CONDITIONAL DEPENDENCE VIA U-STATISTICS PRUNING

Signal Processing & Communications

Ferran de Cabrera[1], Marc Vilà-Insa[2] and Jaume Riba[3]

Signal Processing and Communications Group (SPCOM)

Universitat Politècnica de Catalunya (UPC)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Departament de Teoria del Senyal i Comunicacions

Highlights

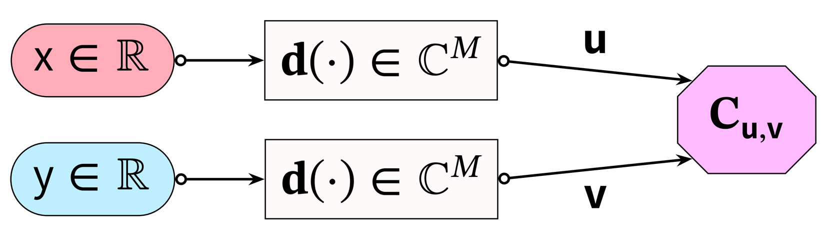
- **Problem to solve:** Measuring conditional dependence between two random phenomena when a confounder has a potential influence on the amount of information between them.
- **Proposed solution:** Novel measure based on incomplete unbiased statistics of degree two.
  - Allows to re-interpret independence as uncorrelatedness on a finite-dimensional feature space.
  - Prunes data according to observations of the confounder itself.
  - Avoids inverting ill-conditioned autocorrelation matrices.

Background

- Measuring conditional dependence is an important task in a variety of areas related to data science.
- Empirically estimating conditional dependence is challenging: known methods in the literature require inferring conditional marginal distributions.
  - Numerical issues related to matrix inversions and regularization.
  - Complexity scales cubically with the number of samples.
- **Proposal:** *conditional Hilbert-Schmidt independence criterion (C-HSIC)*
  - Translates statistical dependence into correlation: mapping data onto finite-dimensional spaces based on steering vectors [1].
  - Bypasses matrix inversion: conditioning is achieved by pruning pairwise differences of data under the control of the confounder [2].
  - Embraces the HSIC as a particular case when the U-Statistic is complete.

Marginal dependence as correlation

- Mapping $\mathbf{d}:\mathbb{R}\to\mathbb{C}^M$, $\quad [\mathbf{d}(\diamond)]_m \triangleq \frac{1}{\sqrt[4]{M}}\mathrm{G}\left(\frac{m}{\sqrt{M}}\right)e^{\jmath 2\pi \frac{m}{\sqrt{M}}\diamond}$



$$\lim_{M\to\infty}\|\mathbf{C}_{\mathbf{u},\mathbf{v}}\|_F^2 = 0 \iff \text{x, y independent} \tag{1}$$

- Data set with L i.i.d. samples of x and y: $\begin{cases}\mathbf{U}\triangleq[\mathbf{u}_1,\dots,\mathbf{u}_L]=[\mathbf{d}(x(1)),\dots,\mathbf{d}(x(L))]\\ \mathbf{V}\triangleq[\mathbf{v}_1,\dots,\mathbf{v}_L]=[\mathbf{d}(y(1)),\dots,\mathbf{d}(y(L))]\end{cases}$.
- Marginal dependence measure from empirical covariance matrix:

$$\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}} = \frac{1}{L-1}\mathbf{U}\underbrace{\boxed{\left(\mathbf{I}-{}^{1}\!/\!_{L}\cdot\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)}}_{\text{Data centering matrix}}\mathbf{V}^{\mathrm{H}} \implies \|\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}\|_F^2 = \frac{1}{(L-1)^2}\mathrm{Tr}(\mathbf{P}\mathbf{U}^{\mathrm{H}}\mathbf{U}\mathbf{P}\mathbf{V}^{\mathrm{H}}\mathbf{V}) \tag{2}$$

