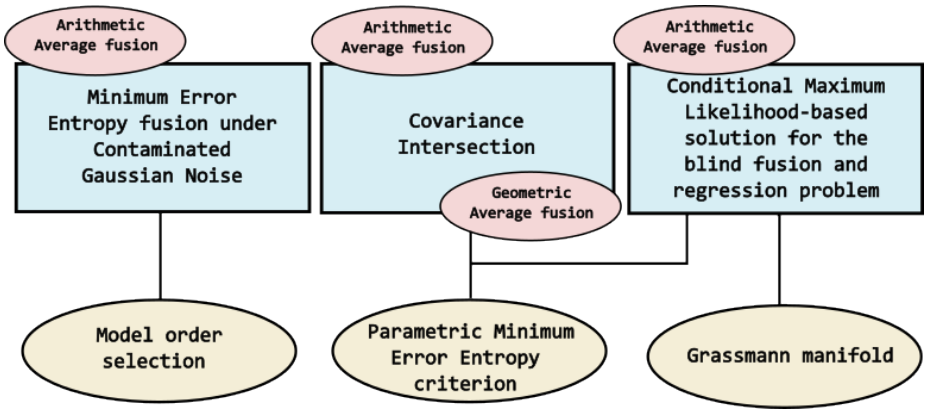


Figure 4.1: Chapter 4 outline.

4.1 General problem statement

All the fusion schemes that are surveyed in this chapter are all based on a simple linear model. Let us consider some phenomenon of interest, denoted as $x(n)$, that is measured by M independent calibrated sensors. The n-th snapshot of M measurements is modeled as follows:

$$\mathbf{y}(n) = x(n)\mathbf{1}_M + \mathbf{w}(n), \tag{4.1}$$

where $\mathbf{w}(n)$ is the noise vector (often considered Gaussian). The intersensor covariance matrix is defined as:

$$\mathrm{E}\left[\mathbf{w}(n)\mathbf{w}^T(n)\right] = \mathbf{Q} \in \mathcal{S}_{++}^M, \tag{4.2}$$

which is assumed to be unknown. The vector that is multiplying $x(n)$ is referred to as the spatial signature and, particularly, it is said that the spatial signature is isotropic when it is composed by an all-ones vectors such as in (4.1). We refer to a set of sensors whose spatial signature is known as calibrated sensors. Note that, for this definition of calibrated sensors, the isotropic spatial signature is not restrictive. In the following lemma, we show the equivalence of (4.1) with a model with an arbitrary but known spatial signature under nominal conditions (Gaussianity of the sensors noise).

Lemma 4.1 (Equivalence of fusion models under known spatial signature)**.** *Let a fusion model be given by:*

$$\mathbf{y}(n) = x(n)\mathbf{a} + \mathbf{w}(n), \tag{4.3}$$

where $\mathbf{a} \in \mathbb{R}^M$ *is its spatial signature such that* $[\mathbf{a}]_m \neq 0$ *for* $m = 1, ..., M$ *and* $\mathbf{w}(n) \sim \mathcal{N}(\mathbf{0}_M, \mathbf{Q})$*. Then, there exists a transformation of* $\mathbf{y}(n)$ *such that:*

$$\mathbf{y}'(n) = x(n)\mathbf{1}_M + \mathbf{w}'(n), \tag{4.4}$$

where $\mathbf{w}'(n) \sim \mathcal{N}(\mathbf{0}_M, \mathbf{\Lambda}_{a^{-1}}\mathbf{Q}\mathbf{\Lambda}_{a^{-1}}^T)$ *and* $\mathbf{\Lambda}_{a^{-1}}$ *is a diagonal matrix such that* $[\mathbf{\Lambda}_{a^{-1}}]_{m,m} = \frac{1}{[\mathbf{a}]_m}$*.*

Proof. Let $\mathbf{b} \in \mathbb{R}^M$ be such that $[\mathbf{b}]_m = \frac{1}{[\mathbf{a}]_m}$ for all $m = 1, ..., M$. Then, we have that:

$$\mathbf{y}'(n) = \mathbf{b} \odot \mathbf{y}(n) = x(n)\mathbf{b} \odot (\mathbf{a} + \mathbf{w}(n)) = x(n)\mathbf{1}_M + \mathbf{b} \odot \mathbf{w}(n), \tag{4.5}$$

which satisfies the isotropic spatial signature condition. The remaining step is to describe the statistical distribution of the new noise vector, i.e. $\mathbf{w}'(n) = \mathbf{b} \odot \mathbf{w}(n)$. Since the original noise vector is Gaussian, $\mathbf{w}'(n)$ is also Gaussian. Thus, we only need to describe the first and second-order moments of the new noise vector. Its expected value is given by:

$$\mathrm{E}[\mathbf{b} \odot \mathbf{w}(n)] = \mathbf{0}_M, \tag{4.6}$$

while its covariance matrix yields:

$$\mathrm{E}[(\mathbf{b} \odot \mathbf{w}(n))(\mathbf{b} \odot \mathbf{w}(n))^T] = \mathrm{E}[(\mathbf{\Lambda}_{a^{-1}}\mathbf{w}(n))(\mathbf{\Lambda}_{a^{-1}}\mathbf{w}(n))^T] =, \tag{4.7a}$$

$$\mathbf{\Lambda}_{a^{-1}}\,\mathrm{E}[\mathbf{w}(n)\mathbf{w}(n)^T]\mathbf{\Lambda}_{a^{-1}}^T = \mathbf{\Lambda}_{a^{-1}}\mathbf{Q}\mathbf{\Lambda}_{a^{-1}}^T, \tag{4.7b}$$

where $\mathbf{\Lambda}_{a^{-1}}$ is a diagonal matrix such that $[\mathbf{\Lambda}_{a^{-1}}]_{m,m} = [\mathbf{b}]_m = \frac{1}{[\mathbf{a}]_m}$. From (4.6) and (4.7), we get the transformed model in (4.4). ■

For simplicity, given the equivalence shown in Lemma 4.1, we prefer models with isotropic spatial signatures, such as the one in (4.1), to a model with a general spatial signature.

The general multisensor fusion problem depicted by (4.1) consists in the retrieval of $x(n)$ from the available measurements contained in $\mathbf{y}(n)$. We exploit the spatial diversity that is present in $\mathbf{y}(n)$ in order to achieve a better estimation of $x(n)$ as compared to the arithmetic mean of the measurements, which may yield a poor performance due to corrupted sensors. The fundamental challenges that arise in this setting are the lack of model knowledge and the presence of anomalies. Essentially, the lack of model knowledge appears when one (or more) parameters that define the model are unknown to the estimation framework. Some particular examples of this challenge are the lack of statistical knowledge, e.g. $\mathbf{Q}$ is unknown, or the mismodeling of the measured phenomena, $x(n)$. The latter case is of interest when the regression task is also considered.

In the following subsections, we survey two key aspects of the stated multisensor fusion problem. On the one hand, we present the *arithmetic* and *geometric average fusion rules* [112], [114], which are the fundamental concepts of data fusion schemes. On the other hand, in order to gain more perspectives on the multisensor fusion problem, we analyze the fusion rule that is obtained when there is full statistical knowledge, which is a benchmark that is set to test the performance of the derived fusion policies.

4.1.1 Arithmetic and Geometric average fusion rules

For the purpose of defining fusion schemes to obtain $x(n)$ from $\mathbf{y}(n)$, two kinds of fusions are often considered in the statistical signal processing literature: the fusion of estimators and the fusion of PDFs. These two techniques play a fundamental role in multisensor data fusion and are based on the *arithmetic average* (AA) and the *geometric average* (GA) fusion policies. As a starting point, the AA and GA fusion of an estimator are defined as follows [112].

Definition 4.2 (Arithmetic Average fusion of estimators)**.** *Let $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M$ be arbitrary estimators. Then, their Arithmetic Average (AA) fusion is given by:*

$$\boldsymbol{\theta}_{AA} = \sum_{m=1}^{M} \omega_m \boldsymbol{\theta}_m, \tag{4.8}$$

where ω_m for $m = 1, ..., M$ are such that $\sum_{m=1}^{M} \omega_m = 1$ and that $\omega_m \in [0, 1] \ \ \forall m = 1, ..., M$. In other words, $\boldsymbol{\theta}_{AA}$ is a convex combination of the available estimators (see (3.23)).

Definition 4.3 (Geometric Average fusion of estimators)**.** *Let $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M$ be arbitrary estimators. Then, their Geometric Average (GA) fusion is given by:*

$$\log(\boldsymbol{\theta}_{GA}) = \sum_{m=1}^{M} \omega_m \log(\boldsymbol{\theta}_m), \tag{4.9}$$

where ω_m are such that $\sum_{m=1}^{M} \omega_m = 1$ and that $\omega_m \in [0, 1] \ \ \forall m = 1, ..., M$. The logarithms in (4.9) are applied element-wise on their input.

Remark 4.1. The geometric average is not a real number in general. Indeed, if any $\boldsymbol{\theta}_m$ is negative, then $\boldsymbol{\theta}_{GA}$ is a complex number.

Remark 4.2. From the Arithmetic Mean-Geometric Mean (AM-GM) inequality [23], it is verified that $\boldsymbol{\theta}_{AA} \geq \boldsymbol{\theta}_{GA}$ in the case of non-negative measurements.

Between the AA and the GA fusion rules, the AA is often preferred to GA for the fusion of a set of estimators because it preserves the unbiasedness. In order to verify this behavior, consider M unbiased estimators of some parameter $\boldsymbol{\theta} \in \mathbb{R}^N$, denoted as $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M$. Then, the AA fusion verifies:

$$\mathrm{E}[\boldsymbol{\theta}_{AA} - \boldsymbol{\theta}] = \sum_{m=1}^{M} \omega_m \, \mathrm{E}[\boldsymbol{\theta}_m] - \boldsymbol{\theta} = \sum_{m=1}^{M} \omega_m \boldsymbol{\theta} - \boldsymbol{\theta} = \mathbf{0}_N, \tag{4.10}$$

which follows immediately from the fact that ω_m for $m = 1, ..., M$ defines a convex combination. Provided that (4.10) is true, it follows from Remark 4.2 that the GA fusion is biased in general. Not only that, but the AA fusion of estimators also offers a potentially better Mean Squared Error (MSE) than the GA fusion of estimators [112]. For the previous reasons, the AA fusion of estimators plays a pivotal role in this thesis.

The advantages of the GA fusion with respect to the AA fusion appear when one is interested in the fusion of PDFs, Probability Hypothesis Density[1] (PHD) or likelihood functions. Indeed, the majority of the information theoretic-based multisensor fusion approaches are founded on the fusion of PDFs, being the Covariance Intersection (CI) principle the staple approach [14], [88], [99]. For completeness, we review the AA and GA fusion of PDFs. The AA fusion of PDFs is defined as follows [112].

Definition 4.4 (Arithmetic Average fusion of PDFs)**.** *Let $f_1(\mathbf{x}|\boldsymbol{\theta}_1), ..., f_M(\mathbf{x}|\boldsymbol{\theta}_M)$ be the representatives of some measurements given in the form of PDFs. Then, their Arithmetic Average (AA) fusion is given by:*

$$f_{AA}(\mathbf{x}|\boldsymbol{\theta}_{AA}) = \sum_{m=1}^{M} \omega_m f_m(\mathbf{x}|\boldsymbol{\theta}_m), \tag{4.11}$$

where ω_m are such that $\sum_{m=1}^{M} \omega_m = 1$ and that $\omega_m \in [0, 1] \ \ \forall m = 1, ..., M$. In other words, $f_{AA}(\mathbf{x}|\boldsymbol{\theta}_{AA})$ is the convex combination of the measurements PDFs.

[1] The main difference between a PHD and a PDF is that the integral of a PHD can be any non-negative value. The remaining axioms of the PDF are also fulfilled by the PHD. The PHD is often considered in detection problems that exploit the multisensor fusion framework.

Remark 4.3. Any mixture model can also be interpreted as an AA fusion of several PDFs.

Although it may seem that the AA fusion of a function is *ad-hoc*, its use is often driven by its desirable properties, especially for the fusion of PHDs [114]. The advantages of the AA fusion of PDFs can be summarized by two different items. From one perspective, the resulting PDF from the AA fusion is biased towards the majority, which is a kind of behavior that provides robustness to an information fusion classification problem since it corrects the *local* (each sensor) errors in favour of the spatial diversity. Additionally, given that the AA fusion can be obtained by a fast procedure (summing several PDFs), it is a reasonable choice for online fusion algorithms [113]. An example of an application of the AA fusion of PDFs in the literature is found in the identification of multiple targets, where the resulting detector is built out of an AA fusion, yielding a *democratic* decision that is shared between several sensors [113], [126]. The main drawback of the AA fusion of PDFs is that, whenever all the sources PDF are all the same, the resulting fused PDF is equal to the sources PDF, yielding no additional gain from this operation.

While the AA fusion of PDFs is known for its robustness, the GA fusion thrives in terms of the performance of the resulting fusion task, especially under nominal conditions (Gaussian PDFs) [102], [112]. The GA fusion of PDFs is defined as follows.

Definition 4.5 (Geometric Average fusion of a function)**.** *Let $f_1(\mathbf{x}|\boldsymbol{\theta}_1), ..., f_M(\mathbf{x}|\boldsymbol{\theta}_M)$ be some representatives of some measurements given in the form of PDFs. Then, their Geometric Average (GA) fusion is given by:*

$$f_{GA}(\mathbf{x}|\boldsymbol{\theta}_{GA}) = \frac{1}{G}\prod_{m=1}^{M} f_m^{\omega_m}(\mathbf{x}|\boldsymbol{\theta}_m), \tag{4.12}$$

where ω_m are such that $\sum_{m=1}^{M}\omega_m = 1$, that $\omega_m \in [0,1] \;\; \forall m = 1, ..., M$ and:

$$G = \int_{-\infty}^{\infty}\prod_{m=1}^{M} f_m^{\omega_m}(\mathbf{x}|\boldsymbol{\theta}_m)\,d\mathbf{x}. \tag{4.13}$$

Remark 4.4. The GA is also known as the Geometric Mean Density (GMD) [14].

In simpler words, the GA fusion of PDFs is the convex combination of log-likelihoods, which can also be interpreted as a sort of convex combination of information. In this regard, the geometric average nature of this kind of fusion implies that the fusion of values with a small density *dominate*, meaning that these values of the sources PDF dictate the shape of the resulting GA fused PDF. The information theoretic interpretation of this phenomenon is that the values with a small density contain a large amount of information and, for this reason, they dominate. This phenomenon does not occur in the AA fusion of PDFs. The GA fusion of PDFs naturally emerges in multisensor fusion schemes [194]. For instance, a scaled geometric fusion appears when one fuses the PDFs of independent random variables. We are interested in the role that GA plays in the CI principle, which, in essence, consists in the GA fusion of Gaussian PDFs. In fact, it is thanks to the GA fusion of PDFs that the CI algorithm is closely related to the Chernoff information and the Rényi entropy of the fused variable [88], [142].

4.1.2 Benchmark fusion policy

For the model given in (4.1), the benchmark fusion policy is defined as the *best* linear combiner based on an AA fusion that retrieves $x(n)$ from $\mathbf{y}(n)$. We define the fused variable as the following linear combination on $\mathbf{y}(n)$:

$$s(n) = \mathbf{f}^T\mathbf{y}(n), \tag{4.14}$$

where $\mathbf{f} \in \mathbb{R}^M$ is such that $\mathbf{f}^T\mathbf{1}_M = 1$. Notice that the previous constraint on $\mathbf{f}$ appears naturally to yield an unbiased fusion. Still, we do not consider the positivity constraint of the AA fusion since it is not known up to this point if it is restrictive for the fusion task. In the following proposition, we study the optimal fusion policy with full statistical knowledge in terms of the minimum variance and the minimum error entropy criteria.

Proposition 4.2 (Equivalence of the MEE and minimum variance criterion fusion under full statistical knowledge)**.** *Let the fusion error random variable be defined as follows:*

$$e(n, \mathbf{f}) = \mathbf{f}^T\mathbf{y}(n) - x(n), \tag{4.15}$$

where $\mathbf{y}(n)$ *is a random process statistically distributed as:*

$$\mathbf{y}(n) \sim \mathcal{N}(x(n)\mathbf{1}_M, \mathbf{Q}). \tag{4.16}$$

Then, the following fusion rule:

$$\mathbf{f}_B = \frac{\mathbf{Q}^{-1}\mathbf{1}_M}{\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{1}_M}, \tag{4.17}$$

is optimal in the minimum variance and minimum error entropy sense. The resulting variance is given by:

$$\gamma_B = \mathrm{E}\left[|e(n, \mathbf{f}_B)|^2\right] = \frac{1}{\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{1}_M}, \tag{4.18}$$

while its respective Rényi differential entropy is:

$$h(e(n, \mathbf{f}_B)) = \frac{M}{2}\left(\log(2\pi) - \log\left(\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{1}_M\right) + \frac{\log(\alpha)}{\alpha - 1}\right). \tag{4.19}$$

The proof of this proposition can be found in Appendix 8.2.1.

Although the previous proposition certifies the conditions in which the equivalence of the MEE and the minimum variance fusion holds, they are rarely found in real applications. Even if those conditions (full statistical knowledge) hold in a practical setting, the resulting fusion scheme is not robust to any kind of model inaccuracies or deviations from the nominal conditions. In spite of that, the fusion rule derived in Proposition 4.2 provides us a reference point for the fusion schemes that are derived in this chapter since it is the best possible performance that a linear fusion scheme can reach. Also, we remark that, although minimizing entropy is equivalent to minimizing the variance in the nominal case, one of the most important aspects that will be highlighted in forthcoming sections is that the approach of minimizing entropy turns out to provide much more robustness in front of model deviations. This constitutes a possible path to pursue robust statistics methods based on information theoretical concepts in the line of this thesis.

4.2 Covariance Intersection

Assuming that there exists a set of (possibly correlated) estimators of a given quantity and their respective estimated uncertainty (variance), the CI principle is a theoretic fusion framework that is aimed at the fusion of the aforementioned estimates. Even though the original derivation of the CI algorithm is founded in the intersection of the locus points of the ellipse defined by the covariance matrix of the estimates, we also review the information theoretic foundations of the CI [88] and relate them to the PMEE criterion described in Definition 2.5. The CI principle is built upon the following assumptions [189]:

1. The resulting fusion of the estimated quantities must produce an improved estimate, i.e. the (estimated) variance of the resulting fusion must be lower than the original variance of the estimations.
2. The fused estimate must be conservative in the sense that the estimation of the fused variance must be higher than the true variance of the fusion. This issue, which indirectly comes from tracking problems that use Kalman filtering to avoid divergence [99], is especially important in those applications where the integrity of the resulting fusion is important to the final user. For example, when the final measurements are to be used to take sensible decisions.
3. The cross-correlations among the estimates are either unknown or they cannot be obtained by the fusion scheme in a practical manner.

The previous assumptions depict the conservative nature of the CI principle. Indeed, we show in this section that this conservative behavior may be too extreme for the fusion problem described in Section 4.1, yielding an unwanted *best sensor selection* policy. Since we want to exploit the diversity present in the multisensor dataset, we show an alternative that achieves the previous goal while keeping the degree of conservativeness of the CI principle.

4.2.1 Derivation of the Covariance Intersection principle

Firstly, we review the derivation of the CI as in its seminal paper [43], [99], [189] and, then, we review an alternative derivation of the CI using the GA fusion of PDFs [88], from where information theoretic arguments naturally emerge. For initial simplicity, we consider the fusion of information of two sources, termed as A and B, whose information is contained on the N-dimensional random vectors given by $\mathbf{x}_a$ and $\mathbf{x}_b$, respectively. These sources are contaminated by noise with unknown statistics, resulting in the following covariance matrices:

$$\mathbf{Q}_{aa} = \mathrm{E}\left[(\mathbf{x}_a - \mathbf{a})(\mathbf{x}_a - \mathbf{a})^T\right], \tag{4.20a}$$

$$\mathbf{Q}_{bb} = \mathrm{E}\left[(\mathbf{x}_b - \mathbf{b})(\mathbf{x}_b - \mathbf{b})^T\right], \tag{4.20b}$$

where $\mathrm{E}[\mathbf{x}_a] = \mathbf{a}$ and $\mathrm{E}[\mathbf{x}_b] = \mathbf{b}$. Their cross-covariances are defined in a similar manner:

$$\mathbf{Q}_{ab} = \mathbf{Q}_{ba} = \mathrm{E}\left[(\mathbf{x}_a - \mathbf{a})(\mathbf{x}_b - \mathbf{b})^T\right] = \mathrm{E}\left[(\mathbf{x}_b - \mathbf{b})(\mathbf{x}_a - \mathbf{a})^T\right], \tag{4.21}$$

which are different from $\mathbf{0}_{N,N}$ in general. Provided that we assumed that the statistics of the measurements noise are unknown, the true values of the previously defined matrices are not available. Instead, consistent estimations of $\mathbf{Q}_{aa}$ and $\mathbf{Q}_{bb}$ are obtainable, which are denoted as $\hat{\mathbf{Q}}_{aa}$ and $\hat{\mathbf{Q}}_{bb}$, respectively. The consistency of those estimators is defined as in [99, Eq. (1)]:

$$\hat{\mathbf{Q}}_{aa} - \mathbf{Q}_{aa} \succeq \mathbf{0}_{N,N}, \tag{4.22a}$$

$$\hat{\mathbf{Q}}_{bb} - \mathbf{Q}_{bb} \succeq \mathbf{0}_{N,N}. \tag{4.22b}$$

In simpler words, the expressions in (4.22) can be interpreted as a possible condition that ensures a minimum degree of robustness of the covariance matrix estimators of A and B. Taking into account the previous definition of consistency, the objective of the CI principle is to find a linear fusion of A and B such that the resulting random variable is consistent. The resulting fused source is denoted as C and its respective random variable is given by the following linear fusion:

$$\mathbf{x}_c = \mathbf{K}_a\mathbf{x}_a + \mathbf{K}_b\mathbf{x}_b, \tag{4.23}$$

where $\mathbf{K}_a$ and $\mathbf{K}_b$ are the weight matrices. The previous weight matrices, $\mathbf{K}_a$ and $\mathbf{K}_b$, are designed in such a way that if $\mathbf{a} = \mathbf{b}$, then $\mathrm{E}[\mathbf{x}_c] = \mathbf{c} = \mathbf{a} = \mathbf{b}$. In other words, we want to maintain the unbiasedness of the fusion. This means that the weight matrices must satisfy:

$$\mathbf{K}_a + \mathbf{K}_b = \mathbf{I}_M. \tag{4.24}$$

The consistency of the fused variable, $\mathbf{x}_c$, is depicted by the following conditions:

$$\hat{\mathbf{Q}}_{aa} - \hat{\mathbf{Q}}_{cc} \succeq \mathbf{0}_{N,N}, \tag{4.25a}$$

$$\hat{\mathbf{Q}}_{bb} - \hat{\mathbf{Q}}_{cc} \succeq \mathbf{0}_{N,N}, \tag{4.25b}$$

$$\hat{\mathbf{Q}}_{cc} - \mathbf{Q}_{cc} \succeq \mathbf{0}_{N,N}, \tag{4.25c}$$

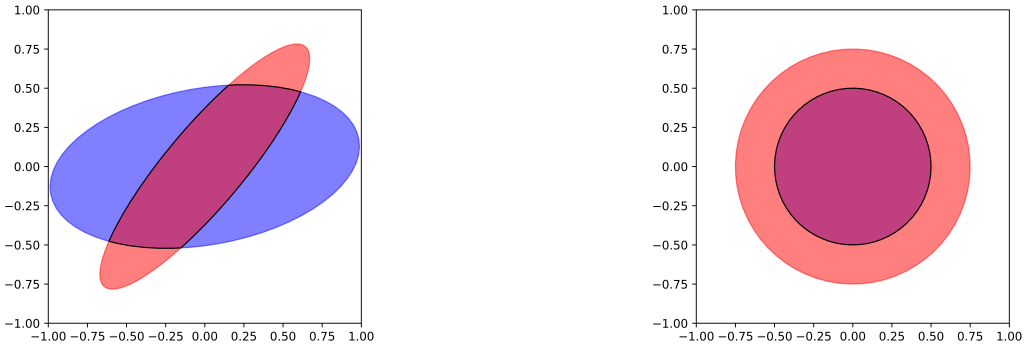
where $\mathbf{Q}_{cc}$ is the true value of the fused variable covariance matrix and $\hat{\mathbf{Q}}_{cc}$ is the estimation of $\mathbf{Q}_{cc}$. To put it in another way, $\hat{\mathbf{Q}}_{cc}$ is a conservative improvement of the original estimated covariances, $\hat{\mathbf{Q}}_{aa}$ and $\hat{\mathbf{Q}}_{bb}$. Given the expression of the fusion in (4.23), the true value of the fusion covariance matrix is given by:

$$\mathbf{Q}_{cc} = \mathrm{E}\left[(\mathbf{x}_c - \mathbf{c})(\mathbf{x}_c - \mathbf{c})^T\right] = \mathbf{K}_a\mathbf{Q}_{aa}\mathbf{K}_a^T + \mathbf{K}_a\mathbf{Q}_{ab}\mathbf{K}_b^T + \mathbf{K}_b\mathbf{Q}_{ba}\mathbf{K}_a^T + \mathbf{K}_b\mathbf{Q}_{bb}\mathbf{K}_b^T. \tag{4.26}$$

In the case where $\mathbf{Q}_{ab}$ and $\mathbf{Q}_{ba}$ (or some consistent estimates) are known, one would be tempted to substitute the consistent estimates of the covariance matrices, $\hat{\mathbf{Q}}_{aa}$ and $\hat{\mathbf{Q}}_{bb}$ (see (4.22)), into (4.26) to obtain the estimate of the fused covariance matrix, $\hat{\mathbf{Q}}_{cc}$. However, even with the assumption that $\mathbf{Q}_{ab} = \mathbf{Q}_{ba} = \mathbf{0}_{N,N}$, ensuring the consistency of the resulting plug-in estimator of $\mathbf{Q}_{cc}$, i.e.:

$$\hat{\mathbf{Q}}'_{cc} = \mathbf{K}_a\hat{\mathbf{Q}}_{aa}\mathbf{K}_a^T + \mathbf{K}_b\hat{\mathbf{Q}}_{bb}\mathbf{K}_b^T, \tag{4.27}$$

is difficult in general [99], [100]. For the previous reason, using $\hat{\mathbf{Q}}'_{cc}$ is not a good strategy in general. In contrast to the approach that yields (4.27), the CI principle exploits the geometrical interpretation of (4.26) to obtain $\mathbf{K}_a$ and $\mathbf{K}_b$ (4.25) [43], [99]. Indeed, equation (4.26) is, essentially, the expression of the intersection between the ellipses defined by $\mathbf{Q}_{aa}$ and $\mathbf{Q}_{bb}$, which is an idea that is intuitively detailed later on by means of Definition 4.6. In order to provide a preliminary intuition to the previous idea, we show a graphical representation of (4.26) in Figure 4.2 for two different cases of the intersection between two ellipses in $\mathbb{R}^2$. To put it simply, the CI algorithm is aimed at finding a third ellipse that encloses the black contoured shapes. In the first case, we depict a *rich* intersection between two ellipses, where we see that the information of both ellipses is needed to obtain the intersection. In contrast, the second case would result in the best sensor selection policy since only the information of the smaller ellipse is used in the CI principle, which is an undesirable case due to the lost spatial diversity.



(a) *Rich* intersection of ellipses.

(b) *Poor* intersection of ellipses.

Figure 4.2: Intersection of ellipses and the CI principle.

For the purpose of providing an intuitive derivation of the CI using the intersection of ellipses idea, let us define the locus points of an ellipse described by a positive definite matrix.

Definition 4.6 (Locus of points of an ellipse)**.** *Consider an arbitrary positive definite matrix,* $\mathbf{C} \in \mathcal{S}^N_{++}$ *and a point* $\boldsymbol{\mu} \in \mathbb{R}^N$*. Then, the ellipse described by these two parameters is given by the following locus of points:*

$$\mathcal{E}_C(\mathbf{Q}, \boldsymbol{\mu}) = \{\mathbf{x} \in \mathbb{R}^N : u(\mathbf{Q}, \mathbf{x} - \boldsymbol{\mu}) \leq C\}, \tag{4.28}$$

where C *is a positive constant that is related to the volume of the ellipse and:*

$$u(\boldsymbol{\Sigma}, \mathbf{z}) = \mathbf{z}^T \boldsymbol{\Sigma}^{-1} \mathbf{z}, \tag{4.29}$$

is the function that defines the contour of the ellipse [189].

Remark 4.5. The points $\mathbf{x} \in \mathbb{R}^N$ that satisfy:

$$u(\mathbf{Q}, \mathbf{x} - \boldsymbol{\mu}) = C, \tag{4.30}$$

can also be interpreted as the set of points that are equally likely in the Gaussian distribution described by $\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q})$. Also, note that, for a given C, the overall volume of the ellipse is proportional to the geometric average of the eigenvalues of the covariance, i.e its determinant.

It is stated in [99] that, for every possible value of $\mathbf{Q}_{ab}$ and $\mathbf{Q}_{ba}$ in (4.26), $\mathcal{E}_C(\mathbf{Q}_{cc}, \mathbf{c})$ always lies in the intersection (assuming that it is non-empty) of $\mathcal{E}_C(\mathbf{Q}_{aa}, \mathbf{a})$ and $\mathcal{E}_C(\mathbf{Q}_{bb}, \mathbf{b})$. Following from the previous idea, the CI principle constructs an ellipse, $\mathcal{E}_C(\hat{\mathbf{Q}}_{cc}, \mathbf{c})$, with a volume as small as possible from the mean and covariance estimates, $\mathbf{c}$ and $\hat{\mathbf{Q}}_{cc}$, such that it encloses the intersection of $\mathcal{E}_C(\hat{\mathbf{Q}}_{aa}, \mathbf{a})$ and $\mathcal{E}_C(\hat{\mathbf{Q}}_{bb}, \mathbf{b})$ without any information about the cross-covariances, $\mathbf{Q}_{ab}$ and $\mathbf{Q}_{ba}$. A consequence of the previous idea is that the resulting estimated covariance, $\hat{\mathbf{Q}}_{cc}$, is consistent as in (4.25). In order to intuitively derive the estimated fused mean and covariance using the CI principle, note that:

$$u(\mathbf{Q}_{bb}, \mathbf{x} - \mathbf{b}) \geq u(\mathbf{Q}_{aa}, \mathbf{x} - \mathbf{a}), \tag{4.31}$$

implies that the value of $\mathbf{x}$ is more likely to occur for a random variable depicted by $\mathcal{N}(\mathbf{a}, \mathbf{Q}_{aa})$ than the one distributed as $\mathcal{N}(\mathbf{b}, \mathbf{Q}_{bb})$ (see Remark 4.5). An equivalent argument is stated in [189], where (4.31) is related to the *sigma contours* (set of points that are equally likely) of a Gaussian distribution. As a result of the previous observation, any feasible estimates of the fused mean and covariance, $\mathbf{c}$ and $\hat{\mathbf{Q}}$, must satisfy [189, Eq. (3)]:

$$\max\Big(u(\hat{\mathbf{Q}}_{aa}, \mathbf{x}-\mathbf{a}), u(\hat{\mathbf{Q}}_{bb}, \mathbf{x}-\mathbf{b})\Big) \geq u(\hat{\mathbf{Q}}_{cc}, \mathbf{x}-\mathbf{c}), \tag{4.32}$$

for all $\mathbf{x} \in \mathbb{R}^M$. In simpler words, the condition in (4.32) ensures that the resulting fusion is an improvement with respect to the original measurements. The CI fusion scheme is obtained after noting that the maximum of two values is lower bounded by their convex combination. Thus, we get the following set of inequalities:

$$\max\Big(u(\hat{\mathbf{Q}}_{aa}, \mathbf{x}-\mathbf{a}), u(\hat{\mathbf{Q}}_{bb}, \mathbf{x}-\mathbf{b})\Big) \geq \omega u(\hat{\mathbf{Q}}_{aa}, \mathbf{x}-\mathbf{a}) + (1-\omega)u(\hat{\mathbf{Q}}_{bb}, \mathbf{x}-\mathbf{b}) = \tag{4.33a}$$

$$\omega u(\hat{\mathbf{Q}}_{aa}^{-1}, \hat{\mathbf{Q}}_{aa}^{-1}(\mathbf{x}-\mathbf{a})) + (1-\omega)u(\hat{\mathbf{Q}}_{bb}^{-1}, \hat{\mathbf{Q}}_{bb}^{-1}(\mathbf{x}-\mathbf{b})) \geq \tag{4.33b}$$

$$u(\omega\hat{\mathbf{Q}}_{aa}^{-1} + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}, \omega\hat{\mathbf{Q}}_{aa}^{-1}(\mathbf{x}-\mathbf{a}) + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}(\mathbf{x}-\mathbf{b})). \tag{4.33c}$$

The justification of the previous expressions is found in [189], which is based on the convexity of $u(\mathbf{P}, \mathbf{z})$ with respect to $\mathbf{P} = \mathbf{Q}^{-1}$ (a precision matrix) and $\mathbf{z}$. Now, we want to get the expression of $(\mathbf{c}, \hat{\mathbf{Q}}_{cc})$ such that $u(\hat{\mathbf{Q}}_{cc}, \mathbf{x}-\mathbf{c})$ is equal to the lower bound shown in (4.33c). The expressions of $\mathbf{c}$ and $\hat{\mathbf{Q}}_{cc}$ are obtained from the following equality (see [189] for more details):

$$u(\omega\hat{\mathbf{Q}}_{aa}^{-1} + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}, \omega\hat{\mathbf{Q}}_{aa}^{-1}(\mathbf{x}-\mathbf{a}) + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}(\mathbf{x}-\mathbf{b})) = (\mathbf{x}-\mathbf{c}_\omega)^T\hat{\mathbf{Q}}_\omega^{-1}(\mathbf{x}-\mathbf{c}_\omega) = \tag{4.34a}$$

$$u(\hat{\mathbf{Q}}_\omega, \mathbf{x}-\mathbf{c}_\omega), \tag{4.34b}$$

where:

$$\hat{\mathbf{Q}}_\omega^{-1} = \omega\hat{\mathbf{Q}}_{aa}^{-1} + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}, \tag{4.35a}$$

$$\mathbf{c}_\omega = \hat{\mathbf{Q}}_\omega\left(\omega\hat{\mathbf{Q}}_{aa}^{-1}\mathbf{a} + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}\mathbf{b}\right), \tag{4.35b}$$

are the resulting covariance matrix and average value of the fusion, respectively, of the CI algorithm. We introduced the subscript ω in $(\mathbf{c}_\omega, \hat{\mathbf{Q}}_\omega)$, which is equivalent to $(\mathbf{c}, \hat{\mathbf{Q}}_{cc})$, to make more evident the dependence of the fused mean and covariance matrix with ω. In this sense, the free parameter, ω, weights the importance assigned to each information source, A and B. It is important to remark that the fused covariance in (4.35a) inherits the consistency of $\hat{\mathbf{Q}}_{aa}$ and $\hat{\mathbf{Q}}_{bb}$ for every possible value of ω (see [99, Appendix A] for the formal proof). An immediate generalization of (4.35) for M different sensors is:

$$\hat{\mathbf{Q}}_\omega^{-1} = \sum_{m=1}^{M} \omega_m\hat{\mathbf{Q}}_m^{-1}, \tag{4.36a}$$

$$\mathbf{c}_\omega = \hat{\mathbf{Q}}_\omega\left(\sum_{m=1}^{M} \omega_m\hat{\mathbf{Q}}_m^{-1}\mathbf{a}_m\right), \tag{4.36b}$$

where $\mathbf{a}_m$ and $\hat{\mathbf{Q}}_m$ for $m = 1, ..., M$ are the expected value and the covariance matrix of the m-th sensor measurements, respectively, and the intersection weights are such that $\sum_{m=1}^{M}\omega_m = 1$ and that $\omega_m \geq 0 \ \ \forall m$. Again, (4.36) consists in a convex combination of the precision matrices and on the weighted combination of measurements. The generalized expression can be obtained by applying recursively the CI rationale between two sensors to the M sensors.

4.2.1.1 GA fusion interpretation of the CI principle

An alternative derivation of the fusion equations given in (4.35) (and (4.36)) comes from the GA fusion of the Gaussian PDFs associated to $\mathbf{x}_a$ and $\mathbf{x}_b$. Let us assume that the measurements of each source are statistically distributed as:

$$\mathbf{x}_a \sim \mathcal{N}(\mathbf{a}, \mathbf{Q}_{aa}), \tag{4.37a}$$

$$\mathbf{x}_b \sim \mathcal{N}(\mathbf{b}, \mathbf{Q}_{bb}), \tag{4.37b}$$

and let us denote their associated PDFs as $p_A(\mathbf{x})$ and $p_B(\mathbf{x})$, respectively. Then, the GA fusion of $p_A(\mathbf{x})$ and $p_B(\mathbf{x})$ is given by:

$$p_C(\mathbf{x}) = \frac{p_A^{\omega}(\mathbf{x}) p_B^{1-\omega}(\mathbf{x})}{\int_{-\infty}^{\infty} p_A^{\omega}(\mathbf{y}) p_B^{1-\omega}(\mathbf{y}) \mathrm{d}\mathbf{y}} = \tag{4.38a}$$

$$\frac{\exp\left(-\frac{\omega}{2}(\mathbf{x}-\mathbf{a})^T \mathbf{Q}_{aa}^{-1}(\mathbf{x}-\mathbf{a}) - \frac{(1-\omega)}{2}(\mathbf{x}-\mathbf{b})^T \hat{\mathbf{Q}}_{bb}^{-1}(\mathbf{x}-\mathbf{b})\right)}{\int_{-\infty}^{\infty} \exp\left(-\frac{\omega}{2}(\mathbf{y}-\mathbf{a})^T \mathbf{Q}_{aa}^{-1}(\mathbf{y}-\mathbf{a}) - \frac{(1-\omega)}{2}(\mathbf{y}-\mathbf{b})^T \hat{\mathbf{Q}}_{bb}^{-1}(\mathbf{y}-\mathbf{b})\right) \mathrm{d}\mathbf{y}}, \tag{4.38b}$$

where (4.38b) is obtained after canceling out the multiplicative constants on the numerator and denominator, and $p_C(\mathbf{x})$ denotes the PDF of the fused random variable. We are interested in the expression of this GA fusion to unveil its connection to the CI principle. Note that the exponents can be further expanded as follows [88]:

$$\frac{\omega}{2}(\mathbf{x}-\mathbf{a})^T \mathbf{Q}_{aa}^{-1}(\mathbf{x}-\mathbf{a}) + \frac{(1-\omega)}{2}(\mathbf{x}-\mathbf{b})^T \hat{\mathbf{Q}}_{bb}^{-1}(\mathbf{x}-\mathbf{b}) = \tag{4.39a}$$

$$\frac{1}{2}\left((\mathbf{x}-\mathbf{c}_\omega)^T \mathbf{Q}_{\omega}^{-1}(\mathbf{x}-\mathbf{c}_\omega) + \omega \mathbf{a}^T \mathbf{Q}_{aa}^{-1}\mathbf{a} + (1-\omega)\mathbf{b}^T \mathbf{Q}_{bb}^{-1}\mathbf{b} - \mathbf{c}_\omega^T \mathbf{Q}_{\omega}^{-1}\mathbf{c}_\omega\right), \tag{4.39b}$$

where $\mathbf{c}_\omega$ and $\mathbf{Q}_\omega$ are defined as in (4.35). Since the terms that do no depend on $\mathbf{x}$ in (4.39b) are common in the numerator and the denominator, they cancel out. The last step to compute the GA fusion is to get the expression of the integral in the denominator after canceling out the terms that are independent of $\mathbf{x}$:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{c}_\omega)^T \mathbf{Q}_{\omega}^{-1}(\mathbf{y}-\mathbf{c}_\omega)\right) \mathrm{d}\mathbf{y} = \sqrt{(2\pi)^M \det(\mathbf{Q}_\omega)}, \tag{4.40}$$

resulting in the final expression of the GA fusion of the measurements PDF:

$$p_C(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M \det(\mathbf{Q}_\omega)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{c}_\omega)^T \mathbf{Q}_{\omega}^{-1}(\mathbf{x}-\mathbf{c}_\omega)\right). \tag{4.41}$$

The conclusion of the previous arguments is that the CI equations in (4.35) are estimating the mean and the covariance matrix of the fused random variable, whose associated PDF is $p_C(\mathbf{x})$, from the estimations of the original sources of information, A and B. Thus, the CI principle assumes that the original measurements are Gaussian. Besides, the GA fusion of two PDFs is closely related to the Chernoff Information, which is defined by the following optimization [142]:

$$C(A,B) = \log \min_{\omega \in [0,1]} \int_{-\infty}^{\infty} p_A^{\omega}(\mathbf{x}) p_B^{1-\omega}(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{4.42}$$

Notice that, essentially, the integrand in (4.42) is a scaled GA fusion of the associated PDFs to A and B. The Chernoff Information has the interpretation of the exponent that yields the best upper bound on the error probability in Bayesian hypothesis testing [141]. For the previous reason, the Chernoff Information is useful to obtain the optimal intersection weights in classification problems whose objective is to exploit the available diversity in the measurements to yield a better probability of error. However, we are interested in estimation rather than classification problems, which is our main detractor to consider the Chernoff Information as an information theoretic criterion to obtain the optimal weights. Instead, in the sequel, we consider entropic measures to obtain these weights, such as the ones shown in [88], and relate them to the PMEE criterion (see Definition 2.5).

4.2.1.2 A minimum entropy interpretation to the optimal intersection weights in the CI principle

In this subsection, we show that the classical approaches to obtain the optimal weights shown in the seminal paper of the CI principle [99], i.e. the minimization of the trace or the determinant of $\hat{\mathbf{Q}}_\omega$, are tightly related to the Rényi entropy of the fused variable [88]. Particularly, these criteria are instances of the PMEE principle defined in Section 2.1.2.1 from Chapter 2. Let the Rényi entropy expression of the fused random variable, denoted as $\mathbf{x}_c$ (see (2.37) and (4.41)), be given by:

$$h_\alpha(\mathbf{x}_c) = \log(\det(\mathbf{Q}_\omega)) + \frac{M}{2}\log(2\pi) + \frac{M}{2}\frac{\log(\alpha)}{\alpha - 1}. \tag{4.43}$$

An intuitive criterion that obtains the intersection weights consists in plugging into (4.43) the estimator of the covariance given in (4.36). In this manner, an instance of the PMEE criterion is obtained straightforwardly to determine the intersection weights that minimize the Rényi entropy of $\mathbf{x}_c$. The resulting criterion from the previous idea yields the following optimization problem:

$$\boldsymbol{\omega}_{det} = \arg\min_{\boldsymbol{\omega}} \log(\det(\hat{\mathbf{Q}}_\omega)) \quad \text{s. t. } \boldsymbol{\omega}^T\mathbf{1}_N = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M, \tag{4.44}$$

where $\boldsymbol{\omega}$ is the vector containing the intersection weights and $\hat{\mathbf{Q}}_\omega$ is defined in (4.36a). Since minimizing the logarithm of a function is equivalent to minimizing the aforementioned function, (4.44) is our preferred determinant-based expression due to its concavity. Note that the alternative expression to (4.44) without the logarithm is commonly considered in the CI literature [99], [163]. An alternative to the determinant-based criteria is the trace minimization of the estimated fused covariance, which is also considered in the literature. The trace minimization criterion can be shown to be closely related to the one in (4.44), as shown in [88]. This connection can be seen from the following upper bound on the determinant of a matrix [23]:

$$\det(\hat{\mathbf{Q}}_\omega) \leq \left(\frac{1}{N}\operatorname{tr}(\hat{\mathbf{Q}}_\omega)\right)^N, \tag{4.45}$$

which is obtained after invoking the inequality of the arithmetic and geometric means (AM-GM inequality) on the eigenvectors of $\hat{\mathbf{Q}}_\omega$. The previous inequality becomes an equality if and only if the eigenvalues of $\hat{\mathbf{Q}}_\omega$ are all equal [23]. Although the right hand side of (4.45) can be used as the final criterion, it is also a majorant function of (4.44). Hence, the resulting optimization is a surrogate function of the determinant, in a similar manner to the majorant functions used in the MM framework. The resulting criterion yields:

$$\boldsymbol{\omega} = \arg\min_{\boldsymbol{\omega}} \operatorname{tr}(\hat{\mathbf{Q}}_\omega) \quad \text{s. t. } \boldsymbol{\omega}^T\mathbf{1}_N = 1, \tag{4.46}$$

where we have applied a monotonically increasing exponential to the right-hand side of (4.45) to obtain the trace cost, while ignoring additive and multiplicative constants. In addition to the surrogate of the Rényi entropy minimization interpretation, the criterion in (4.46) also minimizes the fusion MSE. Note that the trace minimization criterion is often preferred in front of the determinant minimization criterion for its simplicity [99], [141], [163]. Still, the minimization of (4.46) often requires iterative schemes in general fusion policies [163].

Instead of the classical trace minimization shown in (4.46), the MM framework can be exploited to obtain an iterative upper bound on the determinant that yields a better approximation than the one in (4.45). To this end, we aim to approximate (4.44) using first-order majorants (see Subsection 3.3.3.1). Invoking Lemma 3.3 and the ideas presented in Subsection 3.3.3.1 from Chapter 3, we get the following upper bound on the log-determinant in (4.44):

$$\log(\det(\hat{\mathbf{Q}}_\omega)) \leq \log(\det(\hat{\mathbf{Q}}_{\omega_i})) + \operatorname{tr}\left(\mathbf{Z}_{\omega_i}\left(\hat{\mathbf{Q}}_\omega - \hat{\mathbf{Q}}_{\omega_i}\right)\right), \tag{4.47}$$

where $\boldsymbol{\omega}_i$ is the i-th iterate of $\boldsymbol{\omega}$, $\hat{\mathbf{Q}}_{\omega_i}$ is constructed using $\boldsymbol{\omega}_i$ and:

$$\mathbf{Z}_{\omega_i} = \hat{\mathbf{Q}}_{\omega_i}^{-1}. \tag{4.48}$$

In contrast to the trace upper bound of the determinant given in (4.45), which is only tight for a particular eigenvalue profile, (4.47) is continuously approximating the log-determinant at each iteration. This behavior is preferable if one aims to minimize the Rényi entropy of the fused variable since it yields a much better approximation of the entropic measure, where the iterative scheme is the price to pay. We remark that the logarithm applied to the determinant is necessary to obtain (4.47) due to the fact that the determinant is a non-convex function by itself. After ignoring additive constants that do not depend on $\boldsymbol{\omega}$, the resulting iterative criterion yields:

$$\boldsymbol{\omega}_{i+1} = \arg\min_{\boldsymbol{\omega}} \operatorname{tr}\left(\mathbf{Z}_{\omega_i}\hat{\mathbf{Q}}_{\omega}\right) \quad \text{s. t. } \boldsymbol{\omega}^T\mathbf{1}_N = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M, \tag{4.49}$$

where, again, the constraints are imported to ensure the consistency of the resulting fusion.

4.2.2 Multisensor fusion under the perspective of Covariance Intersection

Since the particularization of the CI principle to the multisensor fusion problem depicted in (4.1) is incomplete by itself without an additional estimation scheme for the noise covariances, we review and extend the resulting criterion from the CI algorithm in this context. Given the model in (4.1), each sensor is modelled as:

$$y_m(n) = x(n) + w_m(n), \tag{4.50}$$

where $w_m(n) \sim \mathcal{N}(0, q_m)$ and $q_m > 0$ for $m = 1, ..., M$. Also, let us assume that $\hat{q}_m$ is the consistent estimator of q_m. In this toy example, we do not consider the cross-correlations between sensors because of the rationale of the CI algorithm. Applying the CI rationale to this problem yields the following set of equations:

$$\hat{q}_{\omega}^{-1} = \sum_{m=1}^{M} \omega_m \hat{q}_m^{-1}, \tag{4.51a}$$

$$\hat{x}_{CI}(n) = \hat{q}_{\omega} \sum_{m=1}^{M} \omega_m \hat{q}_m^{-1} y_m(n), \tag{4.51b}$$

where ω_m for $m = 1, ..., M$ are the intersection weights and $\hat{x}_{CI}(n)$ is the estimation of $x(n)$ that results from the fusion operation. In this case, since the estimated covariance is not a matrix, the determinant and trace minimization criteria are equivalent, consisting in the minimization of $\hat{q}_{\omega}$. Emanating from (4.51), the resulting optimization problem that determines the intersection weights is:

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \left(\sum_{m=1}^{M} \omega_m \hat{q}_m^{-1}\right)^{-1} \quad \text{s. t. } \boldsymbol{\omega}^T\mathbf{1}_M = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M, \tag{4.52}$$

where $\boldsymbol{\omega}$ is the vector whose m-th component is ω_m. Note that the previous minimization problem is also equivalent to:

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \sum_{m=1}^{M} \omega_m \hat{q}_m^{-1} \quad \text{s. t. } \boldsymbol{\omega}^T\mathbf{1}_M = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M, \tag{4.53}$$

which is a more preferable expression due to the fact that it is a linear program [29] on $\boldsymbol{\omega}$. In spite of the fact that (4.53) is a quite simple optimization problem, we avoid it because it results in the best sensor selection solution. Hence, it does not exploit the spatial diversity. The previous observation can be verified from the fact that (4.53) is maximized by selecting the sensor with the maximum value of $\hat{q}_m^{-1}$ or, equivalently, the sensor with the minimum variance. This issue results from the fact that the CI principle does not consider the estimation of the cross-covariances, which is a possible feature when the fusion operation is coupled with the regression of the measurements (see Section 4.4).

