A TEST FOR CONDITIONAL CORRELATION BETWEEN RANDOM VECTORS BASED ON WEIGHTED U-STATISTICS

icassp 2022 Singapore

Marc Vilà Insa (*IEEE Student Member*) and **Jaume Riba Sagarra** (*IEEE Senior Member*)

Signal Theory and Communications Department (TSC), Technical University of Catalonia (UPC)

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

Motivation

Statistical graphical models are fundamental tools for representing interrelationships between variables and have found applications in many fields of science and technology. A recurrent problem associated with them is determining whether two variables are correlated when conditioned to a third one, called *confounder*. This task is usually affected by **the curse of dimensionality** [1]: improvements in data acquisition techniques have brought a much faster increase in their dimensionality than the speed at which samples are available. Not only does this phenomenon cause an unbearable rise in computational complexity of classical methods, but it also violates many of their statistical assumptions, making them perform poorly.

This issue motivates the development of detection methods for conditional correlation that are robust to high-dimensional/small-sample regimes.

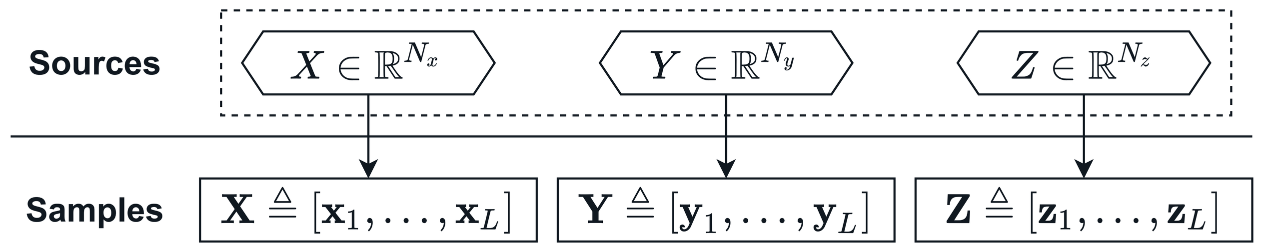
Problem statement

Preliminary definitions

- U and V are generic sources and can represent sources X and Y indistinctly.
- Z is the potential confounder.

Average conditional cross-covariance matrix

$$\mathbf{C}_{UV|Z} \triangleq \mathrm{E}_Z\left[\mathbf{C}_{U,V|Z=\mathbf{z}}\right] = \int_{\mathbb{R}^{N_z}} \mathbf{C}_{UV|Z=\mathbf{z}} dF_Z(\mathbf{z}) \quad (1)$$



The problem studied is the detection of correlation between X and Y conditioned to Z. Its related binary hypothesis test can be defined as:

$$\left.\begin{array}{l}\mathcal{H}_0 : \mathbf{C}_{XY|Z} = \mathbf{0}\\ \mathcal{H}_1 : \mathbf{C}_{XY|Z} \neq \mathbf{0}\end{array}\right\}. \quad (2)$$

Tests for correlation

Consider the *Likelihood Ratio Test (LRT)* associated with the previous problem:

$$\frac{\max_{\mathbf{C}_{WW|Z}} f(\mathbf{W}|\mathbf{C}_{WW|Z})}{\max_{\mathbf{C}_{XX|Z}} f(\mathbf{X}|\mathbf{C}_{XX|Z}) \max_{\mathbf{C}_{YY|Z}} f(\mathbf{Y}|\mathbf{C}_{YY|Z})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda, \quad W \triangleq \begin{bmatrix} X \\ Y \end{bmatrix}. \quad (3)$$

Most approaches for solving it involve determinants or inverses [2], [3], which might become computationally problematic. An alternative test for correlation that avoids these issues is the **RV coefficient**:

$$\mathrm{T_{RV}}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) \triangleq \frac{\|\widehat{\mathbf{C}}_{XY|Z}\|_F^2}{\|\widehat{\mathbf{C}}_{XX|Z}\|_F \|\widehat{\mathbf{C}}_{YY|Z}\|_F}, \quad \mathrm{T_{RV}} \downarrow \Rightarrow \mathcal{H}_0. \quad (4)$$