Unbounded dimension

$$\|\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}\|_F^2 \xrightarrow{M\to\infty} \mathrm{HSIC}\big(\{x(l)\};\{y(l)\}\big) \triangleq \lim_{M\to\infty}\frac{1}{(L-1)^2}\mathrm{Tr}(\mathbf{PKPQ}) \tag{3}$$

$$\left.\begin{array}{l}\mathbf{K}\triangleq \lim\limits_{M\to\infty}\mathbf{U}^{\mathrm{H}}\mathbf{U}\\ \mathbf{Q}\triangleq \lim\limits_{M\to\infty}\mathbf{V}^{\mathrm{H}}\mathbf{V}\end{array}\right\} \implies [\mathbf{K}]_{l,l'} = \lim_{M\to\infty}\mathbf{d}^{\mathrm{H}}(x(l))\mathbf{d}(x(l')) = \boxed{\kappa(x(l)-x(l'))}$$

$$\boxed{\kappa(\diamond)\triangleq\int_{-\infty}^{\infty}\mathrm{G}^2(f)e^{-\jmath 2\pi f\diamond}\,\mathrm{d}f}$$

- *E.g.* Universal Gaussian kernel: $\kappa(\diamond)=\exp\left(-\left(\diamond/(\hat{\sigma}L^{-1/5})\right)^2\right)$

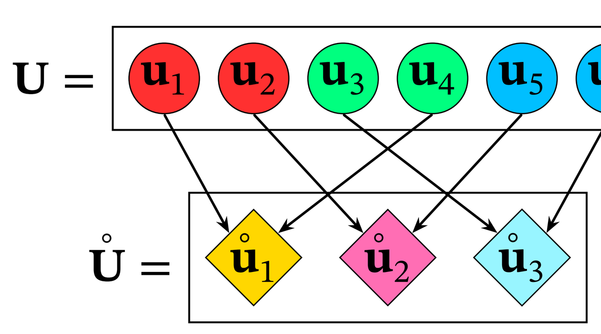
Sample covariance matrix as incomplete U-Statistic

- L samples $\implies K_{\max}\triangleq\frac{L(L-1)}{2}$ unique pairs.
- Orderings $\mathrm{f}_1(k)$ and $\mathrm{f}_2(k)$ for $1\le k\le K\le K_{\max}$.
- Pairwise differences: $\quad \mathring{\mathbf{u}}_k\triangleq\frac{\mathbf{u}_{\mathrm{f}_1(k)}-\mathbf{u}_{\mathrm{f}_2(k)}}{\sqrt{2}},\quad \mathring{\mathbf{v}}_k\triangleq\frac{\mathbf{v}_{\mathrm{f}_1(k)}-\mathbf{v}_{\mathrm{f}_2(k)}}{\sqrt{2}}$.
- Sample cross-covariance between virtual sources $\mathring{\mathbf{u}}$ and $\mathring{\mathbf{v}}$:

$$\left.\begin{array}{l}\mathring{\mathbf{U}}\triangleq[\mathring{\mathbf{u}}_1,\dots,\mathring{\mathbf{u}}_K]\\ \mathring{\mathbf{V}}\triangleq[\mathring{\mathbf{v}}_1,\dots,\mathring{\mathbf{v}}_K]\end{array}\right\}\implies \boxed{\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}}=\frac{1}{K}\mathring{\mathbf{U}}\mathring{\mathbf{V}}^{\mathrm{H}}} \quad \textit{(incomplete U-Statistic)} \tag{4}$$

- $K=K_{\max}\implies \widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}}\equiv\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}$ (complete).

Robustness to pruning



$\frac{L(L-1)}{2}-\left\lfloor\frac{L}{2}\right\rfloor$ unused pairs

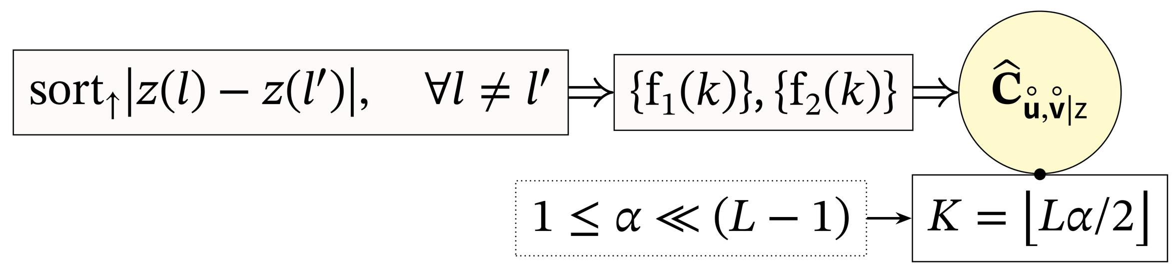
$\implies \left\lfloor\frac{L}{2}\right\rfloor$ i.i.d. pairs ($\mathcal{O}(L)$)