As an alternative, we propose a criterion that does not trivially result in the best sensor selection solution. This new criterion is constructed using a lower bound of the cost function in (4.53) (given

that it is a maximization problem) to derive an MM-like algorithm. In light of this objective, we scale the cost function in (4.53) by a positive constant, φ, yielding the following equivalent expression:

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \sum_{m=1}^{M} \varphi \omega_m \hat{q}_m^{-1} \quad \text{s.t. } \boldsymbol{\omega}^T \mathbf{1}_M = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M. \tag{4.54}$$

Following a similar rationale as in (4.45), notice that each term of the cost function in (4.54) is lower bounded by:

$$\varphi \frac{\omega_m}{\hat{q}_m} \geq \log\left(1 + \varphi \frac{\omega_m}{\hat{q}_m}\right), \tag{4.55}$$

which follows from the log-inequality typically used in the context of information theory:

$$\log(x) \leq x - 1. \tag{4.56}$$

Invoking (4.55) on each term of the cost function in (4.54), we get the following criterion:

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \sum_{m=1}^{M} \log\left(1 + \varphi \frac{\omega_m}{\hat{q}_m}\right) \quad \text{s.t. } \boldsymbol{\omega}^T \mathbf{1}_M = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M, \tag{4.57}$$

which has the same structure as the waterfilling optimization problem in communications theory [186]. Here, we can see the motivation behind the introduced free parameter, φ, which consists in the regulation of the sparsity degree of $\boldsymbol{\omega}$. The smaller the value of φ, the closer the waterfilling formulation is to the original one in (4.54), which results in the best sensor selection fusion policy (the sparsest solution). This also happens in communications over parallel channels at low Signal-to-Noise Ratio (SNR), where the general policy of the waterfilling power allocation strategy collapses to the simple opportunistic policy of transmitting all the information through the best channel. While, in the communications context, the SNR is a physical parameter associated to the channel quality, here an equivalent magnitude, φ, is introduced, having a novel role as a regularization parameter that promotes the anti-sparsity of the final solution. On the other hand, letting $\varphi \to \infty$ is equivalent to the infinite SNR regime in the waterfilling problem. The solution for this case can be retrieved considering the fact that, for $\varphi \to \infty$, the cost function in (4.57) yields:

$$\sum_{m=1}^{M} \log\left(1 + \varphi \frac{\omega_m}{\hat{q}_m}\right) \approx M \log(\varphi) + \sum_{m=1}^{M} \left(\log(\omega_m) - \log(\hat{q}_m)\right), \tag{4.58}$$

whose maximization, given the original constraints, results in $\hat{\boldsymbol{\omega}} = \frac{1}{M}\mathbf{1}_M$, confirming that φ is an anti-sparse regularization parameter. Therefore, the well-known agnostic uniform allocation policy that emerges in communications problems in the high SNR regime is linked under this perspective to the agnostic policy emerging under the context of information fusion.

As a summary, each of the detailed criteria ((4.53) and (4.57)) is targeted towards a specific casuistic of the multisensor problem. While the waterfilling formulation is an alternative way to promote anti-sparsity on the intersection weights, whose main goal is to exploit the available diversity in the dataset, the minimum variance criterion from (4.53) or, equivalently, (4.54) yields an optimal performance under nominal conditions (Gaussian measurements). It is important to remark that the resulting intersection weights for both formulations yield a consistent estimate of $\hat{q}_\omega$. The only issue with the CI algorithm in the considered multisensor problem is to find consistent estimators of q_m since it does not provide of a natural way to estimate these parameters. In a future section, we show a practical way to incorporate this issue into the fusion scheme.

4.3 Minimum Error Entropy fusion under Contaminated Gaussian Noise

In contrast to the approach given by the CI, where the robustness of the resulting fusion is determined by the consistency of the resulting covariance (see (4.22)), in this subsection we contextualize the MEE

principle for a general estimation of parameters in the data fusion problem using a Gaussian Mixture Model (GMM) [131] to provide robustness to the resulting fusion policy. Since the characterization of the differential entropy of a GMM is a difficult task, our proposal consists on the computation of the Rényi entropy of the *contaminated Gaussian model* [188], which is much more tractable than to compute the Rényi entropy of a general GMM. In fact, the contaminated Gaussian model is worst-case limit of the GMM in terms of the modeled contamination. In this manner, we naturally provide robustness to the resulting fusion scheme. As a by-product, it is shown that minimizing Rényi's entropy of the fitting error with entropic index $\alpha \in (1, \infty]$ is naturally linked with model order regularizers in the context of multi-sensor fusion problems. The proposed idea is summarized in our published work [121]. The previous rationale comes as an alternative to known tools that also bypass the challenge of dealing with the Rényi entropy of a GMM, such as upper-bounds [133], approximations [84] or the use of kernel methods [118].

The proposed rationale is structured as follows: in Subsection 4.3.1 we present the contaminated Gaussian model and we show preliminary results on its Rényi entropy. Then, we particularize the previous results to the multisensor fusion and study the resulting cost function in Subsection 4.3.2.

4.3.1 Rényi entropy limit of the contaminated Gaussian model

Let $X \sim \mathcal{N}(0,1)$ be a standard normal random variable whose respective PDF is denoted as $p_X(x)$ and let:

$$p_Y(y; L) = \sum_{l=1}^{L} \frac{\omega_l}{\sqrt{\gamma_l}} p_X\left(\frac{y}{\sqrt{\gamma_l}}\right), \tag{4.59}$$

be a GMM with zero-mean components, where $\sum_{l=1}^{L} \omega_l = 1$ are the mixture weights and $\gamma_l \geq 1$ is the variance of the l-th cluster ordered in an ascending way. One of the motivating points behind the consideration of GMMs is that they are universal approximation tools for smooth densities [131] which can also be interpreted as an AA fusion of several PDFs (see Definition 4.4). Taking (4.59) into account, the contaminated Gaussian model is defined as follows [188].

Definition 4.7 (Univariate contaminated Gaussian random variable)**.** *Let Z be a univariate contaminated Gaussian random variable. Then, it is denoted as $Z \sim c\mathcal{N}(0, u, \varepsilon, v, L)$ and its respective PDF is given by:*

$$p_Z(z) = \frac{1-\varepsilon}{\sqrt{u}} p_X\left(\frac{z}{\sqrt{u}}\right) + \frac{\varepsilon}{\sqrt{v}} p_Y\left(\frac{z}{\sqrt{v}}; L\right), \tag{4.60}$$

where $\varepsilon \in [0,1]$ is the degree of contamination, $0 \leq u < \infty$ is the nominal variance and $\gamma_l v$ is the variance of the l-th contaminated cluster.

Remark 4.6. Considering $L = 1$ and $v\gamma_1 \gg u$ is a common practice for modeling contaminated densities. The necessity of a general L in the linear fusion context is seen in the sequel.

Notice that the previous random variable serves as an alternative to known PDFs, e.g. elliptical distributions [180], to model heavy-tailed populations. With the purpose of deriving a robust criterion targeted to the multisensor fusion problem, we want to obtain a descriptor of the contaminated Gaussian random variable that is bounded for an unbounded contamination. To this end, we study the behavior of the Rényi entropy for the case where $v \to \infty$ in $Z \sim c\mathcal{N}(0, u, \varepsilon, v, L)$ in the following lemma.

Lemma 4.3 (Rényi entropy of the univariate contaminated Gaussian random variable)**.** *Let $Z_k \sim c\mathcal{N}(0, u, \varepsilon, v_k, L)$, with degree of contamination $\varepsilon \in [0,1)$ and nominal variance $0 < u < \infty$. Additionally, let $u < v_1 < ... < v_k < \infty$ with $k \in \mathbb{N}$ be an unbounded monotonically increasing sequence. Then, for $\alpha \in (1, \infty]$, the Rényi entropy of Z_k has the following limit:*

$$\lim_{k\to\infty} h_\alpha(Z_k) = h_\alpha(X) + \frac{1}{2}\log(u) + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{1-\varepsilon}\right). \tag{4.61}$$

Proof. The previous limit requires the solution of the following integral (see equation (2.29) from Definition 2.4):

$$\lim_{k\to\infty} h_\alpha(Z_k) = \lim_{k\to\infty} \frac{1}{1-\alpha}\log\left(\int_{-\infty}^{\infty} p_{Z_k}^{\alpha}(z)\mathrm{d}z\right), \tag{4.62}$$

which can be easily computed by means of the *Lebesgue Dominated Convergence Theorem* (LDCT) [32], [46]. The LDCT is detailed in Appendix 8.2.2 consisting on the interchange between the limit and the integral in (4.62). Firstly, we show that $p_{Z_k}(z)$ satisfies the conditions of the LDCT. The LDCT requires the sequence generated by $\{p^{\alpha}_{Z_k}(z)\}_{k\in\mathbf{N}}$ to be dominated (in the majorization sense) by some Lebesgue integrable function so that the limit can be moved inside the integral. For the purpose of finding a majorizing function for the previous sequence, notice that the PDF of a standard normal random variable is dominated by the following non-integrable function[2]:

$$p_X(x) < \frac{1}{(1+|x|)}, \tag{4.63}$$

implying that the l-th cluster from (4.59) is upper bounded by:

$$\frac{1}{\sqrt{\gamma_l}} p_X\left(\frac{y}{\sqrt{\gamma_l}}\right) < \frac{1}{\sqrt{\gamma_l}} \frac{1}{1+\frac{|y|}{\sqrt{\gamma_l}}} = \frac{1}{\sqrt{\gamma_l}+|y|} < \frac{1}{\sqrt{\gamma_0}+|y|}, \tag{4.64}$$

for $0<\gamma_0<\min_l(\gamma_l)$, where the monotonically decreasing behavior of $(\sqrt{\gamma_l}+|y|)^{-1}$ with γ_l has been used to obtain the last inequality. Imposing (4.64) to all the clusters in (4.60) yields:

$$p_{Z_k}(z) < \frac{1}{\sqrt{u}+|z|}, \tag{4.65}$$

since $u < v_k \min_l(\gamma_l)$ for all $k\in\mathbb{N}$ and $z\in\mathbb{R}$. Additionally, for $\delta\in[0,\infty)$, the function δ^{α} is monotonically increasing with δ for positive α, resulting in the fact that $p^{\alpha}_{Z_k}(z)$ is dominated by:

$$p^{\alpha}_{Z_k}(z) < \left(\frac{1}{\sqrt{v}+|z|}\right)^{\alpha} \quad \forall z\in\mathbb{R}, k\in\mathbb{N}, \tag{4.66}$$

which is a Lebesgue integrable function for $\alpha>1$. The previous rationale ensures that the limit and the integral can be interchanged in (4.62) as long as $\alpha>1$. Now, we have that:

$$\lim_{k\to\infty} p_{Z_k}(z) = \frac{1-\varepsilon}{\sqrt{u}} p_X\left(\frac{z}{\sqrt{u}}\right), \tag{4.67}$$

from which, after using the LDCT and performing the change of variables $x=\frac{z}{\sqrt{u}}$, we obtain the final expression of (4.62):

$$h_{\alpha}(Z) = \frac{1}{1-\alpha}\log\int (1-\varepsilon)^{\alpha}\left(\frac{1}{\sqrt{u}}p_X\left(\frac{z}{\sqrt{u}}\right)\right)^{\alpha}\mathrm{d}z \tag{4.68a}$$

$$= \frac{1}{1-\alpha}\log\int \left(\frac{1}{\sqrt{u}}p_X\left(\frac{z}{\sqrt{u}}\right)\right)^{\alpha}\mathrm{d}z + \frac{\log(1-\varepsilon)^{\alpha}}{1-\alpha} \tag{4.68b}$$

$$= \frac{1}{1-\alpha}\log\int \sqrt{u}\left(\frac{1}{\sqrt{u}}\right)^{\alpha} p^{\alpha}_X(x)\,\mathrm{d}x + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{1-\varepsilon}\right) \tag{4.68c}$$

$$= \frac{1}{1-\alpha}\log\int p^{\alpha}_X(x)\mathrm{d}x + \log(u^{1/2}) + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{1-\varepsilon}\right) \tag{4.68d}$$

$$= h_{\alpha}(X) + \frac{1}{2}\log(u) + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{1-\varepsilon}\right). \tag{4.68e}$$

■

[2]The justification of this upper bound is non-elegant, non-relevant and may hinder the understanding of this Proposition. A numerical experiment (and taking the limit for $x\to\pm\infty$) can confirm this inequality. Additionally, it is noted that it can be extended to many other bounded densities by using a large enough (although finite) scale factor on the right-hand side.

The previous lemma provides a closed-form and bounded information-theoretic descriptor of a distribution with unbounded variance. It is important to remark the bounded nature of the proposed entropic measure because the classical second-order moment would diverge in this worst-case scenario, compromising the utility of the contaminated GMM as a mechanism to achieve a robust estimator. We highlight this fact since the resulting boundedness yields from the property of the entropy of being sensitive to the probability of the events and not on their magnitude. Moreover, note that the additional contribution to the entropy of X due to the unbounded contamination, i.e. the last term in (4.61), is independent of the shape of the contamination density. Indeed, it depends solely on the outlier probability ε and the entropic index α. As a result, the mathematical difficulties associated with the Rényi entropy derivation of a GMM are bypassed. Besides, in a similar manner to the variance, the Shannon entropy, which can be obtained by letting $\alpha \to 1$ in (4.61), also becomes unbounded in the unbounded contamination case. In other words, the Shannon entropy does not provide any degree of conservativeness in front of gross outliers [121]. In contrast, for $\alpha \to \infty$ (often referred to as min-entropy), the last term in (4.61) becomes asymptotically unitary, showing that the min-entropy leads to the most conservative description in the presence of outliers. As a conclusion, we can see here the interest of the use of generalized Rényi entropies in signal processing as proposed in this dissertation since it offers a great ability to provide improved robustness and interpretability as compared to just picking the Shannon entropy, mimicking the communications context. This is in the line of other authors in the recent years which proposed the use of generalized entropies in applications beyond communications, such as in [150] or in [78].

Now, with Lemma 4.3 in mind, we move forward to the linear fusion context, where several measurements are combined by means of an AA fusion. For the purpose of modeling the joint contaminated distribution of several sensors, we generalize the univariate contaminated Gaussian from Definition 4.7 as follows.

Definition 4.8 (Multivariate contaminated Gaussian random variable (for data fusion))**.** *Let* $\mathbf{z}$ *be a multivariate contaminated Gaussian random variable. Then, it is denoted as* $\mathbf{z} \sim c\mathcal{N}(\mathbf{0}_M, \mathbf{Q}, \varepsilon, v\mathbf{I}_M)$*, where* $\mathbf{Q} \in \mathcal{S}_{++}^M$*, and its associated PDF is given by:*

$$p_{\mathbf{z}}(\mathbf{z}) = \frac{1-\varepsilon}{\sqrt{(2\pi)^M \det(\mathbf{Q})}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{Q}^{-1}\mathbf{z}\right) + \frac{\varepsilon}{\sqrt{(2\pi v)^M}} \exp\left(-\frac{1}{2v}\mathbf{z}^T\mathbf{z}\right). \tag{4.69}$$

Remark 4.7. We have already particularized for $L = 1$ since, for the multivariate case, we do not need to consider a general L.

We are interested in the study of the entropy after an AA fusion of a multivariate contaminated Gaussian vector, i.e. $\mathbf{f}^T\mathbf{z}$ subject to $\mathbf{f}^T\mathbf{1}_M = 1$. Note that the number of clusters of the resulting PDF grows exponentially with M due to the successive convolution of binary mixtures associated to linear combinations. We avoid the previous issue using similar arguments to the ones from Lemma 4.3 in the following theorem, where we study the entropy of the fused contaminated Gaussian random variable.

Theorem 4.4 (Rényi entropy of the fused contaminated Gaussian random variable)**.** *Let* $\mathbf{z}_k \sim c\mathcal{N}(\mathbf{0}_M, \mathbf{Q}, \varepsilon, v_k\mathbf{I}_M)$ *be a multivariate contaminated Gaussian vector and let* $z_k = \mathbf{f}^T\mathbf{z}_k$ *be the fused random variable. Additionally, let* $0 < v_k < \infty$ *be an unbounded monotonically increasing sequence. Then, for* $\alpha \in (1, \infty]$*, the Rényi entropy of* z_k *has the following limit:*

$$\lim_{k\to\infty} h_\alpha(z_k) = h_\alpha(X) + \frac{1}{2}\log(\mathbf{f}^T\mathbf{Q}\mathbf{f}) + \frac{\beta}{2}||\mathbf{f}||_0, \tag{4.70}$$

where X *is a standard normal random variable and:*

$$\beta = \frac{2\alpha}{\alpha - 1}\log\left(\frac{1}{1-\varepsilon}\right). \tag{4.71}$$

Proof. Let $n = ||\mathbf{f}||_0 \leq M$ be the number of sensors that are really involved in the fusion. Then, the PDF of z_k becomes a GMM containing 2^n clusters, which accounts for the resulting number of clusters from the n different convolutions of the contaminated Gaussian model. From those clusters, one of them is non-contaminated while the remaining $L' = 2^n - 1$ clusters are contaminated. The non-contaminated cluster has variance $u' = \mathbf{f}^T\mathbf{C}\mathbf{f}$ and its associated mixture weight is $1 - \varepsilon' = (1-\varepsilon)^n$,

which comes from the statistical component-wise independence of the contaminated clusters. The remaining L' clusters have unbounded variance, so z_k becomes a contaminated Gaussian following the general density expression introduced in Definition 4.7. Therefore, invoking Lemma 4.3 with u' and ε', we get (4.70). ■

As opposed to what happens in CS applications (see Subsection 2.1.1 from Chapter 2), where the ℓ_0 norm appears in an heuristic manner, the ℓ_0 emerges in a natural manner in the entropic measure that results from Theorem 4.4. In this way, the previous result may (potentially) serve as the foundation for a information theoretic justification for the utilization of the ℓ_0 norm as a sparse promoting regularizer in signal processing applications. As a by-product of the previous theorem, one is able to formulate an optimization problem that minimizes the entropy given in (4.70). Indeed, the main drive of this new formulation is to explore the robustness provided by the bounded information theoretic descriptor of an unbounded contamination model. This motivates the definition of the Entropic Best Linear Unbiased Estimator (E-BLUE) of the fusion rule, $\mathbf{f}$, which is surveyed in the next subsection.

4.3.2 Entropic Best Linear Unbiased Estimator

In order to model a random variable that meets the requirements of Theorem 4.4, we consider a fusion model that is a particularization of the one given in (4.1). Let a sensor network of M unreliable sensors be:

$$\mathbf{y} = x\mathbf{1}_M + \mathbf{z}, \tag{4.72}$$

where $\mathbf{z} \sim c\mathcal{N}(\mathbf{0}_M, \mathbf{Q}, \varepsilon, v\mathbf{I}_M)$ for $v \to \infty$. For simplicity, we assume that $\mathbf{Q}$ is known from now on. The contaminated Gaussian noise is our approach of describing a highly contaminated scenario. In a similar fashion to previous instances of the fusion model, we consider that the sensors are calibrated, so their spatial signature is known. From the previous model, we define the E-BLUE of x in the following manner.

Definition 4.9 (E-BLUE of the measurements)**.** *Let a set of M measurements be given by the model in* (4.72)*. Also, let the fused error random variable be defined as:*

$$e(\mathbf{f}) = \mathbf{f}^T\mathbf{y} - x. \tag{4.73}$$

Then, the E-BLUE estimator of x is:

$$\hat{x}_{\text{E-BLUE}} = \mathbf{f}_{\text{E-BLUE}}^T\mathbf{y}, \tag{4.74}$$

where:

$$\mathbf{f}_{\text{E-BLUE}} = \arg\min_{\mathbf{f}} h_\alpha(e(\mathbf{f})) \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1. \tag{4.75}$$

The previous constraint is needed to achieve an unbiased estimator of x.

Remark 4.8. Notice that, in contrast to the CI approach (see, for instance, Subsection 4.2.2), we do not import the positivity constraint on $\mathbf{f}$. The rationale behind it is that the benchmark fusion policy in (4.17) can have negative components. Also, we are not interested in keeping the consistency (see (4.22)) of the estimations since the robustness is ensured by the introduction of the contaminated Gaussian model.

A more explicit expression of (4.75) is obtained by means of Theorem 4.4, yielding the following expression:

$$h_\alpha(e(\mathbf{f})) = h_\alpha(\mathbf{f}^T\mathbf{z}) = h_\alpha(X) + \frac{1}{2}\log(\mathbf{f}^T\mathbf{Q}\mathbf{f}) + \frac{\beta}{2}||\mathbf{f}||_0 = \tag{4.76a}$$

$$\frac{1}{2}\log(2\pi) + \frac{\log(\alpha)}{2(\alpha-1)} + \frac{1}{2}\log(\mathbf{f}^T\mathbf{Q}\mathbf{f}) + \frac{\beta}{2}||\mathbf{f}||_0. \tag{4.76b}$$

After ignoring additive and positive multiplicative constants that do not depend on $\mathbf{f}$, the optimization problem that obtains the E-BLUE fusion policy is:

$$\mathbf{f}_{\text{E-BLUE}} = \arg\min_{\mathbf{f}} \log(\mathbf{f}^T\mathbf{Q}\mathbf{f}) + \beta||\mathbf{f}||_0 \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1, \tag{4.77}$$

where β is defined in (4.71). The cost function in (4.77) is non-convex since the first term is a non-convex function (see Example 3.2). Yet, there are two straightforward solutions for the previous formulation. In the uncontaminated case ($\varepsilon = 0$) the E-BLUE solution coincides with the minimum variance criterion fusion (see (4.17) from Proposition 4.2) since $\beta = 0$. On the contrary, for the maximum degree of contamination case ($\varepsilon = 1$) or the Shannon entropy cost for a non-negative degree of contamination ($\alpha \to 1^+$ and $\varepsilon > 0$), β becomes unbounded. This latter case reduces the E-BLUE criterion to the selection of the best sensor policy, i.e. choosing the sensor with maximum signal to noise ratio (SNR), which is defined as:

$$\gamma_m = \frac{1}{[\mathbf{Q}]_{m,m}}. \tag{4.78}$$

Clearly, the interplay between the entropic index of the Rényi entropy and the degree of contamination exhibits an information-theoretic operational interpretation as a precision/reliability trade-off. Furthermore, a natural model order selection criteria appears from the ℓ_0 norm regularization. Yet, (4.77) is a challenging optimization problem in general. In the following subsections we study the optimization problem depicted by (4.77).

4.3.2.1 Uncorrelated case

For illustration purposes, we show in the following paragraphs that the uncorrelated sensors case of the sensor model in (4.72), which is equivalent to consider a diagonal $\mathbf{Q}$, results in an intuitive fusion criteria that is a function of the diagonal elements of $\mathbf{Q}$. Not only that, but a strong link with model order selection ideas appears naturally in this context. Provided that $\mathbf{Q}$ is a diagonal matrix, and without any loss of generality, we assume that $\mathbf{Q}$ is such that:

$$\mathbf{Q} = \text{diag}(\mathbf{q}), \tag{4.79}$$

where $\mathbf{q} \succeq \mathbf{0}_M$ is ordered in a non-decreasing order, meaning that $q_m \leq q_{m+1}$ for $[\mathbf{q}]_m = q_m$ and $m = 1, ..., M$. Moreover, we denote the order of the fusion policy as $n = ||\mathbf{f}||_0$. With the previous considerations, the optimization problem in (4.77) can be further rewritten as:

$$\mathbf{f}_{\text{E-BLUE}}, n_{\text{E-BLUE}} = \arg\min_{\mathbf{f}_n, n} \log(\mathbf{f}_n^T \mathbf{Q}_n \mathbf{f}_n) + \beta n \quad \text{s. t. } \mathbf{f}_n^T \mathbf{1}_n = 1, \tag{4.80}$$

where $\mathbf{f}_n \in \mathbb{R}^n$ and $\mathbf{Q}_n \in \mathbb{R}^{n \times n}$. The new matrix, $\mathbf{Q}_n$, is a reduced version of $\mathbf{Q}$, which is constructed as follows:

$$\mathbf{Q}_n = \text{diag}(\mathbf{q}_n), \tag{4.81}$$

where $\mathbf{q}_n \in \mathbb{R}^n$ is the vector containing the first n components of $\mathbf{q}$ that correspond to the best n sensors. Now, with the aim of solving (4.80), we proceed as follows: firstly, we obtain the optimal value of the cost function for a fixed n with respect to $\mathbf{f}_n$. Then, we minimize with respect to n to obtain the final solution. To this end, for a fixed value of n, the resulting reduced-size E-BLUE solution for $\mathbf{f}$ yields:

$$\mathbf{f}_n = \frac{\mathbf{Q}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T \mathbf{Q}_n^{-1}\mathbf{1}_n}. \tag{4.82}$$

In order to obtain the optimal n, we plug the previous solution into (4.80) to yield:

$$n_{\text{E-BLUE}} = \arg\min_n \underbrace{-\log\left(\mathbf{1}_n^T \mathbf{Q}_n^{-1}\mathbf{1}_n\right)}_{\text{"likelihood" term}} + \underbrace{\beta n}_{\text{penalty}}, \tag{4.83}$$

which is an expression that closely resembles to the information-theoretic model order selection cost given in (2.83). Indeed, the first term in the cost function of (4.83) adopts the role of the negative log likelihood, while the second term naturally arises as a linear penalty with respect to the fusion order, n. Note that, in a similar manner to (2.83) [177], (4.83) is also a result of an information theoretic argument. In fact, it is remarked that it is a novel result in the information theoretic model order

selection framework since it is one of the first approaches that is founded on the Rényi entropy of a random variable. In order to obtain the optimal n, we rewrite (4.83) as:

$$n_{\text{E-BLUE}} = \arg\min_{n} -\log\left(F(n)\right) + \beta n, \tag{4.84}$$

where $F(n)$ is the maximal ratio combining gain, which is defined as:

$$F(n) = \mathbf{1}_n^T \mathbf{Q}_n^{-1} \mathbf{1}_n = \sum_{m=1}^{n} [\mathbf{Q}_n^{-1}]_{m,m} = \sum_{m=1}^{n} \gamma_m, \tag{4.85}$$

where γ_m for $m = 1, ..., M$ is defined in (4.78). Considering that the diagonal elements of $\mathbf{Q}_n$ are ordered in a non-decreasing order, γ_m is ordered in a non-increasing manner. Thus, $-\log\left(F(n)\right)$ is a convex function on n, so the solution of (4.84) is unique. This solution can be found by selecting the minimum n such that the discrete derivative of the cost function in (4.84) becomes positive:

$$-\log\left(F(n+1)\right) + \beta(n+1) - \left(-\log\left(F(n)\right) + \beta n\right) > 0, \tag{4.86a}$$

$$\log\left(\frac{F(n)}{F(n+1)}\right) + \beta > 0, \tag{4.86b}$$

$$\frac{\sum_{m=1}^{n} \gamma_m}{\sum_{j=1}^{n+1} \gamma_j} > e^{-\beta}, \tag{4.86c}$$

$$(1 - e^{-\beta}) \sum_{m=1}^{n} \gamma_m > e^{-\beta} \gamma_{n+1}, \tag{4.86d}$$

$$e^{\beta} - 1 > \frac{\gamma_{n+1}}{\sum_{m=1}^{n} \gamma_m}, \tag{4.86e}$$

yielding the following model order selection criterion:

$$n_{\text{E-BLUE}} = \min_{1 \le n \le M} n \quad \text{s.t.} \quad \frac{\gamma_{n+1}}{\sum_{m=1}^{n} \gamma_m} < e^{\beta} - 1. \tag{4.87}$$

The final solution is obtained after plugging n_{E-BLUE} into (4.82). The interpretation of (4.87) is that the prior information about the reliability of the available sensors has to be used as a threshold on the relative increment of precision that the $(n+1)$-th sensor provides with respect to the precision achieved with n sensors. In other words, the selected number of sensors is not dependent on any scale factor on the variances, but on the shape of its cumulative histogram profile.

Besides, regarding the resulting threshold, there are two limiting cases that are insightful of the final solution. On the one hand, the higher the reliability of the sensor network, the smaller the threshold $e^{\beta} - 1$ and, consequently, more sensors are admitted. On the other hand, for a fixed ε, the threshold is down-weighted according to (4.71) in the case of a high entropic index α. As a result, it can be concluded that the Rényi entropy emerges as a trade-off between precision and reliability in the presented multisensor fusion problem, being the entropic index, α, the parameter that explicitly controls this trade-off. This provides an operational interpretation to the entropic index when the Rényi entropy is used as a cost function to design a fusion rule in the multisensor fusion problem.

For illustration purposes, Figure 4.3 shows how the number of selected sensors is affected by the degree of contamination for different α and different scenarios of heteroskedasticity (heterogeneity of sensor qualities). Clearly, it is verified that the number of accepted sensors is generally larger for an homogeneous set of sensor qualities. Yet, this is not the case for $\alpha \to 1^+$ since the resulting fusion policy tends to be fully sparse (only one non-zero component). As a general rule, the sparsity of the resulting fusion policy increases for larger values of ϵ, which is the case where the sensors are more likely to be contaminated, and for $\alpha \to 1^+$, being the Shannon entropy case.

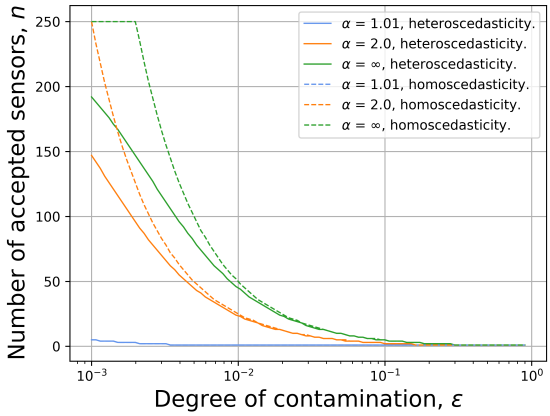


Figure 4.3: Number of selected sensors according to criterion (4.87) for different α and ε. Solid: homoscedasticity, $\gamma_m = \gamma$, $\forall m$; dashed: heteroscedasticity, $\gamma_m = (M - m + 1)\gamma$, where $\gamma > 0$ is any scale factor. The total amount of sensors is $M = 250$.

4.3.2.2 Correlated case

For a general intersensor covariance, $\mathbf{Q}$, we show that the classical ℓ_1 norm relaxation to deal with the ℓ_0 regularized problems in (4.77) is ill-posed. We consider that any alternative formulation of (4.77) is ill-posed if the limiting cases, i.e. $\beta \to 0$ and $\beta \to \infty$, do not align with the analysis shown in the previous subsection. In particular, if the alternative formulation does not yield the minimum variance solution (see Proposition 4.2) and the best sensor solution for $\beta \to 0$ and $\beta \to \infty$, respectively, then it is considered an ill-posed reformulation. In line with the previous survey, we analyze the ℓ_1 relaxation of the ℓ_0 norm in (4.77) and show that it is badly conditioned for $\beta \to \infty$. With this aim, let us consider that $\beta \to \infty$ and let us plug this limit into (4.77). Then, the resulting expression is:

$$\min_{\mathbf{f}} \varepsilon(\mathbf{f}^T\mathbf{Q}\mathbf{f}) + ||\mathbf{f}||_0 \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1, \tag{4.88}$$

where $\varepsilon(\mathbf{f}^T\mathbf{Q}\mathbf{f})$ is a negligible value that ensures that the optimal solution is an all-zeroes vector except for a single component that is equal to 1 corresponding to the sensor with the highest SNR (see (4.78)). For clarity in the exposition, we ignore the contributions of $\varepsilon(\mathbf{f}^T\mathbf{Q}\mathbf{f})$ from now on. If the ℓ_0 norm were relaxed to the ℓ_1 norm, (4.88) would have resulted in:

$$\min_{\mathbf{f}} ||\mathbf{f}||_1 \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1. \tag{4.89}$$

The previous convex program has an infinite amount of solutions, some of which are undesired in the multisensor fusion problem. In fact, solving (4.89) consists in finding the smallest ℓ_1 norm ball such that it is tangent to the constraint set. In order to find the tangent points, let us rewrite (4.89) as follows:

$$\min_{\mathbf{f}} \text{sign}^T(\mathbf{f})\mathbf{f} \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1, \tag{4.90}$$

where $\text{sign}(\mathbf{f})$ denotes the element-wise sign function. Let us consider the level set of the cost function in (4.90):

$$\text{sign}^T(\mathbf{f})\mathbf{f} = c, \tag{4.91}$$

for some positive constant c. Notice that $\text{sign}(\mathbf{f})$ is a vector containing 1 and -1 depending on the sign of the respective entry on $\mathbf{f}$. Thus, every level set of the ℓ_1 norm is composed by 2^M surfaces, which correspond to each possible value of $\text{sign}(\mathbf{f})$, i.e. each possible orthant in $\mathbb{R}^M$. It follows from (4.91) that one of the surfaces of the ℓ_1 level sets is parallel to the constraint set. The previous observation is seen by restricting the search to the positive orthant, i.e. $\mathbf{f} \succeq \mathbf{0}_M$. With the previous constraint, the ℓ_1 level set becomes:

$$\mathbf{1}_M^T\mathbf{f} = c, \tag{4.92}$$

which is a plane that is affine to the constraint set. It is clear from the previous surface that every value of $\mathbf{f}$ that satisfies:

$$\mathbf{1}_M^T \mathbf{f} = 1 \text{ and } \mathbf{f} \succeq \mathbf{0}_M, \tag{4.93}$$

is an optimal solution of (4.90). In Figure 4.4, we show graphically the intuition behind (4.93). The structure of the previous solution suggests that there is no guarantee that any optimization algorithm does not yield the naive fusion rule, i.e. $\hat{\mathbf{f}} = \frac{1}{M}\mathbf{1}_M$, in the limit, which is a contradiction with the initial analysis of this problem. Consequently, this is a sparse-aware signal processing problem where the ℓ_1 norm fails to be an adequate surrogate of an optimization problem that is regularized using the ℓ_0. We remark that it is due to the unitary sum constraint that the previous technique fails in this problem.

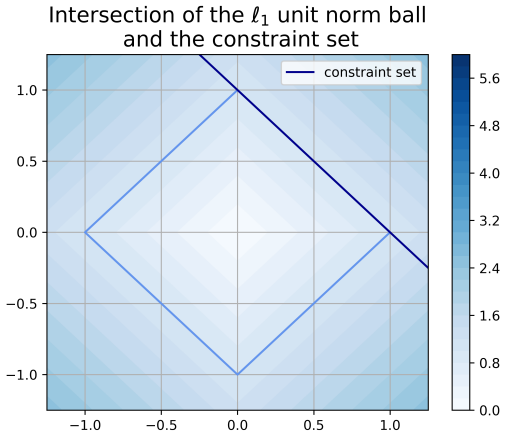


Figure 4.4: Graphical representation of (4.93) in $\mathbb{R}^2$. A segment of the constraint set coincides with one of the faces of the ℓ_1 unit norm ball.

As a conclusion, there is no way to avoid a combinatorial optimization search to solve the general correlation case in (4.77). A combinatorial optimization search algorithm is undesired due to the fact that its computational cost grows exponentially with M, being the main reason why they are avoided if possible.

4.4 Conditional Maximum Likelihood-based solution for the blind fusion and regression problem

In the previous two approaches, we considered that the intersensor covariance, $\mathbf{Q}$, is known. Although the previous assumption is insightful to study the previously detailed criteria, it hinders the practical implementation of the resulting fusion scheme. For this reason, in this section we aim to derive a fusion scheme that not only fuses in a nearly optimal manner the information gathered by the multisensor dataset, but also performs a regression on the measurements. The rationale behind formulating the fusion and regression of the sensor measurements in a joint manner will become clear later on. Just as an anticipation to motivate this fact, it is proven in this section that the inclusion of regression into the fusion operations offers a natural regularization mechanism to the fusion itself. In addition to the previous property, the proposed fusion scheme also enables the possibility of estimating a measure of integrity of the measurements, which is one of the main objectives behind the multimodal data fusion framework [110]. Following from the previous rationale, we are proposing a tight fusion scheme in contrast to the naive fusion-regression mechanisms, which consists in an initial processing of the measurements in a sensor by sensor basis (regression operation) and then fusing the results on a later stage. The idea of tight fusion schemes resonates with the classical tightly-coupled combination in Global Navigation Satellite System (GNSS) receivers, which was proven to outperform the naive loosely-coupled combination (naive fusion-regression scheme). See [47], [153] for a comparison between

loosely-coupled and tightly-coupled schemes. The contributions of this section can also be found in our published works from [122], [124] and are summarized as follows:

1. As previously stated, we propose to jointly tackle the problem of fusion and regression in a sensor network. In this regard, we show that the fusion and regression operations are coupled when there is a lack of statistical knowledge or some violation of the initial assumptions. In particular, we are interested in the case where there is no direct access to a noise variance estimator, which happens when the measured pattern is not known. The previously mentioned lack of statistical knowledge is what motivates the incorporation of the regression into the fusion operation since it is a necessary step to estimate the intersensor covariance. In fact, when the regression of sensors is performed in a first stage before the fusion operation, the error between the data and the regressed data, which contains valuable information about the data integrity, is lost in the fusion process.

   With the aim of jointly processing the fusion and regression tasks, we reformulate the fusion and regression problem in terms of *time invariant* parameters, which is grounded on the coupling of a subspace-based regression technique [49], [91] with an AA fusion of the measurements. One of the main features of this reformulation is that the curse of dimensionality of the proposed model is mitigated.
2. We prove that the maximization of the Conditional Maximum Likelihood (CML) function [134], [160], [164], [166], whose purpose is to estimate the time invariant parameters, yields a particular expression of the PMEE (see Definition 2.5) criterion. Indeed, the CML principle can be seen as an alternative tool to obtain the *parameters* of the PMEE cost.

One of the main advantages of the previous methodology, in contrast to the ideas presented in Section 4.3, is that the resulting cost function admits an efficient implementation of the MM framework. In this regard, the MM implementation constructs an iterative procedure that estimates the intersensor covariance and, thus, it *assesses* the integrity of the measurements. However, the convergence of the resulting algorithm is only ensured for a sufficient sample size, as shown in the sequel.

The previous ideas are structured as follows: Subsection 4.4.1 introduces the blind fusion and regression problem with the aforemetioned joint treatment. In subsections 4.4.2 and 4.4.3, we derive the PMEE criterion for this problem using the CML principle and we present the MM-based algorithm for its solution, respectively. Then, we show the conditions for the convergence of the aforementioned algorithm in Subsection 4.4.4. Finally, in Subsection 4.4.5 we offer solutions to the non-convergence of the MM-based algorithm for small sample sizes, while in Subsection 4.4.6 we show numerical simulations to grasp a better understanding of the whole rationale.

4.4.1 Blind joint fusion and regression problem statement

For the purpose of incorporating the regression task into the multisensor fusion problem in (4.1), we consider a temporally redundant time series. Inspired by the subspace-based regression [49], [91], we model N samples of $x(n)$ as:

$$\mathbf{x} = \mathbf{B}\mathbf{u}, \tag{4.94}$$

where $\mathbf{x} \in \mathbb{R}^N$ contains N temporal samples of $x(n)$, $\mathbf{B} \in \mathbb{R}^{N \times D}$ is the matrix of regressors, $\mathbf{u}$ is a vector of features and $D < N$ is the intrinsic dimension of the model. The main motivation behind (4.94) lies in the fact that many sequences can be well approximated by time series of finite rank [71], where D is often related to the complexity of the time series [128]. In fact, linear models as the ones in (4.94) often appear in the subspace-based regression framework [49], [91], from which the Principal Component Regression [201] is highlighted as the staple approach. While the subspace model of the measured phenomena defines a linear regression model, it is capable of approximating (even with zero modeling errors) non-linear functions. Some toy examples that are insightful to grasp the intuition behind the previous observation are found in periodic and polynomial-like functions. In this regard, a periodic function is perfectly fitted using $D = 1$ and N equal to the period of said function and, also, any polynomial-like function is approximated using a spline basis [12], [105]. Essentially, the considered subspace model in (4.94) is suitable to fit linear and non-linear functions, whose implications are shown in Subsection 4.4.6. Notice that the intrinsic dimension, D, plays a role in the degree of approximation

of the original time series.

Let us assume that there are M calibrated sensors that measure the phenomenon of interest described by (4.94). Then, stacking a batch of N temporal samples of these sensors results in the following compact model:

$$\mathbf{Y} = \mathbf{x}\mathbf{1}_M^T + \mathbf{W}, \tag{4.95}$$

where $\mathbf{Y} \in \mathbb{R}^{N\times M}$. The measurement noise matrix of the previous expression is statistically distributed as follows:

$$\mathbf{W} \sim \mathcal{MN}_{N,M}(\mathbf{0}_{N,M}, \mathbf{I}_N, \mathbf{Q}), \tag{4.96}$$

where the identity matrix accounts for the uncorrelation among the rows (time realizations) of $\mathbf{W}$. For the purpose of providing practical scalability to the processing that is developed from (4.95), we consider K different realization of the aforementioned compact model. Each realization, which is also referred to as a measurement block, is given by:

$$\mathbf{Y}_k = \mathbf{x}_k\mathbf{1}_M^T + \mathbf{W}_k = \mathbf{B}\mathbf{u}_k\mathbf{1}_M^T + \mathbf{W}_k, \tag{4.97}$$

where $\mathbf{u}_k$ is assumed unknown but deterministic (CML framework terminology), accounting for the variability between blocks, while the noise matrices, $\mathbf{W}_k$, are independent and identically distributed between blocks of samples. The partition of the data in K blocks provides a degree of flexibility to the overall data processing mechanism, allowing for a better control of the numerical stability that is added to the natural scalability as the number of available measurements increases. As an example of a particular practical implementation of the block model in (4.97), one can consider the partition of a sequence of samples into blocks of consecutive subsamples, with the possibility of using or not overlapping between blocks, in a similar manner to what it is classically done in spectral estimation methods. Yet, the consideration of overlapping blocks is out of the scope of this dissertation. Essentially, the motivation behind the consideration of the previous block model is that it is needed a certain amount of measurements blocks to be able to identify the matrix of regressors, $\mathbf{B}$, when it is unknown. The latter idea is justified in Subsection 4.4.3.2.

It is important to highlight the fact that the fusion and regression tasks are based on block independent parameters, i.e. $\mathbf{B}$ and $\mathbf{Q}$. Henceforth, we refer to any parameter that is independent of the block as a *time invariant* parameter. This kind of parameters are useful to reformulate the estimation of $\mathbf{x}_k$ with as few parameters as possible. While $\mathbf{B}$ is a clear first choice for a time invariant parameter that is related to the regression of the measurements, we consider a linear fusion policy as in (4.14) to be the respective fusion parameter. We choose a linear fusion policy instead of the intersensor covariance, $\mathbf{Q}$, due to the fact that it is more suited for the MM-based algorithm that is developed in the following subsections, i.e. it yields simpler expressions. In consequence, the resulting fused model is:

$$\mathbf{s}_k(\mathbf{B}, \mathbf{f}) = \mathbf{Y}_k\mathbf{f} = \mathbf{B}\mathbf{u}_k\mathbf{1}_M^T\mathbf{f} + \mathbf{W}_k\mathbf{f} \quad \text{s. t. } \mathbf{1}_M^T\mathbf{f} = 1, \tag{4.98}$$

where, again, the constraint $\mathbf{1}_M^T\mathbf{f} = 1$ is needed to achieve an unbiased fusion, i.e. $\mathrm{E}\left[\mathbf{s}_k(\mathbf{B}, \mathbf{f})\right] = \mathbf{B}\mathbf{u}_k = \mathbf{x}_k$. Although parameterization with respect to $\mathbf{B}$ would, in principle, suffice as to get a parameter dimensionality independent from K, it is important to remark that the incorporation of the redundant fusion vector, $\mathbf{f}$, is one of the key distinctive aspects of the proposed approach. The optimal values of $\mathbf{B}$ and $\mathbf{f}$ are linked when the intersensor covariance is unknown, as expanded later on.

With the previous rationale in mind, the resulting blind sensor fusion and regression problem consists in the determination of the best fusion policy, $\mathbf{f}$, and of the matrix of regressors, $\mathbf{B}$, using the available K realizations of the block model in (4.97). We assume that the intersensor covariance matrix, $\mathbf{Q}$, is unknown, while the intrinsic dimension, D, is considered as a prior information. The aforementioned prior would mean that the complexity of the measured phenomena time series is known [71], [128]. Although the intrinsic dimension estimation lies in the model order selection methodology (see Section 2.3), we consider a fixed value of D for concreteness and to highlight the novelties of the proposed methodology.

4.4.2 Derivation of the PMEE criterion from the CML principle

In this subsection, we show the connections between the CML principle and the PMEE criterion (see Definition 2.5) for the multisensor fusion problem stated in the previous subsection. This connection is

obtained after noting that the CML function of the joint likelihood of the sensor noise matrix yields an equivalent expression to the one from Definition 2.5. The CML function is defined as follows [134], [160], [164], [166].

Definition 4.10 (Conditional Maximum Likelihood function)**.** *Let $f_{\mathbf{Y}}(\mathbf{Y}|\boldsymbol{\theta})$ be the likelihood function of some random variable $\mathbf{Y}$ parameterized by $\boldsymbol{\theta}$. Then, its CML function is given by $f_{\mathbf{Y}}(\mathbf{Y}|\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$. The procedure of obtaining estimator of $\boldsymbol{\theta}$ and plugging it into the likelihood is referred to as* compressing *the likelihood.*

Remark 4.9. We also refer to the log-likelihood function reparameterized by the ML estimator as the CML function.

