UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

On the Majorization-Minimization framework and g-convex optimization

Exploiting diversity using sparse-aware and information theoretic criteria

Carlos Alejandro Lopez Molina

Departament de Teoria del Senyal i de Comunicacions

Acknowledgmenets

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.

Overview

1. Motivation and goals
2. Sparsity, information theoretic measures and subspaces
3. Algorithmic framework
4. Diversity in Data fusion
5. Angular Diversity in Wireless Communications
6. Quantifying Diversity via Mutual Information

Motivation and goals

Main subject of study: Information diversity

The concept of **Information diversity** is explored in this dissertation.

Information diversity

Diversity is the complementary information that emerges when multiple sources are processed jointly potentially improving the accuracy, uncertainty or integrity of the fusion.

Why diversity?

- Motivation from the formal definition of diversity in Wireless comms.
- Intuitive definition of diversity in Multimodal Data Fusion.
- Problem in GNSS (cycle slip detection) where multi satellite processing may be useful.

Where do we study information diversity?

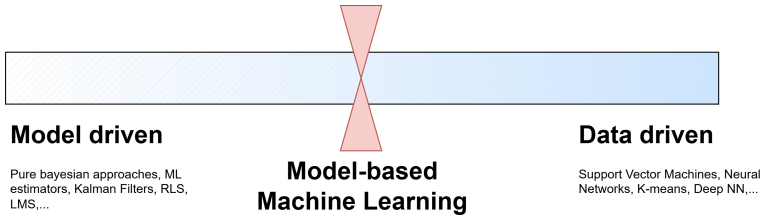
The concept of diversity is studied in the following scenarios:

- **Information/sensor fusion**: Straightforward expression of diversity → Exploiting diversity
- **Wireless Communications**: Hidden expression of diversity in FDD systems → Exploiting diversity
- **Detection of correlation between two sources**: Reformulation into a Mutual Information estimation problem → Quantifying diversity

General tone: Discover information that lies in a lower-dimensional space

How do we study information diversity?

Key Idea: Model-driven Machine Learning



- **Information theoretic criteria**: Robustness and intuition.
- **Sparse-aware techniques**: Useful in practical scenarios.
- **Grassmann manifold**: Structural priors. Tightly related to sparse-aware techniques.
- **Majorization-Minimization optimization framework**: Efficient tool for non-convex and sparse-aware problems.

Sparsity, info. theory and subspaces

Cost functions summary

General idea: How to measure concentration of information?

Sparse-aware cost functions

- ℓ_p norms
  - ℓ_0 norm $\rightarrow$ Ideal sparse regularizer.
  - ℓ_1 norm $\rightarrow$ Tightest convex relaxation of the ℓ_0 norm.
- Functions that can be parameterized on the Grassmann manifold.

$$f(\mathbf{X}) = f(\mathbf{XM}) \tag{1}$$

Information theoretic cost functions

- Information theoretic model-order selection rules
- Rényi Entropy
  - Parametric Minimum Error Entropy criterion.
  - Contaminated Gaussian entropy cost.
- (Shannon) Mutual Information.

Relating sparsity, entropy and subspaces: Sparse modeling

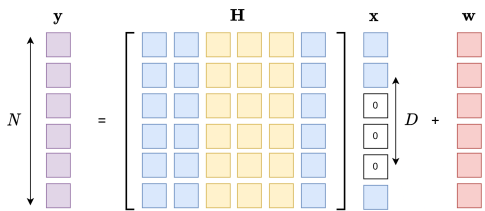
These concepts are related straightforwardly from the following model:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \quad \mathbf{y} \in \mathbb{R}^N, \mathbf{H} \in \mathbb{R}^{N \times D}, \mathbf{x} \in \mathbb{R}^D \tag{2}$$

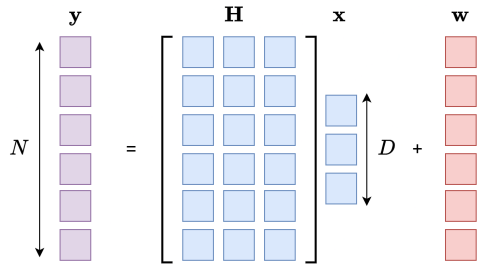
Meaning

- N is the ambient space dimensions.
- $\mathbf{H}\mathbf{x}$ is a sparse model of $\mathbf{y}$:
  - $D < N \rightarrow$ Low-rank subspace (intrinsic dimension).
  - $D > N \rightarrow$ Classical sparse model (size of the dictionary).

Relating sparsity, entropy and subspaces: Classical sparse model vs Subspace model



(a) Classical sparse modeling.



(b) Subspace modeling.

Relating sparsity, entropy and subspaces: On sparsity and entropy

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, $s(\cdot)$ and $h(\cdot)$ satisfy:

- $\uparrow s(\mathbf{p}) \equiv \downarrow h(X)$.
- $\downarrow s(\mathbf{p}) \equiv \uparrow h(X)$.

Parametric Minimum Error Entropy criterion: Definition

Let us consider $\mathbf{X} \sim \mathcal{MN}_{N,M}(\mathbf{0}, \mathbf{K}, \mathbf{Q})$. Then:

$$p_{\mathbf{X}}(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{MN}{2}} \det(\mathbf{Q})^{\frac{N}{2}} \det(\mathbf{K})^{\frac{M}{2}}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{Q}^{-1}\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X})\right) \tag{3}$$

Also, let:

$$h_\alpha(\mathbf{X}) = \frac{MN}{2}\log(2\pi) + \frac{N}{2}\log(\det(\mathbf{Q})) + \frac{M}{2}\log(\det(\mathbf{K})) + \frac{MN}{2}\frac{\log(\alpha)}{\alpha - 1} \tag{4}$$

Parametric Minimum Error Entropy (PMEE) estimators

Assuming two parametric estimations of the covariances, $\hat{\mathbf{K}}(\boldsymbol{\theta})$ and $\hat{\mathbf{Q}}(\boldsymbol{\theta})$, the PMEE is:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{M}\log(\det(\hat{\mathbf{Q}}(\boldsymbol{\theta}))) + \frac{1}{N}\log(\det(\hat{\mathbf{K}}(\boldsymbol{\theta}))) \tag{5}$$

Parametric Minimum Error Entropy criterion: Insights

Intuition behind the PMEE

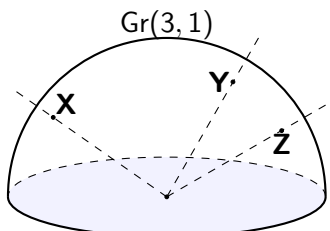
- The PMEE naturally appears in the Conditional Maximum Likelihood (CML) function.
  - CML compresses nuisance parameters.
- The *Minimum Error Entropy* part.
  - Usually a non-parametric cost.
  - Naturally robust. $\rightarrow$ Minimum Determinant criterion
- The *Parametric* part.
  - Ensure almost optimal performance under nominal conditions (Gaussianity).
  - Results in a *tractable* optimization problem.

Grassmann manifold: Definition

The Grassmann manifold

The Grassmann manifold, $\mathrm{Gr}(N, D)$, is defined as the set of D dimensional subspaces with ambient dimension N:

$$\mathrm{Gr}(N, D) = \{[\mathbf{X}] \subset \mathbb{R}^N : [\mathbf{X}] \text{ is a subspace}, \dim([\mathbf{X}]) = D\} \tag{6}$$



Points are represented as:

$$[\mathbf{X}] = \{\mathbf{XR} : \mathbf{X} \in \mathrm{St}(N, D), \forall \mathbf{R} \in \mathrm{O}(D)\} \tag{7}$$

with:

$$\mathrm{St}(N, D) = \{\mathbf{X} \in \mathbb{R}^{N \times D} : \mathbf{X}^T\mathbf{X} = \mathbf{I}\} \tag{8}$$

Grassmann manifold: Principal Angles Between Subspaces

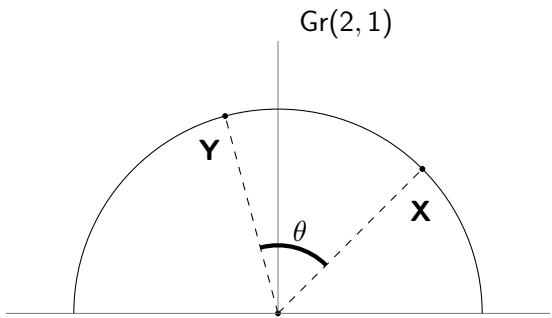
PABS

The PABS between $\mathbf{X}$ and $\mathbf{Y}$ from their SVD:

$$\mathbf{X}^T\mathbf{Y} = \mathbf{U}\cos(\mathbf{\Theta})\mathbf{V}^T \qquad (9)$$

Lemma (Principal alignment)

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

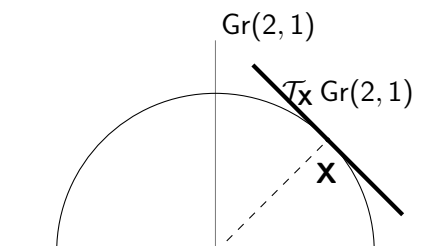


Grassmann manifold: Tangent Space

Grassmann manifold tangent space

Let any $\mathbf{X} \in \mathrm{Gr}(N, D)$, then the tangent space is defined by the following set:

$$\mathcal{T}_{\mathbf{X}} \mathrm{Gr}(N, D) = \{\mathbf{\Delta} \in \mathbb{R}^{N \times D} : \mathbf{X}^T \mathbf{\Delta} = \mathbf{0}_{D \times D}\} \tag{10}$$

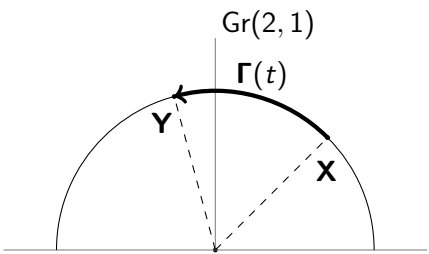


- Descent directions of a function belong to the tangent space!
- Exponential maps relate the Tangent Space with another point.

