Accepted version of the article published in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. DOI: 10.1109/ICASSP40776.2020.9053506. Available at: http://ieeexplore.ieee.org

Estimation of Information in Parallel Gaussian Channels via Model Order Selection

Carlos Alejandro López, Ferran de Cabrera, *Student Member, IEEE*, and Jaume Riba, *Senior Member, IEEE*
Signal Theory and Communications Department, Technical University of Catalonia (SPCOM/UPC)
{carlos.alejandro.lopez, ferran.de.cabrera, jaume.riba}@upc.edu

***Abstract*—We study the problem of estimating the overall mutual information in M independent parallel discrete-time memory-less Gaussian channels from N independent data sample pairs per channel (inputs and outputs). We focus on the case where the number of active channels L is sparse in comparison with the total number of channels ($L \ll M$), for which the direct application of the *maximum likelihood* principle is problematic due to *overfitting*, especially for moderate to small N. For this regime, we show that the bias of the mutual information estimate is reduced by resorting to the *minimum description length* (MDL) principle. As a result, simple pre-processing based on a per-channel threshold on the empirical *squared correlation coefficient* is required with a fixed threshold that monotonically decreases with N as $1 - N^{-1/N}$, for $N \geq 4$. The resulting improvement is shown in terms of the estimated information bias.**

***Index Terms*—Min. Description Length (MDL), Bayesian Info. Criterion (BIC), Locally Most Powerful Invariant Test (LMPIT), Maximum Likelihood (ML), Squared Pearson Coefficient, Mutual Inf. (MI), Generalized Likelihood Ratio Test (GLRT).**

I. Introduction

A general and important problem in the field of multivariate statistical analysis [5] is testing whether two M-dimensional Gaussian vectors are uncorrelated or not. It has been shown in [7] that the Locally Most Powerful Invariant Test (LMPIT) for this problem is given by the Frobenius norm of the *sample coherence matrix*. This fundamental test is given by the sum of squared canonical correlations, which are the squared Pearson coefficients of virtual independent parallel channels given by the canonical coordinates.

In the case of data with unknown statistics, a more challenging problem is estimating the mutual information between two sources and, more generally, the so-called *universal* information measures (see [16] for methods and applications concerned with this field). This line of research finds numerous applications in data science and machine learning. Recently, the problem of estimating information has been linked in [2] with the aforementioned problem of coherence estimation by mapping the bivariate data onto a high-dimensional feature space based on the *empirical characteristic function*. In particular, the Frobenius norm of the coherence matrix computed after this high-dimensional mapping converges with M to the so-called *squared-loss* mutual information [14].

In all the aforementioned applications, the model consisting of parallel-channels with independent information per channel plays an important role in the process of simplification and deep understanding of the original problems. Note that the independence assumption is very useful in modeling many wireless communications scenarios, and it appears in a variety of areas such as radar, multitone transmissions and multi-antenna schemes [15]. As a more direct example, the independence arises as well in the frequency-domain representation of stationary time-series, where the canonical correlations coincide asymptotically (w.r.t. data size) with the squared roots of the magnitude squared coherence spectrum [11].

One of the problems to be faced when working with high-dimensional data is the fact that, in most practical situations, only few L components are correlated among the large amount of M parallel virtual channels. The necessity of detecting the presence of a sparse correlated subset of components emerge naturally in numerous scenarios (see [1], [6] and references therein for a motivation). This means that the high-dimensional data tend to exhibit a low-rank structure irrespective of the application. This paper focuses on that scenario while holding the parallel channel model for simplicity. The purpose is to show that simple sparse-aware detectors and estimators can be obtained via the well-known Minimum Description Length (MDL) principle proposed by Rissanen for model order selection [8] [9], which coincides with the Bayesian Information Criterion (BIC) by Schwarz (see [13] for an excellent overview). We show that the MDL principle applied to the problem of estimating information improves both the LMPIT and the ML estimator.