We are interested in compressing the likelihood function of the measurements in such a way that the resulting CML function is parameterized completely by $\mathbf{B}$ and $\mathbf{f}$. For this purpose, consider the log-likelihood function associated to $\mathbf{Y}_k$, after ignoring additive constants that do not depend on $\mathbf{u}_k$, $\mathbf{Q}$ or $\mathbf{B}$:

$$\ell_{\mathbf{Y}_k}(\mathbf{Y}_k|\mathbf{B},\mathbf{Q},\mathbf{u}_k) = -\frac{N}{2}\left(\log(\det(\mathbf{Q})) + \frac{1}{N}\operatorname{tr}\left(\mathbf{Q}^{-1}(\mathbf{Y}_k - \mathbf{B}\mathbf{u}_k\mathbf{1}_M^T)^T(\mathbf{Y}_k - \mathbf{B}\mathbf{u}_k\mathbf{1}_M^T)\right)\right), \tag{4.99}$$

where the normalization with respect to $\frac{1}{N}$ on the second term is done for convenience. In order to introduce the linear fusion rule, we firstly compress the likelihood by substituting the ML estimator of $\mathbf{u}_k$ derived from (4.98), denoted as $\hat{\mathbf{u}}_k$ (see Appendix 8.2.3):

$$\ell_{\mathbf{Y}_k}(\mathbf{Y}_k|\mathbf{B},\mathbf{Q},\mathbf{f}) = \ell_{\mathbf{Y}_k}(\mathbf{Y}_k|\mathbf{B},\mathbf{Q},\hat{\mathbf{u}}_k) = \tag{4.100a}$$

$$-\frac{N}{2}\left(\log(\det(\mathbf{Q})) + \frac{1}{N}\operatorname{tr}\left(\mathbf{Q}^{-1}(\mathbf{Y}_k - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y}_k\mathbf{f}\mathbf{1}_M^T)^T(\mathbf{Y}_k - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y}_k\mathbf{f}\mathbf{1}_M^T)\right)\right) = \tag{4.100b}$$

$$-\frac{N}{2}\left(\log(\det(\mathbf{Q})) + \frac{1}{N}\operatorname{tr}\left(\mathbf{Q}^{-1}(\mathbf{Y}_k - \mathbf{P}_B\mathbf{Y}_k\mathbf{f}\mathbf{1}_M^T)^T(\mathbf{Y}_k - \mathbf{P}_B\mathbf{Y}_k\mathbf{f}\mathbf{1}_M^T)\right)\right), \tag{4.100c}$$

where $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ is the projection matrix of the subspace spanned by the columns of $\mathbf{B}$. The remaining parameter to fully compress the likelihood is $\mathbf{Q}$. In light of this, we need the joint log-likelihood function of all blocks to obtain an estimation of $\mathbf{Q}$ that includes as much of the available information as possible. Given the statistical independence between the noise matrices of different blocks, the joint log-likelihood is obtained as the summation of the marginal log-likelihood functions of all the blocks:

$$\ell_{\mathbf{Y}}(\mathbf{Y}|\mathbf{B},\mathbf{Q},\mathbf{f}) = \sum_{k=1}^{K}\ell_{\mathbf{Y}_k}(\mathbf{Y}_k|\mathbf{B},\mathbf{Q},\mathbf{f}) = -\frac{KN}{2}\log(\det(\mathbf{Q})) - \frac{N}{2}\sum_{k=1}^{K}\operatorname{tr}\left(\mathbf{Q}^{-1}\hat{\mathbf{C}}_k(\mathbf{B},\mathbf{f})\right), \tag{4.101}$$

where:

$$\hat{\mathbf{C}}_k(\mathbf{B},\mathbf{f}) = \frac{1}{N}(\mathbf{Y}_k - \mathbf{P}_B\mathbf{Y}_k\mathbf{f}\mathbf{1}_M^T)^T(\mathbf{Y}_k - \mathbf{P}_B\mathbf{Y}_k\mathbf{f}\mathbf{1}_M^T), \tag{4.102}$$

is the sample covariance of the k-th block parameterized by $\mathbf{B}$ and $\mathbf{f}$. The ML estimation of $\mathbf{Q}$ is obtained from the maximization of (4.101), which is obtained as follows:

$$\frac{\partial\ell(\mathbf{Y}|\mathbf{B},\mathbf{Q},\mathbf{f})}{\partial\mathbf{Q}} = \mathbf{0}_{M,M}, \tag{4.103a}$$

$$-\frac{KN}{2}\mathbf{Q}^{-1} + \frac{N}{2}\sum_{k=1}^{K}\mathbf{Q}^{-1}\hat{\mathbf{C}}_k(\mathbf{B},\mathbf{f})\mathbf{Q}^{-1} = \mathbf{0}_{M,M}, \tag{4.103b}$$

$$\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f}) = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{C}}_k(\mathbf{B},\mathbf{f}). \tag{4.103c}$$

Plugging (4.103c) into (4.101) yields the fully compressed CML function:

$$\ell_{\mathbf{Y}}(\mathbf{Y}|\mathbf{B},\mathbf{f}) = \sum_{k=1}^{K}\ell_{\mathbf{Y}_k}(\mathbf{Y}_k|\mathbf{B},\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f}),\mathbf{f}) = -\frac{KN}{2}\log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f}))) - \frac{MKN}{2}, \tag{4.104}$$

whose last term can be ignored since it does not depend on $\mathbf{B}$ or $\mathbf{f}$. Finally, the maximization of (4.104) is what defines the final criterion. Changing the maximization by a minimization and ignoring additive constants, the CML criterion is:

$$\hat{\mathbf{B}}, \hat{\mathbf{f}} = \arg\min_{\mathbf{f},\mathbf{B}} \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{B}, \mathbf{f}))) \quad \text{s.t. } \mathbf{f}^T \mathbf{1}_M = 1, \tag{4.105}$$

where the constraint is imported from (4.98) to obtain an unbiased fusion. Notice that the previous criterion is a particularized expression of Definition 2.5 for the statistical model given in (4.96). In fact, the previous rationale shows a possible method to construct the PMEE cost function (for Gaussian measurements only) using the parameters that are natural to the blind fusion and regression problem. As shown earlier, these natural parameters are the matrix of regressors, $\mathbf{B}$, and the AA fusion combiner, $\mathbf{f}$. Optimization problems similar to the previous one have been encountered in other Signal Processing applications. As an example, we refer to [172, Section 3], where the log-determinant of a parameterized covariance matrix has been used to provide robustness to the estimation of the Direction of Arrival (DoA) in the presence of independent interferences.

The cost function in (4.105) can be further parsed by noting that it satisfies the homogeneity condition of the Grassmann manifold (see (2.46)). Thus, this cost function can be equivalently optimized with respect to a Grassmann constrained variable. This additional constraint is proven necessary to derive a globally convergent algorithm. The previous ideas are formalized in the following lemma.

Lemma 4.5 (Log-determinant of a parameterized sample covariance homogeneity condition)**.** *Let $f : \mathbb{R}^M \times \mathbb{R}^{N\times D} \to \mathbb{R}$ be defined as:*

$$f(\mathbf{B}, \mathbf{f}) = \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{B}, \mathbf{f}))). \tag{4.106}$$

Then, f satisfies the homogeneity condition given in (2.46), which is:

$$f(\mathbf{B}, \mathbf{f}) = f(\mathbf{B}\mathbf{U}, \mathbf{f}), \tag{4.107}$$

where $\mathbf{U} \in \mathrm{GL}(D)$. As a result, $f(\mathbf{B}, \mathbf{f})$ can be reparameterized by a Grassmann constrained variable in place of the unconstrained variable, $\mathbf{B}$.

Proof. The homogeneity condition can be verified from the sample covariance estimation. Given the expression in (4.103c), we only need to proof that the homogeneity condition is fulfilled for every $\hat{\mathbf{C}}_k(\mathbf{B}, \mathbf{f})$. Provided that $\hat{\mathbf{C}}_k(\mathbf{B}, \mathbf{f})$ is parameterized by the projection matrix $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ and that:

$$\mathbf{P}_{BU} = \mathbf{B}\mathbf{U}(\mathbf{U}^T\mathbf{B}^T\mathbf{B}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{B}^T = \mathbf{B}\mathbf{U}\mathbf{U}^{-1}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{U}^{-T}\mathbf{U}^T\mathbf{B}^T = \tag{4.108a}$$

$$\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T = \mathbf{P}_B, \tag{4.108b}$$

for every $\mathbf{U} \in \mathrm{GL}(D)$, it is invariant to this linear transformation. As a result, $f(\mathbf{B}, \mathbf{f})$ satisfies (4.107). ■

The previous lemma ensures that one could always choose any representative $\mathbf{H} = \mathbf{B}\mathbf{U}$ such that $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ and that $\mathbf{H}\mathbf{H}^T = \mathbf{P}_H = \mathbf{P}_B$, and still obtain the same value of the cost function. Since the compactness of the sequence generated by an iterative scheme is a necessary condition for its global convergence (see Section 3.3), we reparameterize (4.105) by adding a constraint on the Grassmann manifold on the matrix of regressors. This reparameterization certifies that every iterate of the newly introduced variable belongs to a compact set. Consequently, the optimization problem that solves the described problem in this section is:

$$\hat{\mathbf{H}}, \hat{\mathbf{f}} = \arg\min_{\mathbf{H},\mathbf{f}} \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}))) \quad \text{s.t. } \mathbf{f}^T \mathbf{1}_M = 1,\ \mathbf{H} \in \mathrm{Gr}(N, D), \tag{4.109}$$

where $\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f})$ is now a function of the projection matrix $\mathbf{P}_H = \mathbf{H}\mathbf{H}^T$. Notice that the previous optimization problem links the optimal solution of $\mathbf{H}$ and $\mathbf{f}$ because these variables are not separable in (4.109) and, hereby, the previous criterion couples the fusion and regression tasks.

4.4.3 MM-based algorithm for the blind fusion and regression problem

The final criterion in (4.109) consists of the composition of the log-determinant function (which is concave) with an estimation of the intersensor covariance matrix. While the minimization of a concave function is already a challenging optimization problem [80], the composition with $\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})$ hinders the assessment of the concavity (or g-concavity) of the cost function in (4.109). Unfortunately, we have to renounce to globally optimal points for these reasons. Even so, although the optimization problem in (4.109) is non-convex in general, its structure is suited for the block MM algorithm built on the Grassmann manifold (see Subsection 3.3.4.2). In the following paragraphs, we show the fundamental steps to construct a block MM algorithm to retrieve the stationary points of (4.109).

The first (and most important) step of the block MM algorithm is the construction of the majorant function. We follow a similar rationale to construct the majorant as in the Concave-Convex Procedure (CCP) [204], which is, essentially, founded on the same idea as in the first-order majorants from Subsection 3.3.3.1. Indeed, the majorant function of the log-determinant can be derived from the first-order characterization of concave functions (see Theorem 3.1 for concave functions). The cost function in (4.109) can be upper bounded at the i-th iteration using the Taylor expansion of the log-determinant:

$$\log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f}))) \leq \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}_i,\mathbf{f}_i))) + \operatorname{tr}\left(\mathbf{Z}_i\left(\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f}) - \hat{\mathbf{Q}}_{ML}(\mathbf{H}_i,\mathbf{f}_i)\right)\right), \tag{4.110}$$

where $\mathbf{H}_i$ and $\mathbf{f}_i$ are i-th iterates and:

$$\mathbf{Z}_i = \hat{\mathbf{Q}}_{ML}^{-1}(\mathbf{H}_i,\mathbf{f}_i), \tag{4.111}$$

is the estimation of the intersensor covariance matrix using the i-th iterates. Note that (4.110) holds because $\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})$ is positive semidefinite. The majorant function is obtained after ignoring additive constants that do not depend on $\mathbf{H}$ or $\mathbf{f}$ from the right hand side of (4.110) and, as a result, the majorant has the following form:

$$g(\mathbf{H},\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i) = \operatorname{tr}\left(\mathbf{Z}_i\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})\right), \tag{4.112}$$

from which a more explicit expression is derived in Appendix 8.2.4, yielding:

$$g(\mathbf{H},\mathbf{f}|\mathbf{Z}_i) = \sum_{k=1}^{K}\operatorname{tr}\left(\mathbf{H}^T\mathbf{Y}_k\left(\mathbf{1}_M^T\mathbf{Z}_i\mathbf{1}_M\mathbf{f}\mathbf{f}^T - 2\mathbf{Z}_i\mathbf{1}_M\mathbf{f}^T\right)\mathbf{Y}_k^T\mathbf{H}\right), \tag{4.113}$$

where the dependence with $\mathbf{H}_i$ and $\mathbf{f}_i$ is incorporated in $\mathbf{Z}_i$ for convenience. We refer to [65, Eq. (6)] for an alternative example of this kind of majorants of the log-determinant function. Still, the joint optimization of (4.113) with respect to $\mathbf{H}$ and $\mathbf{f}$ is still challenging due to the coupling between both variables. This is the reason why a block MM algorithm is chosen instead, where $\mathbf{H}$ and $\mathbf{f}$ are the respective blocks of variables. In this way, the respective majorant of each block is obtained from (4.113) as:

$$g_f(\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i,\mathbf{Z}_i) = g(\mathbf{H}_i,\mathbf{f}|\mathbf{Z}_i), \tag{4.114a}$$

$$g_H(\mathbf{H}|\mathbf{H}_i,\mathbf{f}_i,\mathbf{Z}_i) = g(\mathbf{H},\mathbf{f}_i|\mathbf{Z}_i), \tag{4.114b}$$

from where the block MM update equations are:

$$\mathbf{f}_{i+1} = \arg\min_{\mathbf{f}} g_f(\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i,\mathbf{Z}_i) \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1, \tag{4.115a}$$

$$\mathbf{H}_{i+1} = \arg\min_{\mathbf{H}} g_H(\mathbf{H}|\mathbf{H}_i,\mathbf{f}_{i+1},\mathbf{Z}_i) \quad \text{s.t. } \mathbf{H}\in\operatorname{Gr}(N,D), \tag{4.115b}$$

$$\mathbf{Z}_{i+1} = \hat{\mathbf{Q}}_{ML}^{-1}(\mathbf{H}_{i+1},\mathbf{f}_{i+1}). \tag{4.115c}$$

The respective constraints in (4.115a) and (4.115b) are imported from (4.109).

The final steps to fully describe our proposed algorithm are the initialization and the stopping criterion. Regarding the initialization, a primal feasible point for both blocks of variables is necessary

Algorithm 2 MM-based blind data fusion and blind regression algorithm

Initialization: ϵ, I_T, $\mathbf{H}_0$ and $\mathbf{f}_0 = \frac{1}{M}\mathbf{1}_M$
for $i = 1$ to I_T **do**
  $\hat{\mathbf{R}}_i = \frac{1}{K}\sum_{k=1}^{K} \mathbf{Y}_k \mathbf{f}_{i-1}\mathbf{f}_{i-1}^T \mathbf{Y}_k^T$
  Compute the eigendecomposition of $\hat{\mathbf{R}}_i = \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T$
  Set $\mathbf{H}_i$ as the first D columns of $\mathbf{U}_i$
  $\mathbf{P}_{H_i} = \mathbf{H}_i \mathbf{H}_i^T$
  $\mathbf{C}_{k,i} = \frac{1}{N}(\mathbf{Y}_k - \mathbf{P}_{H_i}\mathbf{Y}_k \mathbf{f}\mathbf{1}^T)^T(\mathbf{Y}_k - \mathbf{P}_{H_i}\mathbf{Y}_k \mathbf{f}\mathbf{1}^T)$
  $\mathbf{Z}_i = \left(\frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{C}}_{k,i}\right)^{-1}$
  $\mathbf{f}_i = \frac{\mathbf{Z}_i \mathbf{1}_M}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M}$
  if $(||\mathbf{f}_i - \mathbf{f}_{i-1}||_2 + d_{arc}(\mathbf{H}_i, \mathbf{H}_{i-1}) < \epsilon)$ **then**
    Set $i^* = i$
    Break
  end if
end for
return $\mathbf{H}_{i^*}$ and $\mathbf{f}_{i^*}$

to ensure the convergence. A non-informative selection of $\mathbf{H}_0$ is any realization of a uniform random distribution on $\mathrm{Gr}(N, D)$ since there is no additional information to select a better one. A simple experiment that draws a sample from a uniform distribution on $\mathrm{Gr}(N, D)$ is depicted as follows [157, Section 9.1.1]: generate $\mathbf{X} \in \mathbb{R}^{N \times D}$ with independent and identically distributed $\mathcal{N}(0, 1)$ random variables. Decompose $\mathbf{X}$ as $\mathbf{H}_0\mathbf{R}$ using the QR factorization. Then, $\mathbf{H}_0$ is uniformly distributed on the Grassmann manifold. Other alternatives can also be found in [157, Section 9.1.1]. Concerning the fusion, it is agnostically initialized with a naive fusion rule, i.e. $\mathbf{f}_0 = \frac{1}{M}\mathbf{1}_M$. As for the stopping rule, a measure of how close the solution is to a stationary point is needed, being simultaneously sensitive to both the fusion and regression. This measure is given by:

$$||\mathbf{f}_{i+1} - \mathbf{f}_i||_2 + d_{arc}(\mathbf{H}_{i+1}, \mathbf{H}_i) < \epsilon, \tag{4.116}$$

where ϵ is the tolerance of the stopping criterion. Note that the previous stopping criterion coincides with the diminishing difference condition of convergence of the MM algorithms [92], [123].

The following subsections focus on the detailed solution of the update equations in (4.115). In each discussion, we show the conditions in which the solutions of (4.115a) and (4.115b) at each iteration are unique. The uniqueness of the solution is relevant for the global convergence of the algorithm, as pointed out in Subections 3.3.2 and 3.3.4. In Algorithm 2, we show a summary of the resulting block MM algorithm that generates the sequence $\{\mathbf{H}_i, \mathbf{f}_i\}_{i \in \mathbb{N}}$ from (4.115).

4.4.3.1 Update equation of the fusion rule, f

It is shown in Appendix 8.2.5 that (4.115a) is a Linearly Constrained Quadratic Program (LCQP) on $\mathbf{f}$. Its expression is given by:

$$\mathbf{f}_{i+1} = \arg\min_{\mathbf{f}} \mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M \mathbf{f}^T \mathbf{D}_i \mathbf{f} - 2\mathbf{f}^T \mathbf{D}_i \mathbf{Z}_i \mathbf{1}_M \quad \text{s.t. } \mathbf{f}^T \mathbf{1}_M = 1, \tag{4.117}$$

where $\mathbf{D}_i = \sum_{k=1}^{K} \mathbf{Y}_k^T \mathbf{H}_i \mathbf{H}_i^T \mathbf{Y}_k$. Note that the previous LCQP is convex since $\mathbf{D}_i$ is always a positive semidefinite matrix. The solution of (4.117) is retrieved by finding the stationary point of its Lagrangian, i.e. solving the following equation:

$$\nabla \mathcal{L}_f = \mathbf{D}_i \mathbf{f} - \mathbf{D}_i \frac{\mathbf{Z}_i \mathbf{1}_M}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M} + \frac{\lambda_f}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M} \mathbf{1}_M = \mathbf{0}, \tag{4.118}$$

where $\nabla \mathcal{L}_f$ is the gradient of the Lagrangian of (4.117) and λ_f is its equality constraint Lagrange multiplier. A solution of (4.118) is:

$$\mathbf{f}_{i+1} = \frac{\mathbf{Z}_i \mathbf{1}_M}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M}, \tag{4.119}$$

with $\lambda_f = 0$, as it is already a primal feasible solution because:

$$\mathbf{1}_M^T \mathbf{f}_{i+1} = \frac{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M} = 1. \tag{4.120}$$

It is relevant to remark that (4.119) is not the unique minimizer of (4.117) in general. In fact, (4.118) has a unique solution for a full rank $\mathbf{D}_i$. This is satisfied when there is a sufficient amount of temporal blocks, i.e. $K \geq M$. Yet, the previous solution in (4.119) is preferred in all cases since it transforms $g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_{i+1}, \mathbf{Z}_i)$, the majorant of $\mathbf{H}$, into:

$$g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_{i+1}, \mathbf{Z}_i) = \sum_{k=1}^{K} \operatorname{tr}\left(\mathbf{H}^T \mathbf{Y}_k \left(\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M \mathbf{f}_{i+1} \mathbf{f}_{i+1}^T - 2\mathbf{Z}_i \mathbf{1}_M \mathbf{f}_{i+1}^T\right) \mathbf{Y}_k^T \mathbf{H}\right) = \tag{4.121a}$$

$$\sum_{k=1}^{K} \operatorname{tr}\left(\mathbf{H}^T \mathbf{Y}_k \left(\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M \frac{\mathbf{Z}_i \mathbf{1}_M}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M} \frac{\mathbf{1}_M^T \mathbf{Z}_i}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M} - 2\mathbf{Z}_i \mathbf{1}_M \frac{\mathbf{1}_M^T \mathbf{Z}_i}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M}\right) \mathbf{Y}_k^T \mathbf{H}\right) = \tag{4.121b}$$

$$\sum_{k=1}^{K} \operatorname{tr}\left(\mathbf{H}^T \mathbf{Y}_k \left(-\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M \frac{\mathbf{Z}_i \mathbf{1}_M}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M} \frac{\mathbf{1}_M^T \mathbf{Z}_i}{\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M}\right) \mathbf{Y}_k^T \mathbf{H}\right) = \tag{4.121c}$$

$$-\mathbf{1}_M^T \mathbf{Z}_i \mathbf{1}_M \sum_{k=1}^{K} \operatorname{tr}\left(\mathbf{H}^T \mathbf{Y}_k \mathbf{f}_{i+1} \mathbf{f}_{i+1}^T \mathbf{Y}_k^T \mathbf{H}\right), \tag{4.121d}$$

which is an insightful expression to obtain the solution of (4.115b).

4.4.3.2 Update equation of the regressors subspace, H

Thanks to (4.121), it can be verified that the solution of (4.115b) is retrieved from the classical PCA problem [27], [97]. Indeed, the resulting optimization problem from (4.121) yields:

$$\mathbf{H}_{i+1} = \arg\max_{\mathbf{H}} \operatorname{tr}\left(\mathbf{H}^T \hat{\mathbf{R}}_i \mathbf{H}\right) \quad \text{s. t. } \mathbf{H} \in \operatorname{Gr}(N, D), \tag{4.122}$$

where:

$$\hat{\mathbf{R}}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{Y}_k \mathbf{f}_i \mathbf{f}_i^T \mathbf{Y}_k^T, \tag{4.123}$$

is the sample correlation matrix of the fused variable (see (4.98)). Note that we exchanged the minimum for a maximum in (4.122) to account for the negative multiplicative constant from (4.121). It is worth remarking that the update equation in (4.122) is searching for the signal subspace of the fused variable, whose information is gathered in $\hat{\mathbf{R}}_i$. This subspace estimation is expected to be improved in each iteration since $\mathbf{f}_i$ yields a better fusion for increasing i. While (4.122) can be considered as an heuristic approach to estimate the regressors subspace, it appears naturally from the block MM rationale.

Invoking Theorems 3.7 and 3.8, we get that the optimal value of $\mathbf{H}$ in (4.122) is given by the subspace spanned by the D singular vectors corresponding to the largest D singular values of $\hat{\mathbf{R}}_i$. In addition, given that $\hat{\mathbf{R}}_i$ is positive semidefinite, the previous solution is unique as long as there exists at least D singular values that are different from 0. This condition is ensured for $K \geq N$.

4.4.4 Convergence analysis

In this subsection, we verify that the sequence generated by (4.115), denoted as $\{\mathbf{H}_i, \mathbf{f}_i\}_{i \in \mathbb{N}}$, is a globally convergent one (see Definition 3.23). With this aim, we particularize the assumptions stated in Subsections 3.3.2 and 3.3.4.2 from Theorem 3.11 to these update equations. While (B1), (B2) and (B3) are immediately fulfilled by the construction of the global majorant (see (4.110)), the conditions in which (B6) is fulfilled are already explored in Subsections 4.4.3.1 and 4.4.3.2. The remaining assumptions are analyzed as follows.

(B4) The continuity of the majorants (on all its arguments) is verified from the fact that $g_f(\mathbf{f}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$ and $g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_{i+1}, \mathbf{Z}_i)$ are constructed using compositions of continuous functions of $\mathbf{H}$, $\mathbf{f}$, $\mathbf{H}_i$, $\mathbf{f}_i$ and $\mathbf{Z}_i$.

(B5) It is certified that $g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$ and $g_f(\mathbf{f}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$ (see (4.115)) are g-quasiconcave and quasiconvex functions, respectively. To this end, we have already shown that $g_f(\mathbf{f}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$ is convex in Subsection 4.4.3.1, and thus it is also quasiconvex.

Besides, we know from Theorem 3.8 that $g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$ is locally g-concave in the following Grassmann ball:

$$B_{\frac{\pi}{4}}(\mathbf{U}_{s,i}) = \{\mathbf{X} \in \mathrm{Gr}(N, D) : \mathbf{\Theta}_{\mathbf{X},i} \preceq \phi \mathbf{I}_D\}, \tag{4.124}$$

where $\mathbf{U}_{s,i}$ is the subspace spanned by the columns of the singular vectors corresponding to the largest D singular values, and we denote $\mathbf{\Theta}_{\mathbf{X},i}$ the principal angles between $\mathbf{X}$ and $\mathbf{U}_{s,i}$. While this local behavior is expected since this majorant is a smooth function over a Riemannian manifold (so it cannot be globally g-convex) [28], [44], it is not restrictive for a sufficiently large sample size. Taking into account equation (4.123) and that each iteration of $\mathbf{f}_i$ decreases the log-determinant of the intersensor covariance, $\mathbf{H}_{i+1}$ is obtained from a (relatively) small perturbation of the previous iterate as long as there exists a unique solution of $g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$, as observed numerically in the sequel. In simpler terms, provided that $\hat{\mathbf{R}}_i$ contains the complete information of the signal subspace, which occurs for $K \geq D$, $g_H(\mathbf{H}|\mathbf{H}_i, \mathbf{f}_i, \mathbf{Z}_i)$ behaves locally as a quasiconcave function for all i. Note that the distance upper bound in $B_{\frac{\pi}{4}}(\mathbf{U}_{s,i})$ is quite large from the Grassmann manifold perspective (it is a half-way between orthogonal subspaces).

Now that we assessed the conditions in which the majorants are well-behaved for the block MM algorithm, the remaining condition that verifies the global convergence is the compactness of $\{\mathbf{H}_i, \mathbf{f}_i\}_{i\in\mathbb{N}}$. Although a common approach in the literature to certify the compactness of $\{\mathbf{H}_i, \mathbf{f}_i\}_{i\in\mathbb{N}}$ is to prove that the following sublevel set [92], [162]:

$$\mathcal{S}_f = \left\{\mathbf{f} \in F, \mathbf{H} \in \mathrm{Gr}(N, D) : \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}))) \leq \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}_0, \mathbf{f}_0)))\right\}, \tag{4.125}$$

is compact, where:

$$F = \left\{\mathbf{f} \in \mathbb{R}^M : \mathbf{1}^T\mathbf{f} = 1\right\}, \tag{4.126}$$

we show in Appendix 8.2.6 that $\mathcal{S}_f$ is unbounded and, therefore, it is not compact. Instead, we show that the solutions of the update equations (4.115a) and (4.115b) lie in a compact set, meaning that $\{\mathbf{f}_i\}_{i\in\mathbb{N}}$ and $\{\mathbf{H}_i\}_{i\in\mathbb{N}}$ are compact sequences.

Firstly, we assess the compactness of $\{\mathbf{f}_i\}_{i\in\mathbb{N}}$. Notice that in (4.119) the norm of the numerator is upper bounded by:

$$||\mathbf{Z}_i\mathbf{1}||^2 \leq \lambda_{1,i}^2 M, \tag{4.127}$$

and that the denominator is lower bounded by:

$$0 \leq M\lambda_{M,i} \leq \mathbf{1}^T\mathbf{Z}_i\mathbf{1}, \tag{4.128}$$

where $\lambda_{1,i}$ and $\lambda_{M,i}$ are the maximum and minimum singular values of $\mathbf{Z}_i$, respectively. Mixing the previous two inequalities, the norm of the iterates, $\{\mathbf{f}_i\}_{i\in\mathbb{N}}$, can be upper bounded by:

$$||\mathbf{f}_i||^2 < \frac{\lambda_{1,i}^2 M}{M\lambda_{M,i}} = \frac{\lambda_{1,i}^2}{\lambda_{M,i}}. \tag{4.129}$$

As a result, the iterates $\{\mathbf{f}_i\}_{i\in\mathbb{N}}$ lie in a compact set because of the previous norm upper bound and the closedness of F as long as $\mathbf{Z}_i$ is well-conditioned for all i. This means that the condition number of every $\mathbf{Z}_i$, defined as:

$$\vartheta(\mathbf{Z}_i) = \frac{\lambda_{1,i}}{\lambda_{M,i}}, \tag{4.130}$$

where $\lambda_{1,i}$ and $\lambda_{M,i}$ are the largest and smallest eigenvalues of $\mathbf{Z}_i$, respectively, has a finite value. In other words, $\mathbf{Z}_i$ is a full rank matrix for all i. This latter property is ensured for for $K \geq M$. On the other hand, the Grassmann manifold is known to be a compact subset of the ND-dimensional sphere

[20] and thus the iterates $\{\mathbf{H}_i\}_{i\in\mathbb{N}}$ also lie in a compact set. Notice that without the compactness of the Grassmann manifold, the iterates of the matrix of regressors would not belong to a compact set. This is a clear advantage of the constrained variable, $\mathbf{H}$, in front of the unconstrained one, $\mathbf{B}$.

As a summary, as long as $K \geq \max(M, D)$, the algorithm depicted by (4.115) is convergent to a stationary point of the original problem. The previous lower bound on the sample size ensures that both majorants have unique minimizers and that all the involved matrices are well-behaved. Note that the convergence point is a stationary point (and not the global optimum) of (4.109) [123]. Regarding the convergence rate, it is known that the MM framework suffers from a sublinear convergence speed [87, Section 5]. Yet, this convergence speed is reasonable enough to choose the MM framework in favour of alternative global non-convex optimization algorithms which are known to suffer from a slower convergence rate [144].

4.4.5 Dealing with small sample sizes

In the convergence analysis of Algorithm 2, we have seen that the MM-based algorithm is sensitive to the sample size. However, it may not be possible in every practical scenario to gather a sufficiently large batch of samples. To tackle this issue, we show two different alternatives that alleviate the need of a sufficient sample size. These two approaches are built upon statistical and structural assumptions on the intersensor covariance, yielding robust estimators of $\mathbf{Q}$ that are well-behaved for small sample sizes. In fact, they can also be interpreted as two alternative methods to the one depicted in Subsection 4.4.2 to obtain parameterized estimators of $\mathbf{Q}$, which are then plugged into the PMEE criterion. In the following subsections, we survey these alternatives and show that the resulting algorithms maintain the same structure as the one from Algorithm 2.

4.4.5.1 Diagonal intersensor covariance assumption

Given that the estimation of the intersensor covariance requires the identification of $\frac{M(M+1)}{2}$ free parameters, it is intuitive to think that a reduction of the estimated degrees of freedom must improve the convergence speed. Additionally, it must also require a smaller sample size to achieve its optimal performance. One way of reducing the free parameters on the estimation of the intersensor covariance is to enforce a diagonal structure on $\mathbf{Q}$, which only requires the determination of M free parameters. In the following paragraphs, we derive the estimations of $\mathbf{Q}$ and the majorant of each block of variables with the assumption that this matrix is diagonal. Considering the diagonal assumption, the ML estimation of $\mathbf{Q}$ is obtained using the following partial derivative (see (4.101) and (4.102)):

$$\frac{\partial \ell(\mathbf{Y}|\mathbf{B}, \mathbf{q}, \mathbf{f})}{\partial q_m} = -\frac{KN}{2} q_m^{-1} + \frac{N}{2} \sum_{k=1}^{K} q_m^{-2} [\hat{\mathbf{C}}_k(\mathbf{B}, \mathbf{f})]_{m,m} = 0, \tag{4.131}$$

where $[\mathbf{Q}]_{m,m} = q_m$. Then, the expression of the ML estimation of the diagonal intersensor matrix is:

$$\hat{\mathbf{Q}}_D(\mathbf{B}, \mathbf{f}) = \hat{\mathbf{Q}}_{ML}(\mathbf{B}, \mathbf{f}) \odot \mathbf{I} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathbf{C}}_k(\mathbf{B}, \mathbf{f}) \odot \mathbf{I}. \tag{4.132}$$

From this point, the final criterion under the diagonal assumption can be derived from two different perspectives. The first one, which emanates from the same rationale as in Subsection 4.4.2, consists in plugging (4.132) into the joint likelihood given in (4.101). In other words, we compress the likelihood using the previous estimator of the covariance. In the alternative perspective, we plug (4.132) into the particularized expression of the PMEE criterion using the statistical model of $\mathbf{W}_k$ (see Definition 2.5 and (4.96)). For both perspectives, the resulting optimization problem under the assumption of a diagonal intersensor covariance, after performing the previously shown change of variables ($\mathbf{B} \in \mathbb{R}^{N\times D} \to \mathbf{H} \in \mathrm{Gr}(N, D)$), yields:

$$\hat{\mathbf{H}}, \hat{\mathbf{f}} = \arg\min_{\mathbf{H},\mathbf{f}} \log(\det(\hat{\mathbf{Q}}_D(\mathbf{H}, \mathbf{f}))) \quad \text{s.t. } \mathbf{1}^T\mathbf{f} = \mathbf{1},\ \mathbf{H}^T\mathbf{H} = \mathbf{I}. \tag{4.133}$$

The previous cost function can be rewritten as follows:

$$\log(\det(\hat{\mathbf{Q}}_D(\mathbf{H},\mathbf{f}))) = \sum_{m=1}^{M} \log\left(\sum_{k=1}^{K}\sum_{n=1}^{N} |[\mathbf{Y}_k]_{n,m} - [\mathbf{P}_H]_{n,:}\mathbf{Y}_k\mathbf{f}|^2\right), \tag{4.134}$$

where $[\mathbf{Y}_k]_{n,m}$ is the (n,m)-th entry of $\mathbf{Y}_k$ and $[\mathbf{P}_H]_{n,:}$ denotes the n-th row of $\mathbf{P}_H$. In simpler words, (4.134) is computing implicitly the Geometric Mean Squared Error (GMSE) [130] of the fusion residuals.

The next step is to show that the previous cost admits a majorant that has a similar expression to the one in (4.112), so that the resulting algorithm is equivalent to Algorithm 2. Following the same rationale as in Subsection 4.4.3, the first-order majorant of the cost function in (4.133) is:

$$g_D(\mathbf{H},\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i) = \operatorname{tr}\left(\mathbf{Z}_{D,i}\hat{\mathbf{Q}}_D(\mathbf{H},\mathbf{f})\right), \tag{4.135}$$

where:

$$\mathbf{Z}_{D,i} = \hat{\mathbf{Q}}_D^{-1}(\mathbf{H}_i,\mathbf{f}_i). \tag{4.136}$$

Notice that (4.135) is, fundamentally, the same expression in which Algorithm 2 is based, up to the diagonal correction of $\mathbf{Z}_{D,i}$. In fact, (4.135) also admits the same convergence analysis as shown in Subsection 4.4.4, except from the fact that $\mathbf{Z}_{D,i}$ is well-behaved for all K. The previous property ensures that the only condition that needs to be fulfilled to verify the convergence of the resulting algorithm is $K \geq D$ since it implies that the matrix of regressors can be uniquely identified. Thus, the diagonal assumption on the intersensor matrix suppresses one of the two constraints on the sample size.

4.4.5.2 Bayesian prior on the sample covariance

In contrast to the diagonal assumption, which is a structural prior on the intersensor covariance matrix, we also explore the use of statistical priors on $\mathbf{Q}$ within the Bayesian framework. We consider the Inverse Wishart (IW) prior distribution on $\mathbf{Q}$ as a way to regularize its estimation. The IW distribution is denoted as $\mathcal{W}^{-1}(\mathbf{\Psi},\nu)$, where $\mathbf{\Psi} \in \mathbb{R}^{M\times M}$ is the scale matrix and $\nu \geq M$ are the degrees of freedom of this distribution. The PDF of the IW distribution satisfies [196]:

$$f_{\mathbf{Q},IW}(\mathbf{Q}) \propto \det(\mathbf{Q})^{-\frac{v+M+1}{2}} \exp\left(\frac{1}{2}\operatorname{tr}(\mathbf{\Psi}\mathbf{Q}^{-1})\right), \tag{4.137}$$

while its logarithm, after ignoring additive constants that do not depend on $\mathbf{Q}$, yields:

$$\ell_{\mathbf{Q},IW}(\mathbf{Q}) = \log(f_{\mathbf{Q},IW}(\mathbf{Q})) = -\frac{v+M+1}{2}\log(\det(\mathbf{Q})) - \frac{1}{2}\operatorname{tr}(\mathbf{\Psi}\mathbf{Q}^{-1}), \tag{4.138}$$

where, from now on, we consider $\mathbf{\Psi} = \beta\mathbf{I}_M$. Note that if $\mathbf{Q}$ follows the IW distribution, then $\mathbf{Q}^{-1}$ is a Wishart random matrix [111], [196]. The motivation behind the IW prior is twofold. In one aspect, the IW, with the considered shape parameter, models the fact that the variances cannot be too high nor too small. This means that we intend to indirectly impose better numerical conditions on the ML intersensor covariance estimation from Subsection 4.4.2. Besides, the IW distribution is known to be the conjugate prior [67] of the matrix normal distribution. The conjugate prior of a distribution is a PDF such that the resulting posterior distribution (with respect to some parameter) belongs to the same kind of probability distributions as the likelihood function, e.g. Gaussian posterior and Gaussian likelihood. This implies that the conjugate prior has a similar structure to the likelihood function. In the case of a matrix Gaussian likelihood, this is materialized in the fact that the logarithm of the IW prior (see (4.138)) has a clear dependence with the logarithm of the determinant of $\mathbf{Q}$ and to the scaled trace of the precision matrix, which is an expression that can be easily *mixed* with the matrix Gaussian likelihood. As a result and informally speaking, the conjugate prior is an algebraic convenience in the Bayesian framework.

For the purpose of obtaining the MAP estimation of $\mathbf{Q}$, let us consider the log-posterior function using the IW and the likelihood function from (4.101):

$$\ell_{\mathbf{Q}|\mathbf{Y}}(\mathbf{B},\mathbf{Q},\mathbf{f}) = \ell_{\mathbf{Y}}(\mathbf{Y}|\mathbf{B},\mathbf{Q},\mathbf{f}) + \ell_{\mathbf{Q},IW}(\mathbf{Q}) = \tag{4.139a}$$

$$-\frac{KN}{2}\log(\det(\mathbf{Q}))-\frac{KN}{2}\operatorname{tr}\left(\mathbf{Q}^{-1}\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f})\right)-\frac{v+M+1}{2}\log(\det(\mathbf{Q}))-\frac{\beta}{2}\operatorname{tr}(\mathbf{Q}^{-1}), \quad (4.139b)$$

where $\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f})$ is defined in (4.103c). In Appendix 8.2.7, we show that the previous expression yields:

$$\ell_{\mathbf{Q}|\mathbf{Y}}(\mathbf{B},\mathbf{Q},\mathbf{f})=-\frac{KN}{2}\left(\left(1+\frac{v+M+1}{KN}\right)\log(\det(\mathbf{Q}))+\operatorname{tr}\left(\mathbf{Q}^{-1}\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f})\right)\right), \quad (4.140)$$

where:

$$\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f})=\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f})+\frac{\beta}{KN}\mathbf{I}_M. \quad (4.141)$$

Following a similar procedure to the one in (4.103), the MAP estimation of $\mathbf{Q}$ is:

$$\hat{\mathbf{Q}}_{IW}(\mathbf{B},\mathbf{f})=\left(1+\frac{v+M+1}{KN}\right)^{-1}\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f}), \quad (4.142)$$

from where the *compressed MAP* function is obtained after plugging $\hat{\mathbf{Q}}_{IW}(\mathbf{B},\mathbf{f})$ into $\ell_{\mathbf{Q}|\mathbf{Y}}(\mathbf{B},\mathbf{Q},\mathbf{f})$. The resulting criterion for the multisensor fusion problem is obtained by maximizing the compressed MAP function, which is equivalent to the following optimization problem:

$$\hat{\mathbf{B}},\hat{\mathbf{f}}=\arg\min_{\mathbf{B},\mathbf{f}}\log\left(\det\left(\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f})+\frac{\beta}{KN}\mathbf{I}_M\right)\right)\quad \text{s. t. } \mathbf{f}^T\mathbf{1}_M=1, \quad (4.143)$$

where we ignored additive and multiplicative constants that do not depend on $\mathbf{B}$ or $\mathbf{f}$, and scaled (4.140) by -1 to yield a minimization problem. The multiplicative constant from $\hat{\mathbf{Q}}_{IW}(\mathbf{B},\mathbf{f})$ has been ignored since $\log\det(\alpha\mathbf{M})=\log(\det(\mathbf{M}))+M\log(\alpha)$ for $\mathbf{M}\in\mathbb{R}^{M\times M}$. Again, we prefer to constrain the matrix of regressors in the Grassmann manifold for the previously mentioned arguments (see Subsection 4.4.4). After adding the previous constraint, we get the final criterion:

$$\hat{\mathbf{H}},\hat{\mathbf{f}}=\arg\min_{\mathbf{H},\mathbf{f}}\log\left(\det\left(\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})+\frac{\beta}{KN}\mathbf{I}_M\right)\right)\quad \text{s. t. } \mathbf{f}^T\mathbf{1}_M=1,\mathbf{H}\in\operatorname{Gr}(N,D). \quad (4.144)$$

Clearly, the majorant function of the previous cost function is:

$$g_{IW}(\mathbf{H},\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i)=\operatorname{tr}\left(\mathbf{Z}_{IW,i}\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})\right), \quad (4.145)$$

where:

$$\mathbf{Z}_{IW,i}=\left(\hat{\mathbf{Q}}_{ML}(\mathbf{H}_i,\mathbf{f}_i)+\frac{\beta}{KN}\mathbf{I}_M\right)^{-1}, \quad (4.146)$$

which also admits the same update equations as those from Algorithm 2.

In the previous rationale, we have achieved two things. On the one hand, we have shown that a Tikhonov-like regularization (diagonal loading) is achieved for the Bayesian methods that estimate the covariance matrix using the IW prior. This is specially useful in the proposed algorithm since it is an alternative approach to mitigate the impact of a small sample size on the covariance matrix estimators. In contrast to the diagonal prior solution, the larger the sample size, the smaller the contribution of the regularization term in (4.144). This latter property ensures the asymptotic optimality of the resulting solution for an infinite sample size while keeping the robustness in the small sample size scenario.

It is important to highlight the fact that the ideas shown in this subsection can be mixed with those shown in Subsection 4.4.5.1, yielding a more robust criterion that can also be integrated within Algorithm 2. In this regard, notice that, after combining the ideas that yield $\hat{\mathbf{Q}}_D(\mathbf{B},\mathbf{f})$ and $\hat{\mathbf{Q}}_{IW}(\mathbf{B},\mathbf{f})$ (see (4.132) and (4.142)), the MAP estimation of a diagonally constrained covariance with an IW prior is:

$$\hat{\mathbf{Q}}_R(\mathbf{B},\mathbf{f})=\left(1+\frac{v+M+1}{KN}\right)^{-1}\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f})\odot\mathbf{I}_M, \quad (4.147)$$

where $\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f})$ is defined in (4.141). The estimations of $\mathbf{H}$ and $\mathbf{f}$ that emanate from $\hat{\mathbf{Q}}_R(\mathbf{B},\mathbf{f})$ are obtained from the following optimization problem:

$$\hat{\mathbf{H}},\hat{\mathbf{f}} = \arg\min_{\mathbf{H},\mathbf{f}} \log\left(\det\left(\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})\odot\mathbf{I}_M + \frac{\beta}{KN}\mathbf{I}_M\right)\right) \quad \text{s. t. } \mathbf{f}^T\mathbf{1}_M = 1, \mathbf{H}\in \mathrm{Gr}(N,D). \tag{4.148}$$

The respective majorant function is:

$$g_R(\mathbf{H},\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i) = \mathrm{tr}\left(\mathbf{Z}_{R,i}\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})\right), \tag{4.149}$$

where, in this case:

$$\mathbf{Z}_{R,i} = \left(\hat{\mathbf{Q}}_{ML}(\mathbf{H}_i,\mathbf{f}_i)\odot\mathbf{I}_M + \frac{\beta}{KN}\mathbf{I}_M\right)^{-1}. \tag{4.150}$$

The necessity of the joint consideration of both covariance estimation regularization approaches will become clear in Subsection 4.4.6.3, where it is shown that the IW prior by itself does not provide any advantage in terms of performance, whereas the combination of the IW with the diagonal assumption is the best performing approach in the small sample size regime.

4.4.6 Performance analysis

The purpose of this subsection is to analyze the performance of Algorithm 2 and the two approaches described in Subsection 4.4.5 by means of numerical simulations. Prior to showing the simulation results, we firstly study the expression of the MSE of the fusion utilizing arbitrary estimators of $\mathbf{f}$ and $\mathbf{H}$. Since the performance analysis of the ideas presented in this Section is founded on the MSE of the fusion, we formally define it as follows.

Definition 4.11 (Mean Squared Error (MSE) of the fusion)**.** *Let* $\mathbf{H}$ *and* $\mathbf{f}$ *be any two particular values of the subspace of regressors and the fusion policy, respectively, that meet the constraints from* (4.109)*. Then, the MSE of the fusion is defined as:*

$$\gamma(\mathbf{H},\mathbf{f}) = \mathrm{E}\left[\frac{1}{N}||\mathbf{x}_k - \hat{\mathbf{x}}_k||_2^2\right] = \mathrm{E}\left[\frac{1}{N}||\mathbf{x}_k - \mathbf{P}_H\mathbf{Y}_k\mathbf{f}||_2^2\right], \tag{4.151}$$

where $\mathbf{x}_k$ *and* $\mathbf{Y}_k$ *are defined in* (4.97) *from Subsection 4.4.1, and* $\mathbf{P}_H = \mathbf{H}\mathbf{H}^T$*.*

Remark 4.10. The previous expected value can be computed using T MonteCarlo realizations as follows:

$$\tilde{\gamma}(\mathbf{H},\mathbf{f}) = \frac{1}{TKN}\sum_{t=1}^{T}\sum_{k=1}^{K}||\mathbf{x}_{k,t} - \mathbf{P}_H\mathbf{Y}_{k,t}\mathbf{f}||_2^2, \tag{4.152}$$

where $\mathbf{x}_{k,t}$ and $\mathbf{Y}_{k,t}$ denote the k-th block of the t-th MonteCarlo simulation of the measurements. In the considered simulations, we approximate $\gamma(\mathbf{H},\mathbf{f})$ using (4.152).

It can be verified that (4.151) has an insightful closed-form solution under some mild assumptions on the problem depicted in Subsection 4.4.1. Regarding the measured phenomenon, we consider that each block of measurements is constructed using the model from (4.97). The values of $\mathbf{u}_k$ and $\mathbf{B}$ are obtained from realizations of a random vector and a random orthogonal matrix, respectively. They are such that:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}_D, \gamma_u\mathbf{I}_D), \tag{4.153a}$$

$$\mathbf{B}^T\mathbf{B} = \mathbf{I}_D, \tag{4.153b}$$

where $\mathbf{B}$ is obtained from a uniform distribution on $\mathrm{Gr}(N,D)$ (see the initialization of the algorithm depicted in Section 4.4.3 or [157, Section 9.1.1]). Note that the previous assumptions on $\mathbf{u}_k$ and $\mathbf{B}$ are not restrictive since the resulting model is capable of approximating any signal whose average power is equal to $\frac{D}{N}\gamma_u$. The latter statement is verified from the fact that the linear model in (4.94) can also approximate any non-linear function and that:

$$\frac{1}{N}\mathrm{E}[||\mathbf{x}_k||_2^2] = \frac{1}{N}\mathrm{E}[\mathbf{u}_k^T\mathbf{B}^T\mathbf{B}\mathbf{u}_k] = \gamma_u\frac{D}{N} = \gamma_u\nu, \tag{4.154}$$

where:

$$\nu = \frac{D}{N}, \tag{4.155}$$

is the temporal redundancy coefficient. A result of the previous assumptions is that the expression of the MSE yields (see Appendix 8.2.8 for its derivation):

$$\gamma(\mathbf{H}, \mathbf{f}) = \frac{\gamma_u}{N}\left(D - \operatorname{tr}(\cos^2(\boldsymbol{\Theta}))\right) + \nu \mathbf{f}^T \mathbf{Q} \mathbf{f}, \tag{4.156}$$

where $\boldsymbol{\Theta}$ denotes the matrix containing the principal angles between $\mathbf{B}$ and $\mathbf{H}$. The first term in (4.156) can be further parsed in terms of the Frobenius distance between the projection matrices of $\mathbf{B}$ and $\mathbf{H}$. Provided that $\mathbf{P}_B = \mathbf{B}\mathbf{B}^T$, $\mathbf{P}_H = \mathbf{H}\mathbf{H}^T$, and:

$$D - \operatorname{tr}(\cos^2(\boldsymbol{\Theta})) = \operatorname{tr}(\mathbf{I}_D - \cos^2(\boldsymbol{\Theta})) = \frac{1}{2}\operatorname{tr}(\mathbf{I}_D + \mathbf{I}_D - \cos^2(\boldsymbol{\Theta}) - \cos^2(\boldsymbol{\Theta})) = \tag{4.157a}$$

$$\frac{1}{2}\operatorname{tr}(\mathbf{B}\mathbf{B}^T\mathbf{B}\mathbf{B}^T + \mathbf{H}\mathbf{H}^T\mathbf{H}\mathbf{H}^T - \mathbf{B}\mathbf{B}^T\mathbf{H}\mathbf{H}^T - \mathbf{H}\mathbf{H}^T\mathbf{B}\mathbf{B}^T) = \tag{4.157b}$$

$$\frac{1}{2}\operatorname{tr}(\mathbf{P}_B\mathbf{P}_B^T + \mathbf{P}_H\mathbf{P}_H^T - \mathbf{P}_B\mathbf{P}_H^T - \mathbf{P}_H\mathbf{P}_B^T) = \frac{1}{2}||\mathbf{P}_B - \mathbf{P}_H||_F^2 = d_{proj}^2(\mathbf{B}, \mathbf{H}), \tag{4.157c}$$

which is the projection F-norm of the Grassmann manifold [57, Subsection 4.3], the MSE given in (4.156) is alternatively rewritten as follows:

$$\gamma(\mathbf{H}, \mathbf{f}) = \frac{\gamma_u}{N} d_{proj}^2(\mathbf{B}, \mathbf{H}) + \nu \mathbf{f}^T \mathbf{Q} \mathbf{f}. \tag{4.158}$$

From the previous closed-form expression of the MSE, it is much more clear that each of the two terms in (4.158) can be related to errors in the fusion and in the regression tasks in a decoupled manner. Clearly, the first term in (4.158) indicates how well the measured phenomenon is approximated. By contrast, the second term in (4.158) is governed entirely by the intersensor covariance and, as a by-product, it represents the errors that are caused by the sensors uncertainties. In fact, these terms account for the bias of the regression model and the variance of the resulting fusion, respectively. In order to obtain a lower bound of (4.151), we substitute $\mathbf{B}$ and $\mathbf{f}_B$ (see (4.94) and (4.17)) into (4.158), yielding the CRLB of the joint fusion and regression problem:

$$\gamma(\mathbf{B}, \mathbf{f}_B) = \frac{\nu}{\mathbf{1}_M^T \mathbf{Q}^{-1} \mathbf{1}_M}, \tag{4.159}$$

which is proportional to the one given in (4.18). Note that the previous expression holds since $d_{proj}(\mathbf{B}, \mathbf{B}) = \frac{1}{\sqrt{2}}||\mathbf{P}_B - \mathbf{P}_B||_F = 0$. Looking at the performance bound in (4.159), it is clear that the time redundancy coefficient in the numerator plays an important role on the fusion performance. Actually, the latter effect is totally decoupled from the fusion gain given by the denominator. Moreover, provided that ν is fixed for a given realization, the fusion MSE cannot tend to 0 for an infinite sample size. This is due to the fact that the amount of free parameters grows with K, which is the main drawback of the regression model in (4.97). The latter issue is the reason why a fair comparison with other known fusion schemes is not possible, e.g. the Kalman filter.