Conditional dependence via U-Statistics

- Conditional cross-covariance matrix: $\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathrm{z}}\triangleq\int_{\mathbb{R}}\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathrm{z}=z}\,\mathrm{d}F_{\mathrm{z}}(z)$.
- Virtual random variable: $\mathring{\mathrm{z}}\triangleq\mathrm{z}_1-\mathrm{z}_2$, with independent $\mathrm{z}_1,\mathrm{z}_2\sim\mathrm{z}$.
- Alternative conditioning:

$$\begin{aligned}\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathrm{z}} &= \iint_{\mathbb{R}^2}\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|z_1=z_2}\,\mathrm{d}F_{\mathrm{z}}(z_1)\,\mathrm{d}F_{\mathrm{z}}(z_2) = \iint_{\mathbb{R}^2}\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{z}=0}\,\mathrm{d}F_{\mathrm{z}}(z_1)\,\mathrm{d}F_{\mathrm{z}}(z_2)\\ &= \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{z}=0}\cdot\iint_{\mathbb{R}^2}\mathrm{d}F_{\mathrm{z}}(z_1)\,\mathrm{d}F_{\mathrm{z}}(z_2) = \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{z}=0}\end{aligned} \tag{5}$$

- Data pruning controlled by close samples of z:



Conditional HSIC

- Conditional dependence measure: $\|\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathrm{z}}\|_F^2=\frac{1}{K^2}\mathrm{Tr}\big(\mathring{\mathbf{U}}^{\mathrm{H}}\mathring{\mathbf{U}}\mathring{\mathbf{V}}^{\mathrm{H}}\mathring{\mathbf{V}}\big)$.
- Asymptotically, it approaches the C-HSIC:

$$\|\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathrm{z}}\|_F^2\xrightarrow{M\to\infty}\text{C-HSIC}_\alpha\big(\{x(l)\};\{y(l)\}\big)\triangleq\frac{1}{4K^2}\mathrm{Tr}\big(\breve{\mathbf{K}}\breve{\mathbf{Q}}\big) \tag{6}$$

$$\begin{cases}\breve{\mathbf{K}}\triangleq\mathbf{K}_{1,1}+\mathbf{K}_{2,2}-\mathbf{K}_{1,2}-\mathbf{K}_{1,2}^{\mathrm{H}}\longrightarrow[\mathbf{K}_{\heartsuit,\spadesuit}]_{k,k'}=\kappa\big(x(\mathrm{f}_\heartsuit(k))-x(\mathrm{f}_\spadesuit(k'))\big)\\ \breve{\mathbf{Q}}\triangleq\mathbf{Q}_{1,1}+\mathbf{Q}_{2,2}-\mathbf{Q}_{1,2}-\mathbf{Q}_{1,2}^{\mathrm{H}}\longrightarrow[\mathbf{Q}_{\heartsuit,\spadesuit}]_{k,k'}=\kappa\big(y(\mathrm{f}_\heartsuit(k))-y(\mathrm{f}_\spadesuit(k'))\big)\end{cases} \tag{7}$$

Numerical illustration

$$\mathcal{M}^+:\begin{cases}\mathrm{I}(\mathrm{x};\mathrm{y})\ge 0\\ \mathrm{I}(\mathrm{x};\mathrm{y}|\mathrm{z})=0\end{cases},\qquad \mathcal{M}^-:\begin{cases}\mathrm{I}(\mathrm{x};\mathrm{y})=0\\ \mathrm{I}(\mathrm{x};\mathrm{y}|\mathrm{z})\ge 0\end{cases}. \tag{8}$$

Figure 1. Model $\mathcal{M}^+$.