Prior work on the application of the MDL principle can also be found in [12], [10] in the context of Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) as a means to handle the small sample support problem. While [10] and [12] focus on rank-reduction, the present paper moves the goal to the estimation and detection of information, although focusing on parallel channels for simplicity.

II. Maximum likelihood estimation of mutual information

Let's consider a set of M mutually independent pairs of sequences given by $x_m(n)$ and $y_m(n)$, with $m = 1, 2, \ldots, M$, from which we have N i.i.d. samples, with $n = 1, 2, \ldots, N$. The m-th pair is represented by the N-length 2×1 vector sequence $\mathbf{z}_m(n)$ defined as

$$\mathbf{z}_m(n) = \left[\begin{array}{c} x_m(n) \\ y_m(n) \end{array} \right]. \quad (1)$$

This work is supported by projects TEC2016-76409-C2-1-R (WINTER), Ministerio de Economia y Competividad, Spanish National Research Plan, and 2017 SGR 578 - AGAUR, Catalan Government.

In the case of zero-mean Gaussian signals, the mutual information of the m-th pair is given by

$$I(x_m; y_m) = -\frac{1}{2} \ln \det \mathbf{C}_m, \tag{2}$$

where $\mathbf{C}_m$ is the coherence matrix associated to the m-th channel defined as

$$\mathbf{C}_m = \mathbf{D}_m^{-1/2} \mathbf{R}_m \mathbf{D}_m^{-1/2}, \tag{3}$$

and matrices

$$\mathbf{R}_m = E\left[\mathbf{z}_m(n)\mathbf{z}_m^T(n)\right] \tag{4}$$

$$\mathbf{D}_m = \text{diag}(\mathbf{R}_m) = \left[\begin{array}{cc} v_{x,m} & 0 \\ 0 & v_{y,m} \end{array}\right] \tag{5}$$

are, respectively, the 2 × 2 autocorrelation matrix and a diagonal matrix containing the two, non-zero variances. Note that

$$\mathbf{C}_m = \left[\begin{array}{cc} 1 & \rho_m \\ \rho_m & 1 \end{array}\right], \tag{6}$$

where ρ_m (with $-1 < \rho_m < 1$) is the Pearson coefficient associated to signals $x_m(n)$ and $y_m(n)$. It is clear that the mutual information $I(x_m; y_m)$ depends solely on three free parameters, namely $v_{x,m}$, $v_{y,m}$ and $r_{xy,m}$, being $r_{xy,m}$ the cross correlation between x_m and y_m. We can relate these three parameters as

$$\rho_m = \frac{r_{xy,m}}{\sqrt{v_{x,m} v_{y,m}}}. \tag{7}$$

As the pairs are independent, the overall mutual information is given by the sum of pairwise mutual information values:

$$I(\mathbf{x}; \mathbf{y}) = \sum_{m \in \mathcal{S}_M} I(x_m; y_m) = -\frac{1}{2} \sum_{m \in \mathcal{S}_M} \ln(1 - \rho_m^2) \tag{8}$$

with $\mathcal{S}_M = \{1 : M\}$.

The objective is estimating $I(\mathbf{x}; \mathbf{y})$ from the available data under the prior knowledge that some of them provide no information, that is, it exists some finite $L \leq M$ such that $\rho_m = 0$ for $m \notin \mathcal{S}_L$, where $\mathcal{S}_L$ is a set of integers indexing the active channels, with cardinality $|\mathcal{S}_L| = L$. From the above exposition, it is clear that we are in front of a parametric formulation of the problem of mutual information estimation. Effectively, the finite number of (continuous) parameters of the problem is equal to $3L$, and L is also a (discrete) parameter to be estimated from the data (the model order).

We want to obtain consistency of the resulting estimate as both $N \to \infty$ and $M \to \infty$. To this end, the model order selection of $L \leq M$ is mandatory in order to avoid an uncontrolled number of parameters to be estimated, which would prevent from achieving the desired consistency due to overfitting.