Considering that the fusion setting depicted in Subsection 4.4.1 depends on too many variables, the comparison of different fusion scenarios by means of the measure given in Definition 4.11 is challenging. For the purpose of facilitating the aforementioned assessment, we resort to an alternative measure built on the MSE from Definition 4.11, which is aimed at canceling out the contributions of the multiplicative constants of the CRLB in (4.159). This new measure is defined as follows.

Definition 4.12 (Normalized Fusion Quality Measure (NFQM))**.** *Let the MSE be defined as in Definition 4.11 and let* $\mathbf{H}$ *and* $\mathbf{f}$ *be any two estimators of the fusion and regression time invariant parameters. Provided that* $\mathbf{B}$ *and* $\mathbf{f}_B$ *are the optimal values of the previous two estimators, the Normalized Fusion Quality Measure (NFQM) is defined as:*

$$i_q(\mathbf{H}, \mathbf{f}) = \frac{\gamma(\mathbf{B}, \mathbf{f}_B)}{\gamma(\mathbf{H}, \mathbf{f})}. \tag{4.160}$$

Remark 4.11. $i_q(\mathbf{H}, \mathbf{f})$ is bounded in $[0, 1]$, where 1 is its optimal value. This property can be verified from the fact that the denominator is lower bounded by the numerator.

In the following subsections, we use the NFQM to analyze the numerical performance of the approaches based on Algorithm 2 for the problem depicted in Subsection 4.4.1. For this purpose, we describe the common simulation parameters in the upcoming paragraphs.

We simulate the sensor network depicted in Subsection 4.4.1, i.e. the block model given by $\mathbf{Y}_k = \mathbf{B}\mathbf{u}_k\mathbf{1}^T + \mathbf{W}_k$, using the same assumptions as those that yield (4.158). This means that we set $\mathbf{B}$ as a random sample from a uniform distribution on $\mathrm{Gr}(N, D)$. As for the vector of features, we set $\mathbf{u}_k \sim \mathcal{N}(0, \gamma_u\mathbf{I})$ for a fixed γ_u. Therefore, for each MonteCarlo simulation, we draw samples from the previous distributions to assign values to $\mathbf{B}$ and $\mathbf{u}_k$ for $k = 1, ..., K$. Regarding the sensors noise, we consider two types of sensors in terms of their quality: a subset of sensors is modeled using a low noise power while the remaining sensors are highly contaminated. The previous behavior is described by a vector of variances, $\mathbf{q} \in \mathbb{R}^M$:

$$\mathbf{q} = [\underbrace{q_1, ..., q_1}_{M_g}, \underbrace{q_2, ..., q_2}_{M-M_g}]^T \succ \mathbf{0}_M, \tag{4.161}$$

where there are M_g components with variance $q_1 = 0.1$ and $M - M_g$ sensors with variance $q_2 = 100$. Notice that $\frac{q_2}{q_1}$ is an indicator of the difficulty of the fusion task and, in this case, we set a scenario with great variance of sensor qualities (heteroskedastic sensors) to emphasize the loss in performance of suboptimal fusion policies. For the set of simulations that consider a correlated sensor network, the intersensor covariance matrix is set to be a pentadiagonal matrix. The physical meaning of the pentadiagonal intersensor covariance matrix is that it models the correlation between spatially close sensors. Accordingly, the intersensor cross-covariances are set as follows:

$$[\mathbf{Q}]_{m,m+1} = [\mathbf{Q}]_{m+1,m} = \frac{2}{3}\rho\sqrt{[\mathbf{q}]_m[\mathbf{q}]_{m+1}}, \tag{4.162a}$$

$$[\mathbf{Q}]_{m,m+2} = [\mathbf{Q}]_{m+2,m} = [\mathbf{Q}]_{m+1,m-1} = [\mathbf{Q}]_{m-1,m+1} = \frac{1}{3}\rho\sqrt{[\mathbf{q}]_m[\mathbf{q}]_{m+2}}, \tag{4.162b}$$

where $\rho \in [0, 1]$. Notice that we have set the correlation coefficients to be all positive to depict the most difficult correlation scenario. Indeed, negative correlation benefit naive fusion policies, e.g. the arithmetic mean of the measurements, because of the fact that naively combining negatively correlated sensors may cancel out their respective noise contributions. Additionally, we compute the expected value of the MSE of the fusion using 2000 MonteCarlo simulations and, using the resulting value, we compute the NFQM. For clarity in the exposition, the remaining parameters are set within a particular simulation context specified above each figure.

As for the estimators of the matrix of regressors and the fusion policy, the behavior of $i_q(\mathbf{H}, \mathbf{f})$ is studied for several approaches. The considered approaches are the summarized as follows:

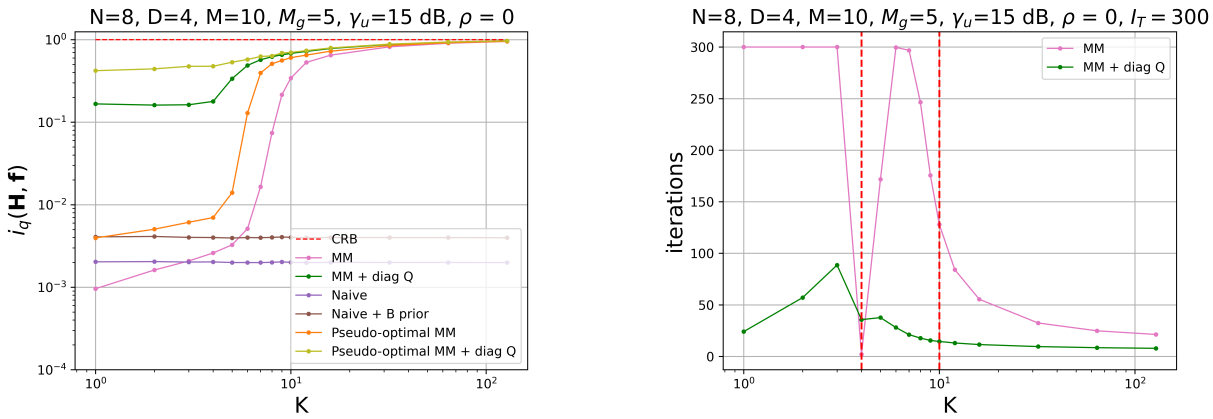
1. The CRLB of the time invariant parameters, implying total knowledge of $\mathbf{Q}$ (or equivalently, $\mathbf{f}_b$) and $\mathbf{B}$. (Labeled as "CRLB")
2. Estimating $\mathbf{f}$ and $\mathbf{H}$ via Algorithm 2. (Labeled as "MM")
3. Estimating $\mathbf{f}$ and $\mathbf{H}$ utilizing the rationale from Subection 4.4.5.1, i.e. assuming that $\mathbf{Q}$ is diagonal. (Labeled as "MM + diag Q")
4. Using the naive fusion rule ($\mathbf{f}_n = \frac{1}{M}\mathbf{1}$) with the assumption that $\mathbf{B}$ is known. (Labeled as "Naive + B prior")
5. A similar estimator to 4) with the additional assumption that $\mathbf{B}$ is unknown, meaning that $\mathbf{P}_B$ is replaced by an identity matrix. (Labeled as "Naive")
6. Estimator of the linear policy and the subspace of regressors that follow the rationale given in Subsection 4.4.5.2. (Labeled as "IW")
7. The approach that results after mixing the ideas provided in subsections 4.4.5.1 and 4.4.5.2, resulting from the majorant in (4.149). (Labeled as "IW + diag Q")
8. Averaging the best 200 realizations (the 10th percentile of (4.160)) of the estimator in 2) in terms of the minimum value of the cost function. (Labeled as "Pseudo-optimal MM")

9. Similarly to the previous estimator, the estimator that averages the 10th percentile of the resulting NFQM of the estimator in 3). (Labeled as "Pseudo-optimal MM + diag Q")

The purpose of the naive estimators is to provide a measure of the difficulty of the fusion problem since these two estimators are the intuitive solution to the sensor fusion problem. In contrast, provided that there is no known alternative (up to the authors knowledge) of finding a global optimal solution to the optimization problem in (4.109), the pseudo-optimal estimators serve as an heuristic way to verify how far the average performance of the MM-based solutions are from the globally optimal estimators.

4.4.6.1 Asymptotic numerical analysis

In this first set of simulations, we are interested in assessing the asymptotic behavior of the previously mentioned estimators to test whether or not they can achieve the CRLB. For this purpose, we want to simulate a challenging scenario for both the fusion and the regression tasks. This means that D is set to be approximately equal to $\frac{N}{2}$, as stated in [175]. The intuition behind the previous value of D is that estimating a subspace of dimension D is equivalent to the estimation of its orthogonal complement of dimension $N - D$. Thus, $D \approx \frac{N}{2}$ is the case where the subspace estimation requires the largest amount of degrees of freedom. In figures 4.5 and 4.6, it is shown the behavior of the NFQM for each of the aforementioned fusion schemes with respect to the number of blocks K in the uncorrelated and correlated sensor network, respectively. The dashed red lines in subfigures 4.5b and 4.6b are placed in $K = D$ and $K = M$ for illustration purposes. Indeed, these two values of K are fundamental in the convergence of the MM algorithm (see Subsection 4.4.4).



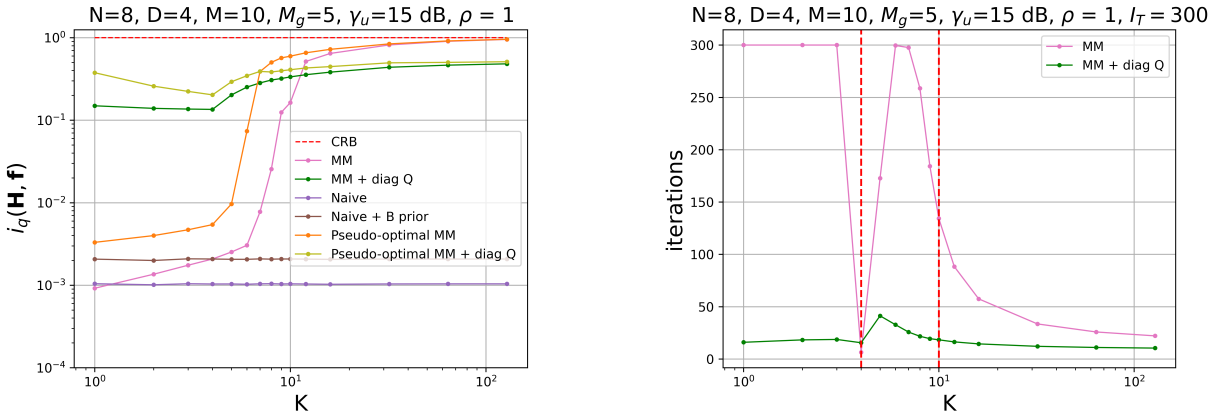
(a) Empiric NFQM

(b) Iterations.

Figure 4.5: Asymptotic behavior of the MM-based estimators in an uncorrelated sensor network.

The main difference between figures 4.5 and 4.6 is found in the asymptotic behavior of the MM-based algorithms with and without the diagonal intersensor covariance assumption. As a matter of fact, it is seen in subfigures 4.5a and 4.6a that, while all the implementations of the MM algorithm tend to the CRLB for the uncorrelated sensor network, the diagonally constrained alternative offers an overall better performance in the small sample size regime. This improved performance is expected because this alternative requires the determination of less parameters. The price to pay is that the diagonally constrained MM-based solution cannot reach the CRLB in the correlated sensor network case (see Subfigure 4.6a) given that its intrinsic assumption does not hold. Regarding the pseudo-optimal estimators, it is shown in the previously mentioned subfigures that our proposed approaches are not far from their pseudo-optimal counterparts. Thus, the convergence to a stationary point offered by the MM-based algorithm is a reasonable solution for the problem at hand. Although, from a rigorous point of view, the presented pseudo-optimal estimators cannot be assured to be the ones that achieve the global optimum, the fact that their performance is very close to their respective local schemes sheds light on the validity of the local and efficient convergence of the MM-based approaches.

On the other hand, the purpose of subfigures 4.5b and 4.6b is to numerically verify the convergence of



(a) Empiric NFQM

(b) Iterations.

Figure 4.6: Asymptotic behavior of the MM-based estimators in a correlated sensor network.

the MM-based algorithms. Surprisingly, considering that the convergence analysis shown in Subsection 4.4.4 ensures that the MM-based algorithm converges to a stationary point of the original problem for $K \geq \max(D, M)$, unexpected results are found in these subfigures. Even though the MM-based algorithms yield terrible results for $K < M$, notice that the MM-based solutions seem to converge for $K = D$ (first dashed red-line). In fact, this phenomenon also occurs for $D \leq K \leq M$ and regardless of whether it is a correlated or uncorrelated sensor network, i.e. the realizations between the dashed lines, since the number of iterations is finite for these values of K in both of the aforementioned subfigures. We conjecture that this phenomenon is caused by the fact that, for $K \geq D$, the update equation in (4.115b) has a unique solution and, thus, subsequent iterates may be obtained by a small step from the current value of $\mathbf{H}_i$ (convergence in terms of the iterates). It is important to remark that the resulting point is not a stationary point of the original problem. This observation highlights the fact that the convergence of the iterates is not a sufficient condition to yield a globally convergent iterative optimization algorithm (see Definition 3.23).

4.4.6.2 Subspace estimation analysis

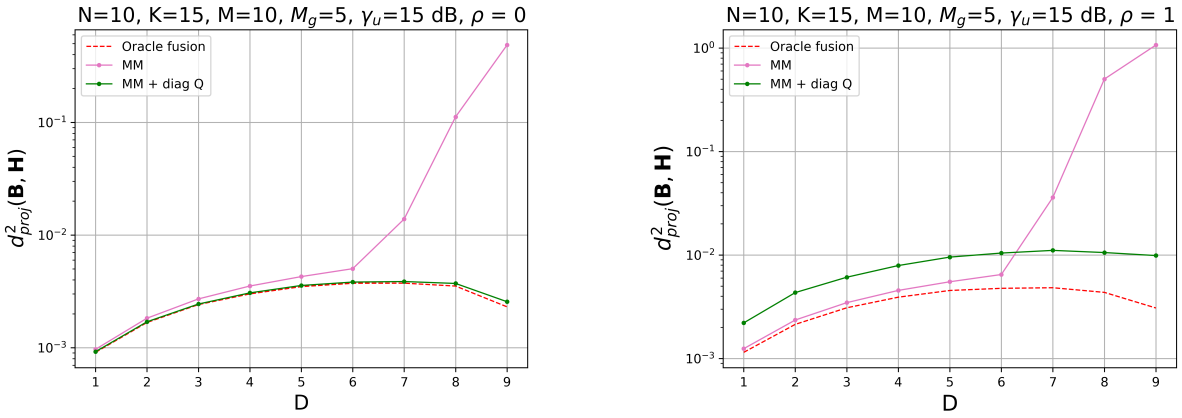
Up to this point, we have verified that the MM-based algorithms are capable of achieving the CRLB for an infinite sample size. Built upon the previous results, we are interested in characterizing the performance of subspace estimators that infer the subspace spanned by the columns of $\mathbf{B}$. The purpose of this assessment is to show that the fusion and regression tasks are coupled for the problem described in Subsection 4.4.1. Particularly, we assess the impact of the fusion on the estimation of the subspace spanned by the columns of $\mathbf{B}$. With the aim at obtaining a benchmark subspace estimator, let us consider the following sample estimator of the fused variable covariance (see (4.98) and (4.123)):

$$\hat{\mathbf{R}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{Y}_k \mathbf{f}_b \mathbf{f}_b^T \mathbf{Y}_k^T, \tag{4.163}$$

where $\mathbf{f}_b$ is the benchmark fusion policy (see (4.17)). Then, the benchmark subspace estimator, $\mathbf{H}_{OF}$, is the one spanned by the columns of the D singular vectors corresponding to the greatest D singular values of $\hat{\mathbf{R}}$. Considering that $\mathbf{H}_{OF}$ is obtained using $\mathbf{f}_b$, we label this subspace estimator as "Oracle fusion" in this set of simulations. In fact, the previous estimator is presumed to be close to the Intrinsic Crámer Rao Bound (ICRB) [174], [175], which is a lower bound on the canonical distance (see 2.77) between the true subspace and its estimation, of the subspace estimation problem.

For this set of simulations, the number of blocks, K, is fixed to 15. This is a sufficiently large sample size such that the MM-based algorithm is globally convergent for $N = 10$ and $M = 15$. The motivation behind only considering the large sample size case is that the impact of the small sample size regime

has already been studied in the previous subsection. As a figure of merit for the subspace estimation, we consider the squared projection F-norm of the Grassmann manifold, $d^2_{proj}(\mathbf{B}, \mathbf{H})$ (see (4.157)). In this manner, only the contributions of the regression errors in the MSE (see (4.151)) are studied.



(a) Uncorrelated sensor network.

(b) Correlated sensor network.

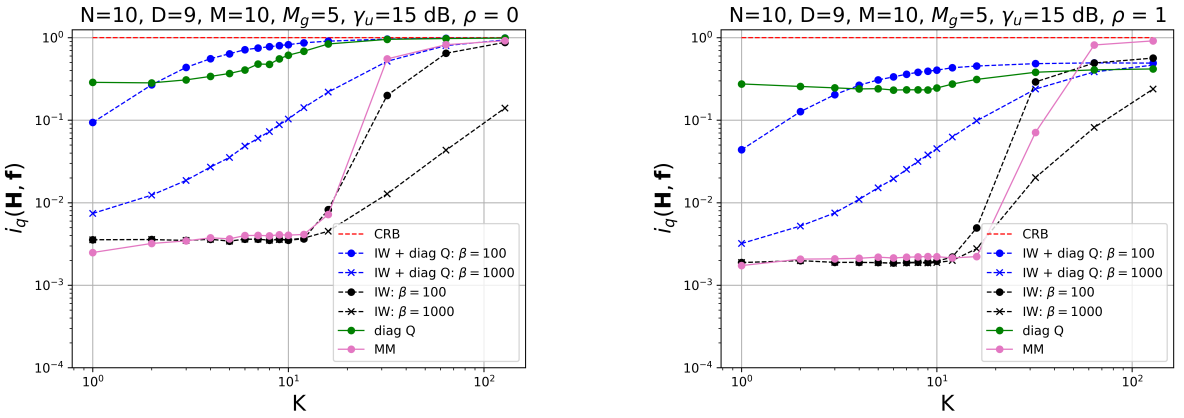
Figure 4.7: Testing the MM-based subspace estimators with respect to the squared projection F-norm of the Grassmann manifold, $d^2_{proj}(\mathbf{B}, \mathbf{H})$.

In Figure 4.7, we plot the results of the aforementioned experiment on a correlated and uncorrelated sensor networks. The general tone of the subspace estimators is that, although it was predicted that the estimation error of a subspace of dimensions D and $N - D$ should be equivalent, the projection F-norm between the true subspace and its estimations has an additional contribution that increases with D in the multisensor fusion and regression problem, as compared to the ICRB given in [174], [175]. Interestingly, after comparing Figure 4.7 with the expression of the MSE (4.158), the additional contribution to the subspace error that increases with D can be explained by the fact that the CRLB of the fusion errors are weighted by $\nu = \frac{D}{N}$. This implies that the fusion errors increase linearly with D, which also affects the subspace of regressors estimation due to the coupling of the fusion and regression in Algorithm 2. This coupling of the fusion and regression operations is further highlighted by the comparison of subfigures 4.7a and 4.7b, where it is seen that the diagonally constrained MM-based solution has a higher subspace error in the correlated sensor network case. Again, this is caused by the fact that the intrinsic assumptions of the diagonally constrained solution does not hold in the correlated sensor network. Thus, the diagonally constrained solution has increased fusion errors which also affect the matrix of regressors subspace estimation.

4.4.6.3 Testing the small sample size approaches

The purpose of this subsection is to compare the regularization approaches described in Subsection 4.4.5. In this regard, we analyze the asymptotic behavior of the IW and the diagonally constrained solutions in a correlated and uncorrelated sensor networks. In light of this, this set of simulations considers the same parameters as those used in Subsection 4.4.6.1, with the exception of the intrinsic dimension, D, which is set as $D = N - 1$. The reasoning behind this value of D is that it depicts the most difficult fusion scenario, as seen from the fact that the second term in (4.159) (the CRLB), which corresponds to the fusion errors, increases linearly with D. We are interested to depict a challenging fusion scenario since the approaches detailed in Subsection 4.4.5 are targeted towards regularizing the intersensor covariance estimation. Thus, the potential gains of the regularized approaches are emphasized in challenging fusion settings.

It is observed in Figure 4.8 that the approaches that consider the diagonal assumption of the intersensor covariance are the better performing ones in the small sample size regime. Yet, it is also verified from both subfigures that the addition of the IW prior to the diagonal assumption results in the best regularized approach for the challenging fusion scenario. Besides, it is also evidenced that the



(a) Uncorrelated sensor network.

(b) Correlated sensor network.

Figure 4.8: Asymptotic performance of the approaches that are targeted towards the small sample size regime.

IW prior by itself does not provide of a sufficient regularization to the intersensor covariance estimation. Actually, the IW prior by itself worsens the performance of the *plain* MM-based solution. As for the regularizing parameter β, it is important to remark that $\beta = 100$ must be the optimal one since, for this value, the IW prior models the fact that every sensor has the variance of the simulated highly contaminated sensors (see (4.161) with $q_2 = 100$) since the considered IW prior assumes that the intersensor covariance is $\beta\mathbf{I}$.

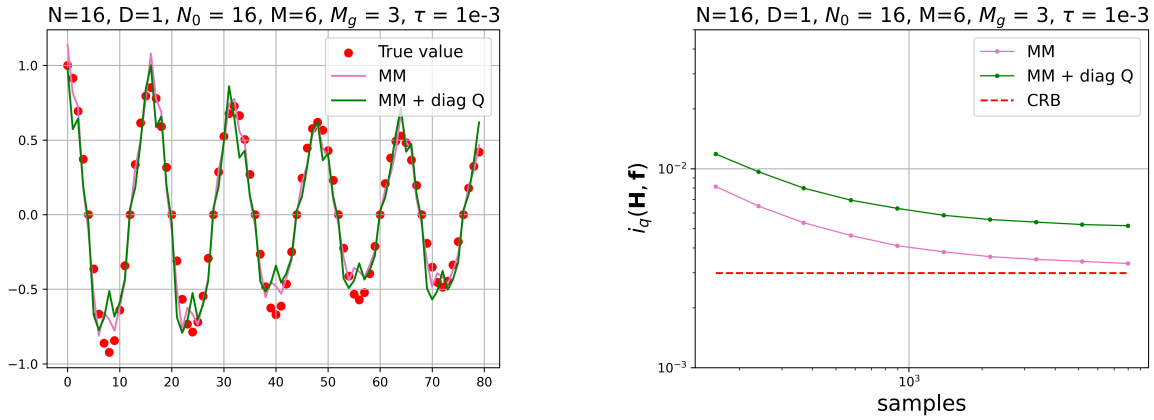
4.4.6.4 Practical example

In this subsection, we show how well the model depicted in (4.94) can approximate a non-linear function using a toy example. For this purpose, we consider a measured phenomenon that has the following form:

$$x(n) = \exp(-\tau n)\cos\left(2\pi\frac{n}{N_0}\right) \quad \text{for } n > 0, \tag{4.164}$$

where $\tau > 0$ is the damping parameter and N_0 is the period of the damped cosine. Similarly to the previous subsection, the remaining simulation parameters are stated on top of each figure, where we aided the MM approaches with the automatic choice of dimensionality approach from [135] to fix the value of D. In order to integrate the proposed fusion and regression scheme into the temporal series described by (4.164), we partition the available measurements in chunks of N samples and feed them into Algorithm 2. For simplicity, we only consider the correlated scenario and the "MM" and "MM + diag Q" estimators of $\mathbf{H}$ and $\mathbf{f}$. Also, we use the MSE of the measurements, $\gamma(\mathbf{H}, \mathbf{f})$, in front of the NFQM from Definition 4.12 for a clearer exposition of the regression analysis, given that we are only interested in showcasing the performance of the MM-based solutions and not on the comparison of different estimators.

While in Subfigure 4.9b it is observed how the MSE of the regression evolves as a function of the number of samples, a snapshot of the subspace-based non-linear regression is shown in Subfigure 4.9a as an example. Regarding the asymptotic behavior, it is observed that the subspace-based regression MSE does not tend to 0 as the number of samples tends to infinity. The reasoning behind the previous result is that the number of parameters in the subspace-based regression model in (4.97) increases with the number of available blocks. Therefore, the ratio between number of parameters and the number of samples is always constant and equal to ν (see (4.155)). In fact, it is shown that only the MM estimator, which is the one that considers the correlation between sensors, reaches the CRLB in the limit, further confirming that the proposed subspace-based regression cannot be an efficient estimator due to the consideration of K data blocks. On the other hand, it is demonstrated in Figure 4.9a that the proposed approaches are capable of approximating a non-linear function without any additional modification of the signal model. Indeed, the previous feature is an advantage with respect to other



(a) Regression. **(b)** MSE.

Figure 4.9: Dampened sinusoid toy example.

known approaches for the fusion and regression problem, such as the Kalman filter, which requires the complete knowledge of the non-linear function. Yet, we remark that we fixed N to be equal to the periodic component of the measured phenomena, being a reasonable prior information in a practical implementation of a non-linear regressor of this kind.

4.5 Concluding remarks

This chapter was devoted to the study of fusion schemes that are capable to exploit the spatial diversity available in the multisensor fusion problem. While we reviewed and extended the CI principle to the considered problem, we also proposed two new fusion schemes with different motivations behind them. Although the fusion schemes provided by the CI algorithm and the E-BLUE are impractical due to the fact that there is no natural way to incorporate the estimation of the noise uncertainties into these schemes, they provide insightful operational meanings to information theoretic descriptors. For instance, the waterfilling formulation that obtains the optimal intersection weights in the CI fusion scheme results in an alternative way of determining the sparsity of the fusion policy using a positive constant. Yet, the implications of the waterfilling formulation are a future line of research. Alternatively, in the E-BLUE fusion scheme, we derived a natural interpretation of the entropic index of the Rényi entropy, which provides a trade-off between the reliability and precision of the final fusion.

On the other perspective, while the MM implementation of the CML fusion scheme did not provide any meaningful connection to other signal processing domains (as compared to the other two fusion schemes), it motivated the derivation of the MM framework on the Grassmann manifold from the previous chapter. In fact, the resulting algorithm is an example of a practical implementation of the block MM framework on the Grassmann manifold.

The next chapter explores the diversity that is found in the angular domain on the Uplink-Downlink Covariance Conversion problem in MIMO communications, which is motivated (in part) by the ideas explored in this chapter.

Chapter 5

Exploiting angular diversity in the Covariance Conversion problem

Continuing the ideas from the previous chapter, we want to explore other forms of diversity that appear in signal processing applications. Particularly, we want to study the so-called *angular diversity* that appears in the Covariance Conversion problem from wireless communication channels.

In order to enable the two-way communication between two terminals in a wireless communication channel, there is an inherent need to avoid the collisions between the transmissions of both terminals. Among all the possible alternatives that enable the two-way communication between two terminals, there are two basic ones, which are the Time and Frequency Division Duplex schemes (TDD and FDD, respectively) [186]. Those schemes are founded on the definition of two independent channels, denoted as the *Uplink* (UL) and *Downlink* (DL) channels, which are multiplexed using different techniques by each communication scheme. For instance, TDD communications use the same carrier frequency to both channels and assign different time slots for each channel. Hence, UL and DL channels are reciprocal as long as the transmission period is shorter than the channel coherence time. Even if the transmission period is larger, the Channel State Information (CSI) obtained in the UL channel can be used in the DL channel with minimal performance loss and training overhead. This is especially true for statistical CSI, which is expected to vary slowly with time [186]. In contrast, FDD schemes allow both terminals to transmit information simultaneously through two independent channels that are allocated in different carrier frequencies. In this way, FDD schemes are capable to achieve a smaller latency as compared to TDD schemes, at expense of the channel reciprocity.

The Covariance Conversion (CC) technique [39], [98], [136] exploits the properties of wireless channels that are independent of the carrier frequency in a way that it can provide channel reciprocity to an FDD scheme. In particular, known approaches of the CC method exploit the fact that the distribution of the signal power in the angular domain is reciprocal to both channels, which is the reason why in this problem it is said that there exist an angular diversity.

While this chapter is motivated by the aforementioned wireless communication problem, the scope of this chapter is to tackle this problem from an algorithmic point of view. Indeed, we show that the angular diversity arises in a different manner as compared to the diversity studied in Chapter 4. What is more, since the angular diversity admits in a natural manner the incorporation of sparse-aware ideas, we explore the use ℓ_1 regularizers in the exploitation of this kind of diversity. With the previous ideas in mind, we distance ourselves (as much as possible) from the communications model in order to focus on the abstract ideas behind the angular diversity. The ideas presented in this chapter can also be found on our published work in [125] and is structured as follows: in Section 5.1 we depict the FDD setting in which the CC methodology thrives and, additionally, we formally define the CC problem. Next, we derive our proposal to solve the CC problem and detail the implications of the angular diversity in our formulations in Section 5.2. We conclude in Section 5.3 with some numerical results that showcase the capabilities of our proposals and compare them with other approaches to the CC problem.

5.1 Problem statement

We consider that two terminals, referred to as User Equipment (UE) and Base Station (BS), are connected thanks to a full duplex channel via an FDD scheme. As previously mentioned, this is done by conveying the information through two independent channels, denoted as channel 1 and channel 2. The associated carrier wavelengths for each channel are λ_1 and λ_2, respectively. For simplicity, even though the presented ideas can be generalized to MIMO schemes, we assume that the channel 1 (BS to UE) is a MISO configuration. Hence, channel 2 (UE to BS) is a SIMO configuration. After denoting the BS number of antennas as M and remarking the fact that the UE has only 1 antenna, channel 1 is modeled as follows:

$$y_1(n) = \mathbf{c}^H \mathbf{h}_1(n) x_1(n) + w_1(n), \tag{5.1}$$

where $y_1(n)$ is the received MISO signal, $\mathbf{h}_1(n) \in \mathbb{C}^M$ is channel 1 weight vector, $\mathbf{c}$ is the channel precoding at the BS, $x(n) \in \mathbb{C}$ are the transmitted symbols and $w_1(n)$ is an additive Gaussian noise process. Similarly, channel 2 (UE to BS) is depicted as follows:

$$\mathbf{y}_2(n) = \mathbf{h}_2(n) x_2(n) + \mathbf{w}_2(n), \tag{5.2}$$

where $\mathbf{y}_2(n) \in \mathbb{C}^M$ is the received SIMO signal, $\mathbf{h}_2(n) \in \mathbb{C}^M$ is channel 2 weight vector, $x(n) \in \mathbb{C}$ are the transmitted symbols and $\mathbf{w}_2(n) \in \mathbb{C}^M$ is an additive Gaussian vector. To simplify the exposition, we denote as channel c any particular channel and c' its complementary channel from this point on, where $c, c' \in \{1, 2\}$. The second-order statistics of the channel vectors are:

$$\mathbf{R}_c = \mathrm{E}\left[\mathbf{h}_c(n) \mathbf{h}_c^H(n)\right] \quad c = 1, 2, \tag{5.3}$$

which can be alternatively described by the following widely accepted model:

$$\mathbf{R}_c = \int_{-\pi}^{\pi} \rho(\theta) \mathbf{a}_c(\theta) \mathbf{a}_c^H(\theta) \mathrm{d}\theta \quad c = 1, 2, \tag{5.4}$$

where $\rho(\theta)$ is the *Angular Power Spectrum* (APS) and $\mathbf{a}_c(\theta)$ is the normalized (unit norm) array response of the BS antennas at the carrier frequency of channel c. Although the original reference that stated the model in (5.4) is [1], we refer to the CC [39], [136], [137] (particularly, [137, Equation (1)]) and estimation of the Direction of Arrival (DoA) [119] literature for a clear description of the previous model. In fact, the considered scenario of MISO/SIMO channels is also considered in [137]. Notice that the model in (5.4) is a generalization of the following expression [86], [115]:

$$\mathbf{R}_c = \sum_{l=1}^{L} \rho_l \mathbf{a}_c(\theta_l) \mathbf{a}_c^H(\theta_l), \tag{5.5}$$

where ρ_l and θ_l are the power and angle of arrival of the signal coming from the l-th path, respectively. The previous expression is considered in DoA problems, e.g. see [119, Eq. (9)]. Clearly, we get (5.4) from (5.5) by letting $L \to \infty$. Equivalently, we can also see (5.5) as a particular case of (5.4) in which the APS is a finite set of Dirac impulses.

In (5.4), $\mathbf{a}_c : [-\pi, \pi] \to \mathbb{C}^M$ depends on the array configuration and the carrier wavelength in general. For instance, the Uniform Linear Array (ULA) response is given by the following expression:

$$\mathbf{a}_c(\theta) = \frac{1}{\sqrt{M}} \left[1, \exp\left(j2\pi \frac{d}{\lambda_c} \sin(\theta)\right), ..., \exp\left(j2\pi \frac{d}{\lambda_c}(M-1) \sin(\theta)\right)\right]^T \quad c = 1, 2, \tag{5.6}$$

where d is the separation between the elements of the BS antenna array. Clearly, $\mathbf{a}_c(\theta)$ is an expression that is dependent on the carrier wavelength. On the contrary, $\rho : [-\pi, \pi] \to \mathbb{R}_+$ is a function that describes how the signal power is distributed along the angular domain [98], [136]. The previous description of the APS suggests that it is a function of the signal scatterers that are distributed between both terminals, i.e. the geometry of the environment. Thus, it is a slow time-varying figure [137] that is invariant to the carrier frequency. The previously mentioned properties of the APS are what

motivates the CC procedure. Not only the consideration of the APS provides the angular diversity to the communications framework (due to the shared information between both channel correlation matrices), but it also opens an alternative to obtain statistical channel reciprocity in FDD systems. The main advantage of statistical CSI over full CSI is that the statistical knowledge of the channel (second-order statistics) is valid for an extended period of time due to the slow time-varying nature of the APS. Just to provide some insights on the use of statistical CSI (thus, motivating the CC methodology), let us briefly introduce the eigen-beamforming idea [66], [210]. In essence, the eigen-beamforming is the procedure that obtains the linear combiners of the previously considered communications channels ($\mathbf{c}$ in the case of channel 1) such that the resulting SNR is optimal on average. Taking channel 1 as an example, the latter idea means that the optimal $\mathbf{c}$ is obtained from the following optimization problem:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} SNR(\mathbf{c}) = \arg\max_{\mathbf{c}} \frac{\mathrm{E}[|\mathbf{c}^H \mathbf{h}_1(n) x(n)|^2]}{\mathrm{E}[|w_1(n)|^2]} = \tag{5.7a}$$

$$\arg\max_{\mathbf{c}} \frac{\mathrm{E}[\mathbf{c}^H \mathbf{h}_1(n) \mathbf{h}_1^H(n) \mathbf{c} |x(n)|^2]}{\mathrm{E}[|w_1(n)|^2]} = \arg\max_{\mathbf{c}} \frac{\mathbf{c}^H \, \mathrm{E}[\mathbf{h}_1(n) \mathbf{h}_1^H(n)] \mathbf{c} \, \mathrm{E}[|x(n)|^2]}{\mathrm{E}[|w_1(n)|^2]} = \tag{5.7b}$$

$$\arg\max_{\mathbf{c}} \frac{\mathbf{c}^H \mathbf{R}_1 \mathbf{c} \, \mathrm{E}[|x(n)|^2]}{\mathrm{E}[|w_1(n)|^2]} = \arg\max_{\mathbf{c}} \mathbf{c}^H \mathbf{R}_1 \mathbf{c}, \tag{5.7c}$$

whose optimal solution is the eigenvector corresponding to the greatest eigenvalue of $\mathbf{R}_1$ (see 3.2.2.1). Clearly, the optimal channel precoding obtained from (5.7c) is useful as long as $\mathbf{R}_1$ is valid. In the case of fixed transmitters and receivers, the temporal variability of $\mathbf{R}_1$ only depends on the APS (see (5.4)), which is valid for a longer period of time than the channel coherence time [186].

5.1.1 Angular Power Spectrum sparsity

As stated in the introduction of this chapter, we are interested in the particularization of the spatial channel model in (5.4) to the cases where the APS is sparse. Firstly, we need to find a definition of a sparse function. The main issue that arises in this task is that it is difficult to find a formal generalization of the definition of sparsity given in Chapter 2 to continuous functions (not to be confused with sparse approximation of a function [34]). Instead, we resort to an intuitive definition of a sparse function. In order to informally define the sparsity of the APS, we consider a model of this function termed the Geometry-based Stochastic Channel Model (GSCM) [136, Section 5]. The GSCM model approximates the APS by superposing the contributions of S independent scatter clusters:

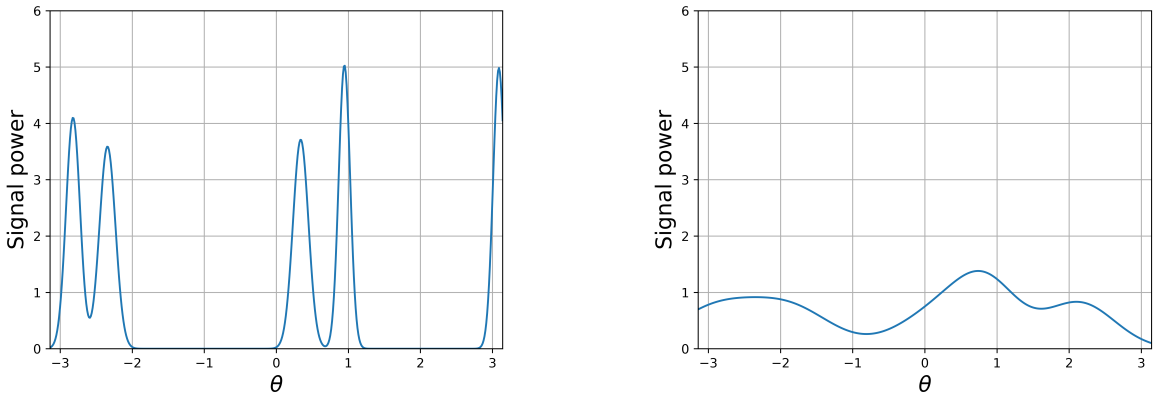
$$\rho(\theta) = \sum_{s=1}^{S} \alpha_s k(\theta, \theta_s, \sigma_s), \tag{5.8}$$

where α_s is a real and positive constant that denotes the s-th cluster weight and $k(\theta, \theta_s, \sigma_s)$ is any appropriate kernel function (real and positive) that is able to model the dispersion of each cluster. In this case, the kernel function is parameterized by θ_s and σ_s. The previous parameters correspond to the s-th cluster center in the angular domain (location parameter) and angular spread (scale parameter), respectively. Any radial basis function, i.e. a function that is such that $f(\theta) = f(|\theta|)$, that vanishes for increasing $|\theta|$ is suitable to model the APS. In practice, the Gaussian kernel is often considered in 5.8, which has the following expression:

$$k(\theta, \theta_s, \sigma_s) = \frac{1}{\sqrt{2\pi\sigma_s}} \exp\left(-\frac{|\theta - \theta_s|^2}{2\sigma_s} \right). \tag{5.9}$$

Notice that (5.8) becomes a GMM by considering (5.9) as the kernel function, which is an idea that resonates with other models that were used in the previous chapter (see, for instance, the contaminated Gaussian model in (4.72)). In this regard, information theoretic measures could be used to assess the sparsity of the APS, although some sort of normalization is expected. Yet, we prefer to consider an intuitive definition of the APS sparsity in this chapter because we want to focus more on the algorithmic perspective of the APS estimation. The APS is said to be sparse if the S clusters in (5.8) are well

differentiated. Note that, while the previous definition is an intuitive one, it explains a phenomenon that occurs in the mmWave and ultra-wideband channels [170]. In these kinds of channels, electromagnetic waves behave in a similar manner to optic waves, which, essentially, is a phenomenon described by a sparse APS. In Figure 5.1, we show graphically two examples of APSs: one that is sparse and another that is non-sparse. The sparse modeling of the APS is achieved for relatively small values of the angular spread in (5.8). To illustrate the magnitude of the angular spreads for each example, in the sparse APS case we draw σ_s from a uniform random variable in $[\frac{0.3\pi}{180}, \frac{0.8\pi}{180}]$ rads, while the angular spread is drawn from $[\frac{3\pi}{180}, \frac{8\pi}{180}]$ rads in the non-sparse case. In contrast to the sparse APS case, where there are angular windows in which $\rho(\theta)$ is very close to zero, the received power is spread through the entire angular domain in the non-sparse case.



(a) Sparse APS. **(b)** Non-sparse APS.

Figure 5.1: Examples of sparse and non-sparse APSs for $S = 5$ and Gaussian kernels.

5.1.2 Covariance Conversion problem statement

In light of the previous ideas, we have all the ingredients to define the Covariance Conversion problem in the FDD setting.

Definition 5.1 (Covariance Conversion (CC) problem in FDD systems)**.** *Let be any two channels of an FDD scheme be modeled by* (5.2) *and* (5.1)*. Also, let their respective channel correlation matrices be* $\mathbf{R}_1$ *and* $\mathbf{R}_2$*. Then, the Covariance Conversion (CC) problem on the previously stated FDD scheme is defined as follows: given an estimation of one the two channel correlation matrices,* $\hat{\mathbf{R}}_c$*, the CC is a procedure that estimates the complementary channel correlation matrix,* $\hat{\mathbf{R}}_{c'}$*.*

Remark 5.1. The previous definition aims to generalize the notation of the Uplink-Downlink Covariance Conversion (UDCC) problem [125]. While there are several alternatives that solve the UDCC problem in FDD systems [98], [115], we focus on the ones that resort on some estimator of $\rho(\theta)$, such as in [136], [137]. The reason for this is that the methods that estimate $\rho(\theta)$ are inherently exploiting the angular diversity of the problem.

Remark 5.2. The estimation of any of the two channel correlation matrices is naturally obtained in the FDD framework. For instance, in the MISO/SIMO schemes described in Section 5.1, the BS can estimate $\mathbf{R}_2$ using the transmissions from the UE in channel 2.

We are interested in the parameters that complicate the CC problem. In this regard, the considered measure of the difficulty of the CC problem is the Frobenius distance between $\mathbf{R}_1$ and $\mathbf{R}_2$. Indeed, an alternative lecture of this distance is that it accounts for the magnitude of the errors that come from the naive utilization of $\mathbf{R}_1$ instead of $\mathbf{R}_2$ (and vice versa) in a statistical CSI framework. In light of the previous intuitions, let us consider the Frobenius distance between $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$||\mathbf{R}_1 - \mathbf{R}_2||_F^2 = \text{tr}\left(\mathbf{R}_1^H\mathbf{R}_1 - 2\Re\left(\mathbf{R}_1^H\mathbf{R}_2\right) + \mathbf{R}_2^H\mathbf{R}_2\right), \tag{5.10}$$

where $\Re(\cdot)$ returns the element-wise real part of the input matrix. Regarding the assessment of the statistical channel reciprocity, the worst-case scenario occurs when (5.10) has a high value. For illustration purposes, we consider a toy example that unveils the cases that result in a higher distance between $\mathbf{R}_1$ and $\mathbf{R}_2$. To this end, we further assume that the BS is constructed using an ULA and that the APS has the following expression:

$$\rho(\theta) = \rho_0 \delta(\theta - \theta_0), \tag{5.11}$$

where ρ_0 is any positive constant, θ_0 is any DoA, and $\delta(\theta)$ is the Dirac delta function. Using the previous toy APS, a much simpler, but insightful, analysis of (5.10) is possible. It is verified that (5.10) yields the following closed-form expression under the previous assumptions (see Appendix 8.3.1 for more details):

$$||\mathbf{R}_1 - \mathbf{R}_2||_F^2 = 2\rho_0^2 \left(1 - \left| \frac{\sin\left(M\pi \frac{d}{\lambda_2} \sin(\theta_0) \left(1 - \frac{\lambda_2}{\lambda_1} \right) \right)}{M \sin\left(\pi \frac{d}{\lambda_2} \sin(\theta_0) \left(1 - \frac{\lambda_2}{\lambda_1} \right) \right)} \right|^2 \right), \tag{5.12}$$

which is a function that has a clear dependence on the ratio between the carrier wavelengths. Firstly, we study the cases where the channel reciprocity holds, i.e. the Frobenius distance between $\mathbf{R}_1$ and $\mathbf{R}_2$ is zero. The first one of them is found when $\theta_0 = 0$ or $\theta_0 = \pi$, which corresponds to the case where there are no phase shifts between array elements because the received/transmitted wavefronts are parallel to the antenna array. In this case, both matrices yield the following expression:

$$\mathbf{R}_1 = \mathbf{R}_2 = \frac{\rho_0}{M} \mathbf{1}_M \mathbf{1}_M^H, \tag{5.13}$$

which describes the fact that all the antennas in the BS are correlated due to the previously mentioned reasons. The remaining case is the trivial one, where both channel carrier frequencies are equal ($\lambda_1 = \lambda_2$). In the previous two cases there is no necessity of the covariance conversion. While in the first one there is no angular diversity, the second one is not an FDD communications setting.

Besides, we show the cases where (5.12) is different from zero by means of an insightful numerical analysis. With this aim, we define the following function, which is inspired by the penalizing multiplicative constant from (5.12):

$$\kappa\left(\theta_0, \frac{\lambda_2}{\lambda_1} \right) = 1 - \left| \frac{\sin\left(\frac{M\pi}{2} \sin(\theta_0) \left(1 - \frac{\lambda_2}{\lambda_1} \right) \right)}{M \sin\left(\frac{\pi}{2} \sin(\theta_0) \left(1 - \frac{\lambda_2}{\lambda_1} \right) \right)} \right|^2, \tag{5.14}$$

where we particularized $d = \frac{\lambda_2}{2}$ for illustration purposes. Note that the magnitude of $||\mathbf{R}_1 - \mathbf{R}_2||_F^2$ is directly proportional to (5.14). In Figure 5.2, the behavior of $\kappa(\theta_0, \frac{\lambda_2}{\lambda_1})$ is observed for different values of θ_0 as a function of the ratio of the channel wavelengths, $\frac{\lambda_2}{\lambda_1}$. It is verified in this figure that, whenever the carrier frequencies differ, the value of the penalizing multiplicative constant increases. In fact, even for reasonable ratios between carrier wavelengths, the value of (5.14) is high enough so that channel reciprocity does not approximately hold. To provide some examples of the practical values of $\frac{\lambda_2}{\lambda_1}$, see, for instance, in [98], [136] where the considered ratios between carrier wavelengths range between 0.9 and 0.95. Likewise, the closer the value of θ_0 is to $\theta_0 = \frac{\pi}{2}$, the higher the value of (5.14). Even though the previous arguments are based on a particularization of the APS (see (5.11)), they can be easily extended to the general case since any general APS is a superposition of infinite terms that are equivalent to (5.11).