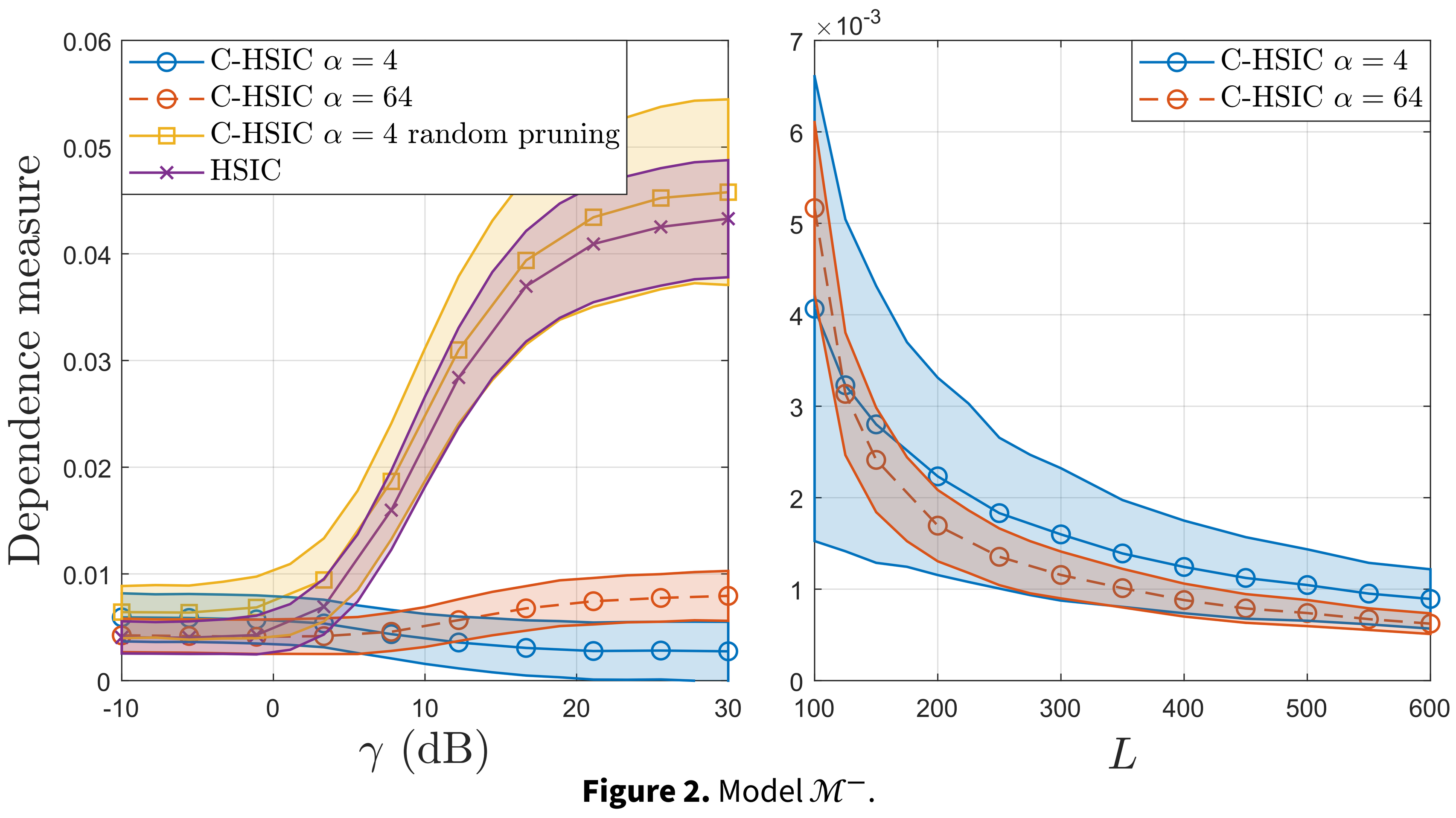
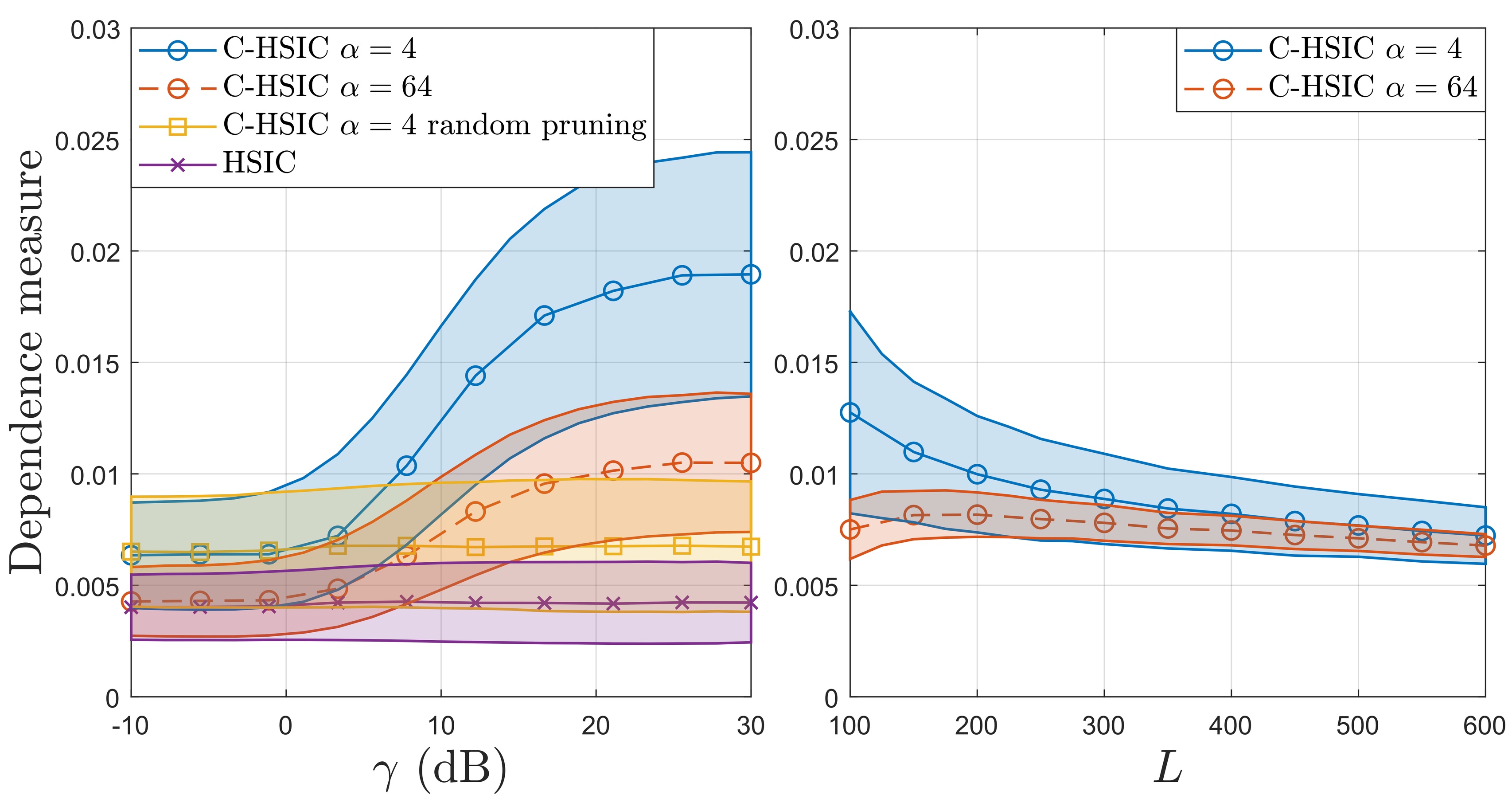