Grassmann manifold: Geodesics

Geodesics

Geodesics are defined as the path with shortest length between any two points in the Grassmannian.



Geodesics

$$\mathbf{\Gamma}(t) = \mathbf{X}\mathbf{V}\cos(\mathbf{\Theta}t)\mathbf{V}^T + \mathbf{U}\sin(\mathbf{\Theta}t)\mathbf{V}^T \tag{11}$$

Aligned Geodesics

$$\mathbf{\Gamma}_a(t) = \mathbf{X}_a\cos(\mathbf{\Theta}t) + \mathbf{\Delta}_a\sin(\mathbf{\Theta}t) \tag{12}$$

Information Theoretic Model-Order Selection: General idea

Information-theoretic Model-Order Selection

$$\hat{D} = \arg\max_{L} \underbrace{\sum_{k=1}^{K} \log(p_{\mathbf{y}}(\mathbf{y}_k|\hat{\boldsymbol{\theta}}_L))}_{\text{likelihood}} - \underbrace{\frac{L}{2}\eta(L,K)}_{\text{penalty}} \tag{13}$$

Criterion	**Penalty**, $\eta(K)$
Bayesian Information Criterion (BIC)	$\ln(K)$
Akaike Information Criterion (AIC)	2
Generalized Information Criterion (GIC)	$\lambda + 1$, for $\lambda > 1$

Validity

- Non-singular Fisher information.
- Asymptotical independence of the Fisher information with respect to $\boldsymbol{\theta}_L$.

Algorithmic framework

General Outline

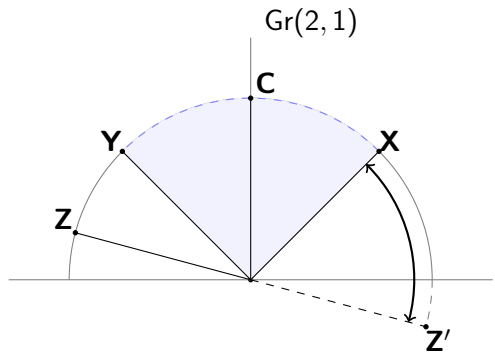
Key contributions of our algorithmic framework

- Particularization of **g-convex optimization** to the Grassmann manifold.
  - What is g-convexity? $\rightarrow$ Implications for the Grassmann manifold.
  - Riemannian Perspective on Principal Component Analysis.
- The **Majorization-Minimization** (MM) framework.
  - Formal introduction of the Grassmann manifold into the MM framework.

G-convex optimization: Grassmann g-convex subsets

Convex sets can get complicated in Riemannian manifolds...

- G-convex sets are convex sets with respect to a geodesic.
- Grassmann ball of "sizes" up to $\phi = \frac{\pi}{4}$ are g-convex.



G-convex balls on the Grassmannian

$$B_\phi(\mathbf{C}) = \{\mathbf{X} \in \mathrm{Gr}(N, D) : \mathbf{\Theta}_\mathbf{X} \prec \phi \mathbf{I}_D\} = \{\mathbf{X} \in \mathrm{Gr}(N, D) : d_{arc}(\mathbf{X}, \mathbf{C}) < \phi\sqrt{D}\} \quad (14)$$

G-convex optimization: G-convexity

On g-convex functions
Convex functions properties are generalized *straightforwardly* to g-convex functions.

$$f(\mathbf{\Gamma}(t)) \leq (1-t)f(\mathbf{X}) + tf(\mathbf{Y}) \quad \text{(15a)}$$

$$f(\mathbf{Y}) \geq f(\mathbf{X}) + \langle \text{grad } f(\mathbf{X}), \mathbf{\Delta} \rangle_{\mathbf{X}} \quad \text{(15b)}$$

$$\text{hess } f(\mathbf{X})[\mathbf{\Delta}, \mathbf{\Delta}] \geq 0 \quad \text{(15c)}$$

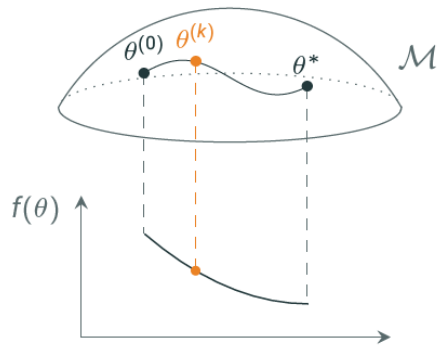


Image extracted from: F. Bouchard, A. Breloy and A. Mian, "Riemannian and information geometry in signal processing and machine learning,". *EUSIPCO 2022 Tutorials*

Riemannian perspective on PCA: Problem statement

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} \frac{1}{2}\text{tr}(\mathbf{X}^T\mathbf{S}\mathbf{X}) \quad \text{s.t. } \mathbf{X} \in \text{Gr}(M, D) \tag{16}$$

with:

$$\mathbf{S} = \mathbf{U}_s\mathbf{D}_s\mathbf{U}_s^T + \mathbf{U}_n\mathbf{D}_n\mathbf{U}_n^T \in \mathbb{R}^{M\times M} \tag{17}$$

Properties

- Can be parameterized with either the Grassmann or the Stiefel manifold.
- PCA cost is **locally** g-concave in $\mathbf{X} \in B_{\frac{\pi}{4}}(\mathbf{U}_s)$ and g-convex in $\mathbf{X} \in B_{\frac{\pi}{4}}(\mathbf{U}_n)$:
  - For $\text{rank}(\mathbf{S}) = D \rightarrow$ Every other stationary point is a **saddle point**. ($\mathbf{D}_n = \mathbf{0}$)
  - For $\text{rank}(\mathbf{S}) > D \rightarrow$ Difficult assessment of the remaining stationary points

Note: Stiefel and Grassmann manifolds can be compared in this problem!

Riemannian perspective on PCA: Toy example settings

Key idea: Use the Riemannian Gradient Descent (RGD) to compare the performance of these two manifolds.
Dimensions: $\mathbf{S} \in \mathbb{R}^{10\times 10}$ and $\mathbf{X} \in \mathbb{R}^{10\times 4}$.

Update equations of RGD in this problem

$$\text{grad } f(\mathbf{X}_i) = \text{proj}_{\mathcal{T}_{\mathbf{X}_i}\mathcal{M}}(\nabla f(\mathbf{X}_i)) = \text{proj}_{\mathcal{T}_{\mathbf{X}_i}\mathcal{M}}(2\mathbf{S}\mathbf{X}_i) \tag{18a}$$

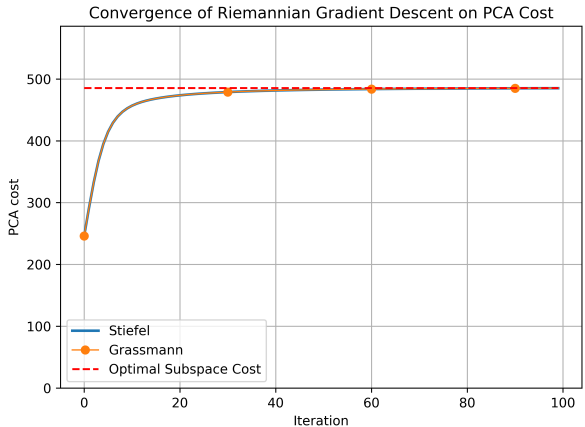
$$\mathbf{X}_{i+1} = \exp_{\mathbf{X}_i}(\eta \,\text{grad } f(\mathbf{X}_i)) \tag{18b}$$

Initialization: $\mathbf{X}_0$ is such that the PABS between $\mathbf{X}_0$ and $\mathbf{X}_{\text{opt}}$ are all exactly ϕ.

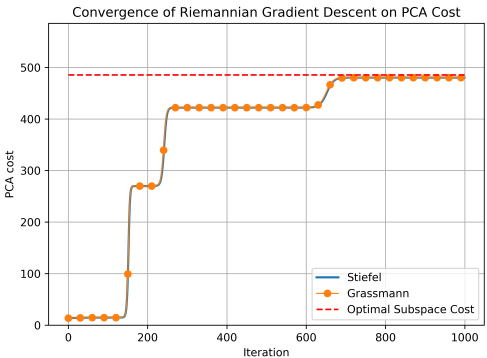
- $\phi = \frac{\pi}{4} \rightarrow$ G-convex set
- $\phi = \frac{\pi}{2} \rightarrow$ Non-g-convex set

Riemannian perspective on PCA: Convergence

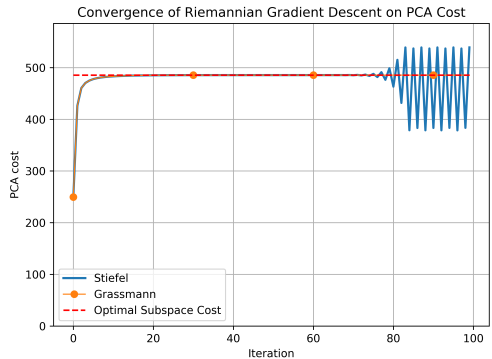
Fixing $\eta = 10^{-3}$ and $\phi = \frac{\pi}{4}$...



Riemannian perspective on PCA: Issues



(a) Non-g-convex set ($\phi = \frac{\pi}{2}$).

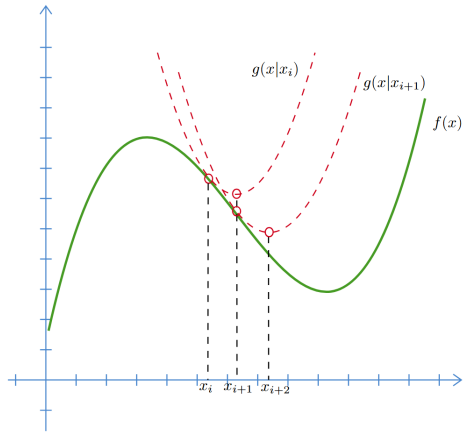


(b) $\eta = 5 \cdot 10^{-3}$.