5.2 Estimation of a quantized sparse APS

Our proposed solution to the CC problem revolves around a numerical approximation of the integral in (5.4). In this regard, we produce an approximation of this integral using a similar expression to the one in (5.5). With this aim, we quantize the angular domain and every associated magnitude.

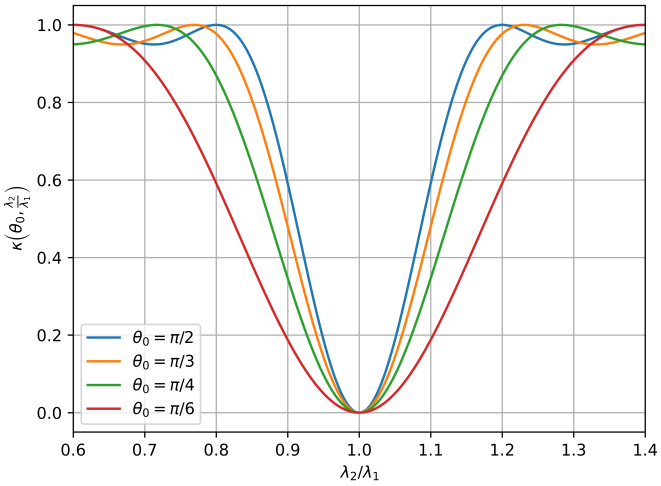


Figure 5.2: Graphical representation of the penalizing multiplicative constant of the Frobenius distance between $\mathbf{R}_1$ and $\mathbf{R}_2$. Different colors correspond to different values of θ_0.

Firstly, we generate a sequence of equispaced elements in $[\theta_l, \theta_u]$, where θ_l and θ_u (with $\theta_l < \theta_u$) are the parameters that describe the sector having relevant propagation paths. This sequence is defined as

$$\theta_n = \theta_l + (n-1)\frac{\theta_u - \theta_l}{N-1} \quad n = 1, ..., N, \tag{5.15}$$

where N is the total number of samples in the quantized angular domain. The previous sequence generates a mapping between the continuous angular domain to the quantized angular domain, which is, essentially, a function that relates $\theta \in [\theta_l, \theta_u]$ to the closest element from $\{\theta_n\}_{n=1,...,N}$. There are some applications that motivate a selection of θ_l and θ_u that is different from $-\pi$ and π (the whole angular domain), respectively. In the case of ULAs, it is known that this array configuration cannot differentiate between θ_0 and its reciprocal angle, $\theta_0 + \pi$. As a result, an efficient choice of θ_l and θ_u for ULAs would be $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, respectively. Additionally, in typical cellular networks sectorization configurations [42], it is wise to choose the visible angles to be in $[-\frac{\pi}{3}, \frac{\pi}{3}]$, given that the received signals are expected to be transmitted/received in this range of DoAs. In all cases, having a narrower expected span of angles results in a better quantization resolution for a fixed complexity, which is accounted by the total number of quantized samples, N.

Using the previous quantization of the angular domain, the numerical approximation of the spatial channel model in (5.4) is given by the following expression:

$$\mathbf{R}_c = \sum_{n=1}^{N} \rho(\theta_n)\mathbf{a}_c(\theta_n)\mathbf{a}_c^H(\theta_n)\Delta\theta \quad c = 1, 2, \tag{5.16}$$

where $\Delta\theta = \frac{\theta_u - \theta_l}{N-1}$ is a multiplicative constant such that (5.16) tends to (5.4) for $N \to \infty$. Note that the previous quantized estimation of the channel correlation matrices can be rewritten in a more compact manner using the vectorization function of a matrix, which is denoted as $\text{vec} : \mathbb{C}^{M \times N} \to \mathbb{C}^{MN}$. This compact expression is given by:

$$\mathbf{r}_c = \mathbf{A}_c\boldsymbol{\rho} \quad c = 1, 2, \tag{5.17}$$

where:

$$\mathbf{r}_c = \text{vec}(\mathbf{R}_c) \quad c = 1, 2, \tag{5.18a}$$

$$\boldsymbol{\rho} = [\rho(\theta_1), ..., \rho(\theta_N)]^T, \tag{5.18b}$$

$$\mathbf{A}_c = \Delta\theta \left[\text{vec}\left(\mathbf{a}_c(\theta_1)\mathbf{a}_c^H(\theta_1)\right), ..., \text{vec}\left(\mathbf{a}_c(\theta_N)\mathbf{a}_c^H(\theta_N)\right)\right] \quad c = 1, 2. \tag{5.18c}$$

An important remark of the previous equations is that the dimensions of $\mathbf{A}_c$ and $\mathbf{r}_c$ are $M^2 \times N$ and M^2, respectively. The system of equations in (5.17) is a more convenient expression for the optimization framework shown in the sequel. In fact, it is thanks to (5.17) that the angular diversity of the CC problem can be highlighted from an algorithmic point of view. In contrast to what occurs in other signal processing problems, the common information between two datasets (channels in this case) does not consist of a shared linear model, i.e. the matrix expression, as it is the case shown in Subsection 4.4.1. Instead, the vectorized second-order statistics of both channels ($\mathbf{r}_c$ in (5.18a)) utilize the same coefficients (a common vector $\boldsymbol{\rho}$) with different matrices of regressors (different matrices $\mathbf{A}_c$ in (5.17)). This is the expression of diversity in this signal processing problem.

The remaining paragraphs are devoted to the estimation of $\boldsymbol{\rho}$. The main issue with the formulation in (5.17) for the estimation of $\boldsymbol{\rho}$ is that, in general, $\mathbf{A}_c$ is not a full-rank matrix. For instance, we have observed numerically that $\mathbf{A}_c$ is such that $\text{rank}(\mathbf{A}_c) < \min(M^2, N)$. This means that, in order to obtain any kind of estimation of $\boldsymbol{\rho}$, we need to resort to Least Squares-like (LS) estimators. Let an estimator of $\mathbf{r}_c$ be $\hat{\mathbf{r}}_c$. Then, the LS-based optimization problem that estimates the APS is:

$$\hat{\boldsymbol{\rho}}_{LS} = \arg\min_{\boldsymbol{\rho}} ||\hat{\mathbf{r}}_c - \mathbf{A}_c\boldsymbol{\rho}||_2^2 \quad \text{s.t. } \Re(\boldsymbol{\rho}) \succeq \mathbf{0}_M, \Im(\boldsymbol{\rho}) = \mathbf{0}_N, \tag{5.19}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ are the element-wise real and imaginary part of its input vector, respectively. Note that the previous constraints are needed to ensure that $\hat{\boldsymbol{\rho}}_{LS}$ retains its physical meaning. While the positivity constraint ensures the angular power distribution interpretation of $\hat{\boldsymbol{\rho}}_{LS}$, the second constraint is needed to mitigate the effects of the noisy entries in $\hat{\mathbf{r}}_c$ caused by its statistical variability. Indeed, if the actual value of $\mathbf{r}_c$ were used, the second constraint would not be needed. The optimal solution of $\hat{\boldsymbol{\rho}}_{LS}$ is easily obtained by deriving the unconstrained LS solution, and then project it to the constraint set, given that it admits a simple projection, yielding:

$$\hat{\boldsymbol{\rho}}_{LS} = \max\left(\Re\left(\mathbf{A}_c^{\dagger}\mathbf{r}_c\right), \mathbf{0}_N\right), \tag{5.20}$$

where the maximum and the real part operations correspond to the projections onto the set depicted by the first and second constraints. Moreover, $\mathbf{A}_c^{\dagger}$ is the Moore-Penrose pseudoinverse of $\mathbf{A}_c$. The previous pseudoinverse is computed from the SVD of $\mathbf{A}_c$ using the following procedure: provided that $\mathbf{A}_c = \mathbf{U}\mathbf{D}\mathbf{V}^H$ is the SVD of $\mathbf{A}_c$, its Moore-Penrose pseudoinverse is given by:

$$\mathbf{A}_c^{\dagger} = \mathbf{V}\mathbf{D}^{\dagger}\mathbf{U}^H, \tag{5.21}$$

where $\mathbf{D}^{\dagger}$ is the pseudoinverse of the matrix containing the singular values of $\mathbf{A}_c$, which consists in the inversion of only the non-zero diagonal entries of $\mathbf{D}$. The previous expression of the pseudoinverse is necessary because its widely known expression, i.e. $\mathbf{A}_c^{\dagger} = (\mathbf{A}_c^H\mathbf{A}_c)^{-1}\mathbf{A}_c^H$, is ill-posed since $\mathbf{A}_c^H\mathbf{A}_c$ is rank deficient.

It is worth mentioning that (5.20) is, essentially, the projection onto the subspace generated by the columns of $\mathbf{A}_c$. This technique is related to the one in [136], where the APS is found as the intersection of several sets that are defined using projections of functions in a Hilbert space (see [136, Eq. (6)]). Still, the difference between both rationales is clearly seen in the conversion operation (and its meaning) of the CC problem from channel c to channel c'. More concretely, in the quantization-based APS estimation, the conversion step is done using the remaining channel equations from (5.17), meaning that if one estimates the APS with $\mathbf{A}_c$ and $\hat{\mathbf{r}}_c$, the converted channel correlation vector is computed as follows:

$$\hat{\mathbf{r}}_{c'} = \mathbf{A}_{c'}\hat{\boldsymbol{\rho}}_{LS}. \tag{5.22}$$

Similarly, the projection method in [136] also solves a system of equations to obtain the coordinates of $\rho(\theta)$ in a Hilbert space, but the involved matrices are a change of basis in said Hilbert space from channel c correlation matrix to channel c' correlation matrix. Since the computation of the change of basis matrices in [136] requires a precomputation of $4M^2$ integrals, the resulting conversion algorithm is much less flexible than the one emanating from (5.17). In fact, not only it is easier to change the array responses in (5.17), as compared to the approach in [136], but it also allows the user to make a compromise between the computational complexity and the desired performance by fixing the number of quantized samples, N.

In the following subsection, we aim to improve the LS solution in (5.20) by informing the optimization formulation of two observations: the APS is sparse (as described in Subsection 5.1.1) and that the estimation of the channel correlation matrix is imperfect. This means that $\boldsymbol{\rho}$ in (5.17) is a sparse vector in the classical sense and that there must be an additional constraint that accounts for the errors in $\hat{\mathbf{r}}_c$. Particularly, we explore the ℓ_1 norm regularization to induce sparsity on $\boldsymbol{\rho}$, in addition to additional constraints that resemble the classical BPD formulation (see (2.18)).

5.2.1 Sparse-aware APS estimation

In order to incorporate the sparse APS assumption and the prior knowledge of an imperfect channel covariance estimation, we resort to the BPD formulation (see (2.18)). Built upon the LS formulation in (5.19), the proposed BPD formulation for the CC problem is:

$$\hat{\boldsymbol{\rho}}_{BPD} = \arg\min_{\boldsymbol{\rho}} ||\boldsymbol{\rho}||_1 \quad \text{s. t. } \; ||\hat{\mathbf{r}}_c - \mathbf{A}_c\boldsymbol{\rho}||_2^2 \leq \varepsilon, \Re(\boldsymbol{\rho}) \succeq \mathbf{0}_N, \Im(\boldsymbol{\rho}) = \mathbf{0}_N, \tag{5.23}$$

where the motivation behind the first constraint is to allow a reasonable amount of error between the prior channel covariance vector, $\hat{\mathbf{r}}_c$, and the linear model given by $\mathbf{A}_c\boldsymbol{\rho}$. Intuitively, the aforementioned constraint allows for an additional degree of flexibility that should robustify the resulting solution. The remaining constraints are imported from (5.19).

It is shown in the remaining of this subsection that the ADMM methodology reviewed in Subsection 3.3.5.1 is suited for the optimization problem in (5.23). Particularly, we revisit the ADMM framework for the sparse regression problem [26] and adapt it to (5.23). For the previous purpose, notice that (5.23) can be rewritten using indicator functions (see Definition 3.25):

$$\hat{\boldsymbol{\rho}}_{BPD} = \arg\min_{\boldsymbol{\rho}} ||\boldsymbol{\rho}||_1 + \mathcal{I}_{B_2(\hat{\mathbf{r}}_c,\varepsilon)}(\mathbf{A}_c\boldsymbol{\rho}) + \mathcal{I}_{\mathbb{R}_+}(\boldsymbol{\rho}), \tag{5.24}$$

where:

$$B_2(\hat{\mathbf{r}}_c, \varepsilon)(\mathbf{x}) = \{\mathbf{x} \in \mathbb{R}^N : ||\mathbf{x} - \hat{\mathbf{r}}_c||_2^2 \leq \varepsilon\}, \tag{5.25}$$

and:

$$\mathbb{R}_+ = \{\mathbf{x} \in \mathbb{R}^N : \Re(\mathbf{x}) \succeq \mathbf{0}_N, \Im(\mathbf{x}) = \mathbf{0}_N\}. \tag{5.26}$$

In order to transform (5.24) into the equivalend standard ADMM formulation shown in (3.179), we introduce two new optimization variables, denoted as $\mathbf{z}_1$ and $\mathbf{z}_2$, and two additional constraints such that the resulting optimization problem is still equivalent to (5.24). With the introduction of these two variables, (5.24) becomes:

$$\min_{\boldsymbol{\rho},\mathbf{z}_1,\mathbf{z}_2} I_{B_2(\hat{\mathbf{r}}_c,\varepsilon)}(\mathbf{z}_1) + I_{\mathbb{R}_+}(\mathbf{z}_2) + ||\mathbf{z}_2||_1 \quad \text{s. t. } \; \mathbf{A}_c\boldsymbol{\rho} = \mathbf{z}_1, \mathbf{z}_2 = \boldsymbol{\rho}. \tag{5.27}$$

The optimization with respect to $\boldsymbol{\rho}$ of the previous optimization problem will become clear with the consideration of the Augmented Lagrangian. Essentially, these new constraints define a change of variables that only affects the respective term (the one with the same variables) of the convex program in (5.27), resulting in an expression that is easier to optimize due to the separability of the cost function. Also, note that the subscript in $\mathbf{z}_1$ and $\mathbf{z}_2$ is not related to the communication channels of the CC problem.

In order to invoke the ADMM framework [30], let us consider the Augmented Lagrangian (see (3.182)) of (5.27):

$$\mathcal{L}_\lambda(\boldsymbol{\rho}, \mathbf{z}_1, \mathbf{z}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = I_{B_2(\hat{\mathbf{r}}_c,\varepsilon)}(\mathbf{z}_1) + \boldsymbol{\mu}_1^T(\mathbf{A}_c\boldsymbol{\rho} - \mathbf{z}_1) + \frac{\lambda}{2}||\mathbf{A}_c\boldsymbol{\rho} - \mathbf{z}_1||_2^2 + \tag{5.28a}$$

$$I_{\mathbb{R}_+}(\mathbf{z}_2) + ||\mathbf{z}_2||_1 + \boldsymbol{\mu}_2^T(\boldsymbol{\rho} - \mathbf{z}_2) + \frac{\lambda}{2}||\boldsymbol{\rho} - \mathbf{z}_2||_2^2, \tag{5.28b}$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^{M_c^2}$ and $\boldsymbol{\mu}_2 \in \mathbb{R}^N$ are the dual variables (Lagrange multipliers) that correspond to the first and second constraints, respectively, and λ is the penalty parameter of the ADMM. Notice that,

thanks to the change of variables, the optimization of the Augmented Lagrangian involves two well defined subproblems, i.e. the optimizing with respect to $\mathbf{z}_1$ or $\mathbf{z}_2$ can be done independently by ignoring the remaining variable. Following the good practices within the ADMM framework, we reformulate the Augmented Lagrangian in terms of the scaled dual variables (see (3.187)), which are denoted as $\mathbf{u}_i = \frac{\boldsymbol{\mu}_i}{\lambda}$ for $i = 1, 2$. In this way, we further rewrite (5.27) with respect to the scaled Lagrange multipliers as follows [30, Section 3.1.1]:

$$\mathcal{L}_\lambda(\boldsymbol{\rho}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{u}_1, \mathbf{u}_2) = I_{B_2(\hat{\mathbf{r}}_c, \varepsilon)}(\mathbf{z}_1) + \frac{\lambda}{2}||\mathbf{A}_c\boldsymbol{\rho} - \mathbf{z}_1 + \mathbf{u}_1||_2^2 - \frac{\lambda}{2}||\mathbf{u}_1||_2^2 + \tag{5.29a}$$

$$I_{\mathbb{R}_+}(\mathbf{z}_2) + ||\mathbf{z}_2||_1 + \frac{\lambda}{2}||\boldsymbol{\rho} - \mathbf{z}_2 + \mathbf{u}_2||_2^2 - \frac{\lambda}{2}||\mathbf{u}_2||_2^2. \tag{5.29b}$$

The ADMM algorithm for this problem consists in the optimization of $\mathcal{L}_\lambda(\boldsymbol{\rho}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{u}_1, \mathbf{u}_2)$ with respect to each block of variables, i.e. $\boldsymbol{\rho}$, $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{u}_1$, and $\mathbf{u}_2$, fixing the remaining ones. While $\mathcal{L}_\lambda(\boldsymbol{\rho}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{u}_1, \mathbf{u}_2)$ must be minimized with respect to $\boldsymbol{\rho}$, $\mathbf{z}_1$ and $\mathbf{z}_2$ (the primal variables), it must be maximized with respect to $\mathbf{u}_1$ and $\mathbf{u}_2$ (the dual variables). The previous idea results in the following update equations:

$$\boldsymbol{\rho}_{k+1} = \arg\min_{\boldsymbol{\rho}} ||\mathbf{A}_c\boldsymbol{\rho} - \mathbf{z}_{1,k} + \mathbf{u}_{1,k}||_2^2 + ||\boldsymbol{\rho} - \mathbf{z}_{2,k} + \mathbf{u}_{2,k}||_2^2, \tag{5.30a}$$

$$\mathbf{z}_{1,k+1} = \arg\min_{\mathbf{z}} I_{B_2(\hat{\mathbf{r}}_c, \varepsilon)}(\mathbf{z}) + \frac{\lambda}{2}||\mathbf{z} - (\mathbf{A}_c\boldsymbol{\rho}_{k+1} + \mathbf{u}_{1,k})||_2^2, \tag{5.30b}$$

$$\mathbf{z}_{2,k+1} = \arg\min_{\mathbf{z}} I_{\mathbb{R}_+}(\mathbf{z}) + ||\mathbf{z}||_1 + \frac{\lambda}{2}||\mathbf{z} - (\boldsymbol{\rho}_{k+1} + \mathbf{u}_{2_k})||_2^2, \tag{5.30c}$$

$$\mathbf{u}_{1,k+1} = \mathbf{u}_{1,k} + \boldsymbol{\rho}_{k+1} - \mathbf{z}_{1,k+1}, \tag{5.30d}$$

$$\mathbf{u}_{2,k+1} = \mathbf{u}_{2,k} + \mathbf{A}_c\boldsymbol{\rho}_{k+1} - \mathbf{z}_{2,k+1}, \tag{5.30e}$$

where the last two update equations are the dual ascent updates of the dual variables. We detail the solution of (5.30a), (5.30b) and (5.30c) in Appendix 8.3.2, which is a highly recommended reading since the solution of the previous update equations is not detailed in [26]. It is also remarked that it is thanks to the proposed reformulation of the original optimization problem that the resulting algorithm becomes, essentially, the alternation between two proximal operators (see (5.30b), (5.30c) and Definition 3.24). As a summary, the closed-form expressions of the previous update equations are:

$$\boldsymbol{\rho}_{k+1} = \left(\mathbf{A}_c^H\mathbf{A}_c + \mathbf{I}_N\right)^{-1}\left(\mathbf{A}_c^H(\mathbf{z}_{1,k} - \mathbf{u}_{1,k}) + \mathbf{z}_{2,k} - \mathbf{u}_{2,k}\right), \tag{5.31a}$$

$$\mathbf{t}_k = \mathbf{A}_c\boldsymbol{\rho}_{k+1} + \mathbf{u}_{1,k}, \tag{5.31b}$$

$$\mathbf{z}_{1,k+1} = \frac{\sqrt{\varepsilon}}{\max(\sqrt{\varepsilon}, ||\mathbf{t}_k - \hat{\mathbf{r}}_c||_2^2)}(\mathbf{t}_k - \hat{\mathbf{r}}_c) + \hat{\mathbf{r}}_c, \tag{5.31c}$$

$$\mathbf{z}_{2,k+1} = \max\left(\mathbf{0}_N, \text{prox}_{\ell_1, \frac{1}{\lambda}}\left(\Re(\boldsymbol{\rho}_{k+1}) + \mathbf{u}_{2_k}\right)\right), \tag{5.31d}$$

$$\mathbf{u}_{1,k+1} = \mathbf{u}_{1,k} + \boldsymbol{\rho}_{k+1} - \mathbf{z}_{1,k+1}, \tag{5.31e}$$

$$\mathbf{u}_{2,k+1} = \mathbf{u}_{2,k} + \mathbf{A}_c\boldsymbol{\rho}_{k+1} - \mathbf{z}_{2,k+1}, \tag{5.31f}$$

where the closed-form expression of $\text{prox}_{\ell_1, \frac{1}{\lambda}}(\cdot)$ can be found in (3.176). It is worth noting that $\mathbf{A}_c^H\mathbf{A}_c$ is naturally regularized in (5.31a) thanks to the Augmented Lagrangian technique. Thus, the previous methodology bypasses the poor numerical conditioning of $\mathbf{A}_c$.

Regarding the initialization of the proposed algorithm, the ADMM does not require a feasible initialization of the dual and primal variables. Therefore, a simple initialization of the optimization variables can be simply:

$$\mathbf{z}_{1,0} = \mathbf{0}_{M^2}, \tag{5.32a}$$

$$\mathbf{u}_{1,0} = \mathbf{0}_{M^2}, \tag{5.32b}$$

$$\mathbf{z}_{2,0} = \mathbf{0}_N, \tag{5.32c}$$

$$\mathbf{u}_{2,0} = \mathbf{0}_N, \tag{5.32d}$$

which is partly motivated by the running sum of residuals interpretation of the dual variables [30] for $\mathbf{u}_{1,0}$ and $\mathbf{u}_{2,0}$. In principle, one may also choose any alternative initialization and still converge to the optimal solution since everything is convex in this context. Besides, for a better use of computational resources, we also define a stopping criterion. We choose a relative difference between consecutive iterates for simplicity, which is given by the following expression:

$$\frac{||\boldsymbol{\rho}_{k+1} - \boldsymbol{\rho}_k||_2^2}{||\boldsymbol{\rho}_k||_2^2} \leq \xi, \tag{5.33}$$

where $\xi > 0$ is a user defined parameter. The previous stopping criterion is simply a condition that tests the convergence of the ADMM iterative algorithm in terms of the primal variables, which is inspired by (3.93). The ideas detailed in this subsection are summarized in Algorithm 3.

Algorithm 3 Sparse-aware APS estimation based on the ADMM algorithm

Initialization: ξ, I_T, $\lambda > 0$, $\varepsilon > 0$, $\mathbf{u}_{1,0} = \mathbf{0}_{M_c^2}$, $\mathbf{z}_{1,0} = \mathbf{0}_{M_c^2}$, $\mathbf{u}_{2,0} = \mathbf{0}_N$ and $\mathbf{z}_{2,0} = \mathbf{0}_N$.

for $i = 1$ to I_T **do**
  $\boldsymbol{\rho}_{k+1} = \left(\mathbf{A}_c^H \mathbf{A}_c + \mathbf{I}_N\right)^{-1} \left(\mathbf{A}_c^H(\mathbf{z}_{1,k} - \mathbf{u}_{1,k}) + \mathbf{z}_{2,k} - \mathbf{u}_{2,k}\right)$.
  $\mathbf{t}_k = \mathbf{A}_c \boldsymbol{\rho}_{k+1} + \mathbf{u}_{1,k}$.
  $\mathbf{z}_{1,k+1} = \frac{\sqrt{\varepsilon}}{\max(\sqrt{\varepsilon}, ||\mathbf{t}_k - \hat{\mathbf{r}}_c||_2^2)}(\mathbf{t}_k - \hat{\mathbf{r}}_c) + \hat{\mathbf{r}}_c$.
  $\mathbf{z}_{2,k+1} = \max\left(\mathbf{0}_N, \text{prox}_{\ell_1, \frac{1}{\lambda}}\left(\Re(\boldsymbol{\rho}_{k+1}) + \mathbf{u}_{2_k}\right)\right)$.
  $\mathbf{u}_{1,k+1} = \mathbf{u}_{1,k} + \boldsymbol{\rho}_{k+1} - \mathbf{z}_{1,k+1}$.
  $\mathbf{u}_{2,k+1} = \mathbf{u}_{2,k} + \mathbf{A}_c \boldsymbol{\rho}_{k+1} - \mathbf{z}_{2,k+1}$.
  if $\left(\frac{||\boldsymbol{\rho}_{k+1} - \boldsymbol{\rho}_k||_2^2}{||\boldsymbol{\rho}_k||_2^2} \leq \xi\right)$ **then**
    Set $i^* = i$
    Break
  end if
end for
return $\hat{\boldsymbol{\rho}}_{BPD} = \boldsymbol{\rho}_{i^*}$

5.3 Numerical results

The purpose of this section is to evaluate the performance of the sparse-aware estimator of the APS given in Subection 5.2.1. For this purpose, we simulate a scenario where the BS performs the CC procedure to obtain $\mathbf{R}_1$ from $\mathbf{R}_2$. This means that, initially, the BS obtains an estimation of $\mathbf{R}_2$ from K independent uses of channel 2 (see (5.2)). The transmitted symbols through channel 2 are such that:

$$x(n) \sim \mathcal{CN}(0, \sigma^2), \tag{5.34}$$

where σ^2 is a known value that fixes the signal power. The previous assumption implies that the covariance matrix of the received SIMO signal is:

$$\mathrm{E}[\mathbf{y}_2(n)\mathbf{y}_2^H(n)] = \mathrm{E}[|x(n)|^2]\,\mathrm{E}[\mathbf{h}_2(n)\mathbf{h}_2^H(n)] + \mathrm{E}[\mathbf{w}_2(n)\mathbf{w}_2^H(n)] = \sigma^2\mathbf{R}_2 + \mathbf{C}_2, \tag{5.35}$$

where $\mathbf{C}_2$ is the covariance matrix of the noise vector, $\mathbf{w}_2(n)$. For simplicity, we also assume that $\mathbf{C}_2$ is known. Inspired by the previous expression, we consider the following ML estimation of $\mathbf{R}_2$ obtained from the K transmissions:

$$\hat{\mathbf{R}}_{2,K} = \frac{\frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_2(n)\mathbf{y}_2^H(n) - \mathbf{C}_2}{\sigma^2}. \tag{5.36}$$

With the aim of facilitating the analysis of the CC procedure and considering the previous argument, we set $\sigma^2 = 1$ and $\mathbf{C}_2 = \mathbf{I}_M$. Henceforth, we also assume that there are $K = 1000$ training samples, which is the same value as the one considered in [136], to obtain the sample correlation matrix in (5.36).

Besides, we consider an ULA in the BS. This means that the array responses, which are used to perform the conversion steps and to generate the channel covariances, are constructed as in (5.6) with $d = \frac{\lambda_2}{2}$. Additionally, the ratio of the carrier wavelengths, $\gamma = \frac{\lambda_2}{\lambda_1}$, is set to $\gamma = 0.9$ since it is a widely tested value of this figure in the CC literature [98], [136]. The remaining ingredient of the simulation setting is the APS. We describe the APS using the expression given in (5.8) particularized for the Gaussian kernel in (5.9) with the sparse assumption. In this regard, we also consider the same parameters as those that were used to simulate a sparse APS in Figure 5.1 from Subsection 5.1.1. This means that we assume that the number of clusters is $S = 5$ and that the location and scale parameters, θ_s and σ_s, of the Gaussian kernels are uniformly drawn from $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $[\frac{0.3\pi}{180}, \frac{0.8\pi}{180}]$, respectively. The rationale behind drawing θ_s from $[-\frac{\pi}{2}, \frac{\pi}{2}]$ is that these values of θ_s coincide with the relevant angles of the ULAs. Also, we set the values of α_s in (5.8) such that:

$$\sum_{s=1}^{S} \alpha_s = 1. \tag{5.37}$$

As for the performance measure, we consider the normalized Frobenius distance, which is defined as follows.

Definition 5.2 (Normalized Frobenius distance between two matrices)**.** *Let the true value of a channel correlation matrix and an arbitrary estimation of it be denoted as* $\mathbf{R}$ *and* $\hat{\mathbf{R}}$*, respectively. Then, the normalized Frobenius distance between the previous two matrices is defined as:*

$$d_{NF}(\mathbf{R}, \hat{\mathbf{R}}) = \frac{||\mathbf{R} - \hat{\mathbf{R}}||_F^2}{||\mathbf{R}||_F^2}, \tag{5.38}$$

which is equivalent to:

$$d_{NF}(\mathbf{R}, \hat{\mathbf{R}}) = \frac{||\mathbf{r} - \hat{\mathbf{r}}||_2^2}{||\mathbf{r}||_2^2}, \tag{5.39}$$

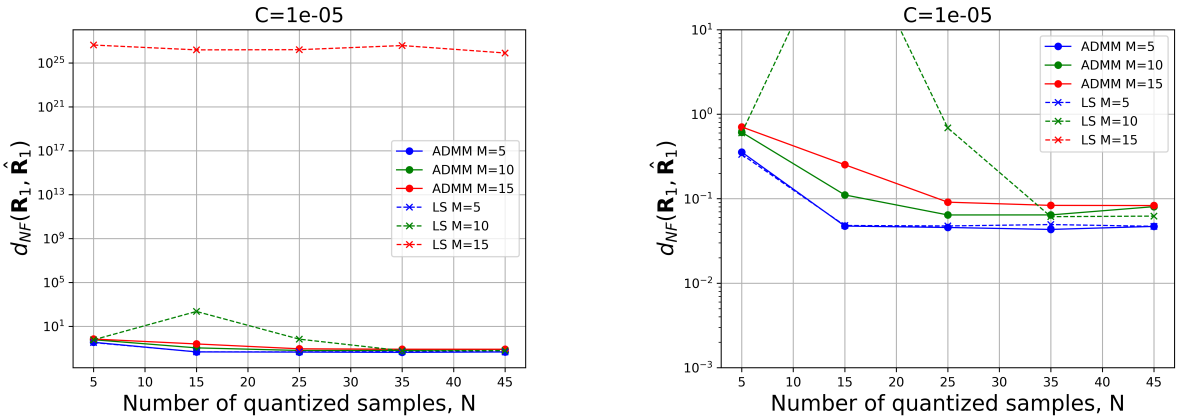
where $\mathbf{r} = \text{vec}(\mathbf{R})$ *and* $\hat{\mathbf{r}} = \text{vec}(\hat{\mathbf{R}})$.

Remark 5.3. In order to obtain the expected value of the previous measure, we resort to $T = 300$ Monte Carlo simulations.

The previous measure of a distance between two matrices is used to assess the performance of the conversion step of the CC framework. In light of this, we use the normalized distance between the tested covariance and the true value of channel covariance matrix of channel 1, $\mathbf{R}_1$. We mainly consider three different alternatives, in addition to our proposed solution from Algorithm 3, to obtain $\mathbf{R}_1$ from $\hat{\mathbf{R}}_2$. As a reference of the difficulty of the CC problem, we consider the naive utilization of $\hat{\mathbf{R}}_{2,K}$, which follows from the rationale shown in Figure 5.2 (see Subsection 5.1.2). Secondly, we resort to the Hilbert projection approach from [136], [137] as a state-of-the-art method of the CC problem. The motivation behind choosing the Hilbert projection approach as a comparison is that it is the other known solution to the CC problem that solves it by implicitly estimating the APS. Thirdly, we consider $\hat{\mathbf{R}}_{1,K}$ as the *optimal* approach to the CC problem. As for the sparse-aware APS estimator, in order to simplify the practical implementation of the ADMM-based algorithm for the following numerical experiments, we consider that ε (see (5.23) and Algorithm 3) has the following expression:

$$\varepsilon = C||\hat{\mathbf{r}}_c||_2^2, \tag{5.40}$$

where C is any positive constant. The motivation behind (5.40) is to yield a constraint in (5.23) that is also a normalized Frobenius distance. In this way, it is easier to choose a value for C that has an intuitive meaning, i.e. the allowed relative error in the APS estimation. Additionally, we consider that $\lambda = 5$ in Algorithm 3.



(a) Without zoom.

(b) Zoom.

Figure 5.3: Normalized Frobenius norm, $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$, as a function of the number of quantized samples.

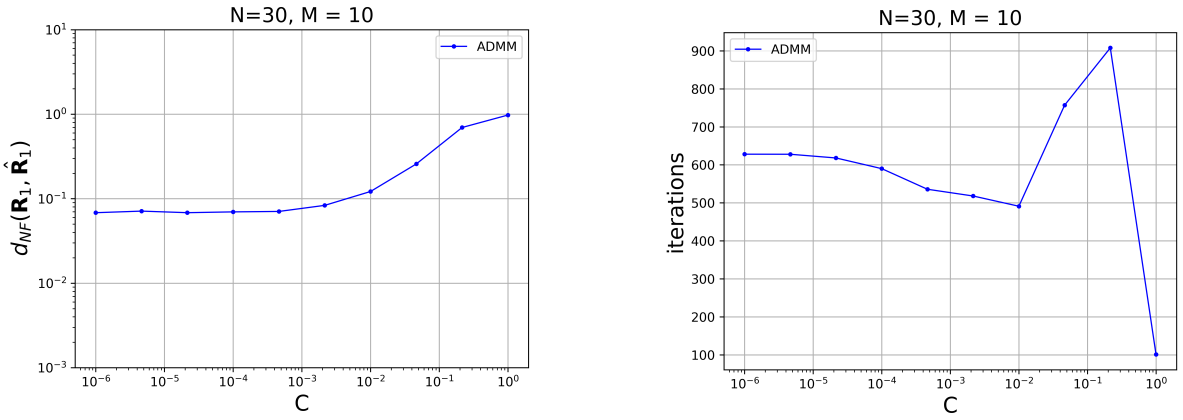
5.3.1 Testing the number of quantized samples

The purpose of this numerical experiment is to characterize the behavior of $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$ with respect to N. Particularly, we want to provide an intuitive rule of thumb for fixing N in Algorithm 3. In Figure 5.3, we show the normalized Frobenius distance between the converted channel covariance and its true value for several values of M (number of BS antennas) for the ADMM-based and the LS (see (5.20)) solutions. Indeed, the LS solution is considered in this numerical experiment for illustration purposes.

While it is clear that there is a minimum value of N such that the ADMM-based solver yields its best performance, e.g. $N = 15$ is sufficient for $M = 5$, and $N \approx 30$ for $M = 10$, there are some cases where the LS solution diverges. For instance, although the LS seems to converge for a certain amount of quantized samples when the BS antenna array has a small number of antennas, it diverges in the large antenna array case for all values of N as shown in Subfigure 5.3a. This is the reason why we discourage the use of the LS solution in the subsequent numerical experiments. On the other hand, with the exception of the increased computational complexity, choosing a sufficiently large number of quantized samples, N, ensures the optimal performance of the ADMM-based APS estimator. Indeed, provided that we only consider BS antenna arrays that have at most 20 elements in the subsequent numerical experiments, setting $N = 50$ is a reasonably large value for the quantization operation.

5.3.2 Testing the fitting constraint parameter

In a similar manner to the previous numerical experiment, we are interested in a procedure to determine the user-defined parameter that regulates the fitting constraint in (5.23), i.e. ε. In light of this goal, we depict the performance of Algorithm 3 in Figure 5.4 with respect to the fitting constraint, ε, for fixed values of N and M. The main conclusion of Subfigure 5.4a is that, for a given scenario, there is a minimum value of C (see (5.40)) that ensures the optimal performance of Algorithm 3. Yet, it is important to remark that, although small values of C may result in an unfeasible convex program in (5.23), the proposed algorithm is capable to obtain the closest optimal solution, which is a valuable behavior in a practical setting. On the other perspective, Subfigure 5.4b suggests that one may fine-tune C to yield a better computational complexity. On this matter, it is seen in the previously mentioned figure that the optimal amount of iterations is obtained for $C = 10^{-2}$, yielding a reasonable performance. Yet, we recommend to fix C to a small value since the potential complexity gain does not compensate the loss in performance.

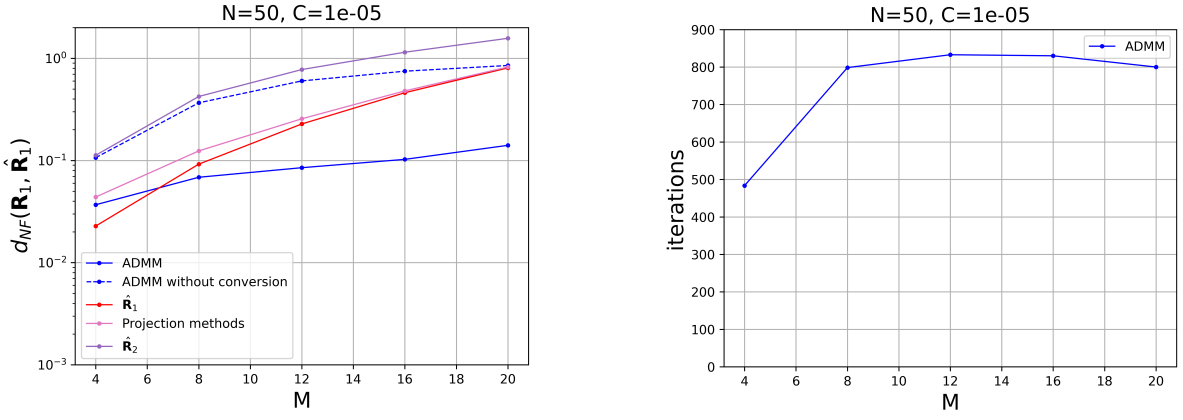


(a) $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$. **(b)** Iterations.

Figure 5.4: Normalized Frobenius norm, $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$, as a function of ε (see (5.40)).

5.3.3 Performance of the CC approaches with an increasing number of antennas

While the previous two subsections were devoted to the comprehension of the parameters that regulate the performance and complexity of Algorithm 3, in this subsection we focus our analysis on the comparison with the aforementioned three approaches to the CC problem, which are the projection methods approach from [136], [137] and the sample covariances of each channel, $\hat{\mathbf{R}}_{1,K}$ and $\hat{\mathbf{R}}_{2,K}$ (see (5.36)). In this case, $\hat{\mathbf{R}}_{1,K}$ is obtained using (5.36) by K independent transmissions from the UE to the BS (see (5.2)) using the carrier frequency of channel 1. With the aim of comparing the aforementioned approaches with our proposed solution in Algorithm 3, we simulate different communications channels where the number of antenna arrays on the BS is modified. In this manner, we can assess the performance of all the considered approaches to the CC problem as a function of M. The aforementioned numerical experiment is depicted in Figure 5.5, where we fixed the values of N and C to $N = 50$ and $C = 10^{-5}$ in light of the analysis shown in the previous two subsections.



(a) $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$. **(b)** Iterations.

Figure 5.5: Normalized Frobenius norm, $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$, as a function of M for different CC approaches.

The main conclusion of this numerical experiment is that our proposed sparse-aware solution ("ADMM" in these figures) is the better performing algorithm for the CC problem, especially in the medium and large antenna array regime ($M \geq 8$). Interestingly, the ADMM-based covariance conversion

also surpasses the channel 1 sample covariance estimator. The previous fact can be explained by two ideas. Not only the simulated scenario is aligned to the implicit assumptions of the sparse-aware APS estimation, i.e. the APS satisfies the properties described in Subsection 5.1.1, but the conversion step of the sparse-aware CC approach can be understood as a parametric estimation, which comes in contraposition to the non-parametric nature of channel 1 sample covariance estimation, $\hat{\mathbf{R}}_{1,K}$. It is worth noting that a parametric estimator is capable of outperforming a non-parametric estimator as long as the data is compatible with the implicit assumptions of the parametric model [212]. In this case, the data satisfies the implicit assumptions of the sparse-aware APS estimation. The latter ideas are further confirmed by the sparse-aware approach without the conversion step ("ADMM without conversion" in Subfigure 5.5a). Yet, the price to pay for the increased performance is an increased computational complexity as compared to the remaining methods. This fact is observed in Subfigure 5.5b, where the number of iterations for each value of M of the ADMM-based algorithm is depicted. It is seen in the aforementioned subfigure that the computational complexity is high due to the increased number of antennas, whose contribution is in the order of M^2 (see (5.17)), and to the elevated number of iterations, which increases with M until $M \approx 8$. The increased overall complexity is the main drawback of our proposed sparse-aware CC procedure since the main bulk of operations is performed during the conversion step. In contrast, the computational cost of the Hilbert space projection method [136], [137] is concentrated on the precomputation (integrals) of the change of basis matrices in the Hilbert space defined by the array responses.

5.4 Concluding remarks

This chapter, in conjunction to the previous one, completes the exploration of the diversity in signal processing applications that is considered within the context of this dissertation. Regarding this topic, while the algorithmic perspective on diversity found in Chapter 4 was known in signal processing, the one explored in this chapter, which is depicted by the reformulation of the angular diversity given in (5.17), is new (up to the authors knowledge). In fact, this expression of diversity, which consists of two different system of equations that share the same coefficients and not the matrix coefficients, is relevant within the context of the CC methodology since both system of equations are needed for the conversion steps. We are aware that the previous phenomenon may be seen as *niche* by other practitioners.

The common factor between this chapter and Chapter 4 is that we assumed that the considered datasets had an implicit diversity. In order to provide the complete picture of the information fusion problem, the next chapter studies an algorithm that quantifies the amount of diversity between two random Gaussian vectors, completing the cycle of this dissertation.

Chapter 6

Discovering diversity via model-order selection rules

In the previous two chapters, we explored two signal processing applications where the diversity present in the data is exploited. Indeed, the implicit assumption we made in both problems surveyed in chapters 4 and 5 was that the considered datasets had some sort of diversity. However, this is not always the case in practical scenarios. For the previous reason, there is often a need to assess whether any two datasets have some sort of relationship. Motivated by this casuistic, the goal of this chapter is to study the estimation of the degree of relationship between two datasets. An example of this kind of measure is considered in [159] to detect whether two M-dimensional random vectors are correlated, being a fundamental problem in multivariate statistical analysis [25], [129].

As an alternative approach and inspired by the information theoretic paradigm of this dissertation, we propose to estimate the MI between two datasets, consisting of two random Gaussian vectors (for simplicity), as a way to quantify the amount of diversity shared between those datasets. The motivation behind the consideration of MI is two fold. Firstly, the MI is a measure that is invariant to homeomorphisms of the involved random variables (see Lemma 2.3), which is a property that we leverage in this chapter to transform a general problem into a simpler one. Also, the MI is the underlying principle of the *information theoretic coherence* [157, Section 11.4]. As shown in a later paragraph, this measure is much more suited to the task of quantifying diversity than the MI itself. This measure is defined as follows:

$$\rho_{\mathrm{IT}}(X,Y) = \sqrt{1-\exp(-I(X;Y))}, \tag{6.1}$$

where X and Y are the tested variables and $I(X;Y)$ is their mutual information. It is easy to show that the value of $\rho_{\mathrm{IT}}(X,Y)$ is more interpretable in a wider range of scenarios than the MI. Its better interpretability comes from the fact that the information theoretic coherence is bounded in $[0,1]$ as opposed to the MI, which is unbounded. In a similar fashion to the ideas presented in Section 4.3, a bounded descriptor is preferred for a better numerical conditioning of the resulting algorithms (infinites cannot implemented with ease). In addition, the limiting values of $\rho_{\mathrm{IT}}(X,Y)$ are naturally related to the diversity quantification task. For instance, we know that $I(X;Y) \to \infty$ implies a functional relationship between X and Y. Accordingly, $\rho_{\mathrm{IT}}(X,Y)$ is equal to 1 in this case. In contrast, independent random variables result in $I(X;Y) = 0$, which ensures that $\rho_{\mathrm{IT}}(X,Y) = 0$. For Gaussian random vectors, $\rho_{\mathrm{IT}}(X,Y)$ coincides with the Pearson correlation coefficient of X and Y, implying that the MI is closely related to the Locally Most Powerful Invariant Test (LMPIT) of the correlation detection problem [159].

Besides, one of the problems to be faced when working with large datasets, i.e. high-dimensional data, is the fact that only few components are correlated in most practical situations. The necessity of detecting the presence of a sparse correlated subset of components emerges naturally in numerous scenarios (see [10], [158] and references therein for a motivation). In other words, high-dimensional data tend to exhibit a low-rank structure, in the sense of limited number of dependent nodes, regardless

of the application. In the particular case of the MI estimation, it is shown in this chapter that the ML estimation of the MI suffers from an additional bias coming from unwanted contributions of the non-correlated components in the low-rank scenario. In order to mitigate this issue, we incorporate the ideas presented in Section 2.3 from Chapter 2 to regularize the MI estimation.

The ideas presented in this Chapter are summarized in our published work [120] and are structured as follows. In Section 6.1, we describe the problem of estimating the MI of two sequences and, more specifically, we state the nominal conditions of this problem. Next, the details of the ML estimation of the MI are surveyed in Section 6.2. Finally, our proposed regularization approach to the MI estimation problem is detailed in Section 6.4 and tested in Section 6.5.

6.1 Problem statement: Mutual Information of two sequences

Let us consider N independent and identically distributed samples of M pairs of zero-mean Gaussian sequences and let us denote the sequences that constitute the m-th pair as $x_m(n)$ and $y_m(n)$ with $m = 1, ..., M$ and $n = 1, ..., N$. Provided that we want to quantify the degree of dependence between $\mathbf{x}(n)$ and $\mathbf{y}(n)$, where $[\mathbf{x}(n)]_m = x_m(n)$ and $[\mathbf{y}(n)]_m = y_m(n)$ for $m = 1, ..., M$, the Pearson correlation coefficient (or, simply, the correlation coefficient) between $x_m(n)$ and $y_m(n)$ is given by:

$$\rho_m = \frac{\mathrm{E}[x_m(n)(y_m(n))]}{\sqrt{\mathrm{E}[|x_m(n)|^2]\,\mathrm{E}[|y_m(n)|^2]}}, \tag{6.2}$$

which is such that $-1 \leq \rho_m \leq 1$ due to the Cauchy-Schwarz inequality. Note that the previous expression is only valid for zero-mean sequences. For initial simplicity, we assume that all sequences are mutually independent and that $x_m(n)$ and $y_m(n)$ are Gaussian random variables with unit-variance. The mutual independence of the sequences implies that the cross-covariance between $\mathbf{x}(n)$ and $\mathbf{y}(n)$, which is given by (zero-mean sequences):

$$\mathrm{E}[\mathbf{x}(n)\mathbf{y}^H(n)] = \mathrm{E}[\mathbf{y}(n)\mathbf{x}^H(n)] = \mathbf{C}_{xy}, \tag{6.3}$$

is diagonal. There are known scenarios that are modeled by mutually independent sequences. To name a few examples, any application that considers parallel Gaussian channels [38], [120] implicitly assumes mutual independence. This assumption is also found in the analysis of the spectra of two random stationary processes since different periodogram frequencies act as independent random variables (see [64], [198] for examples). Nevertheless, while the independence and unit-variance assumptions may seem too restrictive at first, we prove in Section 6.4 that any setting that considers general variances and/or non-independent pairs of sequences can be mapped to the previously depicted scenario.

In light of the parallel Gaussian channels interpretation, it is convenient to gather $x_m(n)$ and $y_m(n)$ as follows:

$$\mathbf{z}_m(n) = \begin{bmatrix} x_m(n) \\ y_m(n) \end{bmatrix}. \tag{6.4}$$

In this manner, we can model the Gaussian channel between $x_m(n)$ and $y_m(n)$ using $\mathbf{z}_m(n)$. Henceforth, we refer to the m-th pair of sequences gathered in $\mathbf{z}_m(n)$ as the m-th Gaussian channel. With the considered assumptions, $\mathbf{z}_m(n)$ is statistically distributed as:

$$\mathbf{z}_m(n) \sim \mathcal{N}(\mathbf{0}_2, \mathbf{C}_m), \tag{6.5}$$

where:

$$\mathbf{C}_m = \begin{bmatrix} 1 & \rho_m \\ \rho_m & 1 \end{bmatrix}. \tag{6.6}$$

The mutual independence between pairs of sequences implies that $\mathbf{z}_m(n)$ and $\mathbf{z}_{m'}(n)$ are independent for $m \neq m'$. Now, we focus on the problem of discovering the diversity that is present in the joint dataset described by $\mathbf{x}(n)$ and $\mathbf{y}(n)$. Although the correlation coefficients given in (6.2) can be used to obtain a measure of the correlation between $\mathbf{x}(n)$ and $\mathbf{y}(n)$ (for example, the sum of the squared coefficients [159]), we resort to the MI of $\mathbf{x}(n)$ and $\mathbf{y}(n)$ as a soft measure of the dependence of these two vectors, which is mainly motivated by (6.1). The following lemma revisits the MI of two Gaussian random variables.

