Accepted version of the article published in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022. DOI: 10.1109/ICASSP43922.2022.9747042. Available at: http://ieeexplore.ieee.org

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

Marc Vilà, Student Member, IEEE, and *Jaume Riba,* Senior Member, IEEE

Signal Theory and Communications Department, Technical University of Catalonia (SPCOM/UPC)
{marc.vila.insa, jaume.riba}@upc.edu

ABSTRACT

This article explores U-Statistics as a tool for testing conditional correlation between two multivariate sources with respect to a potential confounder. The proposed approach is effectively an instance of *weighted U-Statistics* and does not impose any statistical model on the processed data, in contrast to other well-known techniques that assume Gaussianity. By avoiding determinants and inverses, the method presented displays promising robustness in small-sample regimes. Its performance is evaluated numerically through its *MSE* and *ROC* curves.

***Index Terms*—** Conditional Correlation, U-Statistics, Order Statistics, Data-Driven Signal Processing, Small Sample Regime

1. INTRODUCTION

Statistical graphical models are a fundamental tool in studying the structure of interrelationships among sets of variables using graphs. Over the past years, they have found application in a variety of fields of science and technology, such as sensor networks [1], cybersecurity [2] or image processing [3]. *Gaussian Graphical Models (GGMs)* are of particular interest due to several desirable properties, the most outstanding one being that studying statistical dependence within them is equivalent to studying correlation [4].

As data acquisition techniques improve, the amount of information to be processed grows dramatically. The super-linear rise in computational complexity of many classical estimation methods can render them intractable for demanding applications. The prevalence of high dimensional signals also presents more fundamental challenges: in practice, data dimensionality increases much faster than the number of available observations, which is often referred to as *the curse of dimensionality* [5]. It violates many statistical assumptions of classical methods, making them perform poorly under such circumstances.

Especially sensitive to high-dimensional data with small sample sizes [6] is the estimation of *precision matrices*, *i.e.* inverse covariance matrices. This problem is very relevant in GGMs, since zero components in their precision matrices imply conditional independence between variables. Many approaches to solving it have been proposed over the years, one of the most remarkable ones being the imposition of an ℓ_1-norm constraint, known as *graphical LASSO* [7].

In this paper we deal with a closely related problem: detecting conditional uncorrelatedness between multivariate sources. We base our test on the well-known *RV coefficient* [8] computed from the covariance matrices of the variables conditioned to a potential confounder. Classical techniques, which rely on assuming Gaussian data, estimate these matrices by taking the *Schur complement* of the confounder covariance matrix [9], making them *model-driven* and vulnerable to limited sample sizes. Instead, we depart from such approaches and develop a method based on weighted *unbiased statistics (U-Statistics)* [10]. By letting the statistics of the confounder dictate the samples to be processed, our approach becomes more *data-driven* and agnostic to underlying statistical models. Furthermore, it avoids matrix inversions and effectively bypasses the curse of dimensionality. Finally, we subject our method to various numerical simulations, to test whether it displays favorable performance and robustness as a test for conditional correlation.

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

2. PROBLEM STATEMENT

Consider a set of three sources $X \in \mathbb{R}^{N_x}$, $Y \in \mathbb{R}^{N_y}$ and $Z \in \mathbb{R}^{N_z}$. L i.i.d. samples are available from each one of them, denoted and grouped as $\mathbf{X} \triangleq [\mathbf{x}_1, \ldots \mathbf{x}_L]$, $\mathbf{Y} \triangleq [\mathbf{y}_1, \ldots \mathbf{y}_L]$ and $\mathbf{Z} \triangleq [\mathbf{z}_1, \ldots \mathbf{z}_L]$, respectively. We define the statistical mean and covariance of any given random vectors U and V as $\boldsymbol{\mu}_U \triangleq \mathrm{E}[\mathbf{u}_l]$, $\boldsymbol{\mu}_V \triangleq \mathrm{E}[\mathbf{v}_l]$ and $\mathbf{C}_{UV} \triangleq \mathrm{E}\left[\mathbf{u}_l \mathbf{v}_l^T\right] - \boldsymbol{\mu}_U \boldsymbol{\mu}_V^T$.

