entropy

MDPI

Article

Computing the Partial Correlation of ICA Models for Non-Gaussian Graph Signal Processing

Jordi Belda, Luis Vergara *, Gonzalo Safont and Addisson Salazar

Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, 46022 València, Spain; jorbelva@upvnet.upv.es (J.B.); gonsaar@upvnet.upv.es (G.S.); asalazar@dcom.upv.es (A.S.)
* Correspondence: lvergara@dcom.upv.es; Tel.: +34-963-877-308

Received: 23 November 2018; Accepted: 24 December 2018; Published: 29 December 2018

Abstract: Conventional partial correlation coefficients (PCC) were extended to the non-Gaussian case, in particular to independent component analysis (ICA) models of the observed multivariate samples. Thus, the usual methods that define the pairwise connections of a graph from the precision matrix were correspondingly extended. The basic concept involved replacing the implicit linear estimation of conventional PCC with a nonlinear estimation (conditional mean) assuming ICA. Thus, it is better eliminated the correlation between a given pair of nodes induced by the rest of nodes, and hence the specific connectivity weights can be better estimated. Some synthetic and real data examples illustrate the approach in a graph signal processing context.

Keywords: partial correlation; independent component analysis; graph signal processing

1. Introduction

1.1. Background

The partial correlation coefficient (PCC) [1] is a classical concept that has relevance in a variety of statistical signal processing problems. Essentially, PCC measures the correlation between two random variables conditioned to other observed random variables. Interest in it has recently increased due to the emergence of graph signal processing (GSP) [2–4]. One key aspect of GSP is defining the graph connectivity. Although this can be done considering the natural interactions from the context where the graph signal is defined (e.g., time or space proximity between two nodes), it is desirable to develop formal statistical methods; that is, given a set of multivariate samples where every sample component is assigned to a node of the graph, the graph connectivity which best describes the implicit dependences between any two nodes can be learned. Thus, PCCs are appropriate candidates to define the connectivity as the effect of the rest of nodes is removed from the pairwise correlation. Actually, the PCC is formally interpreted as the correlation between the residuals obtained after optimal estimation of the values of the two involved nodes from the rest of nodes. Optimality is in the sense of minimum linear mean square error (LMSE). Fortunately, it is not necessary to make an explicit estimation, as the PCCs can be computed from the so-called precision matrix (inverse of the covariance matrix). Thus, many efforts have focused onto estimating the precision matrix both in GSP [5,6] and in statistics [7–12]. However, the minimum LMSE estimator is optimum only if Gaussianity can be assumed. Accordingly, we can say that methods based on the precision matrix are suboptimal in non-Gaussian scenarios. The concept of PCC is extended to a Gaussian mixture model (GMM) in [13]. Apart from this work, and to our knowledge, there have been no other attempts to consider non-Gaussian models in graph connectivity learning.

1.2. New Contributions and Paper Organization

In this work we consider the partial correlation computation under a non-Gaussian model, in particular, a model of independent component analysis (ICA). ICA [14–17] is a consolidated technique which has found a myriad of applications in statistical signal processing (e.g., blind source separation [18–25]) and pattern recognition (see [26–31] and references therein). From the perspective of this work, ICA is a model which incorporates non-Gaussianity through some independent variables (sources), which are linearly mixed to create the observed samples. This makes it highly versatile and allows for the modeling of non-Gaussian multivariate densities.

In the next Section we define a new partial correlation coefficient: ICA-PCC. The basic concept is to replace the implicit linear estimation of conventional PCC by a nonlinear estimation (conditional mean) assuming an underlying ICA model. Then, a general formula is presented to compute the residual covariance matrix from where the ICA-PCCs are to be computed. An essential part of this formula is a diagonal matrix having entries equal to the mean-square-errors of estimating the sources of the ICA model. Then, in Section 3, a practical method is presented to estimate such a matrix from the ICA model parameters. Finally, Section 4 includes some simulations to illustrate the improved estimation of the partial correlation by ICA-PCC in non-Gaussian scenarios. A real data example with EEG multichannel highly non-Gaussian signals is also included to quantify changes in brain connectivity between normal and abnormal states of a patient during sleep.

2. The Partial Correlation of ICA Models

2.1. Statement of the Problem

Let $\mathbf{x} = [x_1 \dots x_N]^T$ be the observation vector having covariance matrix $E[\mathbf{x}\mathbf{x}^T] = \mathbf{C}_{xx}$. We assume that $\mathbf{x}$ obeys an ICA model, then

$$\mathbf{x} = \mathbf{U}\mathbf{s} \quad \mathbf{s} = \mathbf{W}\mathbf{x} \tag{1}$$

where $\mathbf{s} = [s_1 \dots s_N]^T$ is a vector of independent sources and $\mathbf{U}$ is a square and invertible mixing matrix ($\mathbf{W} = \mathbf{U}^{-1}$ is the de-mixing matrix). The sources are considered standardized (zero mean and unit variance), otherwise they may have different non-Gaussian marginal densities, which factorize the joint probability density function (pdf) $p(\mathbf{s}) = p(s_1) \cdot \dots \cdot p(s_N)$. Notice that

$$E[\mathbf{s}] = 0 \quad \mathbf{C}_{ss} = E\left[\mathbf{s}\mathbf{s}^T\right] = \mathbf{I} \quad E[\mathbf{x}] = 0 \quad \mathbf{C}_{xx} = E\left[\mathbf{x}\mathbf{x}^T\right] = E\left[\mathbf{U}\mathbf{s}\mathbf{s}^T\mathbf{U}^T\right] = \mathbf{U}\mathbf{U}^T \tag{2}$$

Every component of $\mathbf{x}$ is assigned to every node of a graph $G\{V, E, \mathbf{A}\}$, where V is the set of N nodes, E is the set of edges connecting the nodes and $\mathbf{A}$ is the adjacency matrix. The generic element a_{nm} is the weight (assumed real and nonnegative) corresponding to the edge connecting node m to node n. We will consider undirected graphs, so $a_{nm} = a_{mn}$. The problem is to learn $\mathbf{A}$ from an available set of observation vectors. PCCs are reasonable candidates as they can measure the correlation between two nodes removing the effect of the rest of nodes. Moreover, PCCs can be computed from the precision matrix $\mathbf{Q}_{xx} = \mathbf{C}_{xx}^{-1}$ in the form

$$\rho_{nm}^{PCC} = -\frac{q_{nm}}{\sqrt{q_{nn}q_{mm}}} \tag{3}$$

where q_{nm} is the nm element of matrix $\mathbf{Q}$ and ρ_{nm}^{PCC} is the PCC of nodes n and m. Equation (3) could be used for any underlying joint probability density $p(\mathbf{x})$, however it is optimal only for the Gaussian case. This is because the formal definition of ρ_{nm}^{PCC} is given by

$$\rho_{nm}^{PCC} = \frac{E[(x_n - L[x_n/\mathbf{x}_{-nm}])(x_m - L[x_m/\mathbf{x}_{-nm}])]}{\sqrt{E\left[(x_n - L[x_n/\mathbf{x}_{-nm}])^2\right]}\sqrt{E\left[(x_m - L[x_m/\mathbf{x}_{-nm}])^2\right]}} \tag{4}$$

where $\mathbf{x}_{-nm}$ is the vector formed by all the samples of $\mathbf{x}$ except x_n and x_m, and $L[x_n/\mathbf{x}_{-nm}]$, $L[x_m/\mathbf{x}_{-nm}]$ are respectively the minimum LMSE estimates of x_n and x_m from $\mathbf{x}_{-nm}$. However, optimum removal of the effect of $\mathbf{x}_{-nm}$ implies the use of the conditional means $E[x_n/\mathbf{x}_{-nm}]$ and $E[x_m/\mathbf{x}_{-nm}]$, which respectively coincide with $L[x_n/\mathbf{x}_{-nm}]$ and $L[x_m/\mathbf{x}_{-nm}]$ only when $p(\mathbf{x})$ is multivariate Gaussian. Thus, in the non-Gaussian case, conventional PCC does not precisely capture the partial correlation, and so, the graph connectivity. In [13] a generalized PCC (GPCC) is defined in the form

$$\rho_{nm}^{GPCC} = \frac{E[(x_n - E[x_n/\mathbf{x}_{-nm}])(x_m - E[x_m/\mathbf{x}_{-nm}])]}{\sqrt{E\left[(x_n - E[x_n/\mathbf{x}_{-nm}])^2\right]}\sqrt{E\left[(x_m - E[x_m/\mathbf{x}_{-nm}])^2\right]}} \tag{5}$$

where the conditional mean $E[x_n/\mathbf{x}_{-nm}]$ depends on the specific model assumed for $p(\mathbf{x})$. In this paper we consider the ICA model (1). Then, the corresponding partial correlation coefficient is called ICA-PCC, and it is represented by $\rho_{nm}^{ICA-PCC}$ to be specific with respect to the general definition (5).