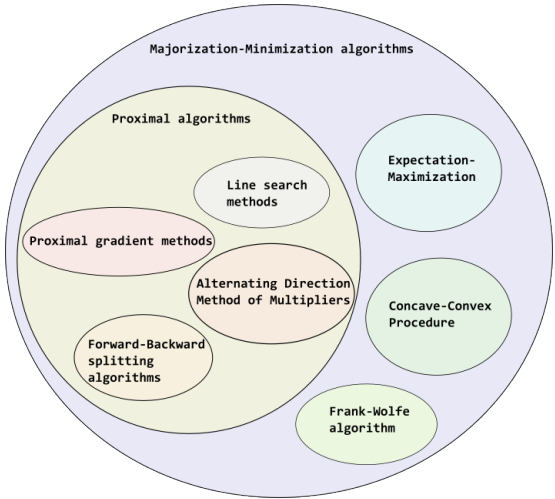
Majorization-Minimization framework: General Insights

Key insights into the MM framework

- Consists in the sequential optimization of surrogate functions.
- Suited for **non-convex** functions.
- Enables **block optimization** in a natural way.



Majorization-Minimization framework: Well-known MM algorithms



Majorization-Minimization framework: MM + Grassmann

- MM + Grassmann manifold converges to $f(\mathbf{X}^*)$:

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} f(\mathbf{G}) \quad \text{s.t. } \mathbf{G} \in \mathrm{Gr}(N, D) \tag{19}$$

  - Surrogate functions majorize the original cost.
  - Derivatives of surrogate and original cost coincide.

- Block MM + Grassmann manifold:

$$\hat{\mathbf{G}}, \hat{\mathbf{c}} = \arg\min_{\mathbf{G},\mathbf{c}} f(\mathbf{G}, \mathbf{c}) \quad \text{s.t. } \mathbf{G} \in \mathrm{Gr}(N, D), \mathbf{c} \in \mathcal{C} \tag{20}$$

  - Convergence with respect to optimization variables is necessary to reach a stationary point.
  - Additional assumptions are required:
    * Quasiconvex majorants $\rightarrow$ G-quasiconvex majorants
    * Unique minimizer of majorants.

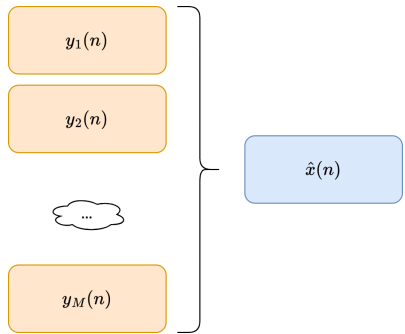
Diversity in Data fusion

Data fusion: Problem statement

How can we accurately estimate a shared latent signal from multiple noisy and correlated data sources?

Main challenges of data fusion

- A single unreliable information source can contaminate the latent signal estimation. (!)
- Correlated uncertainties. Model integrity may be lost.



Data fusion: General model definition

General model

$$\mathbf{y}(n) = x(n)\mathbf{1}_M + \mathbf{w}(n) \tag{21}$$

with:

$$\mathrm{E}\left[\mathbf{w}(n)\mathbf{w}^T(n)\right] = \mathbf{Q} \in \mathcal{S}_{++}^M \tag{22}$$

Note: An isotropic spatial signature is not restrictive!

Optimal fusion scheme

Under nominal conditions...

$$\mathbf{w}(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}) \tag{23}$$

we have

$$\mathbf{f}_B = \frac{\mathbf{Q}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{Q}^{-1}\mathbf{1}} \implies \hat{x}_B(n) = \mathbf{f}_B^T\mathbf{y}(n) \tag{24}$$

Data fusion: What are the simplest forms of information fusion?

What?

- **Parameters**: Estimators, measurements, figures...
- **Distributions** $\rightarrow$ Fusing information

How? Convex combinations!

- Arithmetic Average (AA) fusion
- Geometric Average (GA) fusion

Data fusion: Fusion of parameters (ν-fusion)

AA fusion

$$\boldsymbol{\theta}_{AA} = \sum_{m=1}^{M} \omega_m \boldsymbol{\theta}_m \qquad (25)$$

- Can reach optimal MSE under nominal conditions.
- Unbiased.

Note: $\sum_{m=1}^{M} \omega_m = 1$.

GA fusion

$$\log(\boldsymbol{\theta}_{GA}) = \sum_{m=1}^{M} \omega_m \log(\boldsymbol{\theta}_m) \qquad (26)$$

- Can yield complex numbers. (!)
- Biased. (!)

Data fusion: Fusion of distributions (f-fusion)

AA fusion

$$f_{AA}(\mathbf{x}|\boldsymbol{\theta}_{AA}) = \sum_{m=1}^{M} \omega_m f_m(\mathbf{x}|\boldsymbol{\theta}_m) \quad (27)$$

- More robust in classification problems.
- Diversity may not be exploited. (!)
- Mixture model.

Note: $\sum_{m=1}^{M} \omega_m = 1$.

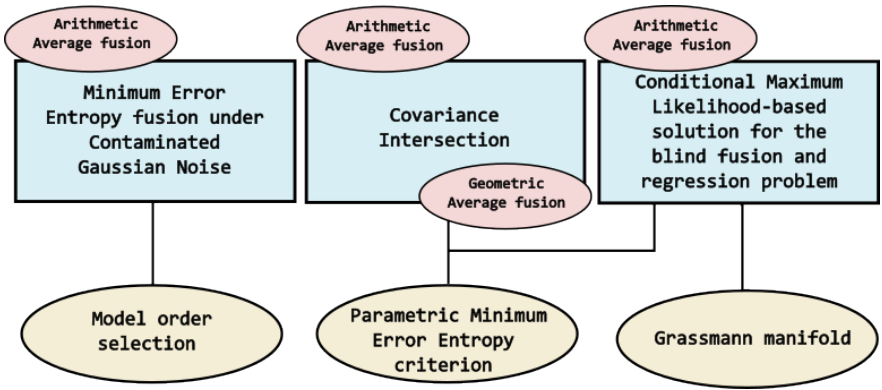
GA fusion

$$f_{GA}(\mathbf{x}|\boldsymbol{\theta}_{GA}) = \frac{1}{G} \prod_{m=1}^{M} f_m^{\omega_m}(\mathbf{x}|\boldsymbol{\theta}_m) \quad (28)$$

with $G = \int_{-\infty}^{\infty} \prod_{m=1}^{M} f_m^{\omega_m}(\mathbf{x}|\boldsymbol{\theta}_m) \mathrm{d}\mathbf{x}$

- Convex combination of information.
- Generalization of the combination of independent measures.
- Low densities are enforced.

Data fusion: Studied sensor fusion schemes



Covariance Intersection: General statement

Assume two sources of information, A and B. The CI constructs a linear fusion (new source, C) of the following form:

$$\hat{\mathbf{Q}}_{\omega}^{-1} = \omega \hat{\mathbf{Q}}_{aa}^{-1} + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1} \tag{29a}$$

$$\mathbf{c}_{\omega} = \hat{\mathbf{Q}}_{\omega}\left(\omega \hat{\mathbf{Q}}_{aa}^{-1}\mathbf{a} + (1-\omega)\hat{\mathbf{Q}}_{bb}^{-1}\mathbf{b}\right) \tag{29b}$$

The resulting fusion is *consistent* $\forall\omega$:

$$\hat{\mathbf{Q}}_{aa} - \hat{\mathbf{Q}}_{\omega} \succeq \mathbf{0} \tag{30a}$$

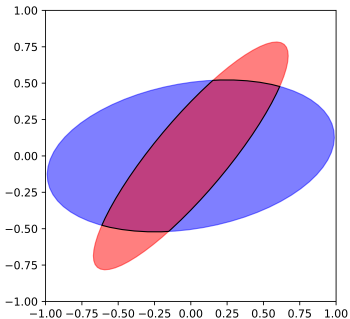
$$\hat{\mathbf{Q}}_{bb} - \hat{\mathbf{Q}}_{\omega} \succeq \mathbf{0} \tag{30b}$$

$$\hat{\mathbf{Q}}_{\omega} - \mathbf{Q}_{\omega} \succeq \mathbf{0} \tag{30c}$$

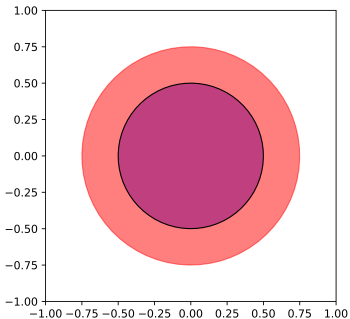
Note: Consistency of $\hat{\mathbf{Q}}_{aa}$ and $\hat{\mathbf{Q}}_{bb}$ mean that they majorize $\mathbf{Q}_{aa}$ and $\mathbf{Q}_{bb}$.

Covariance Intersection: Intuition

Constructing the associated ellipses from the estimated covariances...



(a) *Rich* intersection of ellipses.



(b) *Poor* intersection of ellipses.

Covariance Intersection: Sensor fusion problem

Recall our sensor model:

$$y_m(n) = x(n) + w_m(n) \quad m = 1, ..., M \tag{31}$$

CI equations (Derived from a GA fusion of PDFs)

$$\hat{q}_\omega^{-1} = \sum_{m=1}^{M} \omega_m \hat{q}_m^{-1} \tag{32a}$$

$$\hat{x}_{CI}(n) = \hat{q}_\omega \sum_{m=1}^{M} \omega_m \hat{q}_m^{-1} y_m(n) \tag{32b}$$

Covariance Intersection: Proposed criterion

The minimum variance/determinant criterion yields:

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \sum_{m=1}^{M} \omega_m \hat{q}_m^{-1} \quad \text{s.t. } \boldsymbol{\omega}^T \mathbf{1}_M = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M \tag{33}$$

Issue: We want to avoid sparse solutions. (!)
Solution (waterfilling): Leverage the majorants idea and introduce an antisparse parameter φ.