The problem studied in this article is the detection of correlation between X and Y conditioned to Z, the potential *confounder*. Its related binary hypothesis test can be defined as:

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{C}_{XY|Z} = \mathbf{0} \\ \mathcal{H}_1 &: \mathbf{C}_{XY|Z} \neq \mathbf{0}, \end{aligned} \tag{1}$$

where $\mathbf{C}_{XY|Z}$ is the *average conditional cross-covariance matrix* [11] between sources X and Y with respect to Z, computed as

$$\mathbf{C}_{XY|Z} \triangleq \mathrm{E}_Z\left[\mathbf{C}_{X,Y|Z=\mathbf{z}}\right] = \int_{\mathbb{R}^{N_z}} \mathbf{C}_{XY|Z=\mathbf{z}} dF_Z(\mathbf{z}). \tag{2}$$

$\mathbf{C}_{XY|Z=\mathbf{z}}$ is the conditional covariance matrix between X and Y for any value $\mathbf{z}$ of Z, and $F_Z(\mathbf{z})$ is its cumulative distribution function.

In this work we have only considered two variables and a single confounder for space limitations and ease of notation. However, the derivations and results obtained are directly valid for any amount of variables and confounders with arbitrary dimensions.

3. TESTS FOR CORRELATION

Before dealing with (1), consider the same binary hypothesis test without conditioning on Z. The *Generalized Likelihood-Ratio Test (GLRT)* related to it is stated as follows:

$$\frac{\max_{\mathbf{C}_{WW}} f(\mathbf{W}|\mathbf{C}_{WW})}{\max_{\mathbf{C}_{XX}} f(\mathbf{X}|\mathbf{C}_{XX}) \max_{\mathbf{C}_{YY}} f(\mathbf{Y}|\mathbf{C}_{YY})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda, \tag{3}$$

where $W \triangleq [X^T \, Y^T]^T$. Assuming jointly Gaussian sources, the solution to (3) results in the *Hadamard Ratio Test (HRT)* [12]:

$$\mathrm{T_{HAD}}(\mathbf{X},\mathbf{Y}) \triangleq \frac{\det[\widehat{\mathbf{C}}_{WW}]}{\det[\widehat{\mathbf{C}}_{XX}]\det[\widehat{\mathbf{C}}_{YY}]} \in [0,1]. \tag{4}$$

Values close to 1 indicate uncorrelatedness between X and Y. For two generic sources U and V, $\widehat{\boldsymbol{\mu}}_U$, $\widehat{\boldsymbol{\mu}}_V$ and $\widehat{\mathbf{C}}_{UV}$ denote the sample means and covariance, computed as

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_U &\triangleq \tfrac{1}{L}\textstyle\sum_{l=1}^{L}\mathbf{u}_l\,, \quad \widehat{\boldsymbol{\mu}}_V \triangleq \tfrac{1}{L}\textstyle\sum_{l=1}^{L}\mathbf{v}_l \\ \widehat{\mathbf{C}}_{UV} &\triangleq \tfrac{1}{L-1}\textstyle\sum_{l=1}^{L}(\mathbf{u}_l-\widehat{\boldsymbol{\mu}}_U)(\mathbf{v}_l-\widehat{\boldsymbol{\mu}}_V)^T. \end{aligned} \tag{5}$$

Obtaining the determinants required for $\mathrm{T_{HAD}}$ may become computationally problematic for high-dimensional variables and/or small sample sizes, not only for their complexity but also because of the potentially large condition number of the matrices involved.

Using various properties of structured matrices, (4) can be rewritten as $\mathrm{T_{HAD}}(\mathbf{X},\mathbf{Y}) = \det[\mathbf{I}_{N_x} - \widehat{\mathbf{C}}\widehat{\mathbf{C}}^T]$, where $\widehat{\mathbf{C}} \triangleq \widehat{\mathbf{C}}_{XX}^{-1/2}\widehat{\mathbf{C}}_{XY}\widehat{\mathbf{C}}_{YY}^{-1/2}$ is usually referred to as *sample coherence matrix* [13] between X and Y. Taking the 1st order Taylor approximation of minus the natural log of this new expression provides a simplified test for the small correlation regime: $\mathrm{T_{FRO}}(\mathbf{X},\mathbf{Y}) \triangleq \|\widehat{\mathbf{C}}\|_F^2$, which accepts uncorrelatedness for values close to 0. It can be shown [13] that $\mathrm{T_{FRO}}$ is the *locally most powerful invariant test (LMPIT)* for correlation between Gaussian vectors. While it avoids the computation of determinants, it still requires the inversion of matrices for the sample coherence, which is computationally non-ideal.