Lemma 6.1 (MI of two Gaussian random variables)**.** *Let X and Y be two zero-mean Gaussian random variables, whose correlation coefficient is given by ρ (see* (6.2)*). Then their MI is given by:*

$$I(X;Y) = -\frac{1}{2}\log(1-\rho^2), \tag{6.7}$$

Remark 6.1. The expression in (6.7) can be rewritten with respect to the *coherence matrix* of the random vector that gathers X and Y (see (6.4)). By denoting $\mathbf{z} = [X, Y]^T$, the desired expression is:

$$I(X;Y) = -\frac{1}{2}\log\left(\det\left(\boldsymbol{\Xi}\right)\right), \tag{6.8}$$

where:

$$\boldsymbol{\Xi} = \mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{D}^{-\frac{1}{2}}, \tag{6.9}$$

is the coherence matrix associated to X and Y and:

$$\mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right] = \mathbf{C}, \tag{6.10a}$$

$$\mathbf{D} = \mathbf{C} \odot \mathbf{I_2}, \tag{6.10b}$$

are the covariance matrix of $\mathbf{z}$ and the diagonal matrix containing the non-zero variances of X and Y, respectively. Note that $\mathbf{D} = \mathbf{I}_2$ for unit-variance random variables, so their coherence matrix is equivalent to their covariance matrix. Expressions (6.7) and (6.8) are proven in Appendix 8.4.1.

The previous lemma, in addition to the mutual independence of the pairs of random variables, results in the fact that the overall MI is given by the sum of the pairwise MI:

$$I(\mathbf{x}(n);\mathbf{y}(n)) = \sum_{m=1}^{M} I(x_m(n);y_m(n)) = -\frac{1}{2}\sum_{m=1}^{M}\log(1-\rho_m^2). \tag{6.11}$$

Clearly, the motivation behind the assumption of mutual independence is that the overall MI results in a compact and intuitive expression. For the purpose of contextualizing the estimation of the MI problem into the sparse-aware methodology, we consider the additional prior knowledge that some of the parallel channels provide no-information. This means that the correlation coefficients of these channels are equal to 0. In order to formally introduce the previous idea, let us denote the set of indices of all the channels as follows:

$$\mathcal{S}_M = \{m \in \mathbb{N} : 1 \leq m \leq M\}. \tag{6.12}$$

Accordingly, the set of active channels, being the ones that contribute with non-zero information, is depicted as follows:

$$\mathcal{S}_D = \{d \in \mathcal{S}_M : |\rho_d| > 0\}, \tag{6.13}$$

where $\mathrm{card}(\mathcal{S}_D) = D < M$ is the total number of active channels. This assumption implies that the diagonal cross-correlation matrix in (6.3) has only D non-zero diagonal elements. The motivation behind the consideration of a sparse set of active channels is two fold. On the one hand, one of the problems that arises when the available data is high-dimensional is that only a few number of components are correlated among the large amount of parallel virtual channels. In essence, any high-dimensional dataset tends to exhibit a low-rank structure irrespective of the particular application (see [10], [158] and references therein). On the other hand, we want to obtain a consistent estimator of the MI as both $N \to \infty$ and $M \to \infty$. This issue is critical since, as shown in Section 6.3, the Maximum Likelihood MI (ML-MI) estimator is naive in the sense that it is biased due to unwanted contributions corresponding to samples from inactive channels. To this end, the sparse assumption on the active channels is a natural way to regularize the number of parameters to be estimated. Hence, the effects of overfitting are mitigated.

Taking into account the depicted scenario, the objective of the remainder of this chapter is to estimate $I(\mathbf{x}(n);\mathbf{y}(n))$ from the available data under the assumption that there exists a sparse set of active channels. Henceforth, we gather the data in the following matrices:

$$\mathbf{X} = [\mathbf{x}(1), ..., \mathbf{x}(N)], \tag{6.14a}$$

$$\mathbf{Y} = [\mathbf{y}(1), ..., \mathbf{y}(N)]. \tag{6.14b}$$

Notice that, provided that the proposed estimation problem is founded in (6.11), we are considering a parametric estimation of the MI, which is motivated by the same arguments as those that promote the use of the PMEE (see Definition 2.5).

6.2 Maximum Likelihood estimation of the Mutual Information

In this section, we show the preliminaries of the ML-MI estimator. In particular, we revisit the CML principle (see Definition 4.10) to estimate the Pearson correlation coefficient of the M independent Gaussian channels with the prior assumption that some of them are equal to zero, i.e. their contribution to the overall MI is equal to 0. We show at the end of this section that the ML-MI estimator is obtained in a natural manner from the CML function. With the previous goal mind, let us consider the log-likelihood of the multichannel dataset described by (6.14) and (6.5):

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}|\{\mathbf{C}_m\}_{m\in\mathcal{S}_M}) = \sum_{m=1}^{M} \ell_{\mathbf{Z}_m}(\mathbf{Z}_m|\mathbf{C}_m), \tag{6.15}$$

where:

$$\mathbf{C}_m = \mathrm{E}[\mathbf{z}_m(n)\mathbf{z}_m^T(n)], \tag{6.16a}$$

$$\mathbf{Z}_m = [\mathbf{z}_m(1), ..., \mathbf{z}_m(N)] \quad m = 1, ..., M, \tag{6.16b}$$

and $\ell_{\mathbf{Z}_m}(\mathbf{Z}_m|\mathbf{C}_m)$ is the log-likelihood function associated to the m-th virtual channel. The expanded expression of $\ell_{\mathbf{Z}_m}(\mathbf{Z}_m|\mathbf{C}_m)$, after ignoring additive constants that do not depend on $\mathbf{Z}_m$ or $\mathbf{C}_m$, is:

$$\ell_{\mathbf{Z}_m}(\mathbf{Z}_m|\mathbf{C}_m) = -\frac{1}{2}\sum_{n=1}^{N}\left(\log(\det(\mathbf{C}_m)) + \mathbf{z}_m^T(n)\mathbf{C}_m^{-1}\mathbf{z}_m(n)\right). \tag{6.17}$$

For convenience, we rewrite the previous function as follows:

$$\ell_{\mathbf{Z}_m}(\mathbf{Z}_m|\mathbf{C}_m) = -\frac{N}{2}\left(\log(\det(\mathbf{C}_m)) + \mathrm{tr}(\mathbf{C}_m^{-1}\hat{\mathbf{C}}_m)\right), \tag{6.18}$$

where $\hat{\mathbf{C}}_m$ is the m-th channel sample covariance matrix:

$$\hat{\mathbf{C}}_m = \frac{1}{N}\mathbf{Z}_m\mathbf{Z}_m^T. \tag{6.19}$$

Following the same ideas that yield (4.103c) in Chapter 4, the sample covariance matrix in (6.19) is known to be the maximizer of (6.18) for the active channels. In contrast, the ML estimator of $\mathbf{C}_m$ for $m \notin \mathcal{S}_D$ is:

$$\hat{\mathbf{D}}_m = \frac{1}{N}\hat{\mathbf{R}}_m \odot \mathbf{I}_2 = \frac{1}{N}\mathbf{Z}_m\mathbf{Z}_m^T \odot \mathbf{I}_2, \tag{6.20}$$

since the covariance matrix is diagonal (see (6.6) for $\rho_m = 0$) for the channels that do not contribute to the overall MI. In fact, the optimality of (6.20) is obtained from the same arguments as those that yield (4.131) in Chapter 4. The next step is to compress $\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}|\{\mathbf{C}_m\}_{m\in\mathcal{S}_M})$ by plugging into this function the estimators given in (6.19) and (6.20). This procedure results in the following CML function:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X}, \mathbf{Y}|\mathcal{S}_D) = -\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\left(\log(\det(\hat{\mathbf{C}}_d)) + \mathrm{tr}(\hat{\mathbf{C}}_d^{-1}\hat{\mathbf{C}}_d)\right) + \sum_{m\notin\mathcal{S}_D}\left(\log(\det(\hat{\mathbf{D}}_m)) + \mathrm{tr}(\hat{\mathbf{D}}_m^{-1}\hat{\mathbf{C}}_m)\right)\right), \tag{6.21}$$

where we introduced $\mathcal{S}_D$ instead of $\{\mathbf{C}_m\}_{m\in\mathcal{S}_M}$ to economize the notation and to remark the sparse assumption on the active channels. The final step to retrieve the ML-MI estimator is to rewrite (6.21) with respect to the coherence matrices of the active channels. In this regard, note that some of the terms in (6.21) yield:

$$\mathrm{tr}(\hat{\mathbf{C}}_d^{-1}\hat{\mathbf{C}}_d) = 2 \quad d \in \mathcal{S}_D, \tag{6.22a}$$

$$\mathrm{tr}(\hat{\mathbf{D}}_m^{-1}\hat{\mathbf{C}}_m) = \mathrm{tr}(\hat{\mathbf{D}}_m^{-\frac{1}{2}}\hat{\mathbf{C}}_m\hat{\mathbf{D}}_m^{-\frac{1}{2}}) = \mathrm{tr}(\hat{\mathbf{\Xi}}_m) = \mathrm{tr}\left(\begin{bmatrix} 1 & \hat{\rho}_m \\ \hat{\rho}_m & 1 \end{bmatrix}\right) = 2 \quad m \notin \mathcal{S}_D, \tag{6.22b}$$

where $\hat{\mathbf{\Xi}}_m = \hat{\mathbf{D}}_m^{-\frac{1}{2}}\hat{\mathbf{C}}_m\hat{\mathbf{D}}_m^{-\frac{1}{2}}$ and $\hat{\rho}_m$ are the estimated coherence matrix, which is referred to as the sample coherence matrix from now on, and the estimated Pearson correlation coefficient of the m-th channel, respectively. Taking into account the previous two equations, (6.21) becomes:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D) = -\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\log(\det(\hat{\mathbf{C}}_d)) + 2D + \sum_{m\notin\mathcal{S}_D}\log(\det(\hat{\mathbf{D}}_m)) + 2(M-D)\right) = \tag{6.23a}$$

$$-\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\log(\det(\hat{\mathbf{C}}_d)) + \sum_{m\notin\mathcal{S}_D}\log(\det(\hat{\mathbf{D}}_m)) + 2M\right). \tag{6.23b}$$

It is possible to further parse the previous expression after noting that:

$$\det(\hat{\mathbf{C}}_d) = \det(\hat{\mathbf{D}}_d^{\frac{1}{2}}\hat{\mathbf{\Xi}}_d\hat{\mathbf{D}}_d^{\frac{1}{2}}) = \det(\hat{\mathbf{D}}_d\hat{\mathbf{\Xi}}_d). \tag{6.24}$$

Then:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D) = -\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\log(\det(\hat{\mathbf{D}}_d\hat{\mathbf{\Xi}}_d) + \sum_{m\notin\mathcal{S}_D}\log(\det(\hat{\mathbf{D}}_m)) + 2M\right) = \tag{6.25a}$$

$$-\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\Big(\log(\det(\hat{\mathbf{\Xi}}_d) + \log(\det(\hat{\mathbf{D}}_d))\Big) + \sum_{m\notin\mathcal{S}_D}\log(\det(\hat{\mathbf{D}}_m)) + 2M\right) = \tag{6.25b}$$

$$-\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\log(\det(\hat{\mathbf{\Xi}}_d)) + \sum_{m=1}^{M}\log(\det(\hat{\mathbf{D}}_m)) + 2M\right). \tag{6.25c}$$

Finally, we get a much clearer expression of the CML function after gathering the additive constants that do not depend on D nor on the estimated correlation coefficients of the active channels:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D) = -\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\log(\det(\hat{\mathbf{\Xi}}_d)) + \text{constants}\right) = -\frac{N}{2}\left(\sum_{d\in\mathcal{S}_D}\log(1-\hat{\rho}_d^2) + \text{constants}\right). \tag{6.26}$$

After a careful comparison of (6.11) and (6.26), we rewrite the CML function as follows:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D) = N\hat{I}_{ML}(\mathbf{x};\mathbf{y}|\mathcal{S}_D) + \text{constants}, \tag{6.27}$$

where:

$$\hat{I}_{ML}(\mathbf{x};\mathbf{y}|\mathcal{S}_D) = -\frac{1}{2}\sum_{d\in\mathcal{S}_D}\log(\det(\hat{\mathbf{\Xi}}_d)) = -\frac{1}{2}\sum_{d\in\mathcal{S}_D}\log(1-\hat{\rho}_d^2), \tag{6.28}$$

is the ML-MI estimator between $\mathbf{x}(n)$ and $\mathbf{y}(n)$ with the sparse assumption on the Gaussian channels. The main issue with the previous estimator is that the indices corresponding to the active channels, $\mathcal{S}_D$, are not known. For this reason, we incorporate the model-order selection framework from Section 2.3 in Chapter 2 to determine the active channels. In fact, the expression given in (6.27) is a fundamental step that has to be done before the introduction of the information theoretic model-order selection rules. Note that, up to this point, we have only used the prior information that the channels are mutually independent (we have not yet introduced the unit-variance assumption).

6.3 Regularized mutual information estimation via model-order selection

With the objective of introducing the general information theoretic model-order optimization problem from (2.83) in Chapter 2, we have to assess whether the depicted MI estimation setting satisfies the conditions in which (2.83) is valid. Particularly, we have to verify the constraints on the Fisher information matrix specified in Section 2.3 with respect to the vector of parameters, which contains the correlation coefficients of the active channels in this case. In essence, the Fisher information associated to the correlation coefficients must be non-singular and must satisfy (2.102) (only for the BIC rule). For clarity in the exposition, we consider that the set of active channels is such that:

$$\mathcal{S}_D = \{d \in \mathbb{N} : 1 \leq d \leq D\}, \tag{6.29}$$

and that the correlation coefficients are gathered in the following vector:

$$\boldsymbol{\rho}_D = \begin{bmatrix} \rho_1 \\ ... \\ \rho_D \end{bmatrix}. \tag{6.30}$$

With the aim of obtaining the Fisher information matrix, we now enforce the unit-variance constraint of the random variables. With the unit-variance assumption, and after ignoring additive constants that do not depend on $\boldsymbol{\rho}_D$ in (6.15) (see also (6.18)), the joint log-likelihood function of the M independent channels is written as follows:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D) = -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\log(\det(\mathbf{C}_m)) + \mathbf{z}_m^T(n)\mathbf{C}_m^{-1}\mathbf{z}_m(n)\right), \tag{6.31}$$

which can be further expanded after noting that:

$$\mathbf{C}_d^{-1} = \begin{bmatrix} 1 & \rho_d \\ \rho_d & 1 \end{bmatrix}^{-1} = \frac{1}{1-\rho_d^2}\begin{bmatrix} 1 & -\rho_d \\ -\rho_d & 1 \end{bmatrix} \quad d \in \mathcal{S}_D, \tag{6.32a}$$

$$\mathbf{C}_m^{-1} = \mathbf{I}_2 \quad m \notin \mathcal{S}_D. \tag{6.32b}$$

The previous particularization of the covariance matrices implies that, under the nominal conditions, the expanded expression of the joint log-likelihood of the M independent channels is:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D) = -\frac{1}{2}\sum_{d\in\mathcal{S}_D}\sum_{n=1}^{N}\left(\log(1-\rho_d^2) + \frac{1}{1-\rho_d^2}\left(x_d^2(n) + y_d^2(n) - 2\rho_d x_d(n) y_d(n)\right)\right) - \frac{1}{2}\sum_{m\notin\mathcal{S}_D}\sum_{n=1}^{N}(x_m^2(n) + y_m^2(n)), \tag{6.33}$$

Notice that the unit-variance assumption results in an expression of the log-likelihood function of the Gaussian channels that is completely parameterized by $\boldsymbol{\rho}_D$. This property results in a simple derivation (and assessment) of the Fisher information matrix of the problem, being the main motivation behind the consideration of the unit-variance assumption. Fortunately, this assumption is proven to be non-restrictive in Section 6.4. It is important to remark that, if general variances were considered, they would need to be accounted in the Fisher information as additional parameters. Thus, this case would result in a more complicated Fisher matrix.

Invoking Definition 2.15, the Fisher information matrix with respect to $\boldsymbol{\rho}_D$ is constructed using the second-order derivatives of $\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D)$. In this regard, the second order cross-derivatives of (6.33) are:

$$\frac{\partial^2 \ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\rho}_D)}{\partial\rho_d \partial\rho_{d'}} = 0, \tag{6.34}$$

for $d \neq d'$. The previous expression follows from the fact that (6.33) is separable in terms of the correlation coefficients. Moreover, the second derivative with respect to ρ_d is given by the following expression:

$$\frac{\partial^2 \ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_D)}{(\partial \rho_d)^2} = -N\left(\frac{1+3\rho_d^2}{(1-\rho_d^2)^3}(\hat{v}_{x,d}+\hat{v}_{y,d}) - \frac{2\rho_d(3+\rho_d^2)}{(1-\rho_d^2)^3}\hat{\rho}_d - \frac{1+\rho_d^2}{(1-\rho_d^2)^2}\right), \tag{6.35}$$

where:

$$\hat{v}_{x,d} = \frac{1}{N}\sum_{n=1}^{N} x_d^2(n), \tag{6.36a}$$

$$\hat{v}_{y,d} = \frac{1}{N}\sum_{n=1}^{N} y_d^2(n), \tag{6.36b}$$

$$\hat{\rho}_d = \frac{1}{N}\sum_{n=1}^{N} x_d(n)y_d(n), \tag{6.36c}$$

are the sample estimations of the variance of $x_d(n)$ and $y_d(n)$, and of their correlation coefficient, respectively. Taking into account that (recall the zero-mean and unit-variance assumptions):

$$\hat{v}_{x,d} = \frac{1}{N}\,\mathrm{E}\left[\sum_{n=1}^{N} x_d^2(n)\right] = 1, \tag{6.37a}$$

$$\mathrm{E}[\hat{v}_{y,d}] = \frac{1}{N}\,\mathrm{E}\left[\sum_{n=1}^{N} y_d^2(n)\right] = 1, \tag{6.37b}$$

$$\hat{\rho}_d = \frac{1}{N}\,\mathrm{E}\left[\sum_{n=1}^{N} x_d(n)y_d(n)\right] = \rho_d, \tag{6.37c}$$

the expected value of (6.35) is simply:

$$\mathrm{E}\left[\frac{\partial^2 \ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\rho}_D)}{(\partial \rho_d)^2}\right] = -N\frac{(1+\rho_d^2)}{(1-\rho_d^2)^2}. \tag{6.38}$$

In light of (6.34) and (6.35), we get that the Fisher information matrix with respect to $\boldsymbol{\rho}_D$ is diagonal, whose diagonal entries are:

$$[\mathbf{F}(\boldsymbol{\rho}_D)]_{d,d} = -\,\mathrm{E}\left[\frac{\partial^2 \ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\rho}_D)}{(\partial \rho_d)^2}\right] = N\frac{(1+\rho_d^2)}{(1-\rho_d^2)^2}. \tag{6.39}$$

Note that the previous expression of the Fisher information is intuitive. On the one hand, the diagonality of the Fisher information matrix is due to the statistical independence of the Gaussian channels. On the other hand, insightful observations are retrieved from the CRLB interpretation of the Fisher information matrix. By noting that the CRLB of each entry of $\boldsymbol{\rho}_D$ is given by the d-th diagonal entry of the Fisher information matrix inverse:

$$\mathrm{E}\left[|\rho_d - \hat{\rho}_d|^2\right] \geq [\mathbf{F}^{-1}(\boldsymbol{\rho}_D)]_{d,d} = \frac{(1-\rho_d^2)^2}{N(1+\rho_d^2)}, \tag{6.40}$$

we get that:

$$[\mathbf{F}^{-1}(\boldsymbol{\rho}_D)]_{d,d} \to 0, \tag{6.41}$$

as $N \to \infty$ and as $|\rho_d| \to 1$. While the first condition ($N \to \infty$) for (6.41) implies that any efficient estimator of ρ_d is consistent such as the one from (6.36c), the second one ($|\rho_d| \to 1$) is not surprising as this would mean that $y_d(n) = \pm x_d(n)$. Thus, no estimator would be needed to infer that $\rho_d = \pm 1$ when $y_d(n) = \pm x_d(n)$.

From (6.39), we can assess whether the information matrix of this problem satisfies the conditions described in Section 2.3. Clearly, the non-singularity of $\mathbf{F}(\boldsymbol{\rho}_D)$ is certified from the fact that it is a diagonal matrix with non-zero diagonal elements. The latter statement is true, provided that $|\rho_d| \neq 1$. Unfortunately, the Fisher information in (6.39) does not fulfill (2.102) since:

$$\frac{1}{N}[\mathbf{F}(\boldsymbol{\rho}_D)]_{d,d} \xrightarrow{N\to\infty} \frac{(1+\rho_d^2)}{(1-\rho_d^2)^2}, \tag{6.42}$$

is a function that cannot be omitted from the BIC optimization problem (see the rationale after (2.100)). Nevertheless, for the sake of deriving a simple (and intuitive) model-order selection rule, we still omit the term that corresponds to the log-determinant of the Fisher information matrix in the BIC (see (2.100) and (2.101)). Notice that ignoring this term in the BIC is necessary to obtain an intuitive model-order selection rule consisting in making independent decisions for each channel, in a similar manner to what is obtained for the other approaches, as described in a later paragraph (see (6.49)). In order to motivate the resulting suboptimal BIC (after ignoring the log-determinant term), we show a graphical representation of (6.42) in Figure 6.1. In this figure, it is evidenced that the limit in (6.42) is approximately constant for small values of ρ_d, converting the proposed suboptimal rule into a locally optimum one. Thus, it is reasonable to ignore the log-determinant term contribution (see (2.100) and (2.101)) in the BIC selection problem because the channels with the smallest values of ρ_d constitute the main difficulty of this problem. The previous analysis justifies the fact that in Section 6.5 the suboptimal BIC is the best-performing approach to the MI estimation task in terms of the estimation bias, and also justifies the results that we detailed in [120], where we did not provide the previously shown Fisher information analysis of the BIC.

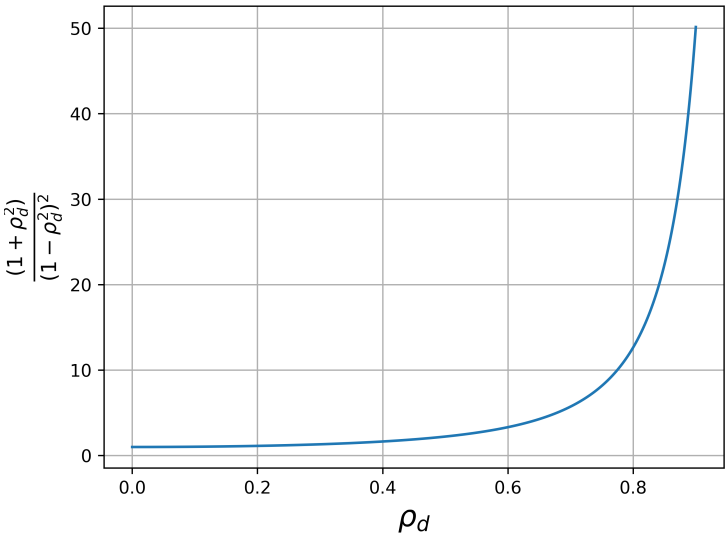


Figure 6.1: Graphical representation of (6.42) as a function of ρ_d.

As a result of the previous ideas, the selection rules shown in Table 2.1 and the general information theoretic model-order selection problem of (2.83) are considered from now on. The particularization of these ideas to the setting studied in this chapter consists on the following optimization problem:

$$\hat{D} = \arg\max_{L} \hat{I}_{ML}(\mathbf{x};\mathbf{y}|\mathcal{S}_L) - \frac{L\eta(N)}{2N} \quad \text{s. t. } L \in \{1,2,...,M\} = \tag{6.43a}$$

$$\arg\max_{L} -\frac{1}{2}\sum_{l=1}^{L}\log(1-\hat{\rho}_l^2) - \frac{L\eta(N)}{2N} \quad \text{s. t. } L \in \{1,2,...,M\}, \tag{6.43b}$$

where we have imported the CML cost from (6.27). In the previous expression, $\mathcal{S}_L \subseteq \mathcal{S}_M$ denotes the set of indices corresponding to the first L entries of $\hat{\boldsymbol{\rho}}_M$, which is the vector containing the estimated correlation coefficients of all channels, and $\eta(N)$ can be any of the approaches given in Table 2.1. Also, $\hat{\boldsymbol{\rho}}_L$ is the vector that contains the first L components of $\hat{\boldsymbol{\rho}}_M$ and $\hat{\rho}_l$ is its corresponding l-th component. Without any loss of generality, we now assume that the square of the estimated Pearson correlation coefficients contained in $\hat{\boldsymbol{\rho}}_M$ and $\hat{\boldsymbol{\rho}}_L$ are ordered in a descending manner (this becomes irrelevant later on), i.e.:

$$\hat{\rho}_l^2 \geq \hat{\rho}_{l'}^2, \tag{6.44}$$

for $l' > l$. Thanks to the descending ordering of the Pearson correlation coefficient, the first term in the log-likelihood function of (6.43) is non-decreasing with L. This is verified from the fact that (6.44) implies:

$$-\log(1-\rho_l^2) \geq -\log(1-\rho_{l'}^2), \tag{6.45}$$

for $l' > l$.

The overfitting tendency of the ML-MI estimator can be deduced from the optimization problem in (6.43). In this regard, notice that the summation in (6.43) is made of non-negative terms. Therefore, the optimal value of (6.43) would yield $\hat{D} = M$ without the penalty term. For this reason, the penalty term in (6.43) decreases linearly with L, so it is possible to compensate the tendency towards overfitting of the ML-MI estimator by penalizing the value of the overall cost for an increasing number of tested channels, L.

With the aim of deriving the optimal solution of (6.43), let us define:

$$h(L) = -\frac{1}{2}\sum_{l=1}^{L}\log(1-\hat{\rho}_l^2) - \frac{L\eta(N)}{2N}. \tag{6.46}$$

Then, the optimal value of $\hat{D}$ in (6.43) is given by the first value of L such that the discrete derivative of $h(L)$ with respect to L is negative:

$$\Delta h(L) = h(L+1) - h(L) = -\frac{1}{2}\log(1-\hat{\rho}_{L+1}^2) - \frac{\eta(N)}{2N}. \tag{6.47}$$

Thus, the expression of the non-active channels is:

$$-\frac{1}{2}\log(1-\hat{\rho}_{L+1}^2) - \frac{\eta(N)}{2N} \leq 0, \tag{6.48a}$$

$$-\frac{1}{2}\log(1-\hat{\rho}_{L+1}^2) \leq \frac{\eta(N)}{2N}, \tag{6.48b}$$

$$\log(1-\hat{\rho}_{L+1}^2) \geq -\frac{\eta(N)}{N}, \tag{6.48c}$$

$$1-\hat{\rho}_{L+1}^2 \geq \exp\left(-\frac{\eta(N)}{N}\right), \tag{6.48d}$$

$$\hat{\rho}_{L+1}^2 \leq 1-\exp\left(-\frac{\eta(N)}{N}\right). \tag{6.48e}$$

Equivalently, the active channels are those that fulfill:

$$\hat{\rho}_L^2 \geq 1-\exp\left(-\frac{\eta(N)}{N}\right). \tag{6.49}$$

The previous two expressions ((6.48e) and (6.49)) imply making independent decisions for each channel. Therefore, the ordering of the estimated Pearson correlation coefficients is not required. A relevant observation on the final threshold per channel in (6.49) is that it tends to 0 as $N \to \infty$ for all the information theoretic model-order selection rules given in Table 2.1. The interpretation of this limiting case is that the resulting model-order selection criteria lean towards accepting every single channel as

an active one. This behavior is expected and acceptable due to the fact that the ML estimation of all the correlation coefficients, $\hat{\boldsymbol{\rho}}_M$, is consistent for an infinite sample size (see (6.41)).

As a summary, the resulting regularized (via model-order selection) MI estimator is given by:

$$\hat{I}_R(\mathbf{x};\mathbf{y}|\mathcal{S}_{\eta(N)}) = -\frac{1}{2}\sum_{m=1}^{M}\log(1-\hat{\rho}_m^2\mathcal{I}_{\mathcal{S}_{\eta(N)}}(\hat{\rho}_m)), \tag{6.50}$$

where $\hat{\rho}_m = [\hat{\boldsymbol{\rho}}_M]_m$ and $\mathcal{I}_{\mathcal{S}_{\eta(N)}}(\cdot)$ is the indicator function (see Definition 3.25) of the set of detected active channels, which is defined as follows:

$$\mathcal{S}_{\eta(N)} = \left\{\rho \in [0,1] : \rho^2 \geq 1 - \exp\left(-\frac{\eta(N)}{N}\right)\right\}. \tag{6.51}$$

6.4 Dealing with non-parallel datasets via Informative Canonical Correlation Analysis

The only issue with the regularized estimator of the MI in (6.50) is that it implicitly assumes that the channels are mutually independent and built on unit-variance random variables. Fortunately, we show in this section that any problem with general Gaussian channels, i.e. correlated channels and with arbitrary variances, can be mapped to the nominal case by a linear transformation of the data. The tool that converts a problem with general Gaussian channels into a nominal one is the Canonical Correlation Analysis (CCA) and, more specifically, its empirical variant [11], [140], [148], [176]. We review the CCA methodology by following an insightful route with the previous aim in the following paragraphs.

As previously stated, the goal of CCA is to find linear transformations of two random vectors such that their cross-covariance is diagonal (see (6.3)) and whose variance is upper bounded by 1. The resulting vectors from the aforementioned linear transformations are denoted in the following manner:

$$\mathbf{u}(n) = \mathbf{A}\mathbf{x}(n), \tag{6.52a}$$

$$\mathbf{v}(n) = \mathbf{B}\mathbf{y}(n), \tag{6.52b}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $M \times M$ matrices that depict the linear transformations and, $\mathbf{x}(n)$ and $\mathbf{y}(n)$ are general M-dimensional Gaussian vectors, i.e. they are mutually dependent and their components have non-unitary variances. The previous linear transformations must be such that the cross-covariance of $\mathbf{u}(n)$ and $\mathbf{v}(n)$:

$$\mathrm{E}[\mathbf{u}(n)\mathbf{v}^T(n)] = \mathrm{E}[\mathbf{v}(n)\mathbf{u}^T(n)] = \mathbf{C}_{uv}, \tag{6.53}$$

is diagonal with entries bounded in $[-1,1]$. An intuitive strategy to obtain $\mathbf{A}$ and $\mathbf{B}$ is to reformulate them as general whitener matrices:

$$\mathbf{A} = \mathbf{F}^T\mathbf{C}_x^{-\frac{1}{2}}, \tag{6.54a}$$

$$\mathbf{B} = \mathbf{G}^T\mathbf{C}_y^{-\frac{1}{2}}, \tag{6.54b}$$

where $\mathbf{F},\mathbf{G} \in \mathrm{O}(M)$ and:

$$\mathbf{C}_x = \mathrm{E}[\mathbf{x}(n)\mathbf{x}^T(n)], \tag{6.55a}$$

$$\mathbf{C}_y = \mathrm{E}[\mathbf{y}(n)\mathbf{y}^T(n)], \tag{6.55b}$$

are the covariance (the original vectors have zero-mean) matrices of $\mathbf{x}(n)$ and $\mathbf{y}(n)$. In (6.54), the square-root of the covariance matrices ensure that the resulting vectors after the linear transformations (see (6.52)) are of unit variance, while the rotation matrices span every possible way of whitening the original vectors. Actually, the rotation matrices are constructed so that the parallel Gaussian setting is achieved for $\mathbf{u}(n)$ and $\mathbf{v}(n)$. This means that $[\mathbf{u}(n)]_m$ and $[\mathbf{v}(n)]_m$ are correlated random variables with unit variance, while the m-th pair of components, $[\mathbf{u}(n)]_m$ and $[\mathbf{v}(n)]_m$, is mutually independent of the

k-th pair of components, $[\mathbf{u}(n)]_k$ and $[\mathbf{v}(n)]_k$. With the intention of obtaining the rotation matrices, $\mathbf{F}$ and $\mathbf{G}$, note that, under the transformations stated in (6.54), $\mathbf{C}_{uv}$ has the following expression:

$$\mathbf{C}_{uv} = \mathrm{E}[\mathbf{u}(n)\mathbf{v}^H(n)] = \mathbf{F}\mathbf{C}_x^{-\frac{1}{2}}\,\mathrm{E}[\mathbf{x}(n)\mathbf{y}^T(n)]\mathbf{C}_y^{-\frac{1}{2}}\mathbf{G}^T = \mathbf{F}^T\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}}\mathbf{G}. \tag{6.56}$$

Also, let us denote the SVD of $\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}}$ as:

$$\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}} = \mathbf{F}_*\mathbf{\Lambda}\mathbf{G}_*^T, \tag{6.57}$$

where the diagonal elements of $\mathbf{\Lambda}$, also referred to as the canonical correlations, verify:

$$0 \leq [\mathbf{\Lambda}]_{m,m} \leq 1, \tag{6.58}$$

since $\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}}$ is a coherence matrix [159]. Then, the requirement of a diagonal $\mathbf{C}_{uv}$ is achieved by setting $\mathbf{F} = \mathbf{F}_*$ and $\mathbf{G} = \mathbf{G}_*$ since:

$$\mathbf{C}_{uv} = \mathbf{F}_*^T\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}}\mathbf{G}_* = \mathbf{F}_*^T\mathbf{F}_*\mathbf{\Lambda}\mathbf{G}_*^T\mathbf{G}_* = \mathbf{\Lambda}. \tag{6.59}$$

Even so, it is worth highlighting the fact that we could have set $\mathbf{F} = -\mathbf{F}_*$ or $\mathbf{G} = -\mathbf{G}_*$, and still guarantee that $\mathbf{u}(n)$ and $\mathbf{v}(n)$ are mutually independent.

In conclusion, the SVD of the coherence matrix, $\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}}$, gives both the rotations and the generalized correlation coefficients. Notice that the procedure that yields the rotation matrices is reminiscent of Lemma 2.4. Thus, the CCA can be understood as an *alignment* (in terms of correlation) of the data. What is more, we remark that the squared correlation coefficient of the m-th channel, which is constructed using the m-th component of $\mathbf{u}(n)$ and $\mathbf{v}(n)$, is obtained from the diagonal components of $\mathbf{\Lambda}$ (all positive due to the definition of the SVD) [159]:

$$\rho_m^2 = [\mathbf{\Lambda}]_{m,m}^2. \tag{6.60}$$

Provided that the MI of $\mathbf{u}(n)$ and $\mathbf{v}(n)$ has the same value as the MI of $\mathbf{x}(n)$ and $\mathbf{y}(n)$ (see Lemma 2.3), the regularized MI estimator of general Gaussian vectors yields after plugging (6.60) into (6.50). Hence, we proved that the methodology described in Section 6.3 is also valid for general Gaussian vectors. However, it is important to mention that $\mathbf{C}_x$, $\mathbf{C}_y$ and $\mathbf{C}_{xy}$ are unknown matrices in practice. In light of this, we need to resort to the empirical CCA methodology [11], [140], [148], [176] when general Gaussian vectors are considered.

6.4.1 Empirical CCA

Following from the rationale shown in Section 6.4, resorting to the empirical variant of the CCA is mandatory since the covariances and cross-covariances of $\mathbf{x}(n)$ and $\mathbf{y}(n)$ are not available. In the sequel, we show that the empirical CCA implicitly estimates these matrices from the available data. We review the ideas presented in [11] and [148] without the incorporation of the Informative CCA rationale. With the assumption that $N > M$, let us consider the SVD of the data matrices (see (6.14)):

$$\mathbf{X} = \mathbf{U}_x\mathbf{D}_x\mathbf{V}_x^T, \tag{6.61a}$$

$$\mathbf{Y} = \mathbf{U}_y\mathbf{D}_y\mathbf{V}_y^T, \tag{6.61b}$$

where $\mathbf{U}_x, \mathbf{U}_y \in \mathbb{R}^{M\times M}$, $\mathbf{D}_x, \mathbf{D}_y \in \mathbb{R}^{M\times M}$ and $\mathbf{V}_x, \mathbf{V}_y \in \mathbb{R}^{N\times M}$. Also, let the sample covariances be:

$$\hat{\mathbf{C}}_x = \frac{1}{N}\mathbf{X}\mathbf{X}^T = \frac{1}{N}\mathbf{U}_x\mathbf{D}_x\underbrace{\mathbf{V}_x^T\mathbf{V}_x}_{\mathbf{I}_M}\mathbf{D}_x\mathbf{U}_x^T = \frac{1}{N}\mathbf{U}_x\mathbf{D}_x^2\mathbf{U}_x^T, \tag{6.62a}$$

$$\hat{\mathbf{C}}_y = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T = \frac{1}{N}\mathbf{U}_y\mathbf{D}_y\underbrace{\mathbf{V}_y^T\mathbf{V}_y}_{\mathbf{I}_M}\mathbf{D}_y\mathbf{U}_y^T = \frac{1}{N}\mathbf{U}_y\mathbf{D}_y^2\mathbf{U}_y^T, \tag{6.62b}$$

$$\hat{\mathbf{C}}_{xy} = \frac{1}{N}\mathbf{X}\mathbf{Y}^T = \frac{1}{N}\mathbf{U}_x\mathbf{D}_x\mathbf{V}_x^T\mathbf{V}_y\mathbf{D}_y\mathbf{U}_y^T. \tag{6.62c}$$

Then, the *sample* coherence matrix [11], [176] is given by:

$$\hat{\mathbf{C}}_x^{-\frac{1}{2}}\hat{\mathbf{C}}_{xy}\hat{\mathbf{C}}_y^{-\frac{1}{2}} = \mathbf{U}_x\mathbf{D}_x^{-1}\underbrace{\mathbf{U}_x^T\mathbf{U}_x}_{\mathbf{I}_M}\mathbf{D}_x\mathbf{V}_x^T\mathbf{V}_y\mathbf{D}_y\underbrace{\mathbf{U}_y^T\mathbf{U}_y}_{\mathbf{I}_M}\mathbf{D}_y^{-1}\mathbf{U}_y^T = \mathbf{U}_x\mathbf{V}_x^T\mathbf{V}_y\mathbf{U}_y^T. \tag{6.63}$$

It can be deduced from the previous expression that the canonical correlations of $\mathbf{x}(n)$ and $\mathbf{y}(n)$ can be retrieved from the SVD of $\mathbf{V}_x^T\mathbf{V}_y$. In fact, considering that these matrices are orthogonally constrained (see the constraint in (2.52)), the canonical correlations can also be understood as the principal angles between the subspaces spanned by the rows of $\mathbf{X}$ and $\mathbf{Y}$. This interpretation is fundamental to uncover a hidden lower bound on the amount of data samples such that the estimated canonical correlations are well-behaved [148, Subsection III-A]. In this regard, notice that the intersection between $\mathbf{V}_x$ and $\mathbf{V}_y$ is of dimension equal to (at least) $T = 2M - N$ for $N < 2M$. Thus, there would be a minimum of T singular values of $\mathbf{V}_x^T\mathbf{V}_y$ that are deterministically equal to 1. The previous phenomenon is undesired since it would result in an infinite MI, even in those cases where there is no functional relationship between some of the channels. For this reason, we have to assume from now on that $N > 2M$ when general Gaussian vectors are considered. Otherwise, the Informative CCA methodology [11], [140], which is out of the scope of this chapter, would be required.

The main conclusion of the previous ideas is that we can estimate the generalized correlation coefficients without the implicit computation of the linear transformations in (6.52). In fact, we only need to estimate the correlation coefficients, i.e. the singular values of $\mathbf{V}_x^T\mathbf{V}_y$, and plug them into (6.50) to estimate the MI.

6.5 Numerical results

The purpose of the remaining of this chapter is to test the MI estimation approach detailed in Section 6.3. Particularly, we are interested to assess whether the proposed approach is capable of mitigating the tendency towards overfitting of the naive ML-MI estimator. For this reason, we are interested in the assessment of the bias of (6.50) and compare it to the ML-MI. Accordingly, the bias of the MI estimators is computed using 1000 Monte Carlo simulations.

The numerical experiments consist in two scenarios. The first one constructs two multichannel datasets, given by $\mathbf{x}(n)$ and $\mathbf{y}(n)$, using the nominal conditions described in Section 6.1 without the unit-variance assumption. The reason behind not considering the unit-variance assumption is that it is not possible to obtain a simple estimator of the correlation coefficients under these conditions, e.g. (6.36c), may take a value greater than 1. In other words, the ML estimation of the correlation coefficient implicitly estimates the variance of the involved random variables and, hence, simulating unit and general variances random variables is equivalent from the estimation of the correlation coefficients problem perspective. Yet, this scenario serves to test the performance of the proposed estimator when its implicit assumptions hold. In the remaining scenario, we construct the same multichannel dataset as in the previous one, yet the correlation coefficients are estimated by means of the empirical CC rationale from Subsection 6.4.1. In other words, we simulate the same kind of data for both scenarios, but the correlation coefficients are estimated using a different methodology.

Regarding the penalty term from the cost function in (6.43), we summarize the considered values of $\eta(N)$ in Table 6.1. The reasoning behind the consideration of three different versions of the GIC is to provide of a more general picture of this information theoretic criterion since there is no universal rule to select its user defined parameter in a practical scenario. Besides, the corrected BIC takes into account the fact that the variances of each channel must be estimated, in contrast to the BIC, which follows straightforwardly from the rationale given in Section 6.3.

All the scenarios are simulated using the following simple statistical model:

$$x_m(n) = \sigma_{x,m}z_{m,1}(n), \tag{6.64a}$$

$$y_m(n) = \sigma_{y,m}z_{m,2}(n), \tag{6.64b}$$

Criterion	**Penalty**, $\eta(N)$
AIC	2
Conservative GIC (c-GIC)	$1+\lambda_c = 10$
Balanced GIC (b-GIC)	$1+\lambda_b = 20$
Aggressive GIC (a-GIC)	$1+\lambda_a = 50$
BIC	$\eta(N) = \ln(N)$
Corrected BIC	$\eta(N) = 3\ln(N)$

Table 6.1: Considered information theoretic criteria for the model-order selection of the active channels (see Table 2.1).

where:

$$z_{m,2}(n) = \rho_m z_{m,1}(n) + \sqrt{1-\rho_m^2} z_{m,3}(n). \tag{6.65}$$

In the previous equations, $z_{m,1}(n)$ and $z_{m,3}(n)$ are standard normal random variables, while $\sigma_{x,m}$ and $\sigma_{y,m}$ are the standard deviations of $x_m(n)$ and $y_m(n)$, respectively. Additionally, $z_{m,1}(n)$ and $z_{m,3}(n)$ are mutually independent with each other and with their respective variables from other channels, $z_{m',1}(n)$ and $z_{m',3}(n)$ with $m \neq m'$. It is easy to verify that the correlation coefficient of the simulated $x_m(n)$ and $y_m(n)$ is ρ_m. This statement is verified from the following fact:

$$\mathrm{E}[x_m(n)y_m(n)] = \sigma_{x,m}\sigma_{y,m}\,\mathrm{E}\left[z_{m,1}(n)\left(\rho_m z_{m,1}(n) + \sqrt{1-\rho_m^2} z_{m,3}(n)\right)\right] = \sigma_{x,m}\sigma_{y,m}\rho_m. \tag{6.66}$$

Also:

$$\mathrm{E}[x_m(n)y_{m'}(n)] = 0, \tag{6.67}$$

since:

$$\mathrm{E}[z_{m,1}(n)z_{m',1}(n)] = \mathrm{E}[z_{m,1}(n)z_{m',3}(n)] = 0, \tag{6.68}$$

for all $m \neq m'$. Moreover, we set the correlation coefficient of the active channels such that the MI of each channel exhibits a linearly decreasing dependence with respect to the channel index, $d = 0, ..., D-1$. This means that the MI associated to the d-th active channel is given by:

$$I_d = C_I\left(1 - \frac{d}{D}\right) = \frac{1}{2}\log(1-\rho_d^2), \tag{6.69}$$

where C_I is a constant that is computed to fix the overall MI to a desired value. Provided that the overall MI is:

$$I_{\mathrm{tot}} = -\frac{1}{2}\sum_{d=0}^{D-1} I_d, \tag{6.70}$$

the value of C_I is:

$$C_I = 2\frac{I_{\mathrm{tot}}}{D+1}. \tag{6.71}$$

Consequently, the Pearson correlation coefficient of the d-th channel are:

$$\rho_d = \tau_d\sqrt{1-\exp\left(-4\frac{I_{tot}}{D}\left(\frac{D-d}{D+1}\right)\right)} \quad d = 0, ..., D-1, \tag{6.72}$$

where τ_d can be either 1 or -1. For simplicity, we set $\tau_d = 1$ for all values of D. The remaining parameters are stated on top of each figure.

6.5.1 Mutually independent datasets

In this set of simulations, we evaluate the performance of the MI estimators under the nominal conditions depicted in Section 6.1 without the unit-variance assumption. In Figure 6.2, we showcase the bias of the naive ML-MI estimator, in addition to the regularized MI estimation approaches as a function

of M and N. These two parameters define the difficulty of the MI estimation problem. Besides, the correlation coefficients are estimated from the coherence matrix that results from the following matrices:

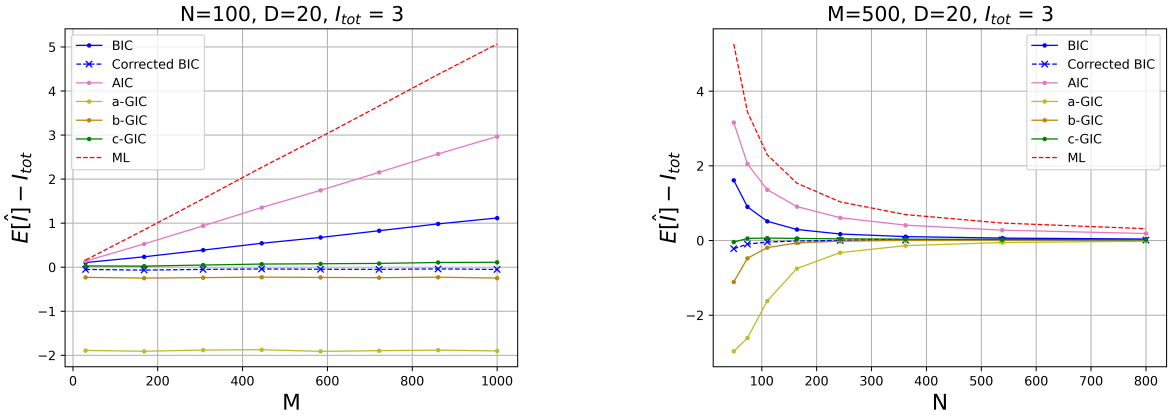
$$\tilde{\mathbf{C}}_x = \frac{1}{N}\mathbf{X}\mathbf{X}^T \odot \mathbf{I}_M, \tag{6.73a}$$

$$\tilde{\mathbf{C}}_y = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T \odot \mathbf{I}_M, \tag{6.73b}$$

$$\tilde{\mathbf{C}}_{xy} = \frac{1}{N}\mathbf{X}\mathbf{Y}^T \odot \mathbf{I}_M, \tag{6.73c}$$

which are given by:

$$\hat{\boldsymbol{\rho}} = \text{diag}(\tilde{\mathbf{C}}_x^{-\frac{1}{2}}\tilde{\mathbf{C}}_{xy}\tilde{\mathbf{C}}_y^{-\frac{1}{2}}). \tag{6.74}$$



(a) As a function of the number of channels (M) with fixed N.

(b) As a function of the number of samples (N) with fixed M.

Figure 6.2: Bias of the MI estimators under nominal conditions.

The conclusions that were already anticipated in Section 6.3 are observed in Figure 6.2. On the one hand, the tendency towards overfitting of the ML estimator is evidenced in Subfigure 6.2a since its bias increases with the greatest slope (as compared to the other alternatives) with respect to M. On the other hand, as verified in Subfigure 6.2b, all estimators are asymptotically unbiased, provided that the threshold of the active channels and the variance of the estimated correlation coefficients tend to 0 for $N \to \infty$ (see (6.49) and Table 6.1). In these subfigures, the best-performing approaches are those that exhibit the smallest slope in Subfigure 6.2a and the fastest convergence to zero-bias in Subfigure 6.2b. Clearly, the c-GIC and the corrected BIC are the alternatives that achieve the best performance. Although the c-GIC is a close contestant, the corrected BIC is preferred due to the fact that it does not require the determination of any user-defined parameter, which is the main drawback of the GIC alternatives.