The ML estimation of $\mathbf{R}_m$ can be formulated as follows. The log-likelihood function associated to the overall multichannel data staked at the columns of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}_m$, and conditioned to the complete set $\{\mathbf{R}_m\}_{m \in \mathcal{S}_M}$, is given by:

$$\ln p(\mathbf{X}; \mathbf{Y} | \{\mathbf{R}_m\}_{m \in \mathcal{S}_M}) = \sum_{m \in \mathcal{S}_M} \ln p(\mathbf{Z}_m | \mathbf{R}_m), \tag{9}$$

where the log-likelihood function associated to the m-th channel is:

$$\ln p(\mathbf{Z}_m | \mathbf{R}_m) = -\frac{1}{2} \sum_{n=1}^{N} \left(\ln\left(2\pi \det \mathbf{R}_m\right) + \mathbf{z}_m^T(n) \mathbf{R}_m^{-1} \mathbf{z}_m(n) \right). \tag{10}$$

Equivalently:

$$\ln p(\mathbf{Z}_m | \mathbf{R}_m) = -\frac{N}{2} \left(\ln\left(2\pi \det \mathbf{R}_m\right) + \text{tr}(\mathbf{R}_m^{-1} \hat{\mathbf{R}}_m) \right), \tag{11}$$

where $\hat{\mathbf{R}}_m$, for $m \in \mathcal{S}_L$, is the per-channel sample covariance matrix:

$$\hat{\mathbf{R}}_m = \frac{1}{N} \mathbf{Z}_m \mathbf{Z}_m^T, \tag{12}$$

which is known [4] to be the maximizer of $\ln p(\mathbf{Z}_m | \mathbf{R}_m)$ with respect to $\mathbf{R}_m$ (that is, the sample covariance matrix is the ML estimate of the covariance matrix for Gaussian signals). For $m \notin \mathcal{S}_L$, however, the ML estimate of $\hat{\mathbf{R}}_m$ is

$$\hat{\mathbf{R}}_m = \frac{1}{N} \text{diag}(\mathbf{Z}_m \mathbf{Z}_m^T) = \hat{\mathbf{D}}_m \tag{13}$$

as we have the prior-knowledge that $\rho_m = 0$, although nothing is known about the variances. Substituting the ML estimates on the log-likelihood function yields the log-likelihood function conditioned only to the active set knowledge:

$$\begin{aligned} &\ln p(\mathbf{X}; \mathbf{Y} | \mathcal{S}_L) = \\ &-\frac{N}{2} \left(\sum_{m \in \mathcal{S}_L} \ln\left(2\pi \det \hat{\mathbf{R}}_m\right) + 2 \sum_{m \in \mathcal{S}_L} \text{tr}(\hat{\mathbf{R}}_m^{-1} \hat{\mathbf{R}}_m) \right. \\ &\left. + \sum_{m \notin \mathcal{S}_L} \ln\left(2\pi \det \hat{\mathbf{D}}_m\right) + \sum_{m \notin \mathcal{S}_L} \text{tr}(\hat{\mathbf{D}}_m^{-1} \hat{\mathbf{R}}_m) \right). \end{aligned} \tag{14}$$

Note that

$$\text{tr}(\hat{\mathbf{R}}_m^{-1} \hat{\mathbf{R}}_m) = \text{tr}(\mathbf{I}_2) = 2, \tag{15}$$

and

$$\text{tr}(\hat{\mathbf{D}}_m^{-1} \hat{\mathbf{R}}_m) = \text{tr}(\hat{\mathbf{D}}_m^{-1/2} \hat{\mathbf{R}}_m \hat{\mathbf{D}}_m^{-1/2}) = \text{tr}(\mathbf{C}_m) = 2. \tag{16}$$