The *RV coefficient* is an alternative test for correlation values close to 0 which generalizes the *squared Pearson coefficient* for univariate random variables. It is defined as

$$\mathrm{T_{RV}}(\mathbf{X},\mathbf{Y}) \triangleq \frac{\|\widehat{\mathbf{C}}_{XY}\|_F^2}{\|\widehat{\mathbf{C}}_{XX}\|_F\|\widehat{\mathbf{C}}_{YY}\|_F}. \tag{6}$$

Although it is an *ad hoc* measure of correlation, it exhibits desirable properties for the purposes of this paper, such as keeping the scale invariance of $\mathrm{T_{HAD}}$ and $\mathrm{T_{FRO}}$ without the need of computing neither determinants nor inverses. This advantage over the other tests allows its usage in challenging high-dimensional/sample-limited settings.

Considering now the conditional case, let Z be the potential confounder of the correlation between X and Y. The problem of detecting this conditional uncorrelatedness can be cast as accepting $\mathcal{H}_0$ for small values of $\mathrm{T_{RV}}$ computed with the corresponding conditional versions of the sample covariance matrices: $\widehat{\mathbf{C}}_{XY|Z}$, $\widehat{\mathbf{C}}_{XX|Z}$ and $\widehat{\mathbf{C}}_{YY|Z}$. The cross-covariance matrix between jointly normally distributed U and V conditioned to Z is obtained from the *Schur complement* of the confounder covariance matrix [9]:

$$\widehat{\mathbf{C}}_{UV|Z} \triangleq \widehat{\mathbf{C}}_{UV} - \widehat{\mathbf{C}}_{UZ}\widehat{\mathbf{C}}_{ZZ}^{-1}\widehat{\mathbf{C}}_{ZV}. \tag{7}$$

Once again, a matrix inversion is needed, bringing back the same issues present with $\mathrm{T_{HAD}}$ and $\mathrm{T_{FRO}}$. In the following sections we present an alternative method of computing the conditional version of $\mathrm{T_{RV}}$. Aside from avoiding matrix inversions, it abides by the definition of conditional covariance without assuming jointly Gaussian variables, making it agnostic to any statistical properties of data.

4. U-STATISTICS TEST FOR CONDITIONAL CORRELATION

In this section, we present a novel test for conditional correlation. We introduce well-known ideas about *U-Statistics* [10] in Section 4.1, and build our contributions upon them in Sections 4.2 and 4.3.

4.1. Covariance estimation using U-Statistics

The covariance matrix between U and V admits an unbiased estimation based on the pairwise differences between samples [14]:

$$\widehat{\mathbf{C}}_{UV} = \tfrac{2}{L(L-1)}\textstyle\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\mathring{\mathbf{u}}_{i,j}\mathring{\mathbf{v}}_{i,j}^T, \tag{8}$$

where the normalized sample differences are defined as

$$\mathring{\mathbf{u}}_{i,j} \triangleq (\mathbf{u}_i - \mathbf{u}_j)/\sqrt{2} \quad , \quad \mathring{\mathbf{v}}_{i,j} \triangleq (\mathbf{v}_i - \mathbf{v}_j)/\sqrt{2}. \tag{9}$$

Notice the estimation with U-Statistics is based on the differences between i.i.d. pairs of data from U and V for which the sample mean is not required, unlike in (5). The structure of the double sum in (8) is an unbiased estimate of $\mathbf{C}_{UV} \equiv \mathrm{E}[\mathring{\mathbf{u}}_{l,l'}\mathring{\mathbf{v}}_{l,l'}^T]$ by using $L(L-1)/2$ instances of $\mathring{\mathbf{u}}$ and $\mathring{\mathbf{v}}$, the number of distinct pairs of $\mathbf{u}$ and $\mathbf{v}$ samples. This implies that only $\lfloor L/2 \rfloor$ of them are independent (see Figure 1), allowing to work with an *incomplete* version of (8):

$$\widehat{\mathbf{C}}'_{UV} = \tfrac{1}{\lfloor L/2 \rfloor}\textstyle\sum_{i=1}^{\lfloor L/2 \rfloor}\mathring{\mathbf{u}}_{i,i+\lfloor L/2 \rfloor}\mathring{\mathbf{v}}_{i,i+\lfloor L/2 \rfloor}^T. \tag{10}$$