With the aim of complementing the previous analysis, in Figure 6.3 we show the average number of detected channels as a function of the number of samples, N. As expected, with the exception of the AIC and the BIC, the average number of detected channels converge to the true value as $N \to \infty$. This result will be useful as a reference for the following numerical experiment.

6.5.2 Mutually dependent datasets

In this last experiment, we evaluate the practical performance of the ideas described in Subsection 6.4.1. To this end, the only difference with the previous scenario is the fact that the canonical correlations are computed from the SVDs of $\mathbf{X}$ and $\mathbf{Y}$ (see (6.61) and (6.63)). In this manner, the errors that emerge due to the empirical CCA can be easily assessed by a quick comparison with the previous experiment.

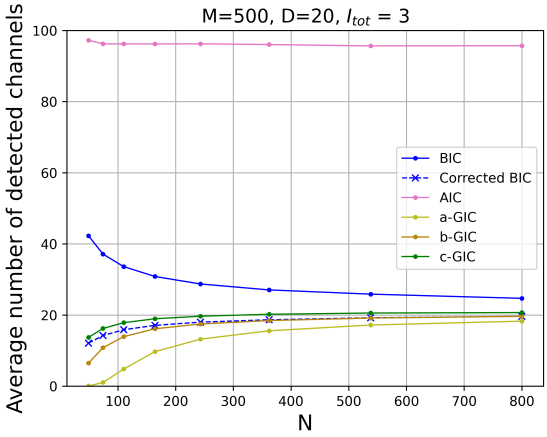
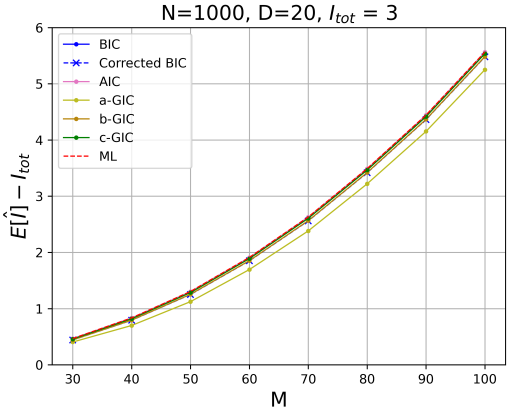
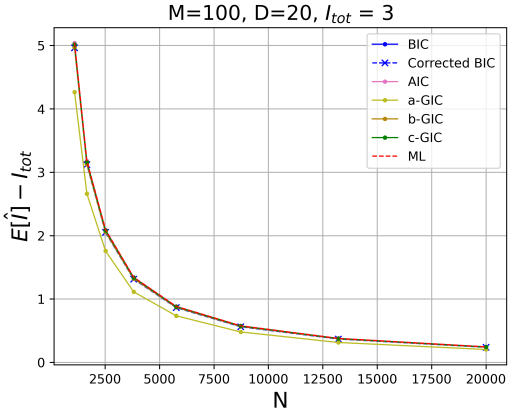


Figure 6.3: Average detected channels under nominal conditions as a function of the number of samples, N. The true value of active channels is $D = 20$.



(a) As a function of the number of channels (M) with fixed N.



(b) As a function of the number of samples (N) with fixed M.

Figure 6.4: Bias of the MI estimators of mutually dependent datasets.

It is evidenced in Figure 6.4, which has a different scale in the x-axis than Figure 6.2, that the bias of all the estimators grows much faster than in the previous experiment. This behavior is expected since this scenario requires the determination of a greater amount of free parameters, i.e. the subspaces spanned by the rows of $\mathbf{X}$ and $\mathbf{Y}$ must be determined. Unfortunately, the regularized MI estimators are not capable of compensating the unwanted bias of the ML-MI estimator. Even the a-GIC, which is the alternative that has the highest penalty, results in a negligible improvement of the ML-MI. This poor performance is also reflected in the average number of detected channels shown in Figure 6.5, where no criterion is capable of detecting the true value of the active channels ($D = 20$).

6.6 Final remarks

In spite of the great results for the mutually independent datasets (even without the unit-variance assumption) experiment, the incorporation of the empirical CCA methodology into our MI estimation framework did not achieve the expected results. Yet, the methodology described in this chapter still offers a possible alternative of transforming a general MI estimation problem with Gaussian vectors into a more manageable one. It is left as a future work the refinement of the practical implementation

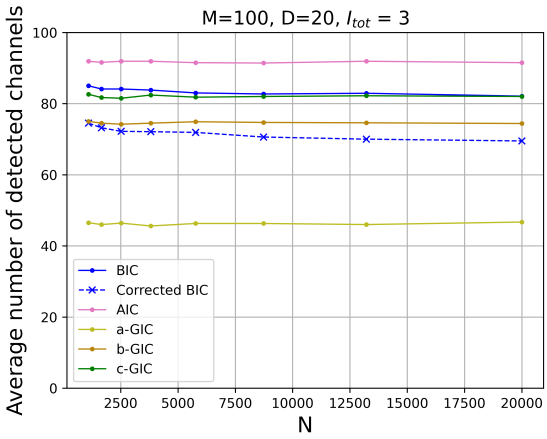


Figure 6.5: Average detected channels of mutually dependent datasets as a function of the number of samples, N. The true value of active channels is $D = 20$.

of the CCA methodology into the methodology shown in this chapter.

As a summary, we showed in this chapter that, whenever the datasets align with the mutual independence assumption (parallel channels interpretation), it is possible to estimate the MI in an efficient manner by making independent decisions for each channel. The resulting measure of diversity would be yielded by plugging the MI estimation into the information theoretic coherence introduced in the beginning of this chapter. A possible reformulation of the diversity quantification problem would be to estimate the information theoretic coherence without resorting to the MI as an intermediate measure, seeing that the unboundedness of the MI is what complicates its estimation. It is conjectured that, thanks to the fact that a coherence is bounded in $[0, 1]$, estimating the information theoretic coherence in a direct manner would have better numerical properties than any MI estimator.

Chapter 7

Conclusions

This dissertation has studied new ways of incorporating and exploiting diversity in signal processing, in addition to the already known results from wireless communications. In the development of these ideas, we discovered new interpretations of well-established problems from the perspective of Information Theory and generalized the concept of sparsity. In the following paragraphs, a short summary of the principal fruits in each chapter is provided with an added value on the original motivation.

Chapter 2 unveiled the link between sparsity, information theoretic measures, the Grassmann manifold and model-order selection rules. Indeed, this connection was shown by a simple linear model, which is referred to with different names depending on the specific field of study. We have shown that, fundamentally, all these approaches for solving inverse problems are equivalent in practice. This chapter can be summarized as the answer to the question: *what kinds of optimization problems are we solving in this dissertation?*

In addition to the review of well-known concepts in numerical optimization, such as convex optimization and the classic Majorization-Minimization (MM) framework, we generalized some of these results to the Grassmann manifold in Chapter 3. Particularly, we generalized the MM framework and its block extensions to the Grassmannian and also provided a novel geometric interpretation of the Principal Component Analysis (PCA) problem, complementing an already known result in the Matrix Factorization literature. However, the only issue with the geometric interpretation of the PCA problem detailed in Subsection 3.2.2.1 is that we were unable to provide a complete assessment of the stationary points of the PCA cost. We only ensured that the stationary points relatively close to the set of eigenvectors corresponding to the largest eigenvalues are local maximum points. Besides, we surveyed in detail the proximal algorithms with an emphasis on the Alternating Direction Method of Multipliers (ADMM) since they were fundamental in Chapter 5. In a similar manner to Chapter 2, Chapter 3 was devoted to respond the following question: *how are we solving the optimization problems encountered throughout this thesis?*

The study of diversity in signal processing applications commenced in Chapter 4. After a brief review on the building blocks of information fusion policies, we explored three different fusion strategies with an information theoretic perspective. Firstly, we analyzed the known Covariance Intersection (CI) principle and its two possible derivations in Section 4.2. We also reviewed the clear link between the minimum determinant criterion to obtain the intersection weights and the Parametric Minimum Error Entropy (PMEE) criterion. Yet, our main result on this topic was highlighting the apparent link, albeit one may consider it as *ad hoc*, of the CI with the waterfilling algorithm for power allocation from wireless communications. Secondly, we derived a bounded information theoretic descriptor of a worst-case contaminated scenario in Section 4.3. Using this descriptor, we were capable of interpreting the ℓ_0 norm regularization from an information theoretic perspective in a setting with unreliable sensors. The main issue with the resulting optimization problem was that it was a challenging non-convex optimization. In fact, we proved that the classical ℓ_1 relaxation of the ℓ_0 norm was not reliable for this context. Lastly, our proposed solution to the blind sensor fusion and regression problem from Section 4.4 naturally related the PMEE criterion, the Conditional Maximum Likelihood (CML) principle

and the block MM algorithm on the Grassmann manifold. In this regard, we showed the potential capabilities of introducing structural priors (low-rank model, diagonal covariance,...) instead of the classical Bayesian priors on a signal processing problem. More specifically, through the use of structural priors, the resulting regression model gained more versatility, in the sense that it can be applied to a wider range of regression problems, and more robustness to model deviations. In addition to this, the convergence analysis of the proposed algorithm also showed that certain majorants are sufficient to ensure the convergence of a general block MM formulation bypassing the necessity of a compact constraint set or the coerciveness of the cost function.

Chapter 5 was devoted to the Covariance Conversion (CC) problem of Frequency Division Duplexing (FDD) schemes from wireless communications. Our attention was drawn towards this problem since we found that some sort of diversity was present in the core of channel covariance estimations of the Uplink and Downlink channels. Interestingly, this kind of diversity was more evident after we reformulated a particular model of the channel covariance matrices via the vectorization function. In addition to diversity, it was possible to define the concept of sparsity in the aforementioned covariance matrices under the mmWave and ultra-wide band conditions, which we exploited to develop a sparse-aware approach to the CC problem. Under this assumption, we showed promising results with a reasonable amount of computational complexity.

Finally, Chapter 6 focused on the Mutual Information (MI) estimation as a means of quantifying the diversity between a set of parallel Gaussian channels under the assumption that some of them provide no information. The main results of this chapter consisted of two focal points. On the one hand, we proved that the ML estimation of the MI tended to overfit the data, meaning that it implicitly assumed that every channel is active. The previous behavior of the ML estimator yielded an unwanted bias. On the other hand, we derived a regularized estimator based on the information theoretic model-order selection rules surveyed in Chapter 2. The model-order selection of the active channels consisted in making simple independent channel decisions, resulting in a computationally efficient regularized MI estimator that eliminated the contributions of the detected non-active channels. This idea prevented the unwanted bias of the ML estimator. In the effort of generalizing the previous estimator to general Gaussian vectors, we reviewed the empirical Canonical Correlation Analysis (CCA) method. Unfortunately, we observed in Subsection 6.5.2 that the MI estimators that resulted from the joint consideration of the Empirical CCA method and the information theoretic model-order selection are unable of attenuating the bias of the ML estimator.

In summary, the key points of this dissertation are:

- Sparsity and minimum entropy, which are equivalent concepts, were related to low-rank subspace models and, hence, to the Grassmann manifold. Taking into account the unveiled connection between these topics, it is always possible to interchange these methodologies.
- The MM framework can be generalized straightforwardly to certain non-convex sets thanks to the geodesically-convex optimization theory, being the results on the Grassmann manifold in this dissertation a particular example of this idea.
- The nexus between seemingly disconnected topics can provide additional insights. For instance, the waterfilling interpretation given in Subsection 4.2.2 suggests a duality between the multisensor fusion and wireless communications problems. Also, the log-determinant cost of a parameterized covariance from Subsection 4.4.2 is also considered in the unstructured interference mitigation of GNSS receivers [172].
- The ℓ_0 norm emerges via minimum entropy arguments of noise variance worst-case scenarios. In these cases, an operational meaning of the entropic index emerges naturally as a precision/reliability trade-off. This idea is conjectured to be generalized to other areas as well.
- Information theoretic measures are a natural way of solving signal processing problems. In this regard, we have shown multiple instances of how the CML principle often results in a parameterized information theoretic cost.
- There is no universal way of exploiting (or discovering) diversity in signal processing [197].

7.1 Future lines of research

In light of the previous outline, we reflect on the potential extensions of this dissertation. First, we take a look at the possible lines of research that emanate from Chapter 3:

- Regarding the convergence proofs from Section 2.2 and, particularly, the convergence proof of the block MM algorithm, they were derived using arguments that only apply to the Grassmann manifold. A clear extension of these ideas would be to generalize the convergence proofs to other Riemannian manifolds. It is remarked that Theorem 3.10 has already been proved for the Stiefel manifold in [31], but this reference does not provide a formal proof for the block extension. We conjecture that the generalization of Theorem 3.11 to the Stiefel manifold and other compact manifolds are straightforward, but not so evident to other general Riemannian manifolds such as the manifold of positive definite matrices.
- In Subsection 3.2.2.1, we provide a Riemannian perspective on the Principal Component Analysis (PCA) problem. In this regard, we characterized the stationary points of the PCA cost function using geometric arguments. Yet, we were only capable of assessing the stationary points that lie on a neighborhood of the eigenvectors corresponding to the greatest eigenvalues. Using Theorem 3.8 as a reference, we believe that it should be possible to characterize all the stationary points of the PCA cost function.

Secondly, we focus on research lines that follow from Chapter 4:

- We showed that the Entropic-Best Linear Unbiased Estimator (E-BLUE) in (4.75) from Subsection 4.3.2 is obtained from a challenging optimization problem. Even the ℓ_1 regularization was unsuited for this formulation in part due to the unbiased fusion constraints. For this reason, a future line of research focuses on solving (4.77) with a computationally efficient algorithm, i.e. avoiding a combinatorial optimization search. Just to provide some insights on this problem, notice that the first term in the cost function of (4.77) is a quasiconvex function (see Example 3.2).
- In the blind fusion and regression problem from Section 4.4, one of the implicit assumptions was that the intrinsic dimension of the linear model in (4.94) was known. A straightforward continuation of this problem would be to eliminate the prior knowledge on the intrinsic dimension. We conjecture that the resulting algorithm from this setting would incorporate known intrinsic dimension detection results (see, for instance, [135]). Nonetheless, a formal incorporation of the intrinsic dimension detection framework into the MM paradigm described in Section 4.4 would be required.

Finally, we analyze the possible extensions of the remaining chapters:

- In Chapter 5, we implemented a sparse regression formulation to solve the CC problem. However, our initial idea was to incorporate information theoretic ideas to this problems, such as the generalized entropy function [83] of the Angular Power Spectrum (APS). For this reason, a potential research line would be to obtain an efficient information theoretic CC algorithm based on generalized entropy functions. In order to provide some insights to this new research line, the main issue that we encountered when we considered entropy functions (non-convex cost) of the APS was the search of a suitable first or second-order majorant (see subsections 3.3.3.1 and 3.3.3.2).
- In the process of finding a practical generalization of the regularized MI estimator from Section 6.3 to mutually dependent datasets, we resorted to the CCA and its empirical variant. To our detriment, the empirical implementation of the CCA did not improve the ML estimation of the MI. Consequently, a straightforward line of research would be to add an additional processing to the empirical CCA framework from Subsection 6.4.1. Alternatively, any generalization of the regularized MI estimator from Section 6.3 to mutually dependent datasets without relying on the CCA framework is also praised.

Chapter 8

Appendix

8.1 Appendices of Chapter 2

8.1.1 Derivation of the generalized Rényi Entropy of a matrix Gaussian distribution

Let us define the following integral:

$$I_\alpha(\mathbf{X}) = \int p_{\mathbf{X}}^{\alpha}(\mathbf{X})\mathrm{d}\mathbf{X}, \tag{8.1}$$

where $\mathbf{X} \sim \mathcal{MN}_{NM}(\mathbf{0}, \frac{1}{\alpha}\mathbf{Q}, \mathbf{K})$ is a random variable whose respective PDF, $p_{\mathbf{X}}(\mathbf{X})$, is given by (2.35). After plugging (2.35) into the previous integral, we obtain the following expression:

$$I_\alpha(\mathbf{X}) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{\alpha MN}{2}} \det(\mathbf{Q})^{\frac{\alpha N}{2}} \det(\mathbf{K})^{\frac{\alpha M}{2}}} \exp\left(-\frac{\alpha}{2}\operatorname{tr}(\mathbf{Q}^{-1}\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X})\right)\mathrm{d}\mathbf{X} = \tag{8.2a}$$

$$\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{\alpha MN}{2}} \det(\mathbf{Q})^{\frac{\alpha N}{2}} \det(\mathbf{K})^{\frac{\alpha M}{2}}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\left(\frac{1}{\alpha}\mathbf{Q}\right)^{-1}\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X}\right)\right)\mathrm{d}\mathbf{X} = \tag{8.2b}$$

$$\frac{C'}{(2\pi)^{\frac{\alpha MN}{2}} \det(\mathbf{Q})^{\frac{\alpha N}{2}} \det(\mathbf{K})^{\frac{\alpha M}{2}}} \int_{-\infty}^{\infty} \frac{1}{C'} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\left(\frac{1}{\alpha}\mathbf{Q}\right)^{-1}\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X}\right)\right)\mathrm{d}\mathbf{X}, \tag{8.2c}$$

where

$$C' = (2\pi)^{\frac{MN}{2}} \det\left(\frac{1}{\alpha}\mathbf{Q}\right)^{\frac{N}{2}} \det(\mathbf{K})^{\frac{M}{2}}. \tag{8.3}$$

Notice that C' is a constant such that the integral in (8.2c) is equal to 1. Thus, $I_\alpha(\mathbf{X})$ yields:

$$I_\alpha(\mathbf{X}) = \frac{(2\pi)^{\frac{MN}{2}} \det\left(\frac{1}{\alpha}\mathbf{Q}\right)^{\frac{N}{2}} \det(\mathbf{K})^{\frac{M}{2}}}{(2\pi)^{\frac{\alpha MN}{2}} \det(\mathbf{Q})^{\frac{\alpha N}{2}} \det(\mathbf{K})^{\frac{\alpha M}{2}}}. \tag{8.4}$$

The desired expression in (2.37) is obtained after taking the logarithm of $I_\alpha(\mathbf{X})$.

8.2 Appendices of Chapter 4

8.2.1 Proof of Proposition 4.2

Firstly, we show that the ML estimation and the MEE criterion for data fusion are equivalent under full statistical knowledge and, secondly, we verify that the efficient estimator[1] coincides with the ML and MEE estimators.

[1]The one that achieves the CRLB

The ML estimator of $x(n)$ is obtained from the maximization of the following PDF (see (4.1)):

$$f_{\mathbf{y}}(\mathbf{y}|x,\mathbf{Q}) = \frac{1}{\sqrt{(2\pi)^M \det(\mathbf{Q})}} \exp\left(-\frac{1}{2}(\mathbf{y}-x\mathbf{1}_M)^T\mathbf{Q}^{-1}(\mathbf{y}-x\mathbf{1}_M)\right), \tag{8.5}$$

which, after taking the minus logarithm and ignoring the additive constants that do not depend on x, results in the following optimization problem:

$$\hat{x}_{ML}(n) = \arg\min_x (\mathbf{y}(n)-x\mathbf{1}_M)^T\mathbf{Q}^{-1}(\mathbf{y}(n)-x\mathbf{1}_M). \tag{8.6}$$

The closed-form solution of (8.6) is given by:

$$\hat{x}_{ML}(n) = \frac{\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{y}(n)}{\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{1}_M} = \mathbf{f}_B^T\mathbf{y}(n), \tag{8.7}$$

where:

$$\mathbf{f}_B = \frac{\mathbf{Q}^{-1}\mathbf{1}}{\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{1}_M}, \tag{8.8}$$

is the optimal linear combiner. In a similar way, we can obtain $\mathbf{f}_B$ from the MEE criterion from the fusion error random variable, $e(n,\mathbf{f})$. Provided that $x(n)$ is deterministic, $e(n,\mathbf{f})$ is a zero-mean Gausian random variable with variance:

$$\mathrm{E}\left[|e(n,\mathbf{f})|^2\right] = \mathrm{E}\left[|\mathbf{f}^T\mathbf{y}(n)-x(n)|^2\right] = \mathrm{E}\left[|\mathbf{f}^T(x(n)\mathbf{1}_M-\mathbf{w}(n)-x(n)|^2\right] = \tag{8.9a}$$

$$\mathrm{E}\left[|\mathbf{f}^T(x(n)\mathbf{1}_M-\mathbf{w}(n))-x(n)|^2\right] = \mathrm{E}\left[(\mathbf{f}^T\mathbf{w}(n))(\mathbf{w}^T(n)\mathbf{f})\right] = \mathbf{f}^T\mathbf{Q}\mathbf{f}. \tag{8.9b}$$

Thus, the Rényi differential entropy of $e(n,\mathbf{f})$ yields (see (2.37) particularized for $N=1$, $M=1$ and the variance calculated in (8.9b)):

$$h(e(n,\mathbf{f})) = \frac{M}{2}\left(\log(2\pi)+\log\left(\mathbf{f}^T\mathbf{Q}\mathbf{f}\right)+\frac{\log(\alpha)}{\alpha-1}\right), \tag{8.10}$$

which results in the following criterion:

$$\hat{\mathbf{f}}_{MEE} = \arg\min_{\mathbf{f}} \log\left(\mathbf{f}^T\mathbf{Q}\mathbf{f}\right) \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1. \tag{8.11}$$

The previous constraint is necessary to ensure an unbiased fusion. Note that the logarithm can be ignored since it is a monotonic function. The solution of (8.11) is easily obtained from its Lagrangian formulation, also yielding $\mathbf{f}_B$

Finally, we prove that $\mathbf{f}_B$ is the linear combiner of the measurements that achieves the CRLB. The CRLB is obtained from the inverse of the Fisher information of $\mathbf{y}(n)$ (see (8.5)):

$$\gamma_{CRLB} = -\frac{1}{\frac{\partial^2 \log(f_{\mathbf{y}}(\mathbf{y}|x,\mathbf{Q}))}{\partial^2 x}} = \frac{1}{\mathbf{1}_M^T\mathbf{Q}^{-1}\mathbf{1}_M}, \tag{8.12}$$

which coincides with the variance obtained in (8.9b) particularized for $\mathbf{f}=\mathbf{f}_B$.

8.2.2 Lebesgue Dominated Convergence Theorem

The LDCT provides the conditions in which limits and integrals are interchangeable [46]. It is stated as follows [32].

Theorem 8.1 (Lebesgue Dominated Convergence Theorem)**.** *Let $\{f_n(x)\}_{n\in\mathbb{N}}$ be a sequence of functions in the space of Lebesgue integrable functions, which is convergent to a Lebesgue integrable function $f:\mathbb{R}\to\mathbb{R}$. Moreover, assume that there exist a Lebesgue integrable function, denoted as $g(x)$, such that $|f_n(x)|\leq g(x)$ for all $n\in\mathbb{N}$ and $x\in\mathbb{R}$. Then:*

$$\lim_{n\to\infty}\int_{-\infty}^{\infty}|f_n(x)-f|dx = 0, \tag{8.13}$$

which also implies:

$$\lim_{n\to\infty}\int_{-\infty}^{\infty}f_n(x)dx = \int_{-\infty}^{\infty}f(x)dx. \tag{8.14}$$

8.2.3 ML estimator of $\mathbf{u}_k$

From (4.98), it is verified that the statistical distribution of $\mathbf{s}_k(\mathbf{B},\mathbf{f})$ is:

$$\mathbf{s}_k(\mathbf{B},\mathbf{f}) \sim \mathcal{N}(\mathbf{B}\mathbf{u}_k, \mathbf{f}^T\mathbf{Q}\mathbf{f}\mathbf{I}_N), \tag{8.15}$$

The covariance matrix of $\mathbf{s}_k(\mathbf{B},\mathbf{f})$ is derived from the fact that the rows of $\mathbf{W}_k$ are statistically independent. Then, the entry-wise variance is given by:

$$\mathrm{E}\left[|[\mathbf{s}_k(\mathbf{B},\mathbf{f})]_n|^2\right] = \mathrm{E}\left[\mathbf{f}^T[\mathbf{W}_k]_{:,m}[\mathbf{W}_k]_{:,m}^T\mathbf{f}\right] = \mathbf{f}^T\mathbf{Q}\mathbf{f}, \tag{8.16}$$

where $[\mathbf{W}_k]_{:,m}$ denotes the m-th column of $\mathbf{W}_k$. From the previous expression we get the covariance matrix given in (8.15). Taking into consideration the statistical model given in (8.15), the estimation of $\mathbf{u}_k$ that minimizes the variance of the fused variable is obtained from the following optimization problem:

$$\hat{\mathbf{u}}_k = \arg\min_{\mathbf{u}} ||\mathbf{s}_k(\mathbf{B},\mathbf{f}) - \mathbf{B}\mathbf{u}||_2^2 \quad \text{s.t. } \mathbf{1}_M^T\mathbf{f} = 1 \equiv \tag{8.17a}$$

$$\arg\min_{\mathbf{u}} ||\mathbf{Y}_k\mathbf{f} - \mathbf{B}\mathbf{u}||_2^2 \quad \text{s.t. } \mathbf{1}_M^T\mathbf{f} = 1, \tag{8.17b}$$

whose closed-form solution is:

$$\hat{\mathbf{u}}_k = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y}_k\mathbf{f} = \mathbf{B}^\dagger\mathbf{Y}_k\mathbf{f}. \tag{8.18}$$

8.2.4 Proof of eq. (4.113)

In order to derive (4.113), we need to firstly expand $\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})$ in (4.112). Let an alternative expression of $\hat{\mathbf{C}}_k(\mathbf{H},\mathbf{f})$ (see (4.102)) be:

$$N\hat{\mathbf{C}}_k(\mathbf{H},\mathbf{f}) = \mathbf{Y}_k^T\mathbf{P}_H^\perp\mathbf{Y}_k + (\mathbf{I} - \mathbf{f}\mathbf{1}^T)^T\mathbf{Y}_k^T\mathbf{P}_H\mathbf{Y}_k(\mathbf{I} - \mathbf{f}\mathbf{1}^T), \tag{8.19}$$

which is easily derived from the fact that $\mathbf{Y}_k = (\mathbf{I} - \mathbf{P}_H)\mathbf{Y}_k + \mathbf{P}_H\mathbf{Y}_k$. Plugging (8.19) into (4.112) yields:

$$\mathrm{tr}\left(\mathbf{Z}_i\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})\right) = \frac{1}{K}\,\mathrm{tr}\left(\mathbf{Z}_i\sum_{k=1}^{K}\hat{\mathbf{C}}_k(\mathbf{H},\mathbf{f})\right) = \tag{8.20a}$$

$$\frac{1}{KN}\sum_{k=1}^{K}\mathrm{tr}\left(\mathbf{Z}_i\left(\mathbf{Y}_k^T(\mathbf{I} - \mathbf{H}\mathbf{H}^T)\mathbf{Y}_k + (\mathbf{I} - \mathbf{f}\mathbf{1}^T)^T\mathbf{Y}_k^T\mathbf{H}\mathbf{H}^T\mathbf{Y}_k(\mathbf{I} - \mathbf{f}\mathbf{1}^T)\right)\right) = \tag{8.20b}$$

$$\frac{1}{KN}\sum_{k=1}^{K}\mathrm{tr}\left(\mathbf{H}^T\left(-\mathbf{Y}_k\mathbf{Z}_i\mathbf{Y}_k^T + \mathbf{Y}_k(\mathbf{I} - \mathbf{f}\mathbf{1}^T)\mathbf{Z}_i(\mathbf{I} - \mathbf{f}\mathbf{1}^T)^T\mathbf{Y}_k^T\right)\mathbf{H}\right) + \mathrm{tr}(\mathbf{Y}_k\mathbf{Z}_i\mathbf{Y}_k^T), \tag{8.20c}$$

whose last additive term can be dropped from the iterative criterion since it does not depend on $\mathbf{H}$ or $\mathbf{f}$. After ignoring this last term and rearranging the remaining ones, we get the following expression:

$$\frac{1}{KN}\sum_{k=1}^{K}\mathrm{tr}\left(\mathbf{H}^T\left(\mathbf{Y}_k((\mathbf{I} - \mathbf{f}\mathbf{1}^T)\mathbf{Z}_i(\mathbf{I} - \mathbf{f}\mathbf{1}^T)^T - \mathbf{Z}_i)\mathbf{Y}_k^T\right)\mathbf{H}\right) = \tag{8.21a}$$

$$\frac{1}{KN}\sum_{k=1}^{K}\mathrm{tr}\left(\mathbf{H}^T\mathbf{Y}_k(\mathbf{1}^T\mathbf{Z}_i\mathbf{1}\mathbf{f}\mathbf{f}^T - 2\mathrm{sym}(\mathbf{Z}_i\mathbf{1}\mathbf{f}^T))\mathbf{Y}_k^T\mathbf{H}\right), \tag{8.21b}$$

where $\mathrm{sym}(\mathbf{A}) = \frac{\mathbf{A}+\mathbf{A}^T}{2}$. We get (4.113) after exploiting the transpose property of the trace, i.e. $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^T)$, in (4.113).

8.2.5 Proof of (4.117)

The majorant of the cost function with respect to $\mathbf{f}$ can be expanded as follows:

$$g_f(\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i,\mathbf{Z}_i) = g(\mathbf{H}_i,\mathbf{f}|\mathbf{Z}_i) = \sum_{k=1}^{K} \mathrm{tr}\left(\mathbf{H}^T\mathbf{Y}_k\left(\mathbf{1}^T\mathbf{Z}_i\mathbf{1}\mathbf{f}\mathbf{f}^T - 2\mathbf{Z}_i\mathbf{1}\mathbf{f}^T\right)\mathbf{Y}_k^T\mathbf{H}\right) = \tag{8.22a}$$

$$\sum_{k=1}^{K} \mathbf{f}^T\mathbf{Y}_k^T\mathbf{H}_i\mathbf{H}_i^T\mathbf{Y}_k\mathbf{f}\mathbf{1}^T\mathbf{Z}_i\mathbf{1} - 2\mathbf{f}^T\mathbf{Y}_k^T\mathbf{H}_i\mathbf{H}_i^T\mathbf{Y}_k\mathbf{Z}_i\mathbf{1} = \tag{8.22b}$$

$$\mathbf{f}^T\left(\sum_{k=1}^{K}\mathbf{Y}_k^T\mathbf{H}_i\mathbf{H}_i^T\mathbf{Y}_k\right)\mathbf{f}\mathbf{1}^T\mathbf{Z}_i\mathbf{1} - 2\mathbf{f}^T\left(\sum_{j=1}^{K}\mathbf{Y}_j^T\mathbf{H}_i\mathbf{H}_i^T\mathbf{Y}_j\right)\mathbf{Z}_i\mathbf{1} = \mathbf{f}^T\mathbf{D}_i\mathbf{f}\mathbf{1}^T\mathbf{Z}_i\mathbf{1} - 2\mathbf{f}^T\mathbf{D}_i\mathbf{Z}_i\mathbf{1}, \tag{8.22c}$$

where we denoted $\sum_{k=1}^{K}\mathbf{Y}_k^T\mathbf{H}_i\mathbf{H}_i^T\mathbf{Y}_k$ as $\mathbf{D}_i$.

8.2.6 Proof of the unboundedness of (4.125)

In light of proving the unboundedness of S, we need to show that $f(\mathbf{H},\mathbf{f}) \to -\infty$ for any feasible $\mathbf{f}$ such that $||\mathbf{f}||_2^2 \to \infty$. Notice that any $\mathbf{f} \in F$ has the following form $\mathbf{f} = \mathbf{P}_1^{\perp}\mathbf{x} + \frac{1}{M}\mathbf{1}_M$, where $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{P}_1^{\perp} = \mathbf{I}_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^T$. Then, provided that $||\mathbf{x}||_2^2 \to \infty \implies ||\mathbf{f}||_2^2 \to \infty$, we get the following limit:

$$\mathbf{C}_k(\mathbf{H},\mathbf{f}) \xrightarrow{||\mathbf{f}||_2^2\to\infty} \mathbf{f}^T\mathbf{Y}_k^T\mathbf{P}_H\mathbf{Y}_k\mathbf{f}\mathbf{1}\mathbf{1}^T, \tag{8.23}$$

which implies that the summation $\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f}) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{C}_k(\mathbf{H},\mathbf{f})$ is proportional to $\mathbf{1}_M\mathbf{1}_M^T$ in the limit. Thus, $\log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f}))) \to -\infty$ due to the rank-one nature of $\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})$ for $||\mathbf{f}||_2^2 \to \infty$.

8.2.7 Proof of equation (4.140)

After rearranging terms in (4.139b), we get:

$$-\frac{KN}{2}\log(\det(\mathbf{Q})) - \frac{KN}{2}\mathrm{tr}\left(\mathbf{Q}^{-1}\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f})\right) - \frac{v+M+1}{2}\log(\det(\mathbf{Q})) - \frac{\beta}{2}\mathrm{tr}(\mathbf{Q}^{-1}) = \tag{8.24a}$$

$$-\frac{1}{2}\left((KN+v+M+1)\log(\det(\mathbf{Q})) + \mathrm{tr}\left(\mathbf{Q}^{-1}\left(KN\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f}) + \beta\mathbf{I}_M\right)\right)\right) = \tag{8.24b}$$

$$-\frac{KN}{2}\left(\left(1+\frac{v+M+1}{KN}\right)\log(\det(\mathbf{Q})) + \mathrm{tr}\left(\mathbf{Q}^{-1}\left(\hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f}) + \frac{\beta}{KN}\mathbf{I}_M\right)\right)\right) = \tag{8.24c}$$

$$-\frac{KN}{2}\left(\left(1+\frac{v+M+1}{KN}\right)\log(\det(\mathbf{Q})) + \mathrm{tr}\left(\mathbf{Q}^{-1}\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f})\right)\right), \tag{8.24d}$$

where:

$$\tilde{\mathbf{Q}}(\mathbf{B},\mathbf{f}) = \hat{\mathbf{Q}}_{ML}(\mathbf{B},\mathbf{f}) + \frac{\beta}{KN}\mathbf{I}_M. \tag{8.25}$$

8.2.8 Proof of equation (4.158)

Assume that $\mathbf{B}$ belongs to $\mathrm{St}(N,D)$, i.e. $\mathbf{B}^T\mathbf{B} = \mathbf{I}_D$, and that the vector of features is statistically distributed as:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}_D, \gamma_u\mathbf{I}_D). \tag{8.26}$$

Then, the closed-form expression of the fusion MSE is derived from the following expression:

$$\gamma(\mathbf{H},\mathbf{f}) = \mathrm{E}\left[\frac{1}{N}||\mathbf{x}_k - \mathbf{P}_H\mathbf{Y}_k\mathbf{f}||_2^2\right] = \mathrm{E}\left[\frac{1}{N}\left(\mathbf{x}_k^T\mathbf{x}_k - 2\mathbf{x}_k^T\mathbf{P}_H\mathbf{Y}_k\mathbf{f} + \mathbf{f}^T\mathbf{Y}_k^T\mathbf{P}_H\mathbf{Y}_k\mathbf{f}\right)\right]. \tag{8.27}$$

After substituting $\mathbf{x}_k$ and $\mathbf{Y}_k$ by $\mathbf{B}\mathbf{u}_k$ and $\mathbf{B}\mathbf{u}_k\mathbf{1}_M^T + \mathbf{W}_k$, respectively, in the previous expression, we get:

$$\mathrm{E}\left[\frac{1}{N}\left(\mathbf{u}_k^T\mathbf{B}^T\mathbf{B}\mathbf{u}_k - 2\mathbf{u}_k^T\mathbf{B}^T\mathbf{P}_H\left(\mathbf{B}\mathbf{u}_k\mathbf{1}_M^T + \mathbf{W}_k\right)\mathbf{f}\right.\right.$$
$$\left.\left.+\mathbf{f}^T\left(\mathbf{B}\mathbf{u}_k\mathbf{1}_M^T + \mathbf{W}_k\right)^T\mathbf{P}_H\left(\mathbf{B}\mathbf{u}_k\mathbf{1}_M^T + \mathbf{W}_k\right)\mathbf{f}\right)\right] = \quad (8.28a)$$

$$\frac{1}{N}\operatorname{tr}\left(\mathbf{B}\,\mathrm{E}\left[\mathbf{u}_k\mathbf{u}_k^T\right]\mathbf{B}^T\right) - \frac{1}{N}\operatorname{tr}\left(\mathbf{P}_H\mathbf{B}\,\mathrm{E}\left[\mathbf{u}_k\mathbf{u}_k^T\right]\mathbf{B}^T\right) + \frac{1}{N}\mathbf{f}^T\,\mathrm{E}\left[\mathbf{W}_k^T\mathbf{P}_H\mathbf{W}_k\right]\mathbf{f}, \quad (8.28b)$$

where we have used the fact that $\mathbf{1}_M^T\mathbf{f} = 1$. The first two terms in (8.28b) are further rewritten taking (8.26) into consideration, yielding:

$$\gamma(\mathbf{H},\mathbf{f}) = \frac{\gamma_u}{N}\operatorname{tr}\left((\mathbf{I}_N - \mathbf{P}_H)\mathbf{B}\mathbf{B}^T\right) + \frac{1}{N}\mathbf{f}^T\,\mathrm{E}\left[\mathbf{W}_k^T\mathbf{P}_H\mathbf{W}_k\right]\mathbf{f}. \quad (8.29)$$

For clarity in the exposition, we deal with each term of (8.29) separately. An insightful expression of the first term can be derived using the principal angles between $\mathbf{H}$ and $\mathbf{B}$ (see Definition 2.11). Denoting the principal angles between $\mathbf{H}$ and $\mathbf{B}$ as $\boldsymbol{\Theta}$, the first term in (8.29) becomes:

$$\frac{\gamma_u}{N}\operatorname{tr}\left((\mathbf{I}_N - \mathbf{P}_H)\mathbf{B}\mathbf{B}^T\right) = \frac{\gamma_u}{N}\operatorname{tr}\left(\mathbf{B}\mathbf{B}^T - \mathbf{H}^T\mathbf{B}\mathbf{B}^T\mathbf{H}\right) = \frac{\gamma_u}{N}\left(D - \operatorname{tr}\left(\cos^2(\boldsymbol{\Theta})\right)\right), \quad (8.30)$$

which holds due to the initial assumptions on $\mathbf{B}$. Besides, we compute the expected value that appears in the second term from (8.29) by considering $\mathbf{V}_k = \mathbf{W}_k\mathbf{H}$ and rewriting the aforementioned covariance as follows:

$$\mathrm{E}\left[\mathbf{W}_k^T\mathbf{P}_H\mathbf{W}_k\right] = \mathrm{E}\left[\mathbf{V}_k\mathbf{V}_k^T\right]. \quad (8.31)$$

Note that the columns of $\mathbf{V}_k$ are independent and identically distributed since $\mathbf{H}$ satisfies $\mathbf{H}^T\mathbf{H} = \mathbf{I}_D$. Then, after the consideration of the previous observation, we get that:

$$\mathrm{E}\left[\mathbf{V}_k\mathbf{V}_k^T\right] = \mathrm{E}\left[\sum_{d=1}^{D}\mathbf{v}_{k,d}\mathbf{v}_{k,d}^T\right] = D\mathbf{Q}, \quad (8.32)$$

where $\mathbf{v}_{k,d}$ denotes the d-th column of $\mathbf{V}_k$. We obtain (4.158) by plugging (8.32) and (8.30) into (8.29).

8.3 Appendices of Chapter 5

8.3.1 Proof of (5.12)

The toy example consists on the assumption of the following APS:

$$\rho(\theta) = \rho_0\delta(\theta - \theta_0), \quad (8.33)$$

and a ULA in the BS. Now, let us consider the original expression of the Frobenius distance between both channel correlation matrices:

$$||\mathbf{R}_1 - \mathbf{R}_2||_F^2 = \operatorname{tr}\left(\mathbf{R}_1^T\mathbf{R}_1 - 2\Re\left(\mathbf{R}_1^H\mathbf{R}_2\right) + \mathbf{R}_2^H\mathbf{R}_2\right). \quad (8.34)$$

Then, obtaining a closed-form expression of the dot product between $\mathbf{R}_1$ and $\mathbf{R}_2$ leads to the desired result. The expression of the aforementioned dot product (under the toy example assumptions) yields:

$$\operatorname{tr}(\mathbf{R}_1^H\mathbf{R}_2) = \operatorname{tr}\left(\left(\rho_0\mathbf{a}_1(\theta_0)\mathbf{a}_1^H(\theta_0)\right)\left(\rho_0\mathbf{a}_2(\theta_0)\mathbf{a}_2^H(\theta_0)\right)\right) = \quad (8.35a)$$

$$\rho_0^2|\mathbf{a}_1^H(\theta_0)\mathbf{a}_2(\theta_0)|^2. \quad (8.35b)$$

Moreover, the dot product between $\mathbf{a}_1(\theta_0)$ and $\mathbf{a}_2(\theta_0)$ yields:

$$\mathbf{a}_1^H(\theta_0)\mathbf{a}_2(\theta_0) = \frac{1}{M}\sum_{m=0}^{M-1}\exp\left(j2\pi md\sin(\theta_0)\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\right) = \quad (8.36a)$$

$$\frac{1}{M}\frac{\exp\left(j2\pi Md\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)-1}{\exp\left(j2\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)-1}= \tag{8.36b}$$

$$\frac{\exp\left(j\pi Md\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}{\exp\left(j\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}\frac{\sin\left(\pi Md\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}{M\sin\left(\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}. \tag{8.36c}$$

After plugging (8.36c) into (8.35b), we get the following closed-form expression of the dot product between $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$\mathrm{tr}(\mathbf{R}_1^H\mathbf{R}_2)=\rho_0^2\left|\frac{\sin\left(M\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}{M\sin\left(\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}\right|^2. \tag{8.37}$$

Besides, the first and last terms in (8.34) yield:

$$\mathrm{tr}(\mathbf{R}_1^H\mathbf{R}_1)=\mathrm{tr}(\mathbf{R}_2^H\mathbf{R}_2)=\rho_0^2, \tag{8.38}$$

which follows from (8.37), after considering that $\lambda_1=\lambda_2$ in both cases. Finally, we obtain the desired expression of the Frobenius distance of said matrices after plugging (8.38) and (8.37) into (8.34):

$$||\mathbf{R}_1-\mathbf{R}_2||_F^2=2\rho_0^2\left(1-\left|\frac{\sin\left(M\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}{M\sin\left(\pi d\sin(\theta_0)\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_1}\right)\right)}\right|^2\right)= \tag{8.39a}$$

$$2\rho_0^2\left(1-\left|\frac{\sin\left(M\pi\frac{d}{\lambda_2}\sin(\theta_0)\left(1-\frac{\lambda_2}{\lambda_1}\right)\right)}{M\sin\left(\pi\frac{d}{\lambda_2}\sin(\theta_0)\left(1-\frac{\lambda_2}{\lambda_1}\right)\right)}\right|^2\right). \tag{8.39b}$$

8.3.2 Solving the ADMM update equations

8.3.2.1 Solution of (5.30a)

Solving (5.30a) is straightforward. Let us consider the aforementioned convex program:

$$\boldsymbol{\rho}_{k+1}=\arg\min_{\boldsymbol{\rho}}||\mathbf{A}_c\boldsymbol{\rho}-\mathbf{z}_{1,k}+\mathbf{u}_{1,k}||_2^2+||\boldsymbol{\rho}-\mathbf{z}_{2,k}+\mathbf{u}_{2,k}||_2^2, \tag{8.40}$$

whose solution consists on the values of $\boldsymbol{\rho}$ that set the gradient of the previous cost function to $\mathbf{0}_N$:

$$2\nabla_{\boldsymbol{\rho}}\left(||\mathbf{A}_c\boldsymbol{\rho}-\mathbf{z}_{1,k}+\mathbf{u}_{1,k}||_2^2+||\boldsymbol{\rho}-\mathbf{z}_{2,k}+\mathbf{u}_{2,k}||_2^2\right)=\mathbf{0}_N, \tag{8.41a}$$

$$\mathbf{A}_c^H(\mathbf{A}_c\boldsymbol{\rho}-\mathbf{z}_{1,k}+\mathbf{u}_{1,k})+\boldsymbol{\rho}-\mathbf{z}_{2,k}+\mathbf{u}_{2,k}=\mathbf{0}_N, \tag{8.41b}$$

$$\left(\mathbf{A}_c^H\mathbf{A}_c+\mathbf{I}_N\right)\boldsymbol{\rho}=\mathbf{A}_c^H(\mathbf{z}_{1,k}-\mathbf{u}_{1,k})+\mathbf{z}_{2,k}-\mathbf{u}_{2,k}, \tag{8.41c}$$

$$\boldsymbol{\rho}_{k+1}=\left(\mathbf{A}_c^H\mathbf{A}_c+\mathbf{I}_N\right)^{-1}\left(\mathbf{A}_c^H(\mathbf{z}_{1,k}-\mathbf{u}_{1,k})+\mathbf{z}_{2,k}-\mathbf{u}_{2,k}\right). \tag{8.41d}$$

Consequently, (8.41d) is the closed-form solution of (5.30a).