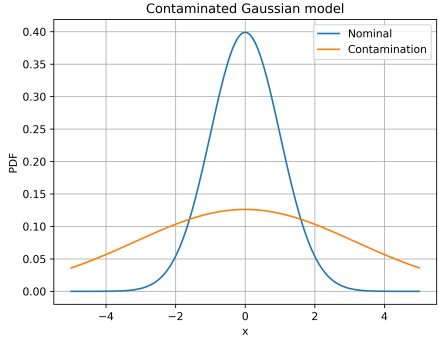
$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \sum_{m=1}^{M} \log\left(1 + \varphi \frac{\omega_m}{\hat{q}_m}\right) \quad \text{s.t. } \boldsymbol{\omega}^T \mathbf{1}_M = 1, \boldsymbol{\omega} \succeq \mathbf{0}_M \tag{34}$$

- High φ (High SNR case) $\rightarrow$ Exploits diversity
- Small φ (Low SNR case) $\rightarrow$ Best sensor policy

Contaminated Gaussian scheme: Preliminaries

Consider the contaminated Gaussian model ($p_X(x)$ is the standard normal distribution):

$$p_Z(z) = \frac{1-\varepsilon}{\sqrt{u}} p_X\left(\frac{z}{\sqrt{u}}\right) + \frac{\varepsilon}{\sqrt{v}} p_X\left(\frac{z}{\sqrt{v}}\right) \tag{35}$$



Idea: Leverage the previous model to yield a robust fusion scheme.

Contaminated Gaussian scheme: Preliminaries 2

Consider a multivariate contaminated Gaussian random variable:

$$p_{\mathbf{z}}(\mathbf{z}) = \frac{1-\varepsilon}{\sqrt{(2\pi)^M \det(\mathbf{Q})}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{Q}^{-1}\mathbf{z}\right) + \frac{\varepsilon}{\sqrt{(2\pi v)^M}} \exp\left(-\frac{1}{2v}\mathbf{z}^T\mathbf{z}\right) \tag{36}$$

and the random variable $g = \mathbf{f}^T\mathbf{z}$.

Rényi Entropy of the fused contaminated Gaussian RV

For $v \to \infty$ and $\alpha > 1$, the Rényi entropy of g is given by:

$$h_\alpha(g) = h_\alpha(X) + \frac{1}{2}\log(\mathbf{f}^T\mathbf{Q}\mathbf{f}) + \frac{\beta}{2}||\mathbf{f}||_0 \tag{37}$$

where X is a standard normal RV and:

$$\beta = \frac{2\alpha}{\alpha-1}\log\left(\frac{1}{1-\varepsilon}\right) \tag{38}$$

Contaminated Gaussian scheme: E-BLUE

Let:

$$\mathbf{y} = x\mathbf{1}_M + \mathbf{z} \tag{39a}$$

$$e(\mathbf{f}) = \mathbf{f}^T\mathbf{y} - x \tag{39b}$$

Entropic-Best Linear Unbiased Estimator

The E-BLUE of x is defined as:

$$\hat{x}_{\text{E-BLUE}} = \mathbf{f}_{\text{E-BLUE}}^T\mathbf{y} \tag{40}$$

where:

$$\mathbf{f}_{\text{E-BLUE}} = \arg\min_{\mathbf{f}} h_\alpha(e(\mathbf{f})) \quad \text{s.t. } \mathbf{f}^T\mathbf{1}_M = 1 \tag{41}$$

Contaminated Gaussian scheme: Obtaining the E-BLUE

A more explicit expression of the E-BLUE is:

$$\mathbf{f}_{\text{E-BLUE}} = \arg\min_{\mathbf{f}} \log(\mathbf{f}^T \mathbf{Q} \mathbf{f}) + \beta ||\mathbf{f}||_0 \quad \text{s.t. } \mathbf{f}^T \mathbf{1}_M = 1 \tag{42}$$

- Constraint ensures an unbiased fusion.
- ℓ_1 relaxation of ℓ_0 fails in general. (!)

The difficulty of this optimization problem depends on $\mathbf{Q}$.

- Diagonal $\mathbf{Q}$ $\implies$ Simple and intuitive optimization problem
- General $\mathbf{Q}$ $\implies$ Combinatorial optimization problem (!)

Contaminated Gaussian scheme: Uncorrelated sensors

The E-BLUE is rewritten as follows:

$$\mathbf{f}_{\text{E-BLUE}} = \arg\min_{\mathbf{f}_n, n} \log(\mathbf{f}_n^T \mathbf{Q}_n \mathbf{f}_n) + \beta n \quad \text{s.t.} \quad \mathbf{f}_n^T \mathbf{1}_n = 1 \tag{43}$$

and we already know that $\mathbf{f}_n = \frac{\mathbf{Q}_n^{-1}\mathbf{1}_n}{\mathbf{1}_n^T \mathbf{Q}_n^{-1}\mathbf{1}_n}$. Then:

$$n_{\text{E-BLUE}} = \arg\min_{n} \underbrace{-\log\left(\mathbf{1}_n^T \mathbf{Q}_n^{-1}\mathbf{1}_n\right)}_{\text{"likelihood" term}} + \underbrace{\beta n}_{\text{penalty}} . \tag{44}$$

Solution:

$$n_{\text{E-BLUE}} = \min_{1 \le n \le M} n \quad \text{s.t.} \quad \frac{\gamma_{n+1}}{\sum_{m=1}^{n} \gamma_m} < e^{\beta} - 1, \gamma_m = \frac{1}{[\mathbf{Q}]_{m,m}} \tag{45}$$

Blind fusion and regression: Preliminaries

By coupling regression and fusion, we unlock latent statistical structure that is inaccessible when these tasks are handled independently $\rightarrow$ Fusion-regression residuals

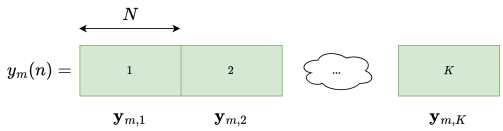
Key contributions

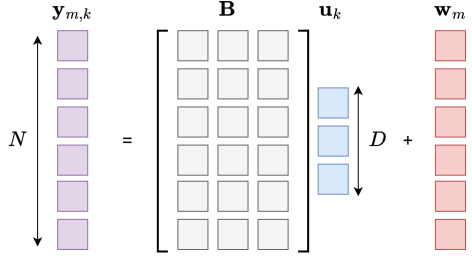
- An AA fusion is coupled with a Subspace-based regression.
- Fusion and regression parameters are determined via the PMEE.
- Grassmann manifold appears in a natural way to ensure convergence.

Blind fusion and regression: Subspace-based regression

Key insights into subspace-based regression

- Assumes some sort of temporal redundancy in $x(n)$.
- Estimates the temporal structure of the phenomenon of interest.
- Needs to partition $y_m(n)$ into K blocks of N samples.
- Temporal structure is encoded in $\mathbf{B} \in \mathbb{R}^{N \times D}$.





Note: Non-linear regression is possible.

Blind fusion and regression: Compact model

After stacking $\mathbf{y}_{m,k}$ for $m = 1, ..., M$, we get:

$$\mathbf{Y}_k = \mathbf{x}_k \mathbf{1}_M^T + \mathbf{W}_k = \mathbf{B}\mathbf{u}_k \mathbf{1}_M^T + \mathbf{W}_k \tag{46}$$

where:

$$\mathbf{W}_k \sim \mathcal{MN}_{N,M}(\mathbf{0}_{N,M}, \mathbf{I}_N, \mathbf{Q}) \tag{47}$$

Goal: Determine $\mathbf{B}$ and $\mathbf{u}_k$ without $\mathbf{Q}$.
Issue: There are too many parameters ($\mathbf{u}_k$ for $k = 1, ..., K$).
Solution: Introduce a new variable that encodes the fusion, $\mathbf{f}$, and *compress* $\mathbf{u}_k$.

Blind fusion and regression: Parametric Minimum Error Entropy (PMEE) criterion

- PMEE can be introduced in this problem from two perspectives:
  - Conditional Maximum Likelihood (CML) principle. Compress $\mathbf{u}_k$ and $\mathbf{Q}$.
  - Parameterize the Rényi entropy of $\mathbf{W}_k$.
- Minimizing the PMEE cost is an optimization problem with respect to fusion and regression parameters.

$$\hat{\mathbf{B}}, \hat{\mathbf{f}} = \arg\min_{\mathbf{f},\mathbf{B}} \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{B}, \mathbf{f}))) \quad \text{s.t.} \quad \mathbf{f}^T \mathbf{1}_M = 1 \quad (!) \tag{48}$$

but we prefer:

$$\hat{\mathbf{H}}, \hat{\mathbf{f}} = \arg\min_{\mathbf{H},\mathbf{f}} \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}))) \quad \text{s.t.} \quad \mathbf{f}^T \mathbf{1}_M = 1, \ \mathbf{H} \in \mathrm{Gr}(N, D) \tag{49}$$

Blind fusion and regression: On the log-determinant

Dissecting $\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f})$, we get:

$$\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}) = \frac{1}{KN} \sum_{k=1}^{K} (\mathbf{Y}_k - \mathbf{P}_H \mathbf{Y}_k \mathbf{f} \mathbf{1}_M^T)^T (\mathbf{Y}_k - \mathbf{P}_H \mathbf{Y}_k \mathbf{f} \mathbf{1}_M^T) \tag{50}$$

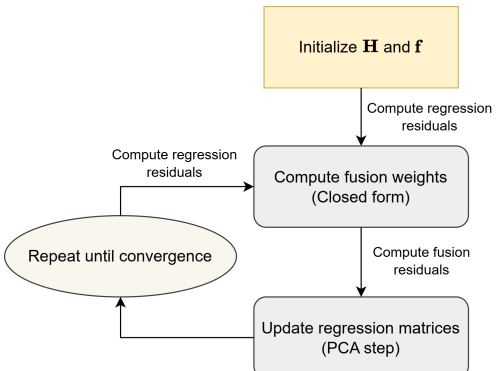
Majorant function Log-determinant is a concave function!

$$\log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}))) \leq \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}_i, \mathbf{f}_i))) + \mathrm{tr}\left(\mathbf{Z}_i \left(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}) - \hat{\mathbf{Q}}_{ML}(\mathbf{H}_i, \mathbf{f}_i)\right)\right) \tag{51}$$

with

$$\mathbf{Z}_i = \hat{\mathbf{Q}}_{ML}^{-1}(\mathbf{H}_i, \mathbf{f}_i) \tag{52}$$

Blind fusion and regression: block MM algorithm



- Based on a first-order majorant.

$$g(\mathbf{H},\mathbf{f}|\mathbf{H}_i,\mathbf{f}_i) = \text{tr}\left(\mathbf{Z}_i\hat{\mathbf{Q}}_{ML}(\mathbf{H},\mathbf{f})\right) \tag{53}$$

- Converges as long as:
  - $\mathbf{H}$ and $\mathbf{f}$ are initialized properly.
  - The number of data blocks is sufficiently large. $K \geq \max(M, D)$.
  - There are techniques to mitigate the effects of a small sample size.