This expression is simply the mean of i.i.d. terms, so its accuracy becomes comparable to that of a standard sample covariance obtained from half of the available data [15]. The number of unused pairs in (10) with respect to the complete U-Statistics (8) is

$$\Delta L = \tfrac{L(L-1)}{2} - \left\lfloor \tfrac{L}{2} \right\rfloor, \tag{11}$$

which increases with $O(L^2)$. Therefore, the higher L is, the higher the amount of pairs that can be omitted in the computation of the incomplete U-Statistics, for some specified degradation in estimation accuracy of $\widehat{\mathbf{C}}'_{UV}$. This detail points at a wide operation margin for discarding redundant data pairs: the amount relevant for the U-Statistics can grow with $O(L)$ instead of $O(L^2)$ with a penalty of 3dB in accuracy, at most. Similar observations can be found in [15].

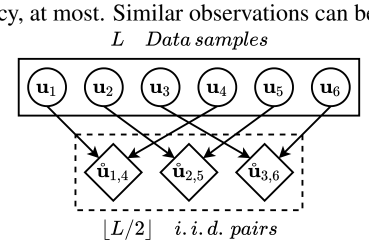


Fig. 1: Independent pairwise sample selection for $L = 6$.

4.2. Weighted U-Statistics for conditional uncorrelatedness

In a similar fashion as in (9), let $\mathring{\mathbf{z}}_{i,j} \triangleq (\mathbf{z}_i - \mathbf{z}_j)/\sqrt{2}$ be the normalized pairwise differences of data samples of the potential confounder Z. Instead of working with variables U, V and Z, consider the following *virtual* random variables:

$$\mathring{U} \triangleq \tfrac{U_1 - U_2}{\sqrt{2}}, \quad \mathring{V} \triangleq \tfrac{V_1 - V_2}{\sqrt{2}}, \quad \mathring{Z} \triangleq \tfrac{Z_1 - Z_2}{\sqrt{2}}, \tag{12}$$

where U_m, V_m and Z_m are random variables with the same statistical properties as U, V and Z, respectively, and independent for different $m = 1, 2$. $\mathring{\mathbf{u}}_{i,j}$, $\mathring{\mathbf{v}}_{i,j}$ and $\mathring{\mathbf{z}}_{i,j}$ are then realizations of those virtual variables. Given that $\mathbf{C}_{UV} \equiv \mathbf{C}_{\mathring{U}\mathring{V}}$, let us recall the definition of the conditional covariance matrix (2). Notice that integrating over all values of Z is equivalent to doing so over the cases in which $\mathring{Z} = \mathbf{0}$ (*i.e.* $Z_1 = Z_2$):

$$\begin{aligned}\mathbf{C}_{UV|Z} &= \int_{\mathbb{R}^{N_z}} \mathbf{C}_{\mathring{U}\mathring{V}|Z=\mathbf{z}}\, dF_Z(\mathbf{z}) \\ &= \iint_{\mathbb{R}^{N_z \times N_z}} \mathbf{C}_{\mathring{U}\mathring{V}|\mathring{Z}(Z_1,Z_2)=\mathbf{0}}\, dF_{Z_1,Z_2}(\mathbf{z}_1,\mathbf{z}_2). \end{aligned} \tag{13}$$

The term inside the integral does not depend on the specific values of $\mathbf{z}_1$ and $\mathbf{z}_2$ but rather on them being equal, so it can be taken outside:

$$\mathbf{C}_{UV|Z} = \mathbf{C}_{\mathring{U}\mathring{V}|\mathring{Z}=\mathbf{0}} \iint_{\mathbb{R}^{N_z \times N_z}} dF_{Z_1,Z_2}(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{C}_{\mathring{U}\mathring{V}|\mathring{Z}=\mathbf{0}}, \quad (14)$$

knowing that $\iint dF_{Z_1,Z_2}(\mathbf{z}_1, \mathbf{z}_2) = 1$. This result implies that conditioning the covariance matrix with respect to a potential confounder Z is equivalent to using only the set of data pairs such that $\mathring{Z} = \mathbf{0}$. Since this is an event of zero probability for continuous random variables, we relax this criterion by employing the pairs for which $\|\mathring{Z}\| < \epsilon$, being $\epsilon > 0$ a design parameter.