2.2. A General Formula for the Residual Covariance

Let us define the vector $\mathbf{x}_{nm} = [x_n \; x_m]^T$. We can express $\mathbf{x}$ in the form

$$\mathbf{x} = \mathbf{T}_{-nm}\mathbf{x}_{-nm} + \mathbf{T}_{nm}\mathbf{x}_{nm} \tag{6}$$

where $\mathbf{T}_{-nm}$ is a matrix of dimension $(N \times (N-2))$ obtained from an $(N \times N)$ identity matrix by dropping the n-th and m-th columns. Similarly, $\mathbf{T}_{nm}$ is a matrix of dimension $(N \times 2)$ obtained from an $(N \times N)$ identity matrix dropping all but the n-th and m-th columns. Let us also define the residual vector $\mathbf{e}_{nm} = [e_n \; e_m]^T e_n = x_n - E[x_n/\mathbf{x}_{-nm}] \quad e_m = x_m - E[x_m/\mathbf{x}_{-nm}]$. Notice that the conditional mean is an unbiased estimator, hence the residuals are zero mean and the residual covariance matrix will be

$$\mathbf{C}_{e_{nm}e_{nm}} = E\left[\mathbf{e}_{nm}\mathbf{e}_{nm}^T\right] = \begin{bmatrix} E\left[e_n^2\right] & E[e_n e_m] \\ E[e_m e_n] & E\left[e_m^2\right] \end{bmatrix} \tag{7}$$

We want to compute the residual covariance matrix so that (5) can be applied. We assume an ICA model. First notice that $\mathbf{e}_{nm} = \mathbf{x}_{nm} - E[\mathbf{x}_{nm}/\mathbf{x}_{-nm}]$, but considering (1) and (6), we may write

$$E[\mathbf{s}/\mathbf{x}_{-nm}] = \mathbf{W}(\mathbf{T}_{-nm}\mathbf{x}_{-nm} + \mathbf{T}_{nm}E[\mathbf{x}_{nm}/\mathbf{x}_{-nm}]) \tag{8}$$

Then, we can solve for $E[\mathbf{x}_{nm}/\mathbf{x}_{-nm}]$

$$E[\mathbf{x}_{nm}/\mathbf{x}_{-nm}] = (\mathbf{W}\mathbf{T}_{nm})^+(E[\mathbf{s}/\mathbf{x}_{-nm}] - \mathbf{W}\mathbf{T}_{-nm}\mathbf{x}_{-nm}) \tag{9}$$

where $(\cdot)^+$ is the Moore-Penrose (left) pseudoinverse. Thus, in (9) we are expressing the conditional mean of $\mathbf{x}_{nm}$ in terms of the conditional mean of the sources and the ICA model parameters. This allows us to derive the following general formula, which in spite of its simplicity requires a rather tedious derivation that can be found in Appendix A

$$\mathbf{C}_{e_{nm}e_{nm}} = (\mathbf{W}\mathbf{T}_{nm})^+\mathbf{M}_{nm}\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \tag{10}$$

where $\mathbf{M}_{nm}$ is an $(N \times N)$ diagonal matrix having in its main diagonal the MSEs of optimally estimating the sources from $\mathbf{x}_{-nm}$, that is,

$$\mathbf{M}_{nm}(i,i) \; = mse_{nmi} = E\left[(s_i - E[s_i/\mathbf{x}_{-nm}])^2\right] \tag{11}$$

Moreover, from (A7) (see Appendix A) we know that $mse_{nmi} = 1 - \mathrm{var}[E[s_i/\mathbf{x}_{-nm}]]$, hence considering that, by definition, mse_{nmi} and var $[\cdot]$ are positive quantities, we conclude that $0 \leq mse_{nmi} \leq 1$.

Notice that (10) is a combination of the contributions of every source to the residual covariance matrix. This can be better seen by expressing (10) in the alternative form

$$\mathbf{C}_{e_{nm}e_{nm}} = \sum_{i=1}^{N} mse_{nmi} \cdot \mathbf{u}_{nmi}^{+} \mathbf{u}_{nmi}^{+T} \tag{12}$$

where $\mathbf{u}_{nmi}^{+}$ is the i-th column of $(\mathbf{WT}_{nm})^{+}$. Notice that $\mathbf{WT}_{nm}$ is a $(N \times 2)$ matrix formed by the n-th and m-th columns of $\mathbf{W}$, i.e., by the coefficients that define the (demixing) contributions of x_n and x_m to $\mathbf{s}$. Thus, $(\mathbf{WT}_{nm})^{+}$ is a $(2 \times N)$ matrix, $\mathbf{u}_{nmi}^{+}$ is a (2×1) vector and $\mathbf{u}_{nmi}^{+} \mathbf{u}_{nmi}^{+T}$ is a (2×2) matrix that can be interpreted as the contribution of source s_i to $\mathbf{C}_{e_{nm}e_{nm}}$. This contribution is weighted by mse_{nmi}. Thus, $mse_{nmi} = 0$ indicates that source s_i is perfectly estimated by $\mathbf{x}_{-nm}$, hence s_i does not contribute to the partial correlation between x_n and x_m. At the other extreme, $mse_{nmi} = 1$ indicates that s_i is independent of $\mathbf{x}_{-nm}$, so it has maximum contribution to the partial correlation between x_n and x_m.

3. Estimating the ICA Partial Correlation Coefficients

We want to estimate $\rho_{nm}^{ICA-PCC}$ by

$$\hat{\rho}_{nm}^{ICA-PCC} = \frac{\hat{E}[e_n e_m]}{\sqrt{\hat{E}\left[(e_n)^2\right]} \sqrt{\hat{E}\left[(e_m)^2\right]}} \tag{13}$$

So, according to (7), we have to estimate $\mathbf{C}_{e_{nm}e_{nm}}$. Considering (10), we need estimates of $\mathbf{W}$ and $\mathbf{M}_{nm}$. Estimates of $\mathbf{W}$, the ICA model parameters, can be obtained using a variety of algorithms [14–17,26–31] so, in the following, we concentrate on the estimation of $\mathbf{M}_{nm}$, i.e., on estimating $mse_{nmi} = E\left[(s_i - E[s_i/\mathbf{x}_{-nm}])^2\right] \quad i = 1 \ldots N$. To compute $E[s_i/\mathbf{x}_{-nm}]$ we will consider a particular form of the Wiener structure, which was proposed in [32], namely:

$$E[s_i/\mathbf{x}_{-nm}] \simeq E\left[s_i/\hat{s}_i^l\right] \tag{14}$$

where $\hat{s}_i^l$ is the LMSE estimator of s_i from $\mathbf{x}_{-nm}$ (we dropped the dependence on nm to ease the notation) and the uni-dimensional conditional mean can be approximated by [32,33]

$$E\left[s_i/\hat{s}_i^l\right] = \sum_{k=1}^{\infty} \frac{1}{k!} E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^k\right] H_k\left(\hat{s}_{in}^l\right) \tag{15}$$

where $H_k(x)$ is the k-th Hermite polynomial and $\hat{s}_{in}^l = \frac{\hat{s}_i^l}{\left(\mathrm{var}\left[\hat{s}_i^l\right]\right)^{\frac{1}{2}}}$ is a standardized Gaussian random variable (this is justified in [32] by using the central limit theorem).