8.3.2.2 Solution of (5.30b)

As for (5.30b), we derive its solution from the following equivalent form:

$$\mathbf{z}_{1,k+1}=\arg\min_{\mathbf{z}}\frac{1}{2}||\mathbf{z}-(\mathbf{A}_c\boldsymbol{\rho}_{k+1}+\mathbf{u}_{1,k})||_2^2\quad\text{s. t. }\ ||\mathbf{z}-\hat{\mathbf{r}}_c||_2^2\leq\varepsilon. \tag{8.42}$$

In other words, the previous expression is obtained after undoing the indicator function step of the ADMM rationale (see (5.27)). For clarity of exposition, we denote $\mathbf{t}_k=\mathbf{A}_c\boldsymbol{\rho}_{k+1}+\mathbf{u}_{1,k}$. The optimization problem in (8.42) is solved by enforcing its KKT optimality conditions [29]:

$$(\mathbf{z}-\hat{\mathbf{r}}_c)-(\mathbf{t}_k-\hat{\mathbf{r}}_c)+\mu(\mathbf{z}-\hat{\mathbf{r}}_c)=\mathbf{0}_{M_c^2}, \tag{8.43a}$$

$$||\mathbf{z}-\mathbf{r}_c||_2^2 \leq \varepsilon, \tag{8.43b}$$

$$\mu \geq 0, \tag{8.43c}$$

$$\mu(||\mathbf{z}-\mathbf{r}_c||_2^2 - \varepsilon) = 0, \tag{8.43d}$$

where μ is the dual variable corresponding to the constaint in (8.42), and the left hand side of (8.43a) is an appropriate expression of the gradient of the problem's Lagrangian. Depending on the value of $\mathbf{t}_k$ with respect to the constraint in (8.42), we have two possible solutions of $\mathbf{z}$ and μ fulfilling (8.43). The first one of them is derived from the following implications

$$\mu = 0 \underbrace{\implies}_{\text{Eq. (8.43a)}} \mathbf{z} = \mathbf{t}_k \underbrace{\implies}_{\text{Eq. (8.43b)}} ||\mathbf{t}_k - \mathbf{r}_c||_2^2 \leq \varepsilon. \tag{8.44}$$

As for the remaining case, any optimal value of μ greater than 0 implies that $||\mathbf{z}-\mathbf{r}_c||_2^2 = \varepsilon$ because of (8.43d). Thus, the value of μ can be obtained by multiplying both sides of (8.43a) by $(\mathbf{z}-\mathbf{r}_c)^T$ and rearranging terms, yielding:

$$\mu = \frac{(\mathbf{t}_k - \hat{\mathbf{r}}_c)^T(\mathbf{z}-\hat{\mathbf{r}}_c)}{\varepsilon} - 1. \tag{8.45}$$

In order to obtain the optimal value of $\mathbf{z}$ for $\mu > 0$, we plug the previous expression of μ into (8.43a) and we get:

$$\frac{(\mathbf{t}_k - \hat{\mathbf{r}}_c)^T(\mathbf{z}-\hat{\mathbf{r}}_c)}{\varepsilon}(\mathbf{z}-\hat{\mathbf{r}}_c) = (\mathbf{t}_k - \hat{\mathbf{r}}_c), \tag{8.46}$$

implying that $(\mathbf{z}-\hat{\mathbf{r}}_c)$ is a scaled version of $(\mathbf{t}_k - \hat{\mathbf{r}}_c)$. Given the previous equation, an intuitive way to obtain the closed-form expression of the optimal value of $\mathbf{z}$ is to note that a consequence of the previous equation is that:

$$\mathbf{z} - \hat{\mathbf{r}}_c = \alpha(\mathbf{t}_k - \hat{\mathbf{r}}_c), \tag{8.47a}$$

$$\mathbf{z} = \hat{\mathbf{r}}_c + \alpha(\mathbf{t}_k - \hat{\mathbf{r}}_c), \tag{8.47b}$$

for some constant α such that the original constraint is satisfied. Consequently, the value of α is obtained by enforcing $||\mathbf{z}-\mathbf{r}_c||_2^2 = \varepsilon$:

$$||\hat{\mathbf{r}}_c + \alpha(\mathbf{t}_k - \hat{\mathbf{r}}_c) - \hat{\mathbf{r}}_c||_2^2 = \varepsilon, \tag{8.48a}$$

$$\alpha = \frac{\sqrt{\varepsilon}}{||\mathbf{t}_k - \hat{\mathbf{r}}_c||_2^2}. \tag{8.48b}$$

As a summary, we combine the solutions of $\mathbf{z}$ and μ obtained in (8.44) and (8.47b), yielding the following closed-form expression of $\mathbf{z}_{1,k+1}$:

$$\mathbf{z}_{1,k+1} = \frac{\sqrt{\varepsilon}}{\max(\sqrt{\varepsilon}, ||\mathbf{t}_k - \hat{\mathbf{r}}_c||_2^2)}(\mathbf{t}_k - \hat{\mathbf{r}}_c) + \hat{\mathbf{r}}_c. \tag{8.49}$$

8.3.2.3 Solution of (5.30c)

Finally, the remaining update equation is solved invoking the proximal operator of the ℓ_1 norm. By undoing the indicator function procedure, we get the following alternative expression of (5.30c):

$$\mathbf{z}_{2,k+1} = \arg\min_{\mathbf{z}} ||\mathbf{z}||_1 + \frac{\lambda}{2}||\mathbf{z} - (\boldsymbol{\rho}_{k+1} + \mathbf{u}_{2_k})||_2^2 \quad \text{s.t. } \Re(\mathbf{z}) \succeq \mathbf{0}_N, \Im(\mathbf{z}) = \mathbf{0}_N, \tag{8.50}$$

which is a constrained version of the proximal operator of the ℓ_1 norm (see Definition 3.24 and Proposition 3.13). The solution of the previous optimization problem mixes the soft-thresholding operator and the projection onto the constraint sets. Provided that the soft-thresholding operator is only defined for real numbers, the previous idea results in the following expression [26]:

$$\mathbf{z}_{2,k+1} = \max\left(\mathbf{0}_N, \text{prox}_{\ell_1, \frac{1}{\lambda}}\left(\Re(\boldsymbol{\rho}_{k+1}) + \mathbf{u}_{2_k}\right)\right). \tag{8.51}$$

8.4 Appendices of Chapter 6

8.4.1 Proof of Lemma 6.1

Provided that:

$$\mathbf{z} = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\mathbf{0}_2, \mathbf{C}), \tag{8.52}$$

the MI of X and Y is obtained straightforwardly from Definition 2.6 and the PDF of $\mathbf{z}$. This means that we have to compute the following integral:

$$I(X;Y) = \iint_{-\infty}^{\infty} p_{XY}(x,y) \log\left(\frac{p_{XY}(\mathbf{x},\mathbf{y})}{p_X(x)p_Y(y)}\right) \mathrm{d}x\mathrm{d}y, \tag{8.53}$$

where $p_{XY}(x,y)$, $p_X(x)$ and $p_Y(y)$ are the joint distribution and marginal distributions of X and Y, respectively. Notice that:

$$p_{XY}(x,y) = p_{\mathbf{z}}(\mathbf{z}) = \frac{1}{\sqrt{2\pi \det(\mathbf{C})}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{C}^{-1}\mathbf{z}\right), \tag{8.54}$$

and:

$$p_X(x)p_Y(y) = \frac{1}{\sqrt{2\pi \det(\mathbf{D})}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{D}^{-1}\mathbf{z}\right), \tag{8.55}$$

where $\mathbf{D} = \mathbf{C} \odot \mathbf{I}_2$. Then, we get:

$$\log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right) = \log\left(\sqrt{\frac{\det(\mathbf{D})}{\det(\mathbf{C})}}\right) - \frac{1}{2}\mathbf{z}^T(\mathbf{C}^{-1} - \mathbf{D}^{-1})\mathbf{z} = -\frac{1}{2}\log(\mathbf{C}\mathbf{D}^{-1}) - \frac{1}{2}\mathbf{z}^T(\mathbf{C}^{-1} - \mathbf{D}^{-1})\mathbf{z}. \tag{8.56}$$

After plugging the previous result into (8.53), we obtain:

$$I(X;Y) = \iint_{-\infty}^{\infty} p_{XY}(x,y)\left(-\frac{1}{2}\log(\mathbf{C}\mathbf{D}^{-1}) - \frac{1}{2}\mathbf{z}^T(\mathbf{C}^{-1} - \mathbf{D}^{-1})\mathbf{z}\right)\mathrm{d}x\mathrm{d}y = \tag{8.57a}$$

$$\mathrm{E}\left[-\frac{1}{2}\log(\mathbf{C}\mathbf{D}^{-1}) - \frac{1}{2}\mathbf{z}^T(\mathbf{C}^{-1} - \mathbf{D}^{-1})\mathbf{z}\right] = -\frac{1}{2}\log(\mathbf{C}\mathbf{D}^{-1}) - \frac{1}{2}\operatorname{tr}\left((\mathbf{C}^{-1} - \mathbf{D}^{-1})\,\mathrm{E}[\mathbf{z}\mathbf{z}^T]\right) = \tag{8.57b}$$

$$-\frac{1}{2}\log(\mathbf{C}\mathbf{D}^{-1}) - \frac{1}{2}\operatorname{tr}\left((\mathbf{C}^{-1} - \mathbf{D}^{-1})\mathbf{C}\right) = -\frac{1}{2}\log(\mathbf{C}\mathbf{D}^{-1}) - \frac{1}{2}\operatorname{tr}\left(\mathbf{I}_2 - \mathbf{C}\mathbf{D}^{-1}\right), \tag{8.57c}$$

where the previous expected value is taken with respect to the PDF of $\mathbf{z}$. The last expression in (8.57c) are further parsed after noting that:

$$\mathbf{C}\mathbf{D}^{-1} = \mathbf{D}^{-\frac{1}{2}}\mathbf{C}\mathbf{D}^{-\frac{1}{2}} = \mathbf{\Xi} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{8.58}$$

which is the coherence matrix of X and Y and ρ is their respective Pearson correlation coefficient, and that:

$$\mathbf{I}_2 - \mathbf{C}\mathbf{D}^{-1} = \begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix}. \tag{8.59}$$

As a result, the second term in (8.57c) is equal to 0, so the desired expression of the MI is:

$$I(X;Y) = -\frac{1}{2}\log(\det(\mathbf{\Xi})) = -\frac{1}{2}\log(1-\rho^2). \tag{8.60}$$

Bibliography

[1] 3GPP, "Spatial channel model for Multiple Input Multiple Output (MIMO) simulations", 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 25.996, Jul. 2020, Version 16.0.0.

[2] M. Abbasi, A. Y. Kruger, and M. Théra, "Gateaux Differentiability Revisited", *Applied Mathematics & Optimization*, vol. 84, no. 3, pp. 3499–3516, Feb. 2021. DOI: 10.1007/s00245-021-09754-y.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre, "Riemannian Geometry of Grassmann Manifolds with a View on Algorithmic Computation", *Acta Applicandae Mathematicae*, vol. 80, no. 2, pp. 199–220, Jan. 2004. DOI: 10.1023/b:acap.0000013855.14971.91.

[4] S. Adhikari, "Matrix Variate Distributions for Probabilistic Structural Dynamics", *AIAA Journal*, vol. 45, no. 7, pp. 1748–1762, Jul. 2007. DOI: 10.2514/1.25512.

[5] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection", in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6. DOI: 10.1109/MLSP.2013.6661898.

[6] P. del Aguila Pla, "Inverse problems in signal processing: Functional optimization, parameter estimation and machine learning", Ph.D. dissertation, KTH Royal Institute of Technology, 2019.

[7] K. Ahn and F. Suarez, "Riemannian Perspective on Matrix Factorization", *ArXiv*, vol. abs/2102.00937, 2021.

[8] H. Akaike, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974. DOI: 10.1109/TAC.1974.1100705.

[9] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy, "Level-set methods for convex optimization", *Mathematical Programming*, vol. 174, no. 1–2, pp. 359–390, Dec. 2018. DOI: 10.1007/s10107-018-1351-8.

[10] E. Arias-Castro, S. Bubeck, and G. Lugosi, "Detection of correlations", *The Annals of Statistics*, vol. 40, no. 1, pp. 412–435, 2012. DOI: 10.1214/11-AOS964.

[11] N. Asendorf and R. R. Nadakuditi, "Improved Detection of Correlated Signals in Low-Rank-Plus-Noise Type Data Sets Using Informative Canonical Correlation Analysis (ICCA)", *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3451–3467, 2017. DOI: 10.1109/TIT.2017.2695601.

[12] D. Aydin and M. S. Tuzemen, "Estimation in Semi-parametric and Additive Regression Using Smoothing and Regression Spline", in *2010 Second International Conference on Computer Research and Development*, 2010, pp. 465–469. DOI: 10.1109/ICCRD.2010.101.

[13] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian Compressive Sensing Using Laplace Priors", *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, 2010. DOI: 10.1109/TIP.2009.2032894.

[14] T. Bailey, S. Julier, and G. Agamennoni, "On conservative fusion of information with unknown non-Gaussian dependence", in *2012 15th International Conference on Information Fusion*, 2012, pp. 1876–1883.

[15] J. Baker, "Strong Convexity Does Not Imply Radial Unboundedness", *The American Mathematical Monthly*, vol. 123, no. 2, pp. 185–188, 2016.

[16] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information", in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 704–711. DOI: 10.1109/ALLERTON.2010.5706976.

[17] R. Baraldi, R. Kumar, and A. Aravkin, "Basis Pursuit Denoise With Nonsmooth Constraints", *IEEE Transactions on Signal Processing*, vol. 67, no. 22, pp. 5811–5823, 2019. DOI: 10.1109/TSP.2019.2946029.

[18] E. Batzies, K. Hüper, L. Machado, and F. S. Leite, "Geometric mean and geodesic regression on Grassmannians", *Linear Algebra and its Applications*, vol. 466, pp. 83–101, 2015. DOI: https://doi.org/10.1016/j.laa.2014.10.003.

[19] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems", *SIAM J. Imaging Sci.*, vol. 2, pp. 183–202, 2009.

[20] T. Bendokat, R. Zimmermann, and P.-A. Absil, "A grassmann manifold handbook: Basic geometry and computational aspects", *Advances in Computational Mathematics*, vol. 50, no. 1, Jan. 2024. DOI: 10.1007/s10444-023-10090-8.

[21] E. van den Berg and M. P. Friedlander, "Probing the Pareto Frontier for Basis Pursuit Solutions", *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2009. DOI: 10.1137/080714488. eprint: https://doi.org/10.1137/080714488.

[22] H. Bhat and N. Kumar, "On the Derivation of the Bayesian Information Criterion", Jan. 2010.

[23] R. Bhatia and F. Kittaneh, "Notes on matrix arithmetic–geometric mean inequalities", *Linear Algebra and its Applications*, vol. 308, no. 1, pp. 203–211, 2000. DOI: https://doi.org/10.1016/S0024-3795(00)00048-3.

[24] S. Bian and L. Zhang, "Overview of Match Pursuit Algorithms and Application Comparison in Image Reconstruction", in *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 2021, pp. 216–221. DOI: 10.1109/IPEC51340.2021.9421295.

[25] J. Bibby, J. Kent, and K. Mardia, "Multivariate analysis", *Academic Press, London*, 1979.

[26] J. M. Bioucas-Dias and M. A. T. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing", in *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2010, pp. 1–4. DOI: 10.1109/WHISPERS.2010.5594963.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[28] R. Bishop and B. O'Neill, "Manifolds of negative curvature", *Transactions of the American Mathematical Society*, vol. 145, pp. 1–49, Nov. 1969. DOI: 10.1090/S0002-9947-1969-0251664-4.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[30] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Found. Trends Mach. Learn.*, vol. 3, pp. 1–122, 2011. DOI: 10.1561/2200000016.

[31] A. Breloy, S. Kumar, Y. Sun, and D. P. Palomar, "Majorization-Minimization on the Stiefel Manifold With Application to Robust Sparse PCA", *IEEE Transactions on Signal Processing*, vol. 69, pp. 1507–1520, 2021. DOI: 10.1109/TSP.2021.3058442.

[32] A. Browder, *Mathematical Analysis: An Introduction*. New York: Springer-Verlag, 1996.

[33] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images", *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009. DOI: 10.1137/060657704. eprint: https://doi.org/10.1137/060657704.

[34] L. Burusheva and V. Temlyakov, *Sparse approximation of individual functions*, 2019. arXiv: 1911.02593.

[35] F. d. Cabrera and J. Riba, "Entropy-Based Non-Data-Aided SNR Estimation", in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 731–735. DOI: 10.1109/IEEECONF44664.2019.9048732.

[36] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems", *Information Sciences*, vol. 328, pp. 26–41, 2016. DOI: https://doi.org/10.1016/j.ins.2015.08.029.

[37] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006. DOI: 10.1109/TIT.2005.862083.

[38] W. Cao, A. Dytso, M. Fauss, G. Feng, and H. V. Poor, "Robust Waterfilling for Approximately Gaussian Inputs", in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9013311.

[39] R. L. G. Cavalcante, L. Miretti, and S. Stańczak, "Error Bounds for FDD Massive MIMO Channel Covariance Conversion with Set-Theoretic Methods", in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–7. DOI: 10.1109/GLOCOM.2018.8647609.

[40] J. E. Cavanaugh, "Unifying the derivations for the Akaike and corrected Akaike information criteria", *Statistics & Probability Letters*, vol. 33, no. 2, pp. 201–208, 1997. DOI: https://doi.org/10.1016/S0167-7152(96)00128-9.

[41] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements", *WIREs Computational Statistics*, vol. 11, no. 3, Mar. 2019. DOI: 10.1002/wics.1460.

[42] G. Chan, "Effects of sectorization on the spectrum efficiency of cellular radio systems", *IEEE Transactions on Vehicular Technology*, vol. 41, no. 3, pp. 217–225, 1992. DOI: 10.1109/25.155968.

[43] L. Chen, P. Arambel, and R. Mehra, "Fusion under unknown correlation - covariance intersection as a special case", in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 2, 2002, 905–912 vol.2. DOI: 10.1109/ICIF.2002.1020908.

[44] S. Chen, S. Ma, A. Man-Cho So, and T. Zhang, "Proximal Gradient Method for Nonsmooth Optimization over the Stiefel Manifold", *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 210–239, 2020. DOI: 10.1137/18M122457X. eprint: https://doi.org/10.1137/18M122457X.

[45] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization", in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1828–1832. DOI: 10.1109/IJCNN.2008.4634046.

[46] A. Choudary and C. P. Niculescu, *Real Analysis on Intervals*, en. New Delhi, India: Springer, Aug. 2016.

[47] G. Cioffi and D. Scaramuzza, "Tightly-coupled Fusion of Global Positional Measurements in Optimization-based Visual-Inertial Odometry", in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5089–5095. DOI: 10.1109/IROS45743.2020.9341697.

[48] M. Danilova, P. E. Dvurechensky, A. V. Gasnikov, *et al.*, "Recent Theoretical Advances in Non-Convex Optimization", *ArXiv*, vol. abs/2012.06188, 2020.

[49] Y. Dar, P. Mayer, L. Luzi, and R. Baraniuk, "Subspace Fitting Meets Regression: The Effects of Supervision and Orthonormality Constraints on Double Descent of Generalization Errors", *ArXiv*, vol. abs/2002.10614, 2020.

[50] F. De Cabrera, "Data-driven information-theoretic tools under a second-order statistics perspective", Ph.D. dissertation, Universitat Politècnica de Catalunya, Jul. 2023.

[51] F. De Cabrera and J. Riba, "Regularized Estimation of Information via Canonical Correlation Analysis on a Finite-Dimensional Feature Space", *IEEE Transactions on Information Theory*, vol. 69, no. 8, pp. 5135–5150, 2023. DOI: 10.1109/TIT.2023.3258182.

[52] F. De Cabrera, J. Riba, and G. Vázquez, "Entropy-based covariance determinant estimation", in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2017, pp. 1–5. DOI: 10.1109/SPAWC.2017.8227751.

[53] P. S. R. Domenico Ciuonzo, *Data Fusion in Wireless Sensor Networks: A statistical signal processing perspective*. Institution of Engineering and Technology, Mar. 2019. DOI: 10.1049/pbce117e.

[54] X. G. Doukopoulos and G. V. Moustakides, "Fast and Stable Subspace Tracking", *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1452–1465, 2008. DOI: 10.1109/TSP.2007.909335.

[55] J. Duchi, *Sequential Convex Programming: Notes for EE364b*, Spring 2018.

[56] A. Eamaz, F. Yeganegi, and M. Soltanalian, "On the Building Blocks of Sparsity Measures", *IEEE Signal Processing Letters*, vol. 29, pp. 2667–2671, 2022. DOI: 10.1109/LSP.2022.3233000.

[57] A. Edelman, T. A. Arias, and S. Smith, "The Geometry of Algorithms with Orthogonality Constraints", *SIAM J. Matrix Anal. Appl.*, vol. 20, pp. 303–353, 1998.

[58] M. Elad, "Sparse and Redundant Representation Modeling—What Next?", *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 922–928, 2012. DOI: 10.1109/LSP.2012.2224655.

[59] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian Antisparse Coding", *IEEE Transactions on Signal Processing*, vol. 65, no. 7, pp. 1660–1672, 2017. DOI: 10.1109/TSP.2016.2645543.

[60] D. Erdogmus and J. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems", *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1780–1786, 2002. DOI: 10.1109/TSP.2002.1011217.

[61] T. van Erven and P. Harremos, "Rényi Divergence and Kullback-Leibler Divergence", *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014. DOI: 10.1109/TIT.2014.2320500.

[62] J. Fan and Y. Zhong, *Optimal Subspace Estimation Using Overidentifying Vectors via Generalized Method of Moments*, 2018. arXiv: 1805.02826 [stat.ME].

[63] J. Fan, T. Hu, Q. Wu, and D.-X. Zhou, "Consistency analysis of an empirical minimum error entropy algorithm", *Applied and Computational Harmonic Analysis*, vol. 41, no. 1, pp. 164–189, 2016, Sparse Representations with Applications in Imaging Science, Data Analysis and Beyond. DOI: https://doi.org/10.1016/j.acha.2014.12.005.

[64] V. Fasen, "Statistical inference of spectral estimation for continuous-time MA processes with finite second moments", *Mathematical Methods of Statistics*, vol. 22, no. 4, pp. 283–309, Oct. 2013. DOI: 10.3103/s1066530713040029.

[65] M. Fazel, H. Hindi, and S. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices", in *Proceedings of the 2003 American Control Conference, 2003.*, vol. 3, 2003, 2156–2162 vol.3. DOI: 10.1109/ACC.2003.1243393.

[66] M. C. Filippou, D. Gesbert, and G. A. Ropokis, "Optimal Combining of Instantaneous and Statistical CSI in the SIMO Interference Channel", in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5. DOI: 10.1109/VTCSpring.2013.6692676.

[67] D. Fink, "A Compendium of Conjugate Priors", 1997.

[68] M. Frank and P. Wolfe, "An algorithm for quadratic programming", *Naval Research Logistics Quarterly*, vol. 3, no. 1–2, pp. 95–110, Mar. 1956. DOI: 10.1002/nav.3800030109.

[69] A. Galántai and C. J. Hegedűs, "Jordan's principal angles in complex vector spaces", *Numerical Linear Algebra with Applications*, vol. 13, no. 7, pp. 589–598, 2006. DOI: 10.1002/nla.491.

[70] K. Gallivan, A. Srivastava, X. Liu, and P. Van Dooren, "Efficient algorithms for inferences on Grassmann manifolds", in *IEEE Workshop on Statistical Signal Processing, 2003*, 2003, pp. 315–318. DOI: 10.1109/SSP.2003.1289408.

[71] J. Gillard and K. Usevich, "Structured low-rank matrix completion for forecasting in time series analysis", *International Journal of Forecasting*, vol. 34, no. 4, pp. 582–597, 2018. DOI: https://doi.org/10.1016/j.ijforecast.2018.03.008.

[72] M. Godavarti and A. O. Hero, "Diversity and degrees of freedom in wireless communications", in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2002, pp. III-2861-III–2864. DOI: 10.1109/ICASSP.2002.5745245.

[73] T. Goldstein, C. Studer, and R. Baraniuk, "A Field Guide to Forward-Backward Splitting with a FASTA Implementation", *ArXiv*, vol. abs/1411.3406, 2014.

[74] M. Grant, S. Boyd, and Y. Ye, "Disciplined Convex Programming", in *Global Optimization: From Theory to Implementation*, L. Liberti and N. Maculan, Eds. Boston, MA: Springer US, 2006, pp. 155–210. DOI: 10.1007/0-387-30528-9_7.

[75] Q. Gu, *Lecture notes in Optimization*, Sep. 2016.

[76] S. Ji-guang, "Perturbation of angles between linear subspaces", *Journal of Computational Mathematics*, vol. 5, no. 1, pp. 58–61, 1987.

[77] S. Han and J. Principe, "A Fixed-Point Minimum Error Entropy Algorithm", in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2006, pp. 167–172. DOI: 10.1109/MLSP.2006.275542.

[78] P. Harremoës, "Interpretations of Rényi entropies and divergences", *Physica A: Statistical Mechanics and its Applications*, vol. 365, no. 1, pp. 57–62, 2006, Fundamental Problems of Modern Statistical Mechanics. DOI: https://doi.org/10.1016/j.physa.2006.01.012.

[79] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The LASSO and Generalizations*. May 2015, pp. 1–337. DOI: 10.1201/b18401.

[80] R. Horst, "On the global minimization of concave functions", *Operations-Research-Spektrum*, vol. 6, no. 4, pp. 195–205, Dec. 1984. DOI: 10.1007/BF01720068.

[81] X.-L. Hu, J. Wen, W. K. Wong, L. Tong, and J. Cui, "On uniqueness of sparse signal recovery", *Signal Processing*, vol. 150, pp. 66–74, Sep. 2018. DOI: 10.1016/j.sigpro.2018.04.002.

[82] S. Huang, D. N. Tran, and T. D. Tran, "Sparse signal recovery based on nonconvex entropy minimization", in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3867–3871. DOI: 10.1109/ICIP.2016.7533084.

[83] S. Huang and T. D. Tran, "Sparse Signal Recovery via Generalized Entropy Functions Minimization", *IEEE Transactions on Signal Processing*, vol. 67, no. 5, pp. 1322–1337, 2019. DOI: 10.1109/TSP.2018.2889951.

[84] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, "On entropy approximation for Gaussian mixture random vectors", in *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2008, pp. 181–188. DOI: 10.1109/MFI.2008.4648062.

[85] M. Hubert, M. Debruyne, and P. J. Rousseeuw, "Minimum covariance determinant and extensions", *WIREs Computational Statistics*, vol. 10, no. 3, Dec. 2017. DOI: 10.1002/wics.1421.

[86] K. Hugl, J. Laurila, and E. Bonek, "Downlink beamforming for frequency division duplex systems", in *Seamless Interconnection for Universal Services. Global Telecommunications Conference. GLOBECOM'99. (Cat. No.99CH37042)*, vol. 4, 1999, 2097–2101 vol.4. DOI: 10.1109/GLOCOM.1999.827574.

[87] D. R. Hunter and K. L. Lange, "A Tutorial on MM Algorithms", *The American Statistician*, vol. 58, pp. 30–37, 2004.

[88] M. Hurley, "An information theoretic justification for covariance intersection and its generalization", in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 1, 2002, 505–511 vol.1. DOI: 10.1109/ICIF.2002.1021196.

[89] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples", *Biometrika*, vol. 76, pp. 297–307, 1989.

[90] H. Iiduka, "Fixed point optimization algorithm and its application to network bandwidth allocation", *Journal of Computational and Applied Mathematics*, vol. 236, no. 7, pp. 1733–1742, 2012. DOI: https://doi.org/10.1016/j.cam.2011.10.004.

[91] A. J. Izenman, "Reduced-rank regression for the multivariate linear model", *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975. DOI: https://doi.org/10.1016/0047-259X(75)90042-1.

[92] M. W. Jacobson and J. A. Fessler, "An Expanded Theoretical Treatment of Iteration-Dependent Majorize-Minimize Algorithms", *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2411–2422, 2007. DOI: 10.1109/TIP.2007.904387.

[93] M. W. Jacobson and J. A. Fessler, "Properties of MM Algorithms on Convex Feasible Sets: Extended Version", 2004.

[94] M. Jaggi, "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization", in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., ser. Proceedings of Machine Learning Research, vol. 28, Atlanta, Georgia, USA: PMLR, Jun. 2013, pp. 427–435.

[95] L. Jing, M. K. Ng, and T. Zeng, "Dictionary Learning-Based Subspace Structure Identification in Spectral Clustering", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 8, pp. 1188–1199, 2013. DOI: 10.1109/TNNLS.2013.2253123.

[96] I. T. Jolliffe, "Principal Component Analysis and Factor Analysis", in *Principal Component Analysis*. New York, NY: Springer New York, 1986, pp. 115–128. DOI: 10.1007/978-1-4757-1904-8_7.

[97] I. Jolliffe and J. Cadima, "Principal Component Analysis: A review and recent developments", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20 150 202, Apr. 2016. DOI: 10.1098/rsta.2015.0202.

[98] M. Jordan, A. Dimofte, X. Gong, and G. Ascheid, "Conversion from Uplink to Downlink Spatio-Temporal Correlation with Cubic Splines", in *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*, 2009, pp. 1–5. DOI: 10.1109/VETECS.2009.5073462.

[99] S. Julier and J. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations", in *Proceedings of the 1997 American Control Conference (Cat. No.97CH36041)*, vol. 4, 1997, 2369–2373 vol.4. DOI: 10.1109/ACC.1997.609105.

[100] S. Julier and J. K. Uhlmann, "Multisensor Data Fusion", in CRC Press, Jun. 2001, ch. General Decentralized Data Fusion with Covariance Intersection (CI).

[101] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*. Pearson Education, 1993.

[102] M. Kayaalp, Y. İnan, E. Telatar, and A. H. Sayed, "On the Arithmetic and Geometric Fusion of Beliefs for Distributed Inference", *IEEE Transactions on Automatic Control*, pp. 1–16, 2023. DOI: 10.1109/TAC.2023.3330405.

[103] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art", *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013. DOI: https://doi.org/10.1016/j.inffus.2011.08.001.

[104] A. V. Knyazev and M. E. Argentati, "Principal Angles between Subspaces in an A-Based Scalar Product: Algorithms and Perturbation Estimates", *SIAM Journal on Scientific Computing*, vol. 23, no. 6, pp. 2008–2040, 2002. DOI: 10.1137/S1064827500377332. eprint: https://doi.org/10.1137/S1064827500377332.

[105] B. Kommineni, S. Basu, and R. Vemuri, "A spline based regression technique on interval valued noisy data", in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 2007, pp. 241–247. DOI: 10.1109/ICMLA.2007.100.

[106] J. Kovacevic and A. Chebira, "Life Beyond Bases: The Advent of Frames (Part I)", *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 86–104, 2007. DOI: 10.1109/MSP.2007.4286567.

[107] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information", *Phys. Rev. E*, vol. 69, p. 066 138, 6 Jun. 2004. DOI: 10.1103/PhysRevE.69.066138.

[108] S. Kullback and R. A. Leibler, "On Information and Sufficiency", *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951. DOI: 10.1214/aoms/1177729694.

[109] S. Lacoste-Julien, "Convergence Rate of Frank-Wolfe for Non-Convex Objectives", *ArXiv*, vol. abs/1607.00345, 2016.

[110] D. Lahat, T. Adali, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects", *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015. DOI: 10.1109/JPROC.2015.2460697.

[111] T. Leonard and J. S. J. Hsu, "Bayesian Inference for a Covariance Matrix", *The Annals of Statistics*, vol. 20, no. 4, pp. 1669–1696, 1992. DOI: 10.1214/aos/1176348885.

[112] T. Li, H. Fan, J. García, and J. M. Corchado, "Second-order statistics analysis and comparison between arithmetic and geometric average fusion: Application to multi-sensor target tracking", *Information Fusion*, vol. 51, pp. 233–243, 2019. DOI: https://doi.org/10.1016/j.inffus.2019.02.009.

[113] T. Li and F. Hlawatsch, "A distributed particle-PHD filter using arithmetic-average fusion of Gaussian mixture parameters", *Information Fusion*, vol. 73, pp. 111–124, 2021. DOI: https://doi.org/10.1016/j.inffus.2021.02.020.

[114] T. Li, X. Wang, Y. Liang, and Q. Pan, "On Arithmetic Average Fusion and Its Application for Distributed Multi-Bernoulli Multitarget Tracking", *IEEE Transactions on Signal Processing*, vol. 68, pp. 2883–2896, 2020. DOI: 10.1109/TSP.2020.2985643.

[115] Y.-C. Liang and F. Chin, "Downlink channel covariance matrix (DCCM) estimation and its applications in wireless DS-CDMA systems", *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 222–232, 2001. DOI: 10.1109/49.914500.

[116] F. Liese and I. Vajda, "On Divergences and Informations in Statistics and Information Theory", *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006. DOI: 10.1109/TIT.2006.881731.

[117] "Line Search Methods", in *Numerical Optimization*. New York, NY: Springer New York, 2006, pp. 30–65. DOI: 10.1007/978-0-387-40065-5_3.

[118] W. Liu, J. C. Príncipe, and S. Haykin, *Kernel Adaptive Filtering. A Comprehensive Introduction*. Wiley, 2010.

[119] Y. Liu, Y. Yan, L. You, W. Wang, and H. Duan, "Spatial Covariance Matrix Reconstruction for DOA Estimation in Hybrid Massive MIMO Systems With Multiple Radio Frequency Chains", *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 12 185–12 190, 2021. DOI: 10.1109/TVT.2021.3113018.

[120] C. A. Lopez, F. de Cabrera, and J. Riba, "Estimation of Information in Parallel Gaussian Channels via Model Order Selection", in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5675–5679. DOI: 10.1109/ICASSP40776.2020.9053506.

[121] C. A. Lopez, F. de Cabrera, and J. Riba, "Minimum Error Entropy Estimation Under Contaminated Gaussian Noise", *IEEE Signal Processing Letters*, vol. 30, pp. 1457–1461, 2023. DOI: 10.1109/LSP.2023.3324295.

[122] C. A. Lopez and J. Riba, "Data Driven Joint Sensor Fusion and Regression Based on Geometric Mean Squared Error", in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095018.

[123] C. A. Lopez and J. Riba, "On the Convergence of Block Majorization-Minimization Algorithms on the Grassmann Manifold", *IEEE Signal Processing Letters*, vol. 31, pp. 1314–1318, 2024. DOI: 10.1109/LSP.2024.3396660.

[124] C. A. Lopez and J. Riba, "Parametric Minimum Error Entropy Criterion: A Case Study in Blind Sensor Fusion and Regression Problems", *IEEE Transactions on Signal Processing*, vol. 72, pp. 5091–5106, 2024. DOI: 10.1109/TSP.2024.3488554.

[125] C. A. Lopez and J. Riba, "Sparse-Aware Approach for Covariance Conversion in FDD Systems", in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1726–1730. DOI: 10.23919/EUSIPCO55093.2022.9909956.

[126] R. Mahler, "Multitarget Bayes filtering via first-order multitarget moments", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003. DOI: 10.1109/TAES.2003.1261119.

[127] A. Mariani, A. Giorgetti, and M. Chiani, "Model Order Selection Based on Information Theoretic Criteria: Design of the Penalty", *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2779–2789, 2015. DOI: 10.1109/TSP.2015.2414900.

[128] I. Markovsky, "Structured low-rank approximation and its applications", *Automatica*, vol. 44, no. 4, pp. 891–909, 2008. DOI: https://doi.org/10.1016/j.automatica.2007.09.011.

[129] A. M. Mathai, S. B. Provost, and H. J. Haubold, *Multivariate Statistical Analysis in the Real and Complex Domains*. Springer International Publishing, 2022. DOI: 10.1007/978-3-030-95864-0.

[130] P. Mayilvahanan, "Estimation of Regression Coefficients Using Geometric Mean of Squared Error for Single Index Linear Regression Model", *International Journal of Artificial Intelligence & Applications*, vol. 7, pp. 75–84, 2016.

[131] V. Maz'ya and G. Schmidt, "On approximate approximations using Gaussian kernels", *IMA Journal of Numerical Analysis*, vol. 16, no. 1, pp. 13–29, 1996. DOI: 10.1093/imanum/16.1.13.

[132] M. T. MCCANN and B. Wohlberg, "Robust and Simple ADMM Penalty Parameter Selection", *IEEE Open Journal of Signal Processing*, vol. 5, pp. 402–420, 2024. DOI: 10.1109/OJSP.2023.3349115.

[133] J. Melbourne, S. Talukdar, S. Bhaban, M. Madiman, and M. V. Salapaka, "The differential entropy of mixtures: New bounds and applications", *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2123–2146, 2022. DOI: 10.1109/TIT.2022.3140661.

[134] X. Mestre and P. Vallet, "On the Resolution Probability of Conditional and Unconditional Maximum Likelihood DoA Estimation", *IEEE Transactions on Signal Processing*, vol. 68, pp. 4656–4671, 2020. DOI: 10.1109/TSP.2020.3015046.

[135] T. Minka, "Automatic Choice of Dimensionality for PCA", in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13, MIT Press, 2000.

[136] L. Miretti, R. L. Cavalcante, and S. Stanczak, "FDD Massive MIMO Channel Spatial Covariance Conversion Using Projection Methods", in *2018 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), 2018, pp. 3609–3613. DOI: 10.1109/ICASSP.2018.8462048.

[137] L. Miretti, R. L. G. Cavalcante, and S. Stańczak, "Channel Covariance Conversion and Modelling Using Infinite Dimensional Hilbert Spaces", *IEEE Transactions on Signal Processing*, vol. 69, pp. 3145–3159, 2021. DOI: 10.1109/TSP.2021.3082461.

[138] M. Mureşan, *A Concrete Approach to Classical Analysis*. Springer New York, 2009. DOI: 10.1007/978-0-387-78933-0.

[139] I. Murray and Z. Ghahramani, "A note on the evidence and Bayesian Occam's razor", Gatsby Computational Neuroscience Unit, University College London, Tech. Rep. GCNU-TR 2005-003, 2005.

[140] R. R. Nadakuditi, "Fundamental finite-sample limit of canonical correlation analysis based detection of correlated high-dimensional signals in white noise", in *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 2011, pp. 397–400. DOI: 10.1109/SSP.2011.5967714.

[141] F. Nielsen, "An Information-Geometric Characterization of Chernoff Information", *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 269–272, 2013. DOI: 10.1109/LSP.2013.2243726.

[142] F. Nielsen, "Chernoff information of exponential families", *ArXiv*, vol. abs/1102.2684, 2011.

[143] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan, "A general analysis of the convergence of ADMM", in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, Lille, France: JMLR.org, 2015, pp. 343–352.

[144] J. B. R. Panos M. Pardalos, "Constrained Global Optimization: Algorithms and Applications", in *Lecture Notes in Computer Science*. Springer-Verlag, 1987, pp. 58–74. DOI: 10.1007/bfb0000041.

[145] N. Parikh and S. Boyd, "Proximal Algorithms", *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014. DOI: 10.1561/2400000003.

[146] G. Pastor, I. Jiménez, R. Jantti, and A. Caamaño, "Mathematics of Sparsity and Entropy: Axioms, Core Functions and Sparse Recovery", *IEEE Transactions on Signal Processing (Draft)*, Jan. 2015.

[147] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, Version 20121115, Sep. 2012.

[148] A. Pezeshki, L. Scharf, M. Azimi-Sadjadi, and M. Lundberg, "Empirical canonical correlation analysis in subspaces", in *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.*, vol. 1, 2004, 994–997 Vol.1. DOI: 10.1109/ACSSC.2004.1399288.

[149] J. Principe and D. Xu, "An introduction to information theoretic learning", in *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, vol. 3, 1999, 1783–1787 vol.3. DOI: 10.1109/IJCNN.1999.832648.

[150] J. C. Principe, *Information theoretic learning* (Information Science and Statistics), en, 2010th ed. New York, NY: Springer, Apr. 2010.

[151] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer New York, 2010. DOI: 10.1007/978-1-4419-1570-2.

[152] M. H. Protter and C. B. Morrey, *A First Course in Real Analysis*. Springer US, 1977. DOI: 10.1007/978-1-4615-9990-6.

[153] T. Qin, S. Cao, J. Pan, and S. Shen, *A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors*, 2019. arXiv: 1901.03642 [cs.CV].

[154] S. Rabanser, L. Neumann, and M. Haltmeier, "Analysis of the Block Coordinate Descent Method for Linear Ill-Posed Problems", *SIAM Journal on Imaging Sciences*, vol. 12, no. 4, pp. 1808–1832, 2019. DOI: 10.1137/19M1243956. eprint: https://doi.org/10.1137/19M1243956.

[155] E. Ram and I. Sason, "On Rényi entropy power inequalities", in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 2289–2293. DOI: 10.1109/ISIT.2016.7541707.

[156] C. Ramírez, V. Kreinovich, and M. Argaez, "Why l1 is a good approximation to l0: A geometric explanation", *Journal of Uncertain Systems*, vol. 7, pp. 203–207, Aug. 2013.

[157] D. Ramírez, I. Santamaría, and L. L. Scharf, *Coherence in Signal Processing and Machine Learning*. Springer, 2022. DOI: https://doi.org/10.1007/978-3-031-13331-2.

[158] D. Ramírez, G. Vázquez-Vilar, R. Lopez-Valcarce, J. Via, and I. Santamaria, "Detection of Rank- P Signals in Cognitive Radio Networks With Uncalibrated Multiple Antennas", *IEEE*

Transactions on Signal Processing, vol. 59, no. 8, pp. 3764–3774, 2011. DOI: 10.1109/TSP.2011.2146779.

[159] D. Ramírez, J. Via, I. Santamaria, and L. L. Scharf, "Locally Most Powerful Invariant Tests for Correlation and Sphericity of Gaussian Vectors", *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2128–2141, 2013. DOI: 10.1109/TIT.2012.2232705.

[160] D. Ramírez, J. Vía, I. Santamaria, and L. Scharf, "Multi-sensor beamsteering based on the asymptotic likelihood for colored signals", in *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 2011, pp. 149–152. DOI: 10.1109/SSP.2011.5967644.

[161] M. Rani, S. B. Dhok, and R. B. Deshmukh, "A Systematic Review of Compressive Sensing: Concepts, Implementations and Applications", *IEEE Access*, vol. 6, pp. 4875–4894, 2018. DOI: 10.1109/ACCESS.2018.2793851.

[162] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization", *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013. DOI: 10.1137/120891009. eprint: https://doi.org/10.1137/120891009.

[163] M. Reinhardt, B. Noack, and U. D. Hanebeck, "Closed-form optimization of covariance intersection for low-dimensional matrices", in *2012 15th International Conference on Information Fusion*, 2012, pp. 1891–1896.

[164] A. Renaux, P. Forster, E. Chaumette, and P. Larzabal, "On the High-SNR Conditional Maximum-Likelihood Estimator Full Statistical Characterization", *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4840–4843, 2006. DOI: 10.1109/TSP.2006.882072.

[165] A. Rényi, "On measures of entropy and information", in *4th Berkeley Symp. Prob. Theory Maths. Stat, Berkeley, CA, USA*, 1961, pp. 547–561.

[166] J. Riba, J. Sala, and G. Vázquez, "Conditional maximum likelihood timing recovery: Estimators and bounds", *IEEE Transactions on Signal Processing*, vol. 49, no. 4, pp. 835–850, 2001. DOI: 10.1109/78.912928.

[167] J. Riba, F. De Cabrera, and J.-M. Juan, "Multi-Satellite Cycle-Slip Detection and Exclusion Using the Noise Subspace of Residual Dynamics", in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2280–2284. DOI: 10.23919/EUSIPCO.2018.8553334.

[168] J. Rissanen, "An Introduction to the MDL Principle", 2005.

[169] S. Ruder, "An overview of gradient descent optimization algorithms", *ArXiv*, vol. abs/1609.04747, 2016.

[170] M. K. Samimi and T. S. Rappaport, "Ultra-wideband statistical channel model for non line of sight millimeter-wave urban channels", in *2014 IEEE Global Communications Conference*, 2014, pp. 3483–3489. DOI: 10.1109/GLOCOM.2014.7037347.

[171] D. F. Schmidt and E. Makalic, "The Consistency of MDL for Linear Regression Models With Increasing Signal-to-Noise Ratio", *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1508–1510, 2012. DOI: 10.1109/TSP.2011.2177833.

[172] G. Seco Granados, "Antenna arrays for multipath and interference mitigation in GNSS receivers", Ph.D. dissertation, Universitat Politècnica de Catalunya. Departament de Teoria del Senyal i Comunicacions, 2000.

[173] C. E. Shannon, "A mathematical theory of communication", *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[174] S. Smith, "Covariance, subspace, and intrinsic Cramer-Rao bounds", *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1610–1630, 2005. DOI: 10.1109/TSP.2005.845428.

[175] S. Smith, "Intrinsic Cramer-Rao bounds and subspace estimation accuracy", in *Proceedings of the 2000 IEEE Sensor Array and Multichannel Signal Processing Workshop. SAM 2000 (Cat. No.00EX410)*, 2000, pp. 489–493. DOI: 10.1109/SAM.2000.878057.

[176] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support", *Signal Processing*, vol. 128, pp. 449–458, 2016. DOI: https://doi.org/10.1016/j.sigpro.2016.05.020.

[177] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules", *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004. DOI: 10.1109/MSP.2004.1311138.

[178] K. Sun and X. A. Sun, "Dual Descent Augmented Lagrangian Method and Alternating Direction Method of Multipliers", *SIAM Journal on Optimization*, vol. 34, no. 2, pp. 1679–1707, 2024. DOI: 10.1137/21M1449099. eprint: https://doi.org/10.1137/21M1449099.

[179] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning", *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017. DOI: 10.1109/TSP.2016.2601299.

[180] Y. Sun, P. Babu, and D. P. Palomar, "Regularized robust estimation of mean and covariance matrix under heavy tails and outliers", in *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2014, pp. 125–128. DOI: 10.1109/SAM.2014.6882356.

[181] R. Tandon and S. Sra, "Sparse nonnegative matrix approximation: New formulations and algorithms", Max Planck Institute for Biological Cybernetics, Tübingen, Germany, Tech. Rep. 193, Sep. 2010.

[182] S. Theodoridis, "Chapter 2 - Probability and Stochastic Processes", in *Machine Learning (Second Edition)*, S. Theodoridis, Ed., Second Edition, Academic Press, 2020, pp. 19–65. DOI: https://doi.org/10.1016/B978-0-12-818803-3.00011-8.

[183] J. A. T. Thomas M. Cover, "Entropy, Relative Entropy, and Mutual Information", in *Elements of Information Theory*. John Wiley & Sons, Ltd, 2005, ch. 2, pp. 13–55. DOI: https://doi.org/10.1002/047174882X.ch2. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/047174882X.ch2.

[184] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[185] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise", *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006. DOI: 10.1109/TIT.2005.864420.

[186] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, May 2005. DOI: 10.1017/cbo9780511807213.

[187] P. Tseng, "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization", *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, Jun. 2001. DOI: 10.1023/a:1017501703105.

[188] J. W. Tukey, "A survey of sampling from contaminated distributions", in *Contributions to Prob. and Statist. (Olkin, I., Ed.)*, 1960, pp. 448–485. DOI: 10.1109/MLSP.2018.8516941.

[189] J. K. Uhlmann, "General data fusion for estimates with unknown cross covariances", in *Defense, Security, and Sensing*, 1996.

[190] C. F. Van Loan, "Generalizing the Singular Value Decomposition", *SIAM Journal on Numerical Analysis*, vol. 13, no. 1, pp. 76–83, 1976. DOI: 10.1137/0713009. eprint: https://doi.org/10.1137/0713009.

[191] M. Vilà, C. A. Lopez, and J. Riba, "Affine Projection Subspace Tracking", in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3705–3709. DOI: 10.1109/ICASSP39728.2021.9415032.

[192] N. K. Vishnoi, "Geodesic Convex Optimization: Differentiation on Manifolds, Geodesics, and Convexity", *ArXiv*, vol. abs/1806.06373, 2018.

[193] D. L. Weakliem, "A Critique of the Bayesian Information Criterion for Model Selection", *Sociological Methods & Research*, vol. 27, no. 3, pp. 359–397, 1999. DOI: 10.1177/0049124199027003002. eprint: https://doi.org/10.1177/0049124199027003002.

[194] J. Wei, F. Luo, S. Chen, and J. Qi, "Robust fusion of GM-PHD filters based on geometric average", *Signal Processing*, vol. 206, p. 108 912, 2023. DOI: https://doi.org/10.1016/j.sigpro.2022.108912.

[195] H. Weidemann and E. Stear, "Entropy analysis of estimating systems", *IEEE Transactions on Information Theory*, vol. 16, no. 3, pp. 264–270, 1970. DOI: 10.1109/TIT.1970.1054444.

[196] Z. Weng and P. M. Djurić, "A Bayesian approach to covariance estimation and data fusion", in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2352–2356.

[197] D. Wolpert and W. Macready, "No free lunch theorems for optimization", *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997. DOI: 10.1109/4235.585893.

[198] P. Wong, "Wavelet decomposition of harmonizable random processes", *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 7–18, 1993. DOI: 10.1109/18.179337.

[199] Y.-C. Wong, "Sectional Curvatures of Grassmann Manifolds", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 60, no. 1, pp. 75–79, 1968.

[200] Z. Wu, S. Peng, B. Chen, H. Zhao, and J. C. Principe, "Proportionate Minimum Error Entropy Algorithm for Sparse System Identification", *Entropy*, vol. 17, no. 9, pp. 5995–6006, 2015. DOI: 10.3390/e17095995.

[201] J. Xu and D. J. Hsu, "On the number of variables to use in principal component regression", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[202] Y. Xu, Z. Li, J. Yang, and D. Zhang, "A Survey of Dictionary Learning Algorithms for Face Recognition", *IEEE Access*, vol. 5, pp. 8502–8514, 2017. DOI: 10.1109/ACCESS.2017.2695239.

[203] T. Yamaguchi, "Locally Geodesically Quasiconvex Functions on Complete Riemannian Manifolds", *Transactions of the American Mathematical Society*, vol. 298, no. 1, pp. 307–330, 1986.

[204] A. L. Yuille and A. Rangarajan, "The Concave-Convex Procedure", *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003. DOI: 10.1162/08997660360581958.

[205] F. M. Zennaro and K. Chen, "Towards understanding sparse filtering: A theoretical perspective", *Neural Networks*, vol. 98, pp. 154–177, 2018. DOI: https://doi.org/10.1016/j.neunet.2017.11.010.

[206] J. Zhan and N. Vaswani, "Robust PCA With Partial Subspace Knowledge", *IEEE Transactions on Signal Processing*, vol. 63, no. 13, pp. 3332–3347, 2015. DOI: 10.1109/TSP.2015.2421485.

[207] J. Zhang, G. Zhu, R. W. Heath, and K. Huang, "Grassmannian Learning: Embedding Geometry Awareness in Shallow and Deep Learning", *ArXiv*, vol. abs/1808.02229, 2018.

[208] L. Zheng and D. Tse, "Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel", *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, 2002. DOI: 10.1109/18.978730.

[209] H. Zhou and Y. Zhang, "EM vs MM: A case study", *Computational Statistics & Data Analysis*, vol. 56, no. 12, pp. 3909–3920, 2012. DOI: https://doi.org/10.1016/j.csda.2012.05.018.

[210] S. Zhou and G. Giannakis, "Optimal transmitter eigen-beamforming and space-time block coding based on channel mean feedback", *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2599–2613, 2002. DOI: 10.1109/TSP.2002.803355.

[211] N. Zorba and A. I. Perez-Neira, "Opportunistic Grassmannian Beamforming for Multiuser and Multiantenna Downlink Communications", *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1174–1178, 2008. DOI: 10.1109/TWC.2008.060972.

[212] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust Estimation in Signal Processing: A Tutorial-Style Treatment of Fundamental Concepts", *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012. DOI: 10.1109/MSP.2012.2183773.