If data is Gaussian, the matrices involved can be obtained from *Schur complements*:

$$\widehat{\mathbf{C}}_{UV|Z} \triangleq \widehat{\mathbf{C}}_{UV} - \widehat{\mathbf{C}}_{UZ}\widehat{\mathbf{C}}_{ZZ}^{-1}\widehat{\mathbf{C}}_{ZV}. \quad (5)$$

A matrix inversion is involved, bringing the same computational problems. We need to find an alternative that avoids such issues and the Gaussianity assumption.

Acknowledgements

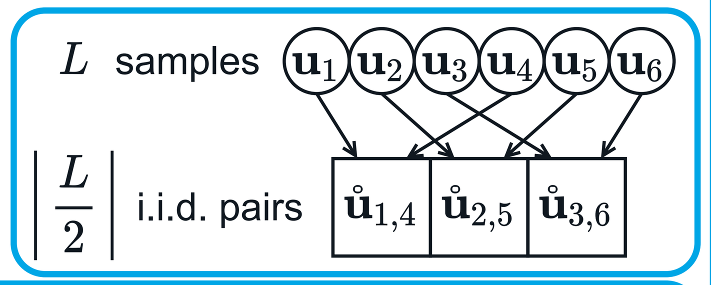
This work has been supported by the Spanish Ministry of Science and Innovation through project RODIN (PID2019-105717RB-C22/MCIN/AEI/10.13039/501100011033).

U-Statistics test for conditional correlation

Covariance estimation using U-Statistics

$$\widehat{\mathbf{C}}_{UV} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \mathring{\mathbf{u}}_{i,j} \mathring{\mathbf{v}}_{i,j}^T, \quad \mathring{\mathbf{u}}_{i,j} \triangleq \frac{\mathbf{u}_i - \mathbf{u}_j}{\sqrt{2}}, \quad \mathring{\mathbf{v}}_{i,j} \triangleq \frac{\mathbf{v}_i - \mathbf{v}_j}{\sqrt{2}} \quad (6)$$

Working with pairs provides an intrinsic redundancy. By discarding the non-independent ones, we can obtain an incomplete U-Covariance matrix with our desired degree of accuracy and reduced complexity.



Incomplete U-Covariance Matrix

$$\widehat{\mathbf{C}}'_{UV} = \frac{1}{\lfloor L/2 \rfloor} \sum_{i=1}^{\lfloor L/2 \rfloor} \mathring{\mathbf{u}}_{i,i+\lfloor L/2 \rfloor} \mathring{\mathbf{v}}_{i,i+\lfloor L/2 \rfloor}^T, \quad \text{Unused pairs: } \Delta L = \frac{L(L-1)}{2} - \left\lfloor \frac{L}{2} \right\rfloor \quad (7)$$

Weighted U-Statistics for conditional uncorrelatedness

***Virtual* random variables:** $U_m \sim U, \quad V_m \sim V, \quad Z_m \sim Z \to$ independent for different $m = 1, 2$

$$\mathring{U} \triangleq \frac{U_1 - U_2}{\sqrt{2}} \to \mathring{\mathbf{u}}_{i,j}, \quad \mathring{V} \triangleq \frac{V_1 - V_2}{\sqrt{2}} \to \mathring{\mathbf{v}}_{i,j}, \quad \mathring{Z} \triangleq \frac{Z_1 - Z_2}{\sqrt{2}} \to \mathring{\mathbf{z}}_{i,j}$$

It is known that $\mathbf{C}_{UV} \equiv \mathbf{C}_{\mathring{U}\mathring{V}}$ [4], so the average conditional covariance matrix expression (1) can be rewritten with it. This allows to obtain an equivalent formulation:

$$\mathbf{C}_{UV|Z} = \int_{\mathbb{R}^{N_z}} \mathbf{C}_{\mathring{U}\mathring{V}|Z=\mathbf{z}} dF_Z(\mathbf{z}) = \iint_{\mathbb{R}^{N_z \times N_z}} \mathbf{C}_{\mathring{U}\mathring{V}|\mathring{Z}(Z_1,Z_2)=\mathbf{0}} dF_{Z_1,Z_2}(\mathbf{z}_1, \mathbf{z}_2) = \boxed{\mathbf{C}_{\mathring{U}\mathring{V}|\mathring{Z}=\mathbf{0}}}. \quad (8)$$