Let us approximate (15) by the first two terms. Taking into account that $H_1(x) = x$ $H_2(x) = x^2 - 1$, we can write

$$E\left[s_i/\hat{s}_i^l\right] = E\left[s_i \cdot \hat{s}_{in}^l\right] \hat{s}_{in}^l + E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right] \left(\frac{\left(\hat{s}_{in}^l\right)^2 - 1}{2}\right) \tag{16}$$

but

$$E\left[s_i \cdot \hat{s}_{in}^l\right] \hat{s}_{in}^l = E\left[s_i \cdot \hat{s}_i^l\right] \frac{\hat{s}_i^l}{\mathrm{var}\left(\hat{s}_i^l\right)} = E\left[\hat{s}_i^l \cdot \hat{s}_i^l\right] \frac{\hat{s}_i^l}{\mathrm{var}\left(\hat{s}_i^l\right)} = \mathrm{var}\left(\hat{s}_i^l\right) \frac{\hat{s}_i^l}{\mathrm{var}\left(\hat{s}_i^l\right)} = \hat{s}_i^l \tag{17}$$

where we consider that $E\left[s_i \cdot \hat{s}_i^l\right] = E\left[\hat{s}_i^l \cdot \hat{s}_i^l\right]$ (due to the orthogonality between the estimation error and the linear estimate). As any LMSE estimator is unbiased, we know that $E\left[\hat{s}_i^l\right] = E[s_i] = 0$. Then, we can express the conditional mean in (16) as the combination of a linear term $\hat{s}_i^l$ plus a nonlinear

term $s_i^{nl} = E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right]\left(\frac{\left(\hat{s}_{in}^l\right)^2-1}{2}\right)$. Let us now compactly express the estimation of $\mathbf{s}$ from $\mathbf{x}_{-nm}$ in the form

$$E[\mathbf{s}/\mathbf{x}_{-nm}] = \hat{\mathbf{s}} = \hat{\mathbf{s}}^l + \hat{\mathbf{s}}^{nl} \tag{18}$$

We can write

$$\begin{aligned}\mathbf{M}_{nm} &= diag\left(E\left[(\mathbf{s}-\hat{\mathbf{s}})(\mathbf{s}-\hat{\mathbf{s}})^T\right]\right) = diag\left(E\left[\left(\left(\mathbf{s}-\hat{\mathbf{s}}^l\right)-\hat{\mathbf{s}}^{nl}\right)\left(\left(\mathbf{s}-\hat{\mathbf{s}}^l\right)-\hat{\mathbf{s}}^{nl}\right)^T\right]\right) \\ &= \mathbf{M}_{nm}^l + diag\left(E\left[\hat{\mathbf{s}}^{nl}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) - 2diag\left(E\left[\left(\mathbf{s}-\hat{\mathbf{s}}^l\right)\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right)\end{aligned}, \tag{19}$$

where $\mathbf{M}^l$ is a diagonal matrix whose elements are the MSEs corresponding to the linear estimation of s_i from $\mathbf{x}_{-nm}$, that is,

$$\mathbf{M}_{nm}^l = diag\left(E\left[\left(\mathbf{s}-\hat{\mathbf{s}}^l\right)\left(\mathbf{s}-\hat{\mathbf{s}}^l\right)^T\right]\right) = diag\left(E\left[\mathbf{s}\mathbf{s}^T\right]\right) - diag\left(E\left[\hat{\mathbf{s}}^l\left(\hat{\mathbf{s}}^l\right)^T\right]\right) \tag{20}$$

In (20), we have considered the orthogonality between the error vector and the estimate vector. However, $\hat{\mathbf{s}}^l$ is the minimum LMSE estimate, so it can be obtained from the Wiener-Hopft equations

$$\hat{\mathbf{s}}^l = \mathbf{C}_{sx_{-nm}}\mathbf{C}_{x_{-nm}x_{-nm}}^{-1}\mathbf{x}_{-nm} \quad \mathbf{C}_{sx_{-nm}} = E\left[\mathbf{s}\mathbf{x}_{-nm}^T\right] \quad \mathbf{C}_{x_{-nm}x_{-nm}} = E\left[\mathbf{x}_{-nm}\mathbf{x}_{-nm}^T\right]. \tag{21}$$

Hence, we can write:

$$\mathbf{M}_{nm}^l = \mathbf{I} - diag\left(E\left[\mathbf{C}_{sx_{-nm}}\mathbf{C}_{x_{-nm}x_{-nm}}^{-1}\mathbf{x}_{-nm}\mathbf{x}_{-nm}{}^T\mathbf{C}_{x_{-nm}x_{-nm}}^{-1}\mathbf{C}_{sx_{-nm}}^T\right]\right) = \mathbf{I} - diag\left(\mathbf{C}_{sx_{-nm}}\mathbf{C}_{x_{-nm}x_{-nm}}^{-1}\mathbf{C}_{sx_{-nm}}^T\right) \tag{22}$$

Taking into account $\mathbf{C}_{sx_{-nm}} = \mathbf{W}\mathbf{C}_{xx}\mathbf{T}_{-nm}$, $\mathbf{C}_{x_{-nm}x_{-nm}} = \mathbf{T}_{-nm}^T\mathbf{C}_{xx}\mathbf{T}_{-nm}$ and $\mathbf{C}_{xx} = \mathbf{W}^{-1}\left(\mathbf{W}^{-1}\right)^T$, we can finally express $\mathbf{M}_{nm}^l$ in terms of the ICA model parameters:

$$\mathbf{M}_{nm}^l = \mathbf{I} - diag\left(\left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{-nm}\left(\mathbf{T}_{-nm}^T\mathbf{W}^{-1}\left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{-nm}\right)^{-1}\mathbf{T}_{-nm}^T\mathbf{W}^{-1}\right). \tag{23}$$

Let us now consider the other two terms in (19). First, notice that $\hat{s}_i^{nl}$ can be interpreted as a linear estimate of s_i from $\left(\hat{s}_{in}^l\right)^2$, because assuming that $\hat{s}_{in}^l$ is a standardized Gaussian random variable, then $\left(\hat{s}_{in}^l\right)^2$ is χ^2 having a mean equal to 1 and variance equal to 2. Hence, we can apply orthogonality again:

$$diag\left(E\left[\left(\mathbf{s}-\hat{\mathbf{s}}^{nl}\right)\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) = 0 \Rightarrow diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) = diag\left(E\left[\hat{\mathbf{s}}^{nl}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right). \tag{24}$$

Consequently, we have

$$diag\left(E\left[\hat{\mathbf{s}}^{nl}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) - 2diag\left(E\left[\left(\mathbf{s}-\hat{\mathbf{s}}^l\right)\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) = -diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) + 2diag\left(E\left[\hat{\mathbf{s}}^l\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right). \tag{25}$$

The second term in (25) is zero, because

$$\left[diag\left(E\left[\hat{\mathbf{s}}^l\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right)\right]_{ii} = E\left[\hat{s}_i^l E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right]\left(\frac{\left(\hat{s}_{in}^l\right)^2-1}{2}\right)\right] = \frac{1}{2}E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right]\left(E\left[\hat{s}_i^l\left(\hat{s}_{in}^l\right)^2\right] - E\left[\hat{s}_i^l\right]\right) \tag{26}$$

where $E\left[\hat{s}_i^l\right] = 0$ and $E\left[\hat{s}_i^l\left(\hat{s}_{in}^l\right)^2\right] = \left(\text{var}\left[\hat{s}_i^l\right]\right)^{\frac{1}{2}} E\left[\left(\hat{s}_{in}^l\right)^3\right] = 0$, because we assume that $\hat{s}_{in}^l$ is Gaussian so its odd moments are zero. Regarding the first term in (25)

$$\begin{array}{l}\left[diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right)\right]_{ii} = E\left[s_i E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right]\left(\frac{\left(\hat{s}_{in}^l\right)^2 - 1}{2}\right)\right] \\ = \frac{1}{2} E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right]\left(E\left[s_i\left(\hat{s}_{in}^l\right)^2\right] - E[s_i]\right) = \frac{1}{2\text{var}^2\left[\hat{s}_i^l\right]} E^2\left[s_i \cdot \left(\hat{s}_i^l\right)^2\right]\end{array} \tag{27}$$