Therefore, we have

$$\begin{aligned} \ln p(\mathbf{X}; \mathbf{Y} | \mathcal{S}_L) = -\frac{N}{2} & \left(\sum_{m \in \mathcal{S}_L} \ln\left(2\pi \det \hat{\mathbf{R}}_m\right) + 2L \right. \\ & \left. + \sum_{m \notin \mathcal{S}_L} \ln\left(2\pi \det \hat{\mathbf{D}}_m\right) + 2(M - L) \right) \end{aligned} \tag{17}$$

$$\begin{aligned} = -\frac{N}{2} & \left(\sum_{m \in \mathcal{S}_L} \ln\left(2\pi \det \hat{\mathbf{R}}_m\right) \right. \\ & \left. + \sum_{m \notin \mathcal{S}_L} \ln\left(2\pi \det \hat{\mathbf{D}}_m\right) + 2M \right) \end{aligned} \tag{18}$$

$$= -\frac{N}{2} \left(\sum_{m \in \mathcal{S}_L} \ln \det \hat{\mathbf{R}}_m + \sum_{m \notin \mathcal{S}_L} \ln \det \hat{\mathbf{D}}_m + c \right), \tag{19}$$

where $c = (2 + \ln(2\pi))\, M$. Since $\det \hat{\mathbf{R}}_m = \det(\hat{\mathbf{D}}_m \hat{\mathbf{C}}_m)$, then

$$\ln p(\mathbf{X}; \mathbf{Y} | \mathcal{S}_L) = -\frac{N}{2} \left(\sum_{m \in \mathcal{S}_L} \ln \det(\hat{\mathbf{D}}_m \hat{\mathbf{C}}_m) + \right.$$

$$+ \sum_{m \notin \mathcal{S}_L} \ln \det \hat{\mathbf{D}}_m + c \Bigg) \tag{20}$$

$$= -\frac{N}{2} \Bigg(\sum_{m \in \mathcal{S}_L} \ln \det \hat{\mathbf{D}}_m + \sum_{m \in \mathcal{S}_L} \ln \det \hat{\mathbf{C}}_m + \sum_{m \notin \mathcal{S}_L} \ln \det \hat{\mathbf{D}}_m + c \Bigg) \tag{21}$$

$$= -\frac{N}{2} \left(\sum_{m \in \mathcal{S}_M} \ln \det \hat{\mathbf{D}}_m + \sum_{m \in \mathcal{S}_L} \ln \det \hat{\mathbf{C}}_m + c \right). \tag{22}$$

Note that the positivity of $\det \hat{\mathbf{D}}_m$ in (22) is ensured with probability 1 because we consider non-null variances, and the positivity of $\det \hat{\mathbf{C}}_m$ is ensured with probability 1 as a result of the Schwarz inequality and the fact that $0 \leq \hat{\rho}_m^2 < 1$. Ignoring additive constants that do not depend on the unknown order L, we have:

$$-\ln p(\mathbf{X}; \mathbf{Y} | \mathcal{S}_L) = \frac{N}{2} \sum_{m \in \mathcal{S}_L} \ln \det \hat{\mathbf{C}}_m + \text{const.} \tag{23}$$

Finally, and more clearly,

$$-\ln p(\mathbf{X}; \mathbf{Y} | \mathcal{S}_L) = -N \hat{I}_{ML}(\mathbf{x}; \mathbf{y} | \mathcal{S}_L) + \text{const}, \tag{24}$$

where, in view of (8) and from the invariance property of ML,

$$\hat{I}_{ML}(\mathbf{x}; \mathbf{y} | \mathcal{S}_L) = -\frac{1}{2} \sum_{m \in \mathcal{S}_L} \ln(1 - \hat{\rho}_m^2) \tag{25}$$

is the ML estimate of mutual information assuming that only L channels within the set $\mathcal{S}_L$ provide non-null information.