Following these ideas, the proposed estimator of the conditional covariance matrix is derived from (8), by only considering pairs of $\mathbf{u}_l$ and $\mathbf{v}_l$ for which their corresponding pairs of $\mathbf{z}_l$ fulfill the presented criterion. It can be calculated with the next expression:

$$\breve{\mathbf{C}}_{UV|Z} \triangleq \frac{\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} \mathring{\mathbf{u}}_{i,j}\mathring{\mathbf{v}}_{i,j}^T I_\epsilon(\|\mathbf{z}_i - \mathbf{z}_j\|)}{\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} I_\epsilon(\|\mathbf{z}_i - \mathbf{z}_j\|)}, \quad (15)$$

where $I_\epsilon(\lambda)$ is an *indicator function* that takes the value of 1 when $0 \leq \lambda \leq \epsilon$ and of 0 otherwise. With $\breve{\mathbf{C}}_{XY|Z}$, $\breve{\mathbf{C}}_{XX|Z}$ and $\breve{\mathbf{C}}_{YY|Z}$ computed as in (15), we can finally obtain an estimate of T_{RV}:

$$\breve{\mathrm{T}}_{\mathrm{RV}}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \frac{\|\breve{\mathbf{C}}_{XY|Z}\|_F^2}{\|\breve{\mathbf{C}}_{XX|Z}\|_F \|\breve{\mathbf{C}}_{YY|Z}\|_F}. \quad (16)$$

Its computation is free from determinants and inverses, thus fulfilling the numerical stability, robustness and complexity requirements set for this paper.

4.3. Order Statistics and efficient computation

While the previous test for conditional correlation displays many desirable properties in algorithmic terms, there is room for improvement. One issue that might arise while configuring it is calibrating the indicator function threshold. Depending on the data being processed, the performance of the detector can be very sensitive to maladjustments of this parameter.

Motivated by the 3dB margin referred to in Section 4.1, we propose an alternative threshold based on the available data themselves. In particular, the amount of data pairs considered in (15) can be set to a fraction of the total number of pairs depending on L:

$$L_p \triangleq \textstyle\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} I_\epsilon(\|\mathbf{z}_i - \mathbf{z}_j\|) = \lceil L\alpha \rceil, \quad (17)$$

such that $\alpha \in [\frac{1}{L}, \frac{L-1}{2}]$ is the *hyper-parameter* controlling this proportion. It can be finely-tuned to only select an amount of data pairs linearly proportional to L, thus obtaining the desired trade-off between estimation bias and variance.

Notice that, for a given α, only the data pairs corresponding to the L_p smallest norms of $\mathring{\mathbf{z}}_{i,j}$ will contribute to the computation of (15). For that reason, it would be very convenient to sort all the values of $\|\mathring{\mathbf{z}}_{i,j}\|$ in ascending order in a vector of size $L(L-1)/2$, called $\mathring{\mathbf{z}}_{\mathrm{sort}}$. Let $q(l) \to (i(l), j(l))$ be a function that returns the pair of indices of $\mathbf{z}$ samples that correspond to each entry l of $\mathring{\mathbf{z}}_{\mathrm{sort}}$. With this order knowledge, the estimator in (15) can be reduced to:

$$\breve{\mathbf{C}}_{UV|Z} = \tfrac{1}{2L_p}\textstyle\sum_{l=1}^{L_p}(\mathbf{u}_{i(l)} - \mathbf{u}_{j(l)})(\mathbf{v}_{i(l)} - \mathbf{v}_{j(l)})^T. \quad (18)$$

Using *Quicksort* [16], the procedure of sorting the confounder sample pairs adds an extra $O(L_p \log L_p)$ operations to the computation of (18), in comparison to (15).