Defining the vector $\hat{\mathbf{s}}^{l(2)} = \left[\left(\hat{s}_1^l\right)^2 \ldots \left(\hat{s}_N^l\right)^2\right]^T$ and taking into account that $\text{var}\left[\hat{s}_i^l\right] = \left[\mathbf{C}_{sx_{-nm}} \mathbf{C}_{x_{-nm}x_{-nm}}^{-1} \mathbf{C}_{sx_{-nm}}^T\right]_{ii}$ we can write

$$diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) = \frac{1}{2} diag^2\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{l(2)}\right)^T\right]\right) diag^{-2}\left(\mathbf{C}_{sx_{-nm}} \mathbf{C}_{x_{-nm}x_{-nm}}^{-1} \mathbf{C}_{sx_{-nm}}^T\right), \tag{28}$$

and considering (21) and (23), (28) can be expressed in terms of the ICA model parameters

$$\begin{array}{l} diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right) = \frac{1}{2} \cdot diag^2\left(\mathbf{W} E\left[\mathbf{x}\left(\left(\left(\mathbf{W}^{-1}\right)^T \mathbf{T}_{-nm}\left(\mathbf{T}_{-nm}^T \mathbf{W}^{-1}\left(\mathbf{W}^{-1}\right)^T \mathbf{T}_{-nm}\right)^{-1} \mathbf{x}_{-nm}\right)^{(2)}\right)^T\right]\right) \\ \cdot diag^{-2}\left(\left(\mathbf{W}^{-1}\right)^T \mathbf{T}_{-nm}\left(\mathbf{T}_{-nm}^T \mathbf{W}^{-1}\left(\mathbf{W}^{-1}\right)^T \mathbf{T}_{-nm}\right)^{-1} \mathbf{T}_{-nm}^T \mathbf{W}^{-1}\right)\end{array} \tag{29}$$

So, in conclusion we can express the matrix $\mathbf{M}_{nm}$ as

$$\mathbf{M}_{nm} = \mathbf{M}_{nm}^l - diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right), \tag{30}$$

where $\mathbf{M}_{nm}^l$ and $diag\left(E\left[\mathbf{s}\left(\hat{\mathbf{s}}^{nl}\right)^T\right]\right)$ can be obtained from (23) and (29), respectively, using estimates $\hat{\mathbf{W}}$ of the model parameters and a sample mean to evaluate the expectation required in (29). Algorithm 1 below describes the estimation procedure.

Algorithm 1: Computing ICA-PCC.

1: **Input: Learning data set** $\mathbf{x}^{(l)}$ $l = 1 \ldots L$
2: **Compute** $\hat{\mathbf{W}}$ from the learning data set (any ICA algorithm is a candidate)
3: **for** $n = 1, 2 \ldots N$
4: **for** $m = n \ldots N$
3: Compute $\hat{\mathbf{M}}_{nm}$ (Equations (23), (29), (30))
4: Compute $\hat{\mathbf{C}}_{e_{nm}e_{nm}} = -\left(\hat{\mathbf{W}}\mathbf{T}_{nm}\right)^+ \hat{\mathbf{M}}_{nm}\left(\left(\hat{\mathbf{W}}\mathbf{T}_{nm}\right)^+\right)^T$
5: Compute $\hat{\rho}_{nm}^{ICA-PCC}$ (Equation (13))
6: Compute $\hat{\rho}_{mn}^{ICA-PCC} = \hat{\rho}_{nm}^{ICA-PCC}$
8: **end for**
9: **end for**
10: **Output** $\hat{\rho}_{mn}^{ICA-PCC}$ $n = 1 \ldots N$ $m = 1 \ldots N$

Equation (30) provides an interesting decomposition of mse_{nmi}. Let mse_{nmi}^l be called the entries of the diagonal matrix $\mathbf{M}_{nm}^l$, then mse_{nmi} can be expressed as mse_{nmi}^l minus a nonnegative term (see Equation (29)), so that $mse_{nmi} \leq mse_{nmi}^l$. The condition $mse_{nmi} = mse_{nmi}^l \Leftrightarrow \mathbf{M}_{nm} = \mathbf{M}_{nm}^l$ holds for the Gaussian case, because then $E\left[s_i \cdot \left(\hat{s}_{in}^l\right)^2\right]$ (see Equation (27)) becomes zero (it is an odd higher-order moment of a multivariate Gaussian variable). In such a case, $E[\mathbf{s}/\mathbf{x}_{-nm}] = \hat{\mathbf{s}}^l$ becomes a linear function

of $\mathbf{x}_{-nm}$ and so the same happens with $E[\mathbf{x}/\mathbf{x}_{-nm}]$ in (9). Hence, the second term in (30) is responsible for the improved reduction in the influence of $\mathbf{x}_{-nm}$ in the estimation of the partial correlation between x_n and x_m, in the non-Gaussian case. Moreover, we should expect similar results for ICA-PCC and PCC for the Gaussian case.

4. Experiments

4.1. Synthetic Data Experiments

In this experiment we evaluated the influence of the training set size in the estimation of $\rho_{nm}^{ICA-PCC}$ as well as comparing the quality of the estimate with the one obtained from the precision matrix. To this aim, we generated synthetic data corresponding to three different ICA models. In the first one, the sources s_i were independent and identically distributed (i.i.d.) random variables having a unit-variance zero-mean uniform pdf. This correspond to an example of sub-Gaussian distribution, as the excess kurtosis is negative, $\kappa - \kappa_G = -1.2$, where κ is the kurtosis, and $\kappa_G = 3$ is the kurtosis of a Gaussian pdf. In the second model, the sources s_i were i.i.d. random variables having a unit-variance zero-mean Laplacian pdf. This is an example of super-Gaussian distribution, as the excess kurtosis is positive $\kappa - \kappa_G = 3$. In the third model, some sources are uniform and the rest are Laplacian. Finally, we also considered the Gaussian case by generating sources having a standard Gaussian pdf. Figure 1 shows the errors corresponding to the estimation of $\rho_{nm}^{ICA-PCC}$ for the four models.

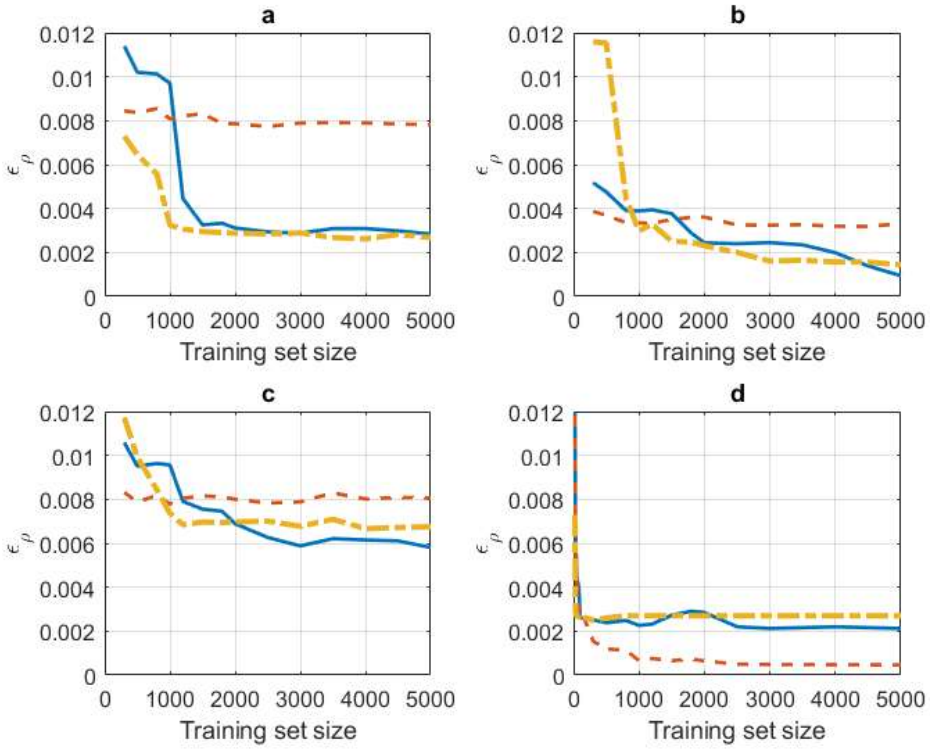

Figure 1. $\in_\rho^{ICA-PCC}$ (blue, Extended-Infomax, yellow, JADE) and $\in_\rho^{PCC}$; (**a**) sub-Gaussian case (**b**) super-Gaussian case (**c**) Mixed (15/5) sub/super-Gaussian case (**d**) Gaussian case.