III. Incorporation of the BIC rule

Assume that L is unknown and should be estimated. Assume for clarity that $\hat{\rho}_m^2 > \hat{\rho}_{m'}^2$ for $m' > m$ (this will become irrelevant later on). Under this assumption, the negative log-likelihood function for selecting L in (24) is a $\cup$ convex, non-increasing function of L, which would then yield $\hat{L} = M$ as the optimal value. This is the well-known problem of model order selection: the ML rule yields to assume maximum complexity of the data. In general, to avoid overfitting, the BIC rule for model order selection incorporates a penalty term of the form $(L/2) \ln N$ to the joint likelihood function conditioned to a given model complexity [13]. Applying the idea to the result obtained in (24), the final function to be minimized against L becomes:

$$\text{BIC}(L) = -\hat{I}_{ML}(\mathbf{x}; \mathbf{y} | \mathcal{S}_L) + \frac{L \ln N}{2N}. \tag{26}$$

Clearly, the right side term in (26) increases linearly with the model complexity L, with a rate that goes to zero as N goes to infinity, such that the selection of active channels becomes more restrictive for small N and more permissive for large N. The minimizer is now

$$\hat{L} = \arg \min_{L = 1, 2, \ldots, M} \text{BIC}(L). \tag{27}$$

In the particular application of the BIC rule in this paper, it is possible to further simplifying the computation of $\hat{L}$ by means of the following argument, which is not possible in other problems. Note that the difference between two consecutive trial values of the BIC indicator is:

$$\triangle(L) = \text{BIC}(L) - \text{BIC}(L-1) = \frac{1}{2} \ln \left(1 - \hat{\rho}_L^2\right) + \frac{\ln N}{2N}. \tag{28}$$

If $\triangle(L) < 0$, we need to keep increasing L to minimize $\text{BIC}(L)$. Otherwise, we stop searching. From (28), this observation implies assigning channel m as active if and only if

$$-\frac{1}{2} \ln(1 - \hat{\rho}_m^2) > \frac{\ln N}{2N}, \tag{29}$$

which yields to M per-channel independent decisions such that m is declared active if and only if

$$\hat{\rho}_m^2 > 1 - N^{-1/N}. \tag{30}$$

Note that since the above rule implies making independent decisions per channel, the ordering of the channels is in fact not required.

Summarizing, the BIC rule applied to the problem of estimating information yields:

$$\hat{I}_{BIC}(\mathbf{x}; \mathbf{y}) = -\frac{1}{2} \sum_{m \in \mathcal{S}_M} \ln \left(1 - \hat{\rho}_m^2 1_{(\hat{\rho}_m^2 > 1 - N^{-1/N})}\right), \tag{31}$$

as the final regularized estimate of information, where 1_a is the indicator function such that $1_a = 1$ if the event a is true and $1_a = 0$ otherwise.

Finally, from the $\hat{I}_{BIC}$ estimator, a GLRT detector [3] of the presence of information can be formulated as declaring the presence of information if

$$\hat{I}_{BIC}(\mathbf{x}; \mathbf{y}) > \gamma_{BIC}, \tag{32}$$

where $\gamma_{BIC} > 0$ is some threshold designed to achieve the specified false alarm probability. This detector is expected to perform better than the GLRT $\left(\hat{I}_{ML}(\mathbf{x}; \mathbf{y} | \mathcal{S}_M) > \gamma_{GLRT}\right)$ and the LMPIT $\left(\sum_{m \in \mathcal{S}_M} \hat{\rho}_m^2 > \gamma_{LMPIT}\right)$.