To obtain (16) with the sorted pairs in a more efficient manner, the cyclic property of the trace operator ($\mathrm{Tr}[\cdot]$) can be exploited. By defining $\mathring{\mathbf{U}} \triangleq \left[\left[\mathbf{u}_{i(1)} - \mathbf{u}_{j(1)}\right], \ldots, \left[\mathbf{u}_{i(L_p)} - \mathbf{u}_{j(L_p)}\right]\right]$ and its associated *Gram matrix* $\mathbf{K}_\mathbf{U} \triangleq \mathring{\mathbf{U}}^T\mathring{\mathbf{U}}$, the conditional correlation detector can be rewritten as

$$\breve{\mathrm{T}}_{\mathrm{RV}}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \frac{\mathrm{Tr}[\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}\mathring{\mathbf{Y}}^T\mathring{\mathbf{Y}}]}{\sqrt{\mathrm{Tr}[\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}]\,\mathrm{Tr}[\mathring{\mathbf{Y}}^T\mathring{\mathbf{Y}}\mathring{\mathbf{Y}}^T\mathring{\mathbf{Y}}]}} \quad (19)$$

$$= \frac{\|\mathbf{K}_\mathbf{X}\mathbf{K}_\mathbf{Y}\|_F^2}{\sqrt{\|\mathbf{K}_\mathbf{X}^2\|_F^2\|\mathbf{K}_\mathbf{Y}^2\|_F^2}}. \quad (20)$$

This result is reminiscent of other expressions common in *Kernel Signal Processing* [17].

5. NUMERICAL RESULTS

In the final section of this paper, we display some numerical simulations to highlight the strengths of our proposed Weighted U-Statistics Method (WUSM). We analyze its performance against the Schur Complement Method (SCM) within two different frameworks: as an estimator of the conditional T_{RV} and as a detector of conditional correlation itself.

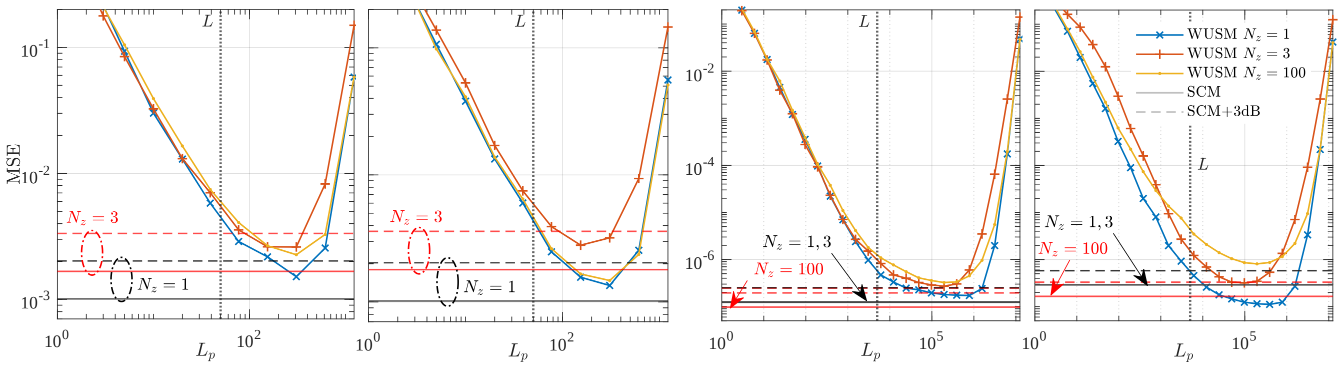
We deal with two zero-mean data models: one with jointly Gaussian variables and the other in which they are distributed as a *Gaussian Copula* [18]. They are different enough as to provide a general sense of the behaviour of our method in arbitrary scenarios. Additionally, they can be fully characterized from second order moments, allowing to generate data that matches a specific covariance matrix.

The tests have been configured as follows. Variables X and Y are scalars ($N_x = N_y = 1$), while confounder Z can be: one-dimensional ($N_z = 1$), to depict the main properties of WUSM in a controlled environment, low-dimensional ($N_z = 3$) and high-dimensional ($N_z = 100$). In terms of data availability, we have considered two scenarios: a sample-limited ($L = 50$) and a regular one ($L = 5000$). The former is of special interest when dealing with the high-dimensional confounder, as expound in Section 1. $M = 500$ tests have been averaged to obtain the results showcased below.