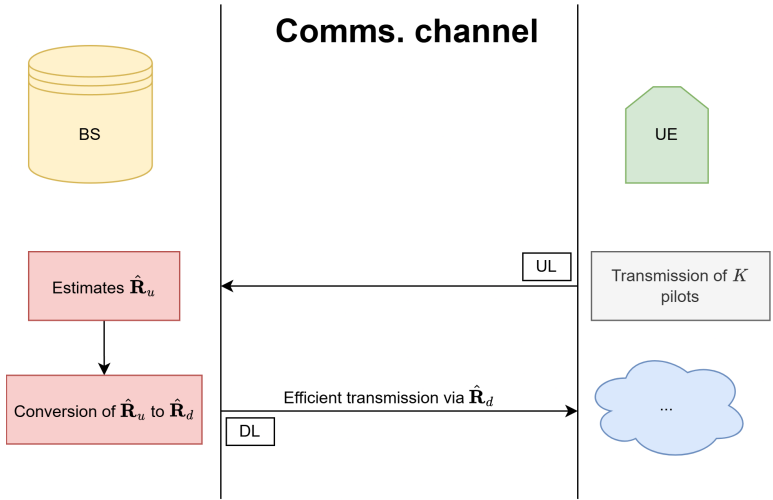
Angular Diversity in Wireless Comms.

Covariance Conversion: Fundamentals in FDD systems

- Channel reciprocity does not hold in FDD systems.
- There is a need to obtain any form of CSI in some comms schemes.
- CC is a possible solution to this problem: Statistical CSI UL $\rightarrow$ Statistical CSI DL.

Key idea: Estimate the common information from one channel and perform a conversion step.

Covariance Conversion: Communications scheme



Covariance Conversion: Model definition

- BS to UE $\rightarrow$ MISO, channel 1 (DL)

$$y_1(n) = \mathbf{c}^H \mathbf{h}_1(n) x_1(n) + w_1(n) \tag{54}$$

- UE to BS $\rightarrow$ SIMO, channel 2 (UL)

$$\mathbf{y}_2(n) = \mathbf{h}_2(n) x_2(n) + \mathbf{w}_2(n) \tag{55}$$

Channel vector second-order statistics model (Saleh-Valenzuela)

$$\mathbf{R}_c = \mathrm{E}\left[\mathbf{h}_c(n)\mathbf{h}_c^H(n)\right] = \int_{-\pi}^{\pi} \rho(\theta)\mathbf{a}_c(\theta)\mathbf{a}_c^H(\theta)\mathrm{d}\theta \quad c = 1, 2 \tag{56}$$

Covariance Conversion: Second-order statistics

$$\mathbf{R}_c = \int_{-\pi}^{\pi} \rho(\theta)\mathbf{a}_c(\theta)\mathbf{a}_c^H(\theta)\mathrm{d}\theta \quad c = 1, 2 \tag{57}$$

Key ideas

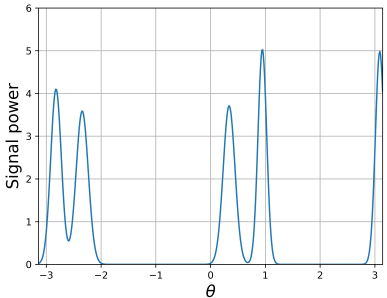
- $\rho(\theta)$ depicts the power distribution along the angular domain.
  - Independent of carrier frequency.
  - Dependent of environment geometry.
- $\mathbf{a}_c(\theta)$ is the array response at f_c.
  - Dependent of array geometry and carrier frequency.
  - Its values do not change over time.

Intuition: Second-order statistics of the channel are a slow time-varying figure due to $\rho(\theta)$.

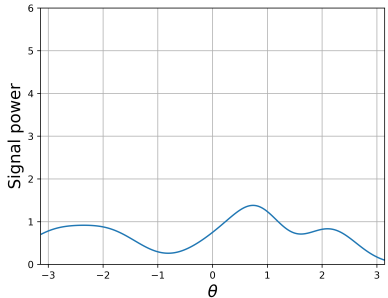
Covariance Conversion: Sparse APS

Let us consider the following APS model:

$$\rho(\theta) = \sum_{s=1}^{S} \alpha_s k(\theta, \theta_s, \sigma_s) = \sum_{s=1}^{S} \frac{\alpha_s}{\sqrt{2\pi\sigma_s}} \exp\left(-\frac{|\theta - \theta_s|^2}{2\sigma_s}\right) \tag{58}$$



(a) Sparse APS.



(b) Non-sparse APS.

Covariance Conversion: Quantized APS

Reformulate the second-order statistics model with respect to:

$$\theta_n = \theta_l + (n-1)\frac{\theta_u - \theta_l}{N-1} \quad n = 1, ..., N \tag{59}$$

where $[\theta_l, \theta_u]$ is visible window. As a result:

$$\mathbf{R}_c \approx \sum_{n=1}^{N} \rho(\theta_n)\mathbf{a}_c(\theta_n)\mathbf{a}_c^H(\theta_n)\Delta\theta \quad c = 1, 2 \tag{60}$$

Key idea: Reformulate the previous matrix into a system of equations.

Covariance Conversion: System of equations formulation

$$\mathbf{r}_c \approx \mathbf{A}_c \boldsymbol{\rho} \quad c = 1, 2 \tag{61}$$

where:

$$\mathbf{r}_c = \text{vec}(\mathbf{R}_c) \quad c = 1, 2 \tag{62a}$$

$$\boldsymbol{\rho} = [\rho(\theta_1), ..., \rho(\theta_N)]^T \tag{62b}$$

$$\mathbf{A}_c = \Delta\theta \left[\text{vec}\left(\mathbf{a}_c(\theta_1) \mathbf{a}_c^H(\theta_1) \right), ..., \text{vec}\left(\mathbf{a}_c(\theta_N) \mathbf{a}_c^H(\theta_N) \right) \right] \quad c = 1, 2 \tag{62c}$$

Example of Conversion equations (Channel 1 $\rightarrow$ Channel 2)

$$\hat{\boldsymbol{\rho}}_{LS} = \arg\min_{\boldsymbol{\rho}} ||\hat{\mathbf{r}}_1 - \mathbf{A}_1 \boldsymbol{\rho}||_2^2 \quad \text{s.t. } \Re(\boldsymbol{\rho}) \succeq \mathbf{0}_M, \Im(\boldsymbol{\rho}) = \mathbf{0}_N \tag{63a}$$

$$\mathbf{R}_2 = \text{vec}^{-1}(\mathbf{A}_2 \hat{\boldsymbol{\rho}}_{LS}) \tag{63b}$$

Covariance Conversion: Sparse-Aware solution (ADMM)

$$\hat{\boldsymbol{\rho}}_{BPD} = \arg\min_{\boldsymbol{\rho}} ||\boldsymbol{\rho}||_1 \quad \text{s.t.} \quad ||\hat{\mathbf{r}}_c - \mathbf{A}_c\boldsymbol{\rho}||_2^2 \leq \varepsilon, \Re(\boldsymbol{\rho}) \succeq \mathbf{0}_N, \Im(\boldsymbol{\rho}) = \mathbf{0}_N \tag{64}$$

MM algorithm (ADMM)

$$\boldsymbol{\rho}_{k+1} = \arg\min_{\boldsymbol{\rho}} ||\mathbf{A}_c\boldsymbol{\rho} - \mathbf{z}_{1,k} + \mathbf{u}_{1,k}||_2^2 + ||\boldsymbol{\rho} - \mathbf{z}_{2,k} + \mathbf{u}_{2,k}||_2^2 \tag{65a}$$

$$\mathbf{z}_{1,k+1} = \arg\min_{\mathbf{z}} \frac{1}{2}||\mathbf{z} - (\mathbf{A}_c\boldsymbol{\rho}_{k+1} + \mathbf{u}_{1,k})||_2^2 \quad \text{s.t.} \quad ||\mathbf{z} - \hat{\mathbf{r}}_c||_2^2 \leq \varepsilon \tag{65b}$$

$$\mathbf{z}_{2,k+1} = \arg\min_{\mathbf{z}} ||\mathbf{z}||_1 + \frac{\lambda}{2}||\mathbf{z} - (\boldsymbol{\rho}_{k+1} + \mathbf{u}_{2_k})||_2^2 \quad \text{s.t.} \quad \Re(\mathbf{z}) \succeq \mathbf{0}_N, \Im(\mathbf{z}) = \mathbf{0}_N \tag{65c}$$

$$\mathbf{u}_{1,k+1} = \mathbf{u}_{1,k} + \boldsymbol{\rho}_{k+1} - \mathbf{z}_{1,k+1} \tag{65d}$$

$$\mathbf{u}_{2,k+1} = \mathbf{u}_{2,k} + \mathbf{A}_c\boldsymbol{\rho}_{k+1} - \mathbf{z}_{2,k+1} \tag{65e}$$

Quantifying Diversity via MI

Model-order MI estimation: Motivation

- The previous problems assume that diversity exists.
  - Common latent signal between sensors.
  - Angular Power Spectrum.
- Diversity often appears in a sparse way.
  - Sparse-aware techniques are required.

Information theoretic coherence

$$\rho_{\mathrm{IT}}(X, Y) = \sqrt{1 - \exp(-I(X; Y))} \tag{66}$$

Key idea: Estimate the MI between X and Y.