Figure 2. Model $\mathcal{M}^-$.



F. de Cabrera, M. Vilà-Insa, and J. Riba, "Conditional dependence via U-Statistics pruning," *IEEE Signal Processing Letters*, vol. 32, pp. 981–985, 2025, ISSN: 1558-2361. DOI: 10.1109/lsp.2025.3539587

References

[1] 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, Aug. 2023, ISSN: 1557-9654. DOI: 10.1109/tit.2023.3258182.

[2] M. Vila and J. Riba, "A test for conditional correlation between random vectors based on weighted U-Statistics," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 5792–5796. DOI: 10.1109/icassp43922.2022.9747042.

Funding

This work was (partially) funded by project MAYTE (PID2022-136512OB-C21) by MICIU/AEI/10.13039/501100011033 and ERDF/EU, grant 2021 SGR 01033, and grant 2022 FI SDUR 00164 by Departament de Recerca i Universitats de la Generalitat de Catalunya.

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES · Cofinanciado por la Unión Europea · AGENCIA ESTATAL DE INVESTIGACIÓN

ORCID: [1]0000-0001-6949-780X, 0000-0002-{[2]7032-1411, [3]5515-8169} EUSIPCO'25 [1]ferrandecabrera@gmail.com, {[2]marc.vila.insa, [3]jaume.riba}@upc.edu