Every curve is an average of 10 curves corresponding to 10 different runs. In every run, an ICA matrix $\mathbf{U} = \mathbf{W}^{-1}$ was randomly selected; every entry was obtained by sampling a standard Gaussian pdf. Then, a varying number of training vectors $\mathbf{x}$ was generated from source vectors $\mathbf{s}$ having independent components sampled from the mentioned marginal pdfs: sub-Gaussian (Figure 1a), super-Gaussian (Figure 1b), mixed of sub/super-Gaussian (Figure 1c) and Gaussian (Figure 1d). The error was computed as

$$\in_{\rho}^{ICA-PCC} = \frac{1}{N^2 - N} \sum_{n=1}^{N} \sum_{m \neq n} \left(\left| \rho_{nm}^{ICA-PCC} \right| - \left| \hat{\rho}_{mn}^{ICA-PCC} \right| \right)^2 \tag{31}$$

and averaged over the 10 runs for every training set size. Notice that $0 \leq \in_{\rho}^{ICA-PCC} \leq 1$, because $\in_{\rho\min}^{ICA-PCC} = 0$, when $\left| \rho_{nm}^{ICA-PCC} \right| = \left| \hat{\rho}_{nm}^{ICA-PCC} \right| \; \forall n \; \forall m \neq n$ and $\in_{\rho\max}^{ICA-PCC} = 1$, when $\left| \rho_{nm}^{ICA-PCC} \right| - \left| \hat{\rho}_{nm}^{ICA-PCC} \right| = \pm 1 \; \forall n \; \forall m \neq n$. In (31), $\rho_{nm}^{ICA-PCC}$ was obtained from Algorithm 1 using the true matrix $\mathbf{W}$ and an extremely large number of instances for the sample mean required to compute the expectation in (29). On the other hand, $\hat{\rho}_{mn}^{ICA-PCC}$ was computed from Algorithm 1 using estimates of $\mathbf{W}$ obtained with the corresponding finite training set. We used the Extended Infomax algorithm described in [34] and the JADE algorithm [35]. Extended Infomax is an extension of the Infomax algorithm [36] used to deal with mixed sub/super-Gaussian sources. It is representative of algorithms that iteratively optimize some defined cost-function like Fast-ICA. JADE is based on matrix computation and diagonalization, so, it is not sensitive to initialization or optimization path problems. The same finite training set was also used to evaluate the expectation in (29). In all cases we considered $N = 20$. For comparison, we also computed the error

$$\in_{\rho}^{PCC} = \frac{1}{N^2 - N} \sum_{n=1}^{N} \sum_{m \neq n} \left(\left| \rho_{nm}^{ICA-PCC} \right| - \left| \hat{\rho}_{mn}^{PCC} \right| \right)^2 \tag{32}$$

which corresponds to the PCCs obtained from empirical estimates $\hat{\mathbf{Q}}$ of the precision matrix as indicated in (3): $\hat{\rho}_{nm}^{PCC} = -\hat{q}_{nm} / \sqrt{\hat{q}_{nn} \hat{q}_{mm}}$. Notice that it is also $0 \leq \in_{\rho}^{PCC} \leq 1$.

Several conclusions may be drawn from Figure 1. First, we can see that in the non-Gaussian cases (a) (b) and (c), PCC cannot decrease the error with increased training set size. This demonstrates the model mismatch due to the implicit Gaussianity of PCC. In these three cases, ICA-PCC methods improve on PCC after a sufficient number of training samples and maintain a decreased error for an increased training set size. The minimum training set size required to improve on PCC depends on the case and on the ICA-PCC method. Thus, this minimum number is smaller in Figure 1a (sub-Gaussians) than in Figure 1b (super-Gaussians), and has an intermediate value in Figure 1c (mixed sub/super Gaussians). However, notice that the non-decreasing error of PCC is much higher in Figure 1a,c, so this suggests that ICA-PCC improvement begins from a smaller training set size. On the other hand, this minimum value is smaller in JADE than in Extended-Infomax. This is due to the different nature of both algorithms. JADE requires a matrix diagonalization, while Extended-Infomax requires iterative learning. However the computational complexity of JADE is much larger, especially as N increases. Regarding convergence for large values of the training set size, we can see that the error level of the mixed case is clearly above the others. This is because, in general all ICA algorithms have more difficulties in estimating the model in the mixed case. Actually, Extended-Infomax was conceived in an effort to deal with the mixed case by incorporating a procedure to estimate the class (sub/super) of every source. This explains the smaller error of Extended-Infomax in Figure 1c with respect to JADE, after a given training set size. For the Gaussian case, PCC yields a very small error, which decreases with increasing training set size. In this case, ICA-PCC is worse than PCC, although the error is reasonably small. Remember that, in the Gaussian case, we expected similar results for both methods, however, the estimation path followed is different: in PCC the precision matrix is directly estimated, while in ICA-PCC the matrices $\mathbf{W}$ and $\mathbf{M}_{nm}$ are estimated in Algorithm 1. This could explain the separation observed in the error curves of Figure 1d. Finally, most ICA algorithms decompose the estimation of $\mathbf{W}$ into two steps: first, estimate a decorrelation matrix, and then, a rotation matrix. When the independent components are Gaussian, any rotation matrix is valid, as all of them are compatible with Gaussianity. However, in the non-Gaussian cases the rotation matrix must be properly estimated for the corresponding non-Gaussian model. This can be interpreted as if a smaller number

of model parameters (entries of the decorrelation matrix) should be actually estimated in the Gaussian case. This explains the faster convergence of the curves in Figure 1d.

4.2. A Real Data Application

We applied the proposed method to quantify the significance of changes in brain connectivity during sleep of patients having disorders like apnea or epilepsy [37]. These disorders are characterized by regular arousal, which are stages of abnormal degraded sleep. The frequency of arousals in a given period of time is related to the seriousness of the pathology. However, the intensity of the arousals may also be relevant for an appropriate diagnosis. Assuming that an arousal is associated to changes in brain connectivity [38], a measure related to the change magnitude may be useful to quantify the significance of the pathology. To this aim, the patient was monitored during sleep by 19 channels of EEG recordings. Every signal channel was segmented into intervals of 2 s and a given feature was computed in every interval and averaged in epochs of 26 s. Each epoch was manually or automatically [22] labelled in two possible states: normal sleep (state 0) or arousal (state 1). Then, associated to every epoch, an observation vector $\mathbf{x}$ was built with one feature extracted from all the channels (the same type of feature for all of them), thus $N = 19$. In this experiment, a total of 2000 epochs were available in every state. Given these data sets, an average measure related to brain connectivity was computed to quantify the importance of brain changes between the two states.

There are many possible definitions of overall connectivity, here, we considered the so called algebraic connectivity [39], which can be computed as the second smallest eigenvalue λ_2 of the graph Laplacian matrix [40] $\mathbf{L} = \mathbf{D} - \mathbf{A}$, being $\mathbf{D}$ a diagonal matrix with entries $d_{nn} = \sum_{m \neq n} a_{nm}$ and $\mathbf{A}$ the adjacency matrix with entries $a_{nm} \geq 0$. The Laplacian matrix is semidefinite positive with the smallest eigenvalue λ_1 equal to zero, then $\lambda_2 \geq 0$. Moreover, it is demonstrated in [39] that $\lambda_2 = N$ for a complete graph (a graph with $a_{nm} = 1 \; \forall n \neq m$). It is also demonstrated in [39] that

$$\lambda_2 \leq \frac{N}{N-1} \min[d_{nn}] \tag{33}$$

Hence, assuming that $0 \leq a_{nm} \leq 1$ (as it will be in our case), the greatest upper bound for λ_2 in (33) corresponds to the complete graph ($a_{nm} = 1 \; \forall n \neq m \; \Rightarrow \; d_{nn} = N - 1 \; \forall n$), therefore $0 \leq \lambda_2 \leq N$. Consequently, we proposed a normalized version of λ_2 to measure the connectivity

$$\varsigma = \frac{\lambda_2}{N} \quad 0 \leq \varsigma \leq 1 \tag{34}$$

The lower bound $\varsigma = 0$ corresponds to a disconnected graph, as it implies an order of multiplicity greater than 1 of the smallest eigenvalue. The upper bound corresponds to a complete graph, which is the one having maximum connectivity under the constraint $0 \leq a_{nm} \leq 1$. We obtained connectivity estimates for every state (0 or 1) and method (ICA-PCC or PCC): $\hat{\varsigma}_0^{ICA-PCC}$, $\hat{\varsigma}_1^{ICA-PCC}$, $\hat{\varsigma}_0^{PCC}$, $\hat{\varsigma}_1^{PCC}$. This was made from (34) with N = 19, after computing the second smallest eigenvalue of the Laplacian matrix, considering that the entries of the associated adjacency matrix are the respective magnitudes of the partial correlation estimates obtained from the training set 0 or 1:

$$\begin{array}{l} a_{nm0}^{ICA-PCC} = \left|\hat{\rho}_{nm0}^{ICA-PCC}\right|, \quad a_{nm1}^{ICA-PCC} = \left|\hat{\rho}_{nm1}^{ICA-PCC}\right| \\ a_{nm0}^{IPCC} = \left|\hat{\rho}_{nm0}^{PCC}\right|, \quad a_{nm1}^{PCC} = \left|\hat{\rho}_{nm1}^{PCC}\right| \end{array} \tag{35}$$