Model-order MI estimation: Problem statement

Nominal conditions

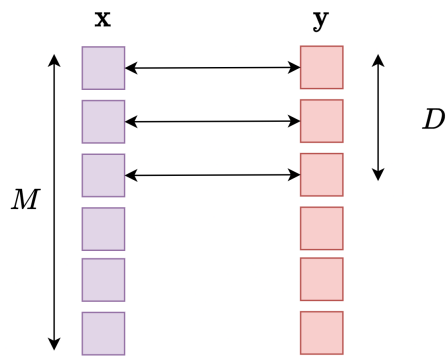
M independent Gaussian channels.

$$\mathbf{z}_m(n) = \begin{bmatrix} x_m(n) \\ y_m(n) \end{bmatrix} \tag{67a}$$

$$\mathbf{z}_m(n) \sim \mathcal{N}(\mathbf{0}_2, \mathbf{C}_m) \tag{67b}$$

$$\mathbf{C}_m = \begin{bmatrix} 1 & \rho_m \\ \rho_m & 1 \end{bmatrix} \tag{67c}$$

Note: Unit-variance and mutually independent sequences.



Model-order MI estimation: Parallel Gaussian channels MI

Mutual independence of sequence pairs implies:

$$I(\mathbf{x}(n);\mathbf{y}(n)) = \sum_{m=1}^{M} I(x_m(n);y_m(n)) = -\frac{1}{2}\sum_{m=1}^{M}\log(1-\rho_m^2) \tag{68}$$

Sparse assumption

Let:

$$\mathcal{S}_M = \{m \in \mathbb{N} : 1 \leq m \leq M\} \tag{69}$$

Then:

$$\mathcal{S}_D = \{d \in \mathcal{S}_M : |\rho_d| > 0\} \tag{70}$$

Issue: $\text{card}(\mathcal{S}_D) = D < M$ is not known in practice!

Model-order MI estimation: Estimating MI

The CML function is a natural estimator of the MI. Starting from:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_M) = -\frac{1}{2}\sum_{n=1}^{N}\Big(\log(\det(\mathbf{C}_m)) + \mathbf{z}_m^T(n)\mathbf{C}_m^{-1}\mathbf{z}_m(n)\Big) \tag{71}$$

we can get to:

$$\ell_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y}|\mathcal{S}_M) = -\frac{N}{2}\left(\sum_{m=1}^{M}\log(1-\hat{\rho}_m^2) + \text{constants}\right) = N\hat{I}_{ML}(\mathbf{x};\mathbf{y}|\mathcal{S}_M) + \text{constants} \tag{72}$$

Issue: $\hat{I}_{ML}(\mathbf{x};\mathbf{y}|\mathcal{S}_M)$ is a biased estimator of the MI.
Solution: Ignore the contributions of the m-th channel if it is an inactive channel.

Model-order MI estimation: Regularized estimation of MI

Model-order selection rules can be leveraged to determine the active channels:

$$\hat{D} = \arg\max_{L} \hat{I}_{ML}(\mathbf{x}; \mathbf{y}|\mathcal{S}_L) - \frac{L\eta(N)}{2N} \quad \text{s. t. } L \in \{1, 2, ..., M\} = \tag{73a}$$

$$\arg\max_{L} -\frac{1}{2}\sum_{l=1}^{L} \log(1 - \hat{\rho}_l^2) - \frac{L\eta(N)}{2N} \quad \text{s. t. } L \in \{1, 2, ..., M\} \tag{73b}$$

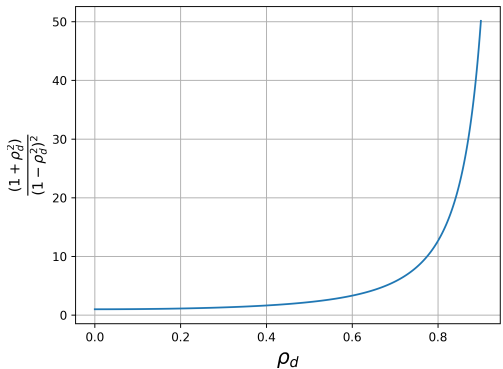
Are the previous opt. problem valid for this problem?

- Fisher information is non-singular ✓→ AIC and GIC are valid
- Fisher information is independent of $\boldsymbol{\rho}$ for infinite sample size (X) → BIC is not valid (approximately valid)

Model-order MI estimation: Why is BIC approximately valid?

The Fisher information satisfies:

$$\frac{1}{N}[\mathbf{F}(\boldsymbol{\rho}_D)]_{d,d} \xrightarrow{N\to\infty} \frac{(1+\rho_d^2)}{(1-\rho_d^2)^2} \tag{74}$$



Model-order MI estimation: Active channel + MI estimator

$$\arg\max_{L} -\frac{1}{2}\sum_{l=1}^{L}\log(1-\hat{\rho}_l^2) - \frac{L\eta(N)}{2N} \quad \text{s. t. } L \in \{1,2,...,M\} \tag{75}$$

Solved by means of a discrete derivative.

Active channel detector and its associated MI estimator

$$\mathcal{S}_{\eta(N)} = \left\{\rho \in [0,1] : \rho^2 \geq 1 - \exp\left(-\frac{\eta(N)}{N}\right)\right\} \tag{76a}$$

$$\hat{I}_R(\mathbf{x};\mathbf{y}|\mathcal{S}_{\eta(N)}) = -\frac{1}{2}\sum_{m=1}^{M}\log(1-\hat{\rho}_m^2\mathcal{I}_{\mathcal{S}_{\eta(N)}}(\hat{\rho}_m)) \tag{76b}$$

Key idea: Parallel channel assumption yields a simple decision rule.

Thanks! Recap

Conference Publications

1. C. A. Lopez, 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.
2. 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.
3. 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.

Journal publications

1. 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.
2. 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.
3. 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.

Takeaways

- Sparsity, entropy and subspaces can model the same phenomena.
- The Grassmann manifold can be incorporated to the MM framework.
  - MM algorithms on the Grassmannian are straightforward.
  - Block MM algorithms require some generalizations.
- Three different fusion schemes were studied.
  - Connections between Waterfilling and Sensor Fusion $\rightarrow$ Covariance Intersection
  - Operational information theoretic interpretation of the ℓ_0 regularization in a worst-case of contamination.
  - Practical fusion scheme by coupling fusion and regression operations.
- Studied another expression of diversity in wireless communications.
- Derived a regularized MI estimator of two Gaussian vectors.
  - Nominal case ✓
  - General case (X)

Supporting slides

G-convex optimization: Gradients and Hessians

Gradients on the Grassmann manifold

The gradient on the Grassmann manifold at a point $\mathbf{X} \in \mathrm{Gr}(N, D)$ of a function $f(\mathbf{X})$ is the unique tangent vector that satisfies:

$$\langle \text{grad } f(\mathbf{X}), \mathbf{\Delta} \rangle_{\mathbf{X}} = \left. \frac{\mathrm{d} f(\mathbf{\Gamma}(t))}{\mathrm{d}t} \right|_{t=0} \quad \forall \mathbf{\Delta} \in \mathcal{T}_{\mathbf{X}} \mathrm{Gr}(N, D) \tag{77}$$

Gradients are computed (in practice) as:

$$\text{grad } f(\mathbf{X}) = (\mathbf{I}_N - \mathbf{X}\mathbf{X}^T) \nabla_{\mathbf{X}} f(\mathbf{X}) \tag{78}$$

Hessians on the Grassmann manifold

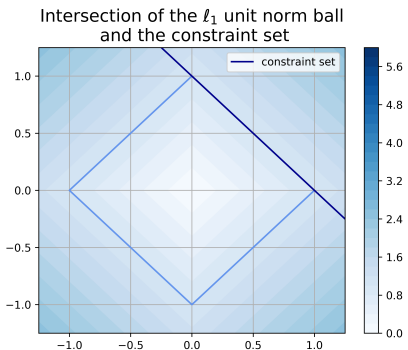
The Hessian of a function $f(\mathbf{X})$ at a point $\mathbf{X}$ on the Grassmann manifold is defined as:

$$\text{hess } f(\mathbf{X})[\mathbf{\Delta}, \mathbf{\Delta}] = \left. \frac{\mathrm{d}^2 f(\mathbf{\Gamma}(t))}{\mathrm{d}t^2} \right|_{t=0} \tag{79}$$

Contaminated Gaussian scheme: Why does the ℓ_1 regularization fail in general?

For $\beta \to \infty$:

$$\min_{\mathbf{f}} \varepsilon(\mathbf{f}^T \mathbf{Q} \mathbf{f}) + ||\mathbf{f}||_0 \quad \text{s.t.} \quad \mathbf{f}^T \mathbf{1}_M = 1 \tag{80}$$

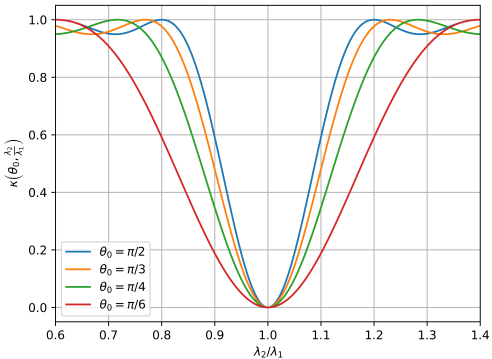


Covariance Conversion: Why is this problem difficult?

Let us compute $||\mathbf{R}_1 - \mathbf{R}_2||_F^2$ with:

$$\rho(\theta) = \rho_0 \delta(\theta - \theta_0) \tag{81}$$

This distance is proportional to...



Model-order MI estimation: What about non-parallel scenarios?