IV. Simulation results

In this section we show the performance of the BIC estimator in (31) in contrast to the ML estimator in (25) for $L = M$. The main phenomena that we are solving is overfitting and thus we will focus on studying the bias of both kinds of estimators. In order to do so, we focus our study on two scenarios: observing the evolution of the bias with M and studying it with N. We consider two modifications of the proposed threshold in (30) that we formulate as

$$\lambda_1 = 1 - N^{-1/N}, \;\; \lambda_2 = 4\lambda_1, \;\; \lambda_3 = \frac{1}{4}\lambda_1. \tag{33}$$

The main goal of introducing these new thresholds is to study the optimality of the found threshold. For instance, we expect λ_2 to be more robust to an increasing number of channels M since it tends to discard more channels than the other thresholds, but to have little robustness to lower values of N since worse estimates of $\hat{\rho}_m^2$ increase the probability of labeling as inactive an actual active channel. On the other hand, λ_3 acts in the opposite sense by having little robustness to M but a better performance at low N.

We also defined the overall Pearson coefficients so that the mutual information is linear in terms of $m = 0, ..., L-1$, the active channel identifier. In particular, the mutual information of each channel, I_m, is such that it satisfies

$$I_m = C\left(1 - \frac{m}{L}\right), \tag{34}$$

where C is an arbitrary constant which can be computed by fixing the overall mutual information as

$$I = \sum_{m \in \mathcal{S}_L} I_m. \tag{35}$$

For the first scenario we have performed two experiments with $L = 20$ active channels and two different values of N and actual mutual information I. These experiments can be seen in Fig. 1, where we show that the ML estimator presents a bias that is directly proportional to the total space dimension, due to the extra bias that comes from the non-active channels, and that by estimating $\hat{I}_{BIC}$ it presents robustness to this effect. However, if we focus on the different variations of λ, we can see that the restrictive threshold λ_2 is compensating the extra bias due to the noise by adding the contribution of fewer channels, but λ_3 presents higher bias as the number of channels increases. Also, note that for the case of λ_3 the estimator starts to approach the ML estimator, which means that the estimator bias tends to be linear with respect to the dimensionality for very small values of the threshold.

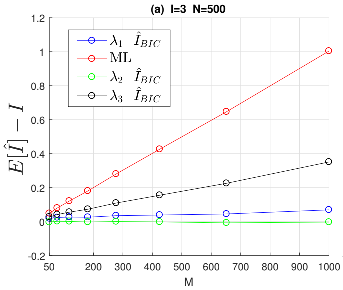

Fig. 1. Bias of the proposed estimator with three different thresholds compared with the ML estimate as a function of M.

On the other hand, the scenario depicted in Fig. 2 is handled in a similar way as the first one, where we fix $L = 20$ and $I = 3$, but we consider an experiment with $M = 1000$ and another with $M = 100$. In this figure we show that by using λ_1 to estimate mutual information, the estimate converges faster to zero than the ML estimator. However, for $M = 100$ we show that being restrictive with λ_3 can be harmful if the environment has a moderated dimensionality, that is when the actual number of active channels is reasonably close to the total number of channels, so the optimal threshold λ_1 or even more permissive ones as λ_2 are encouraged.

Finally, we study the impact of the different thresholds in the overall estimation of channels for $L = 20$. This experiment can be seen in Fig. 3, where we can see that λ_2 performs poorly when the mutual information is too low, and its convergence to the correct number of active channels is much slower than λ_1. Regarding λ_3, it is clear that it always detects more channels than necessary, explaining the increased bias in the previous experiments. To conclude, λ_1 stabilizes the total number of active channels detected to the actual value L, but a limited number of samples N may induce an error in the detection, thus converging to a higher value than the actual one.

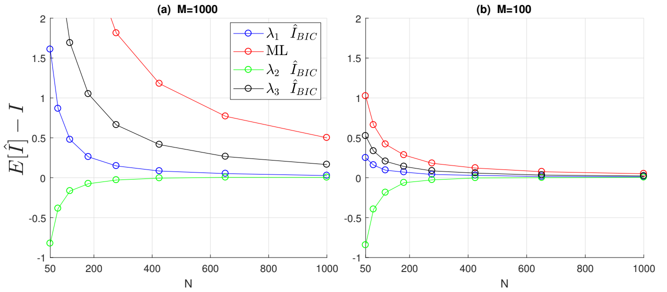

Fig. 2. Bias of the proposed estimator with three different thresholds compared with the ML estimate as a function of N.