Two different features were considered separately. The first is "amplitude" (*Amp*), which is the maximum amplitude in the corresponding 2 s interval, the second is the "alfa-slow-index" (*Asi*), which is the ratio of power in the alpha band (8.0–11 Hz) to the combined power in the delta (0.5–3.5 Hz) and theta (3.5–8.0 Hz) bands. Tables 1 and 2 show the results corresponding to the *Amp* and the *Asi* features, respectively, for 6 different patients. Together with the normalized connectivity, we

included the connectivity variation between states defined as $\Delta^{ICA-PCC} = \left|\hat{\varsigma}_1^{ICA-PCC} - \hat{\varsigma}_0^{ICA-PCC}\right|$ and $\Delta^{PCC} = \left|\hat{\varsigma}_1^{PCC} - \hat{\varsigma}_0^{PCC}\right|$. We also included a kurtosis estimate for every patient and state. This estimate was obtained as a mean of all the 19 empirical kurtosis separately calculated for every component of vector $\mathbf{x}$, i.e., the empirical kurtosis of the marginal distributions of $\mathbf{x}$. Notice that the estimated kurtosis is clearly above the Gaussian reference $\kappa_G = 3$, so Gaussianity assumption does not hold in this case.

Table 1. Results corresponding to the amplitude (*Amp*).

Subj.	κ_0	κ_1	$\hat{\varsigma}_0^{ICA-PCC}$	$\hat{\varsigma}_1^{ICA-PCC}$	$\Delta^{ICA-PCC}$	$\hat{\varsigma}_0^{PCC}$	$\hat{\varsigma}_1^{PCC}$	Δ^{PCC}
S1	6.46	4.58	0.30	0.33	**0.03**	0.03	0.03	0.00
S2	8.05	5.29	0.74	0.39	**0.35**	0.04	0.04	0.00
S3	9.84	6.76	0.57	0.28	**0.29**	0.03	0.02	0.01
S4	9.04	8.87	0.39	0.66	**0.27**	0.04	0.02	0.02
S5	9.61	15.13	0.31	0.44	**0.13**	0.02	0.03	0.01
S6	9.14	13.82	0.24	0.36	**0.12**	0.02	0.02	0.00

Table 2. Results corresponding to the alfa-slow-index (*Asi*).

Subj.	κ_0	κ_1	$\hat{\varsigma}_0^{ICA-PCC}$	$\hat{\varsigma}_1^{ICA-PCC}$	$\Delta^{ICA-PCC}$	$\hat{\varsigma}_0^{PCC}$	$\hat{\varsigma}_1^{PCC}$	Δ^{PCC}
S1	16.32	22.51	0.31	0.51	**0.20**	0.02	0.02	0.00
S2	10.52	9.09	0.34	0.60	**0.26**	0.02	0.02	0.00
S3	9.91	7.05	0.68	0.48	**0.20**	0.02	0.03	0.01
S4	8.39	11.69	0.37	0.74	**0.37**	0.03	0.02	0.01
S5	7.72	13.15	0.22	0.71	**0.49**	0.02	0.03	0.01
S6	11.86	9.24	0.43	0.56	**0.13**	0.02	0.03	0.01

We can see in Tables 1 and 2 that the PCC method yields very small values of connectivity for all subjects and states, therefore, it is not sensitive to possible changes between states. However, ICA-PCC provides larger values of connectivity and significant changes between states. Figures 2 and 3 show the estimated adjacency matrices corresponding to the different subjects, methods and states. We can see that PCC magnitudes are, in general, much lower than ICA-PCC magnitudes, therefore, PCC has more difficulty revealing the interrelations between the different EEG channels due to the brain activity. This may be explained in terms of the residuals e_n and e_m. Notice from (12) that

$$E\left[e_n^2\right] = \sum_{i=1}^{N} mse_{nmi}\left(v_{1nmi}^{+}\right)^2 \quad E\left[e_m^2\right] = \sum_{i=1}^{N} mse_{nmi}\left(v_{2nmi}^{+}\right)^2 \tag{36}$$

where $\left\{v_{1nmi}^{+}\right\}$ and $\left\{v_{2nmi}^{+}\right\}$ are the elements of vectors $\mathbf{v}_{1nm}^{+}$ and $\mathbf{v}_{2nm}^{+}$, respectively, and later, these are the first and second row of $(\mathbf{WT}_{nm})^{+}$, respectively. We showed in Section 3 that PCC should be similar to ICA-PCC for $mse_{nmi} = mse_{nmi}^{l}$, but for non-Gaussian observations $mse_{nmi} < mse_{nmi}^{l}$, so it is deduced from (36) that

$$\begin{aligned} E\left[e_n^2\right] \leq E^l\left[e_n^2\right] &= \sum_{i=1}^{N} mse_{nmi}^{l}\left(v_{1nmi}^{+}\right)^2 \\ E\left[e_m^2\right] \leq E^l\left[e_m^2\right] &= \sum_{i=1}^{N} mse_{nmi}^{l}\left(v_{2nmi}^{+}\right)^2 \end{aligned} \tag{37}$$

where equality holds in the Gaussian case. So, PCC provides overestimated residuals where the actual partial correlation between x_n and x_m may be eventually hidden. This masking effect should increase with the non-Gaussianity character of the observations. In our experiment, the features are highly non-Gaussian as demonstrated by the kurtosis values of Tables 1 and 2. So, when using PCC, the "true" residuals seem to be overestimated by rather uncorrelated residuals that provide a too low estimation of the actual interrelation between the different EEG channels.