5.1. Estimation

The metric of choice to assess the estimation capabilities of WUSM is the *Mean Squared Error (MSE)* of $\breve{\mathrm{T}}_{\mathrm{RV}}$, defined as $\mathrm{MSE}(\breve{\mathrm{T}}_{\mathrm{RV}}) \triangleq \mathrm{var}(\breve{\mathrm{T}}_{\mathrm{RV}}) + \mathrm{bias}^2(\breve{\mathrm{T}}_{\mathrm{RV}})$. We consider data with a covariance matrix for which X and Y are uncorrelated once conditioned to Z ($\mathbf{C}_{XY|Z} \approx 0$ and $\mathrm{T}_{\mathrm{RV}}(X, Y|Z) \approx 0$) and they all three have unit power dimension-wise. This matrix has been carefully selected so that X and Y are noticeably correlated ($\mathrm{T}_{\mathrm{RV}}(X, Y) \geq 0.2$) when unconditioned, in order to observe clear trends in the MSE curve, *i.e.* whether our method is estimating $\mathrm{T}_{\mathrm{RV}}(X, Y|Z)$ or $\mathrm{T}_{\mathrm{RV}}(X, Y)$. Figure 2 contains such curves for various L_p values and confounder dimensions in four different settings, as well as the MSE values obtained with SCM and a 3dB margin above them.

In the small sample regime (Figures 2a and 2b), there are not noticeable differences between data models. In the Gaussian case, SCM reaches the optimal MSE for $N_z = 1$ and $N_z = 3$. Our technique does not reach such a value but presents an operation band of a certain width less than 3dB over the SCM MSE. Notice that, for $N_z = 100$, while WUSM displays a similar behaviour as in the other two cases, the MSE of SCM reaches very high values (outside of the plot range) due to numerical problems caused by the inversion of



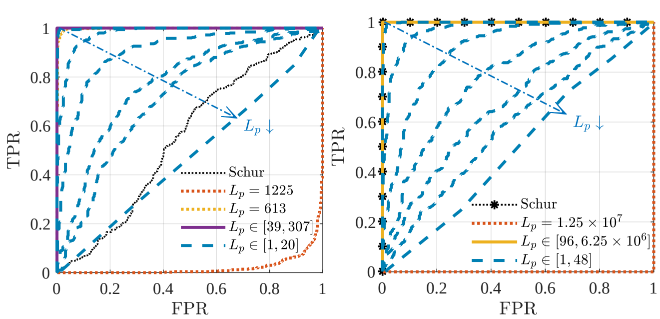
(a) Gaussian, $L = 50$ (b) Copula, $L = 50$ (c) Gaussian, $L = 5000$ (d) Copula, $L = 5000$

Fig. 2: MSE curves of WUSM for various values of L_p against the MSE of SCM.

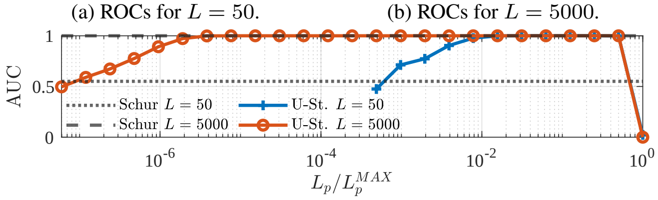
ill-conditioned sample covariance matrices. Robustness against this situation is one of the main strengths of our method.

In the regular sample size regime (Figures 2c and 2d), there is a noticeable increase in the span of the operation bands, roughly proportional to $O(L^2)$. Their floors become wider, making the left side of the bands approach the L pairs threshold. They now include a larger range of L_p configurations that result in similar performance. This trend has been conjectured in Section 4.1 due to the redundancy present in the construction of sample data pairs. In contrast to the sample-limited regime, there is now a clear distinction between the two data models. SCM is based on the premise of Gaussian data, which is not fulfilled by the Copula model. In that case, its MSE is no longer the optimal one, while WUSM can reach lower values for a wide set of L_p configurations due to its more data-driven approach. This phenomenon is especially noticeable for $N_z = 1$.

5.2. Detection



(a) ROCs for $L = 50$. (b) ROCs for $L = 5000$.



(c) AUC of the two previous ROC curves.

Fig. 3: ROC curves and their corresponding AUCs for various L_p.

Due to extension constraints, only a reduced subset of detection test configurations ($N_z = 100$, Copula distribution) will be displayed since there is important overlapping between their results. They are presented in terms of *Receiver Operation Characteristic (ROC)* curves for various L_p and their corresponding *Area Under Curve (AUC)* (Figure 3).