Since $\Pr\{\mathring{Z} = 0\} = 0$ for continuous variables, we relax this criterion by using the pairs of samples such that $0 \leq \|\mathring{Z}\| \leq \epsilon'$. The proposed estimator of conditional covariance can then be computed as:

$$\check{\mathbf{C}}_{UV|Z} \triangleq \frac{\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \mathring{\mathbf{u}}_{i,j} \mathring{\mathbf{v}}_{i,j}^T I_\epsilon(\|\mathbf{z}_i - \mathbf{z}_j\|)}{\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} I_\epsilon(\|\mathbf{z}_i - \mathbf{z}_j\|)}, \quad I_\epsilon(\lambda) \triangleq \begin{cases} 1, & 0 \leq \lambda \leq \epsilon \\ 0, & \text{otherwise} \end{cases}. \quad (9)$$

Order statistics

Calibrating ϵ might be very sensitive to the specific data. Motivated by the previously mentioned redundancy, we present an alternative pair selection method.

For a given ϵ, only the $L_p \in [1, L(L-1)/2]$ data pairs corresponding to the smallest norms of $\mathring{\mathbf{z}}_{i,j}$ will be used in the estimation. For that reason it is very convenient to sort the L_p smallest values of $\|\mathring{\mathbf{z}}_{i,j}\|$ in ascending order in $\mathring{\mathbf{z}}_{\mathrm{sort}}$. Function $q(l) \to (i(l), j(l))$ returns the pair of indices from $\mathbf{z}$ samples that correspond to entry l of $\mathring{\mathbf{z}}_{\mathrm{sort}}$.

$$\left\{\begin{array}{c} \|\mathring{\mathbf{z}}_{1,2}\| \\ \|\mathring{\mathbf{z}}_{1,3}\| \\ \vdots \\ \|\mathring{\mathbf{z}}_{L-1,L}\| \end{array}\right\} \xrightarrow{\text{sort}} \mathring{\mathbf{z}}_{\mathrm{sort}} \triangleq \begin{bmatrix} \|\mathring{\mathbf{z}}_{i(1),j(1)}\| \text{ (min)} \\ \vdots \\ \|\mathring{\mathbf{z}}_{i(l),j(l)}\| \\ \vdots \\ \|\mathring{\mathbf{z}}_{i(L_p),j(L_p)}\| \end{bmatrix} \xrightarrow{q(l)} \left\{\begin{array}{c} (i(1), j(1)) \\ \vdots \\ (i(l), j(l)) \\ \vdots \\ (i(L_p), j(L_p)) \end{array}\right\}. \quad (10)$$

$$\check{\mathbf{C}}_{UV|Z} = \frac{1}{2L_p} \sum_{l=1}^{L_p} (\mathbf{u}_{i(l)} - \mathbf{u}_{j(l)})(\mathbf{v}_{i(l)} - \mathbf{v}_{j(l)})^T$$

$$\Downarrow \quad (11)$$

$$\check{\mathrm{T}}_{\mathrm{RV}}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \frac{\|\check{\mathbf{C}}_{XY|Z}\|_F^2}{\|\check{\mathbf{C}}_{XX|Z}\|_F \|\check{\mathbf{C}}_{YY|Z}\|_F}$$