Key idea: Leverage the invariance with respect to homeomorphisms of the MI $\rightarrow$ Canonical Correlation Analysis (CCA)

Key insights into CCA

Find transformations of the original channel vectors such that the resulting ones fulfill the parallel channels condition:

$$\mathbf{u}(n) = \mathbf{A}\mathbf{x}(n) \tag{82a}$$

$$\mathbf{v}(n) = \mathbf{B}\mathbf{y}(n) \tag{82b}$$

It is known that $\mathbf{A}$ and $\mathbf{B}$ are obtained from the following SVD:

$$\mathbf{C}_x^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_y^{-\frac{1}{2}} = \mathbf{F}_*\mathbf{\Lambda}\mathbf{G}_*^T \tag{83}$$

where:

$$0 \leq [\mathbf{\Lambda}]_{m,m} = |\rho_m| \leq 1 \tag{84}$$

Model-order MI estimation: Empirical CCA

In practice, $\mathbf{C}_x$, $\mathbf{C}_y$ and $\mathbf{C}_{xy}$ are unknown. To solve this problem, let:

$$\mathbf{X} = [\mathbf{x}(1), ..., \mathbf{x}(N)] = \mathbf{U}_x \mathbf{D}_x \mathbf{V}_x^T \tag{85a}$$

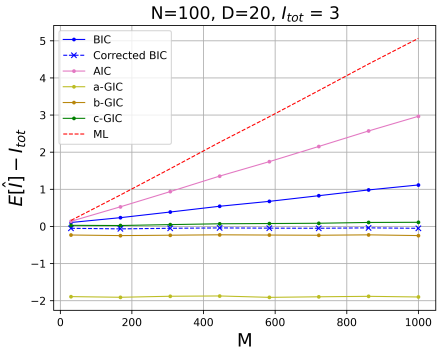
$$\mathbf{Y} = [\mathbf{y}(1), ..., \mathbf{y}(N)] = \mathbf{U}_y \mathbf{D}_y \mathbf{V}_y^T \tag{85b}$$

Then:

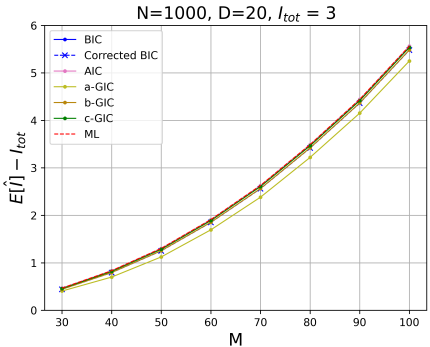
$$\hat{\mathbf{C}}_x^{-\frac{1}{2}} \hat{\mathbf{C}}_{xy} \hat{\mathbf{C}}_y^{-\frac{1}{2}} = \mathbf{U}_x \mathbf{D}_x^{-1} \underbrace{\mathbf{U}_x^T \mathbf{U}_x}_{\mathbf{I}_M} \mathbf{D}_x \mathbf{V}_x^T \mathbf{V}_y \mathbf{D}_y \underbrace{\mathbf{U}_y^T \mathbf{U}_y}_{\mathbf{I}_M} \mathbf{D}_y^{-1} \mathbf{U}_y^T = \mathbf{U}_x \mathbf{V}_x^T \mathbf{V}_y \mathbf{U}_y^T. \tag{86}$$

Key idea: The singular values of $\mathbf{V}_x^T \mathbf{V}_y$ (principal angles) yield the canonical correlations.
Issue: Empirical CCA requires $N > 2M$. N is ambient space and M would be the intrinsic subspace dimension!

Model-order MI estimation: Preliminary results



(a) Nominal case.



(b) General case (Empirical CCA).

Supporting slides: Data fusion results 1

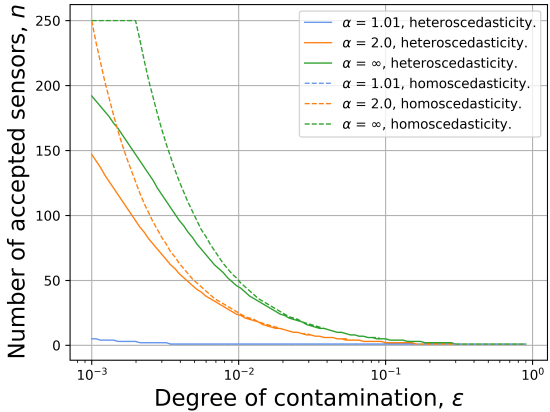
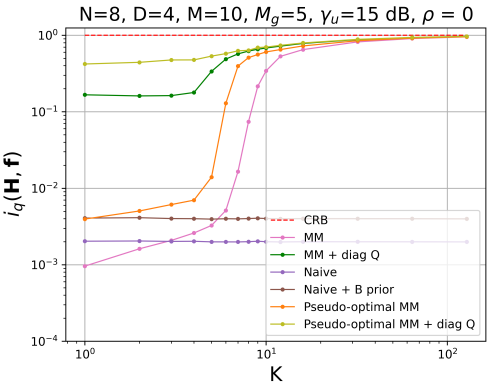
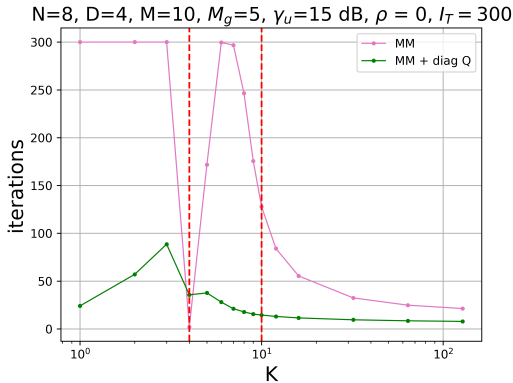


Figure: Number of selected sensors for different α and ε. Solid: homoscedasticity, $\gamma_m = \gamma$, $\forall m$; dashed: heteroscedasticity, $\gamma_m = (M - m + 1)\gamma$, where $\gamma > 0$ is any scale factor. The total amount of sensors is $M = 250$.

Supporting slides: Data fusion results 2



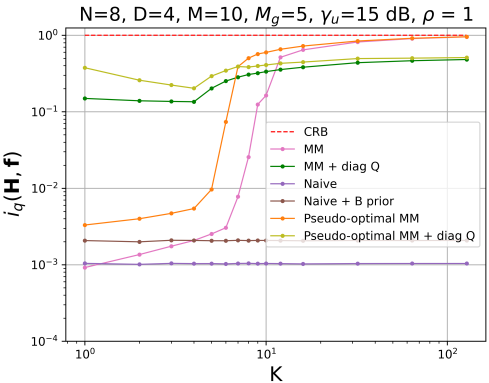
(a) Empiric NFQM



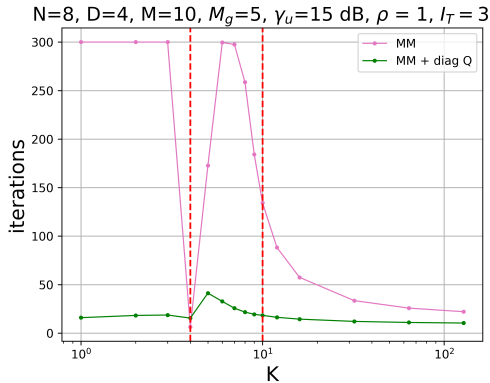
(b) Iterations.

Figure: Asymptotic behavior of the MM-based estimators in an uncorrelated sensor network.

Supporting slides: Data fusion results 3



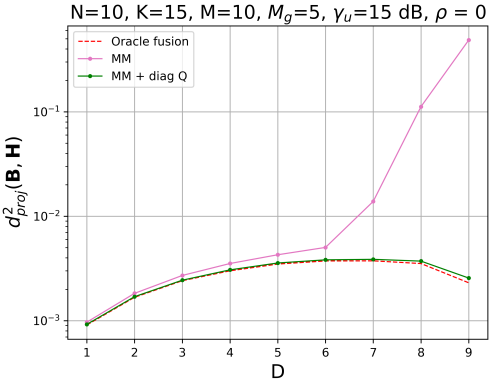
(a) Empiric NFQM



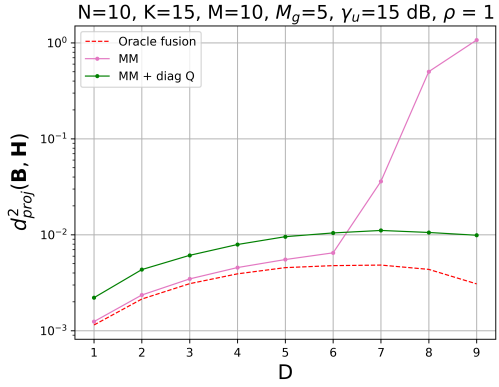
(b) Iterations.

Figure: Asymptotic behavior of the MM-based estimators in a correlated sensor network.

Supporting slides: Data fusion results 4



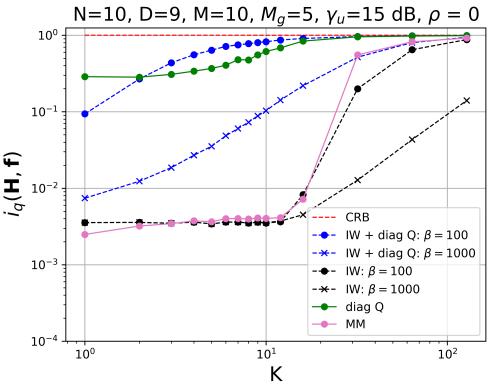
(a) Uncorrelated sensor network.



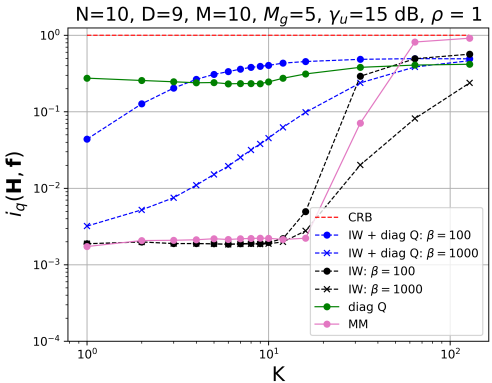
(b) Correlated sensor network.

Figure: Testing the MM-based subspace estimators with respect to the squared projection F-norm of the Grassmann manifold, $d^2_{proj}(\mathbf{B},\mathbf{H})$.