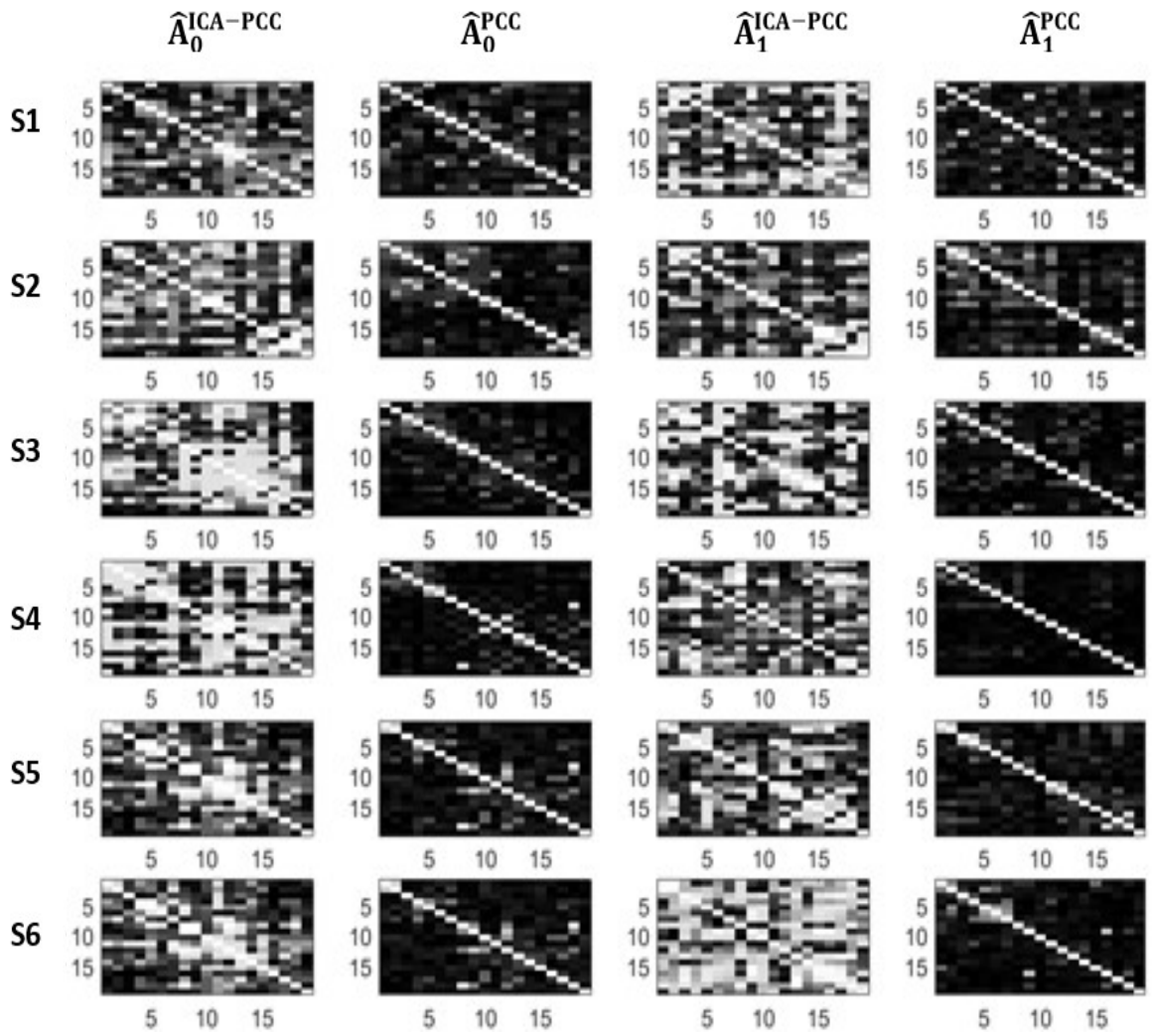

Figure 2. Adjacency matrices corresponding to *Amp*.

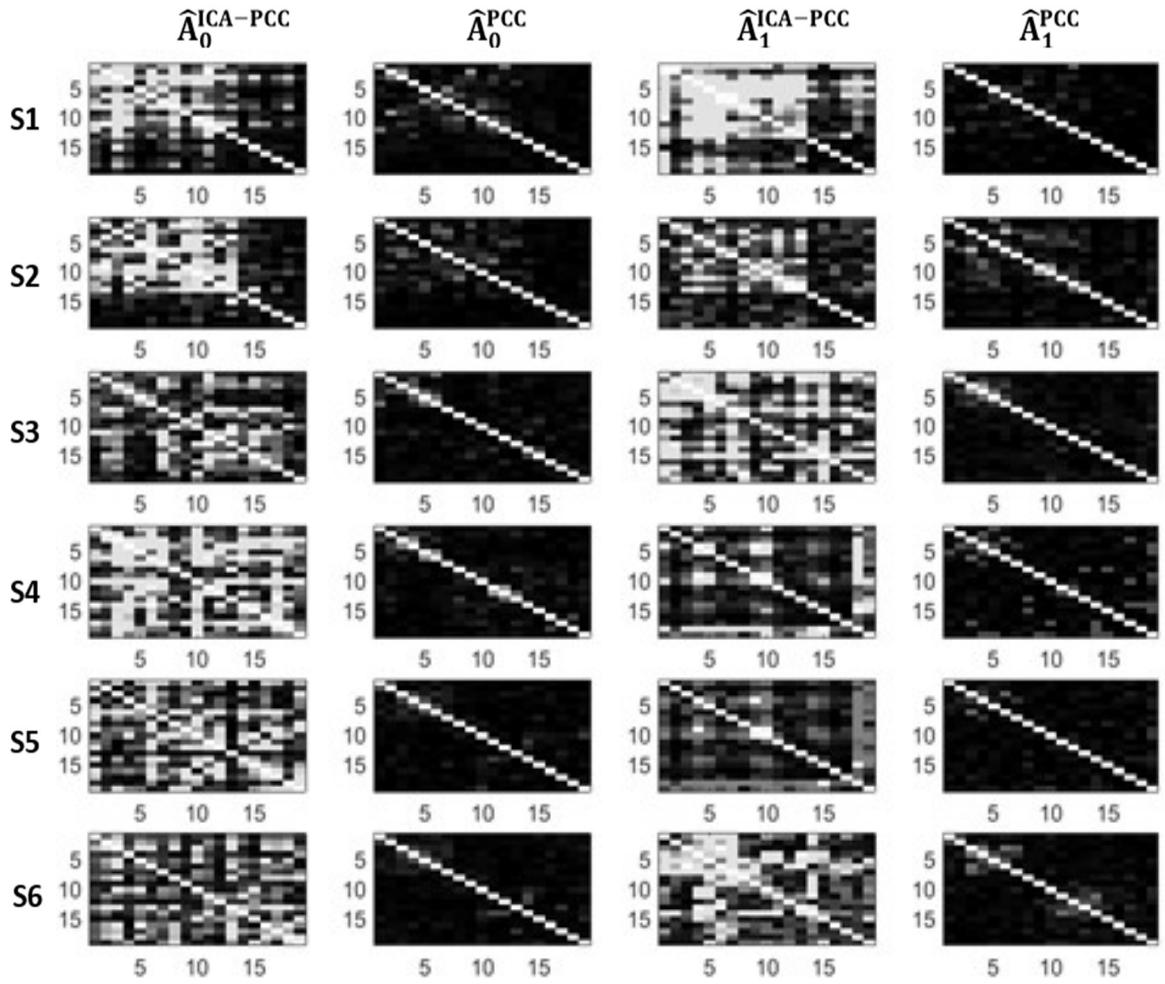

Figure 3. Adjacency matrices corresponding to *Asi*.

5. Conclusions and Extensions

Partial correlations may be used to define the weights of an undirected graph for subsequent graph signal processing. Conventionally, partial correlations are obtained from the precision matrix, but this is optimal only under the Gaussianity assumption. Hence, we have proposed a new method for computing the partial correlation, assuming a non-Gaussian model (ICA). The latter is a versatile model which suits a diversity of non-Gaussian pdfs.

The proposed method requires the computation of the ICA model parameters, which can be made by using any of the many existing algorithms. Two different ICA methods have been considered in the synthetic examples, which may be considered representative of two different kinds of approaches to estimating the ICA model parameters. Both yield similar performance. Computing the mean-square-errors corresponding to the optimal estimation of the sources is also required. Hence, we have proposed a second-order approximation of the conditional mean. Higher orders could be tried at the price of increased complexity.

We have verified, both by simulations and by real data experiments that the new method better captures the pairwise and overall connectivity of the graph compared to the precision matrix in non-Gaussian scenarios. The results could be extended to larger values of N but the training set sizes should be correspondingly increased to keep the quality of the model parameter estimates.

Future extensions of this work can be devised. Some kind of regularization is desirable to emphasize the relevant information provided by the graph connectivity and/or to establish more natural relations between the connected nodes. Thus, sparsity is a common requirement of graph learning (see [41] as a representative example). Considering Equation (10), sparsity could be imposed

by selecting only those sources that significantly contribute to the partial correlation between x_n and x_m, i.e., by soft or hard thresholding on mse_{nmi}. On the other hand, smoothness regularization could be tried in a similar manner to the approach proposed in [42] for the Gaussian case. To this aim, it could be considered that the representation matrix U can be factorized in a correlation matrix multiplied by a rotation (unitary) matrix [43]. Understanding how this rotation relates to the graph connectivity may allow the definition of cost functions, which include some possible smoothness related terms. Other structural constraints [44] could also be compatible with the non-Gaussian model.

Author Contributions: The work reported here was developed in collaboration among all authors. All authors have contributed to the preparation of the manuscript, and have approved it. Conceptualization, L.V. and A.S.; Data curation, J.B. and G.S.; Formal analysis, J.B., L.V. and G.S.; Funding acquisition, L.V. and A.S.; Investigation, J.B., G.S. and A.S.; Software, J.B. and G.S.; Supervision, L.V. and A.S.; Writing—original draft, J.B. and L.V.; Writing—review & editing, J.B. and L.V.