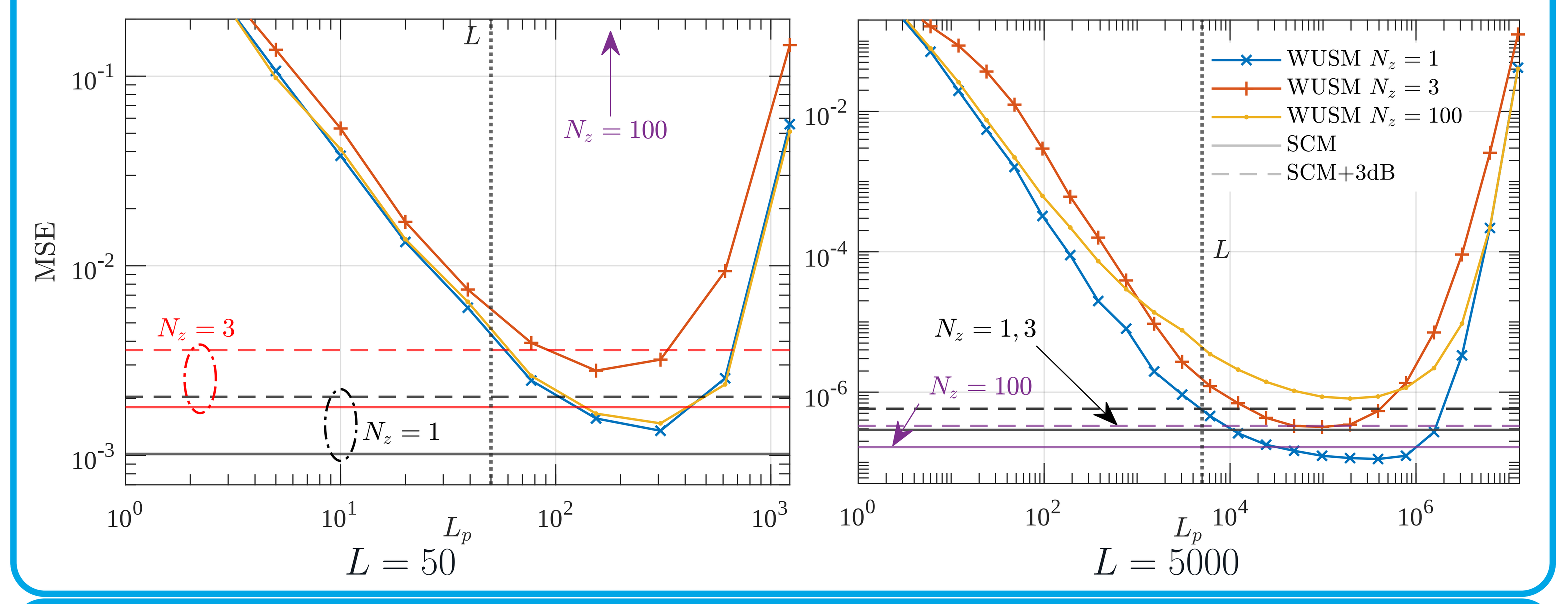
Numerical results

Test settings

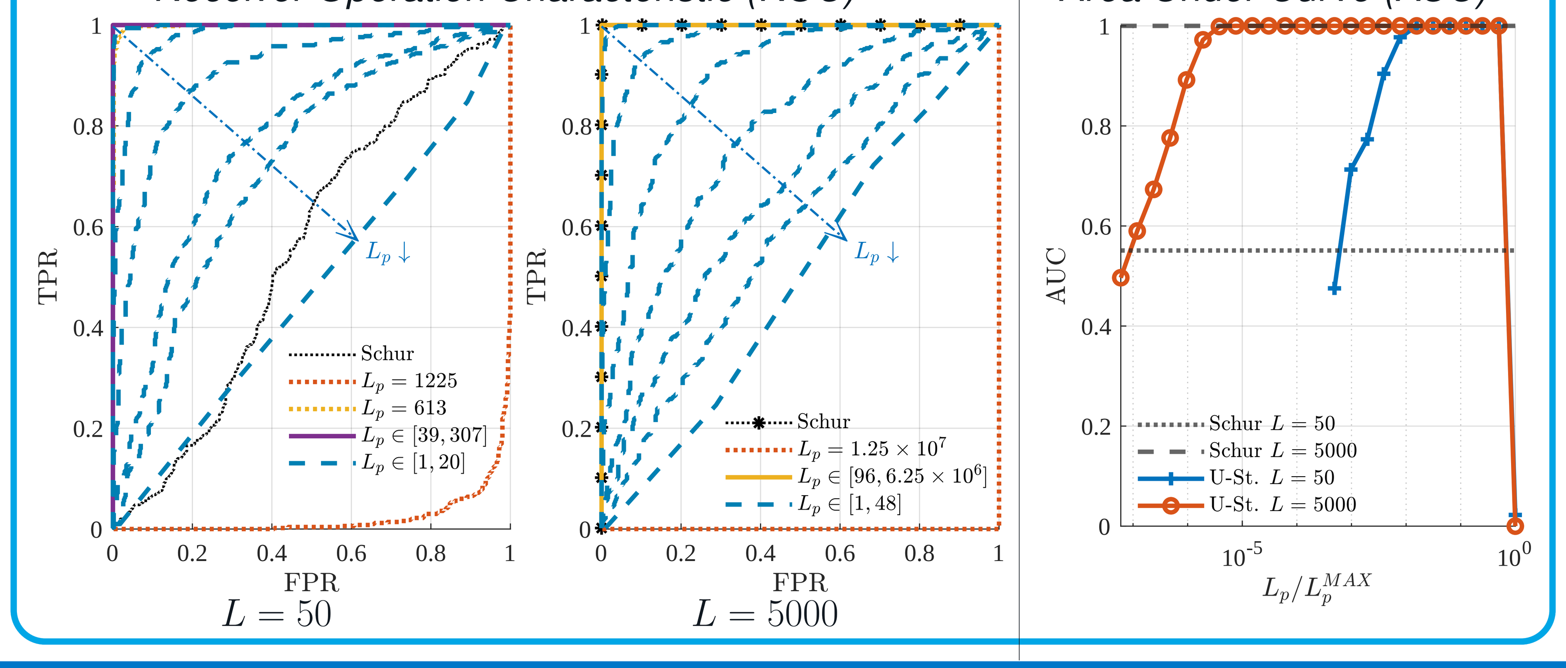
- Nº of averaged tests (M): 500
- Techniques:
  - Weighted U-Stats. Method *(WUSM)*
  - Schur Complement Method *(SCM)*
- Mean: $\mathrm{E}[X] = \mathrm{E}[Y] = \mathrm{E}[Z_n] = 0$
- Power: $\mathrm{E}[X^2] = \mathrm{E}[Y^2] = \mathrm{E}[Z_n^2] = 1$
- RV Coefficient ($\mathrm{T_{RV}}$):
  - $\mathrm{T_{RV}}(X, Y) > 0.2$
  - $\mathrm{T_{RV}}(X, Y|Z) \approx 0$
- Data model: *Gaussian Copula*

Estimation

Mean Squared Error: $\mathrm{MSE}(\check{\mathrm{T}}_{\mathrm{RV}}) \triangleq \mathrm{var}(\check{\mathrm{T}}_{\mathrm{RV}}) + \mathrm{bias}^2(\check{\mathrm{T}}_{\mathrm{RV}})$



Detection ($N_z = 100$)



Future research

Design aspects
- Alternative criteria for sorting (different norms).
- Soft indicator functions and data-driven weighting.

Applications
- Integration of information theoretic methods: moving from correlation to dependence (*characteristic function mapping* [5]).

References

[1] D. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1–32, Jan. 2000.

[2] R. López-Valcarce, G. Vazquez-Vilar, and J. Sala, "Multiantenna spectrum sensing for Cognitive Radio: overcoming noise uncertainty," in *2010 2nd International Workshop on Cognitive Information Processing*, 2010, pp. 310–315.

[3] D. Ramirez, 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.

[4] A. J. Lee, *U-Statistics: Theory and Practice.* Routledge, 2019.

[5] J. Riba and F. de Cabrera, "Regularized Estimation of Information via High Dimensional Canonical Correlation Analysis," 2020, arXiv:2005.02977 [cs.IT].