Supporting slides: Data fusion results 5



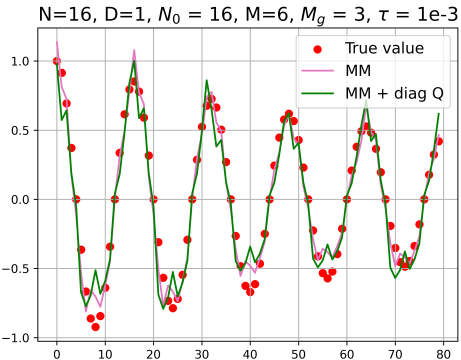
(a) Uncorrelated sensor network.



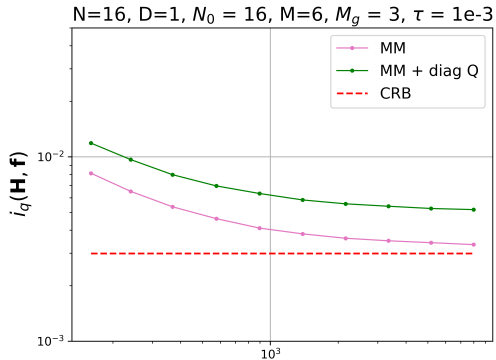
(b) Correlated sensor network.

Figure: Asymptotic performance of the approaches that are targeted towards the small sample size regime.

Supporting slides: Data fusion results 6



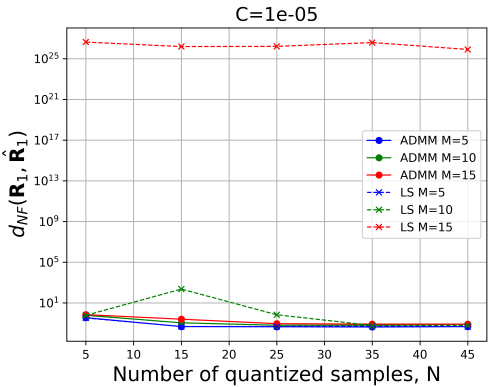
(a) Regression.



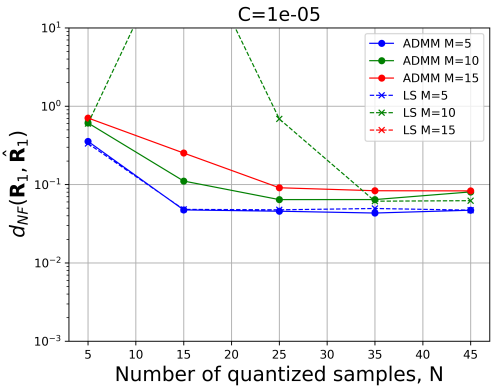
(b) MSE.

Figure: Dampened sinusoid toy example.

Supporting slides: Covariance Conversion results 1



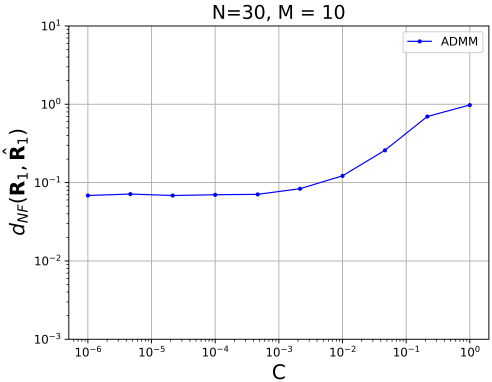
(a) Without zoom.



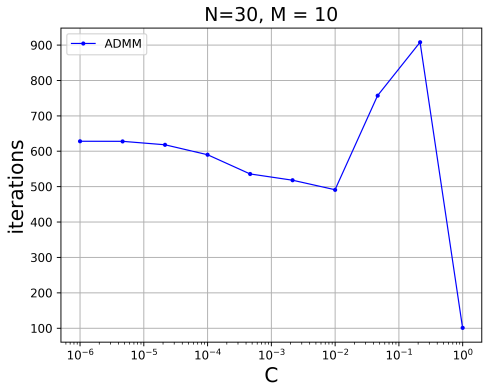
(b) Zoom.

Figure: Normalized Frobenius norm, $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$, as a function of the number of quantized samples.

Supporting slides: Covariance Conversion results 2



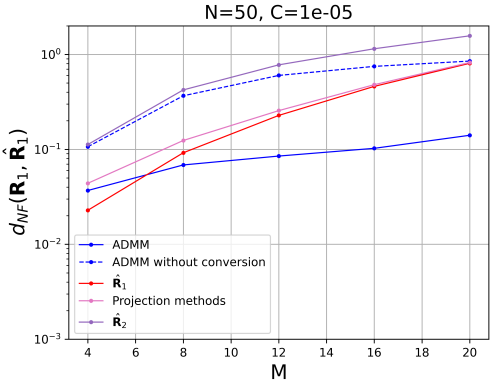
(a) $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$.



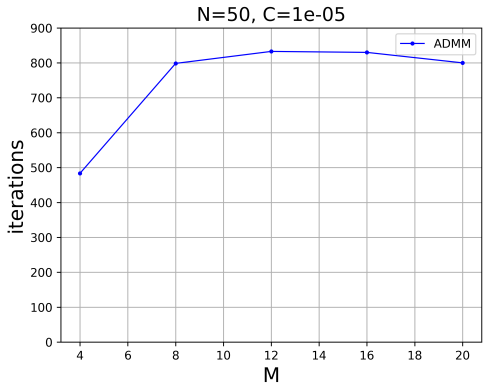
(b) Iterations.

Figure: Normalized Frobenius norm, $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$, as a function of ε.

Supporting slides: Covariance Conversion results 3



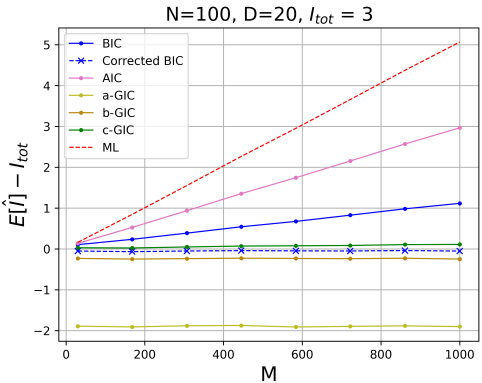
(a) $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$.



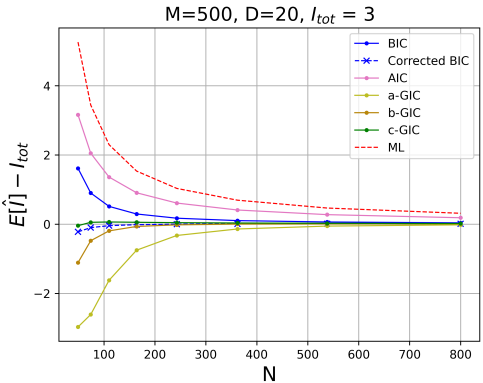
(b) Iterations.

Figure: Normalized Frobenius norm, $d_{NF}(\mathbf{R}_1, \hat{\mathbf{R}}_1)$, as a function of M for different CC approaches.

Supporting slides: MI estimation results 1



(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: Bias of the MI estimators under nominal conditions.

Supporting slides: MI estimation results 2

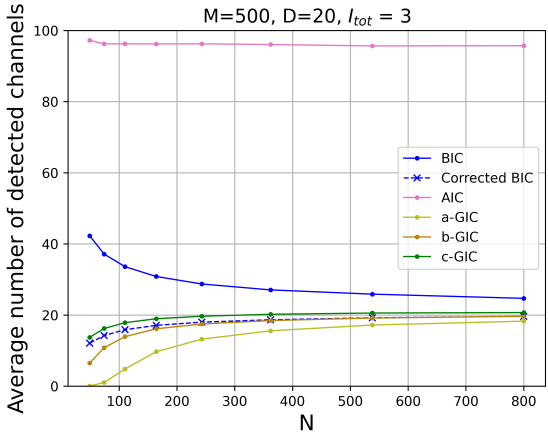
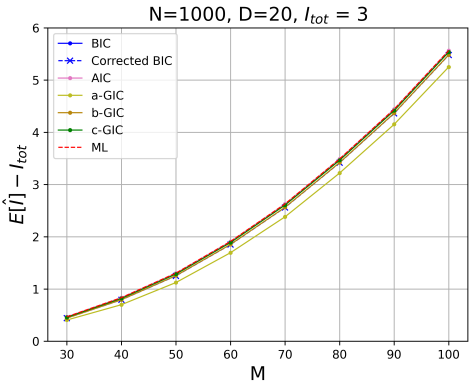
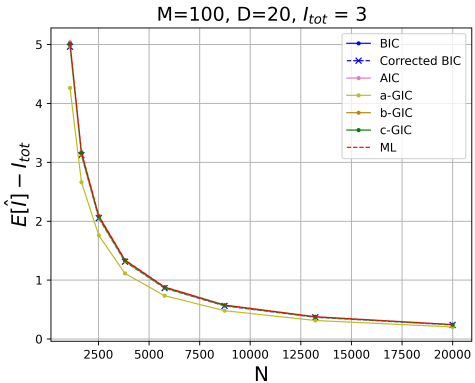


Figure: Average detected channels under nominal conditions as a function of the number of samples, N. The true value of active channels is $D = 20$.

Supporting slides: MI estimation results 3



(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: Bias of the MI estimators of mutually dependent datasets.

Supporting slides: MI estimation results 4

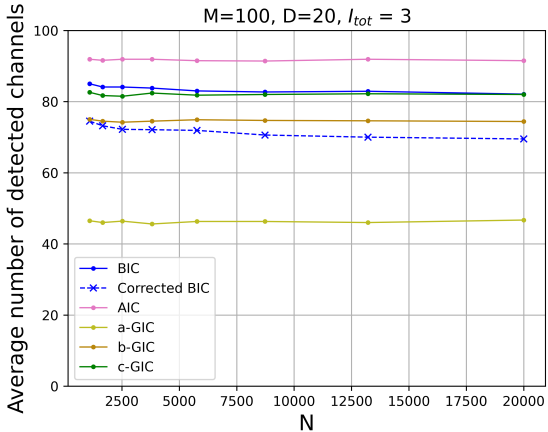


Figure: Average detected channels of mutually dependent datasets as a function of the number of samples, N. The true value of active channels is $D = 20$.