ROC plots in Figures 3a and 3b can be interpreted as follows. X-Axis is the *False Positive Rate (FPR)* and Y-Axis is the *True Positive Rate (TPR)*. When WUSM operates with all available data pairs, it estimates $\mathrm{T}_{\mathrm{RV}}(X, Y)$ and effectively behaves as a correlation detector for the unconditioned case. As commented in Section 5.1, the covariance matrices of data models have been designed with non-negligible correlation when not conditioned on the confounder. For this reason, WUSM behaves as a detector of the opposite hypothesis when it deals with all sample pairs ($L_p^{\mathrm{MAX}} = L(L-1)/2$).

As L_p decreases, the ROC curve of WUSM improves until reaching a near-optimal shape (solid line in the plots). Its detection quality remains at that level for a range of L_p values. After that, when the number of considered pairs becomes too low, the performance of the detector starts to deteriorate progressively. Finally, when L_p is minimum, WUSM becomes nonfunctional as a detector. Unsurprisingly, SCM presents near-optimal detection capabilities in the regular sample size case but very poor performance in the small-sample one.

AUC curves in Figure 3c make very apparent that the operation band of WUSM for detection is very wide in terms of L_p and it roughly scales with $O(L^2)$. This result is very positive for the purposes of this paper since detection is the main functionality of our test for conditional correlation.

6. CONCLUSIONS

In this paper we have developed a test for conditional correlation based on weighted U-Statistics that avoids the demanding operations and Gaussianity requirements of other classical techniques. The resulting detector displays a very robust performance in small-sample/high-dimensional scenarios.

We can outline various lines of research that could potentially stem from this work. In terms of design, there are some aspects of our approach that can be adjusted, such as the ordering of the pairwise differences of samples. We have employed the Euclidean distance because it is a very natural choice for scalar data, but becomes an arbitrary one for multivariate data: establishing the *difference* between samples can be done following different criteria. Another design aspect that can be finely tuned is the selection of sample pairs. Alternative indicator functions can be used to obtain soft selection thresholds. Additionally, they can become data-adaptive by being aware of the sample histogram shape.

As for application-specific adjustments, our test could be generalized to detect conditional dependence under arbitrary data models by applying more advanced information theoretic techniques, such as the *characteristic function mapping* [19].

7. REFERENCES

[1] M. Cetin, L. Chen, J. W. Fisher, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, “Distributed Fusion in Sensor Networks,” *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 42–55, 2006.

[2] A. J. Gibberd and J. D. B. Nelson, “High Dimensional Change-point Detection with a Dynamic Graphical LASSO,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2684–2688.

[3] Q. Li, M. Qiao, W. Bian, and D. Tao, “Conditional Graphical LASSO for Multi-label Image Classification,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2977–2986.

[4] A. Wiesel, Y. C. Eldar, and A. O. Hero III, “Covariance Estimation in Decomposable Gaussian Graphical Models,” *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1482–1492, 2010.

[5] G. Cao, L. R. Bachega, and C. A. Bouman, “The Sparse Matrix Transform for Covariance Estimation and Analysis of High Dimensional Signals,” *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 625–640, 2011.

[6] Y. Song, P. J. Schreier, D. Ramirez, and T. Hasija, “Canonical correlation analysis of high-dimensional data with very small sample support,” 2016, arXiv:1604.02047 [cs.IT].

[7] L. Zhou, L. Wang, and P. Ogunbona, “Discriminative Sparse Inverse Covariance Matrix: Application in Brain Functional Network Classification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3097–3104.

[8] P. Robert and Y. Escoufier, “A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 25, no. 3, pp. 257–265, 1976.

[9] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics*, Wiley, 2008.

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

[11] S. M. Ross, *A First Course in Probability*, Prentice Hall, 1998.

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

[13] 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.

[14] S. Minsker and X. Wei, “Robust Modifications of U-statistics and Applications to Covariance Estimation Problems,” 2018, arXiv:1801.05565 [math.ST].

[15] X. Chen and K. Kato, “Randomized incomplete U-statistics in high dimensions,” 2019, arXiv:1712.00771 [math.ST].

[16] C. A. R. Hoare, “Quicksort,” *The Computer Journal*, vol. 5, no. 1, pp. 10–16, 01 1962.

[17] J. C. Principe, *Information Theoretic Learning*, Springer, 2010.

[18] R. B. Nelsen, *An Introduction to Copulas*, Springer, 2006.

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