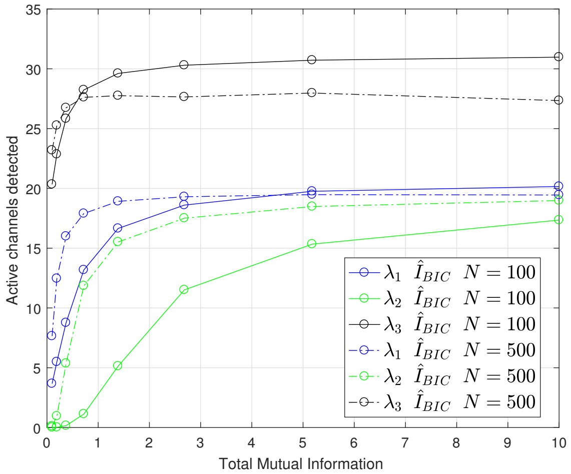

Fig. 3. Active channels detected with three different thresholds as a function of the mutual information value I, with $L = 20$ and $M = 100$.

V. Conclusions

In this paper we have shown that the problem of mutual information estimation for parallel independent channels can be improved by discarding the inactive channels following a given rule. This rule is based on the MDL principle, and it naturally determines an optimal threshold for determining which channels are active and which ones are not. Moreover, if this threshold is not followed, any variation may induce less robustness in terms of bias with respect to the total number of channels M or number of observations N. We have observed that in this particular problem it is possible to achieve a single channel criterion from a multi-channel approach. This feature is of great interest as single channel criteria are computationally less expensive than multi-channel processing.

In view of the proposed ideas, the natural extension of this work would be the case of general $M_x \times M_y$ MIMO channels with low-rank $L < \min(M_x, M_y)$, with Gaussian inputs.

References

[1] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detection of correlations. *The Annals of Statistics*, 40(1), 2012.

[2] F. de Cabrera and J. Riba. Squared-loss mutual information via high-dimension coherence matrix estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5142–5146, 2019.

[3] S. M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*, volume II. Prentice-Hall, New York, 1993.

[4] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, New York, 1998.

[5] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. London, U.K.: Academic, 1995.

[6] D. Ramírez, G. Vázquez-Vilar, R. López-Valcarce, J. Vía, and I. Santamaría. Detection of rank-P signals in cognitive radio networks with uncalibrated multiple antennas. *IEEE Transactions on Signal Processing*, 59(8):3764–3774, Aug 2011.

[7] D. Ramírez, J. Vía, I. Santamaría, and L. L. Scharf. Locally most powerful invariant tests for correlation and sphericity of Gaussian vectors. *IEEE Trans. Inf. Theory*, 59(4):2128–2141, Apr. 2013.

[8] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465 – 471, 1978.

[9] T. Roos, P. Myllymäki, and J. Rissanen. MDL denoising revisited. *IEEE Transactions on Signal Processing*, 57(9):3347–3360, Sep. 2009.

[10] N. J. Roseveare and P. J. Schreier. Model-order selection for analyzing correlation between two data sets using CCA with PCA preprocessing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5684–5687, 2015.

[11] I. Santamaría and J. Vía. Estimation of the magnitude squared coherence spectrum based on reduced-rank canonical coordinates. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 985–988, 2007.

[12] Y. Song, P. Schreier, D. Ramírez, and T. Hasija. Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing*, pages 449–458, Nov. 2016.

[13] P. Stoica and Y. Selén. Model-order selection: a review of information criterion rules. *IEEE Signal Process. Magazine*, July 2004.

[14] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52 (12 pages), 2009.

[15] David Tse. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[16] Q. Wang, S. R. Kulkarni, and S. Verdú. *Universal estimation of Information measures for analog sources*. Number 5:3. Foundations and trends in Communications and Information Theory, 2009.