Funding: This research was funded by Spanish Administration and European Union under grants TEC2014-58438-R and TEC2017-84743-P.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivation of the General Formula

Due to the orthogonality between the error and the conditional mean, we can write

$$\mathbf{C}_{e_{nm}e_{nm}} = E\left[\mathbf{x}_{nm}\mathbf{x}_{nm}^T\right] - E\left[E[\mathbf{x}_{nm}/\mathbf{x}_{-nm}]E^T[\mathbf{x}_{nm}/\mathbf{x}_{-nm}]\right] \tag{A1}$$

Let us define

$$\mathbf{f}(\mathbf{x}_{-nm}) = E[\mathbf{s}/\mathbf{x}_{-nm}] \equiv \mathbf{f} \quad \mathbf{q}(\mathbf{x}_{-nm}) = -\mathbf{W}\mathbf{T}_{-nm}\mathbf{x}_{-nm} = \mathbf{q} \tag{A2}$$

So, from (9), we may write $E[\mathbf{x}_{nm}/\mathbf{x}_{-nm}] = (\mathbf{W}\mathbf{T}_{nm})^+(\mathbf{f}+\mathbf{q})$, hence

$$\begin{aligned}\mathbf{C}_{e_{nm}e_{nm}} = E\left[\mathbf{x}_{nm}\mathbf{x}_{nm}^T\right] - E\left[(\mathbf{W}\mathbf{T}_{nm})^+(\mathbf{f}+\mathbf{q})\left(\mathbf{f}^T+\mathbf{q}^T\right)\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T\right] = \underbrace{E\left[\mathbf{x}_{nm}\mathbf{x}_{nm}^T\right]}_{1} - \underbrace{(\mathbf{W}\mathbf{T}_{nm})^+E\left[\mathbf{q}\mathbf{q}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T}_{2} \\ \underbrace{-(\mathbf{W}\mathbf{T}_{nm})^+E\left[\mathbf{f}\mathbf{f}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T}_{3} \underbrace{-\left((\mathbf{W}\mathbf{T}_{nm})^+E\left[\mathbf{f}\mathbf{q}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T + (\mathbf{W}\mathbf{T}_{nm})^+E^T\left[\mathbf{f}\mathbf{q}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T\right)}_{4}\end{aligned} \tag{A3}$$

We have to compute the 4 terms of (A3).

- *Term 1*

$$E\left[\mathbf{x}_{nm}\mathbf{x}_{nm}^T\right] = \mathbf{C}_{x_{nm}x_{nm}} = \mathbf{T}_{nm}^T\mathbf{C}_{xx}\mathbf{T}_{nm} \tag{A4}$$

- *Term 2*

$$\begin{aligned}E[\mathbf{q}\mathbf{q}^T] = \; & E\left[\mathbf{W}\mathbf{T}_{-nm}\mathbf{x}_{-nm}\mathbf{x}_{-nm}^T\mathbf{T}_{-nm}^T\mathbf{W}^T\right] = \mathbf{W}\mathbf{T}_{-nm}\mathbf{C}_{x_{-nm}x_{-nm}}\mathbf{T}_{-nm}^T\mathbf{W}^T = \mathbf{W}\mathbf{T}_{-nm}\mathbf{T}_{-nm}^T\mathbf{C}_{xx}\mathbf{T}_{-nm}\mathbf{T}_{-nm}^T\mathbf{W}^T \\ & = \mathbf{W}(\mathbf{I}-\mathbf{T}_{nm}\mathbf{T}_{nm}^T)\mathbf{C}_{xx}(\mathbf{I}-\mathbf{T}_{nm}\mathbf{T}_{nm}^T)\mathbf{W}^T) = \mathbf{W}\mathbf{C}_{xx}\mathbf{W}^T + \mathbf{W}\mathbf{T}_{nm}\mathbf{T}_{nm}^T\mathbf{C}_{xx}\mathbf{T}_{nm}\mathbf{T}_{nm}^T\mathbf{W}^T \\ & -\mathbf{W}\mathbf{T}_{nm}\mathbf{T}_{nm}^T\mathbf{C}_{xx}\mathbf{W}^T - \mathbf{W}\mathbf{C}_{xx}\mathbf{T}_{nm}\mathbf{T}_{nm}^T\mathbf{W}^T\end{aligned} \tag{A5}$$

$$\begin{aligned}(\mathbf{W}\mathbf{T}_{nm})^+E[\mathbf{q}\mathbf{q}^T]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T = (\mathbf{W}\mathbf{T}_{nm})^+\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T + \mathbf{T}_{nm}^T\mathbf{C}_{xx}\mathbf{T}_{nm} \\ -\mathbf{T}_{nm}^T\mathbf{W}^{-1}\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T - (\mathbf{W}\mathbf{T}_{nm})^+\left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{nm}\end{aligned}$$

- *Term 3*

$$\begin{array}{l} i \neq j \\ E\left[\mathbf{f}\mathbf{f}^T\right](i,j) = E\left[E[s_i/\mathbf{x}_{-nm}]E\left[s_j/\mathbf{x}_{-nm}\right]\right] \\ = E[E[s_i/\mathbf{x}_{-nm}]]E\left[E\left[s_j/\mathbf{x}_{-nm}\right]\right] = E[s_i]E\left[s_j\right] = 0 \\ i = j \\ E\left[\mathbf{f}\mathbf{f}^T\right](i,i) = E\left[E^2[s_i/\mathbf{x}_{-nm}]\right] = \mathrm{var}[E[s_i/\mathbf{x}_{-nm}]] + E^2[E[s_i/\mathbf{x}_{-nm}]] \\ = \mathrm{var}[E[s_i/\mathbf{x}_{-nm}]] + E^2[s_i] = \mathrm{var}[E[s_i/\mathbf{x}_{-nm}]] \end{array} \tag{A6}$$

Let us express $\mathrm{var}[E[s_i/\mathbf{x}_{-nm}]]$ in terms of mse_i as defined in (A7)

$$\begin{array}{l} mse_{nmi} = E\left[(s_i - E[s_i/\mathbf{x}_{-nm}])^2\right] = E\left[s_i^2\right] + E\left[E^2[s_i/\mathbf{x}_{-nm}]\right] - 2E[s_i E[s_i/\mathbf{x}_{-nm}]] \\ E[(s_i - E[s_i/\mathbf{x}_{-nm}])E[s_i/\mathbf{x}_{-nm}]] = 0 \\ \Rightarrow \quad E[s_i E[s_i/\mathbf{x}_{-nm}]] = E[E[s_i/\mathbf{x}_{-nm}]E[s_i/\mathbf{x}_{-nm}]] \\ \Rightarrow mse_{nmi} = E\left[s_i^2\right] - E\left[E_i^2[s_i/\mathbf{x}_{-nm}]\right] = 1 - \mathrm{var}[E[s_i/\mathbf{x}_{-nm}]] \end{array} \tag{A7}$$

where we have taken into account that the conditional mean is an unbiased estimator. Then, we know that $E\left[\mathbf{f}\mathbf{f}^T\right](i,i) = \mathrm{var}[E[s_i/\mathbf{x}_{-nm}]] = 1 - mse_{nmi}$ and if we define $\mathbf{M}_{nm}$ as a $(N \times N)$ diagonal matrix having in its main diagonal the values mse_{nmi} $i = 1 \ldots N$, we can write

$$(\mathbf{W}\mathbf{T}_{nm})^+ E\left[\mathbf{f}\mathbf{f}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T = (\mathbf{W}\mathbf{T}_{nm})^+(\mathbf{I} - \mathbf{M}_{nm})\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \tag{A8}$$

- *Term 4*

$$\begin{array}{l} E\left[\mathbf{f}\mathbf{q}^T\right] = -E\left[\mathbf{f}\mathbf{x}_{-nm}^T\mathbf{T}_{-nm}^T\mathbf{W}^T\right] \\ E\left[E[s_i/\mathbf{x}_{-nm}]x_j\right] \overbrace{=}^{x_j \in \{\mathbf{x}_{-nm}\}} \int_{x_{-nm}} x_j E[s_i/x_{-nm}]p(\mathbf{x}_{-nm})d\mathbf{x}_{-nm} = \int_{x_{-nm}} x_j\left[\int_{s_i} s_i p(s_i/x_{-nm})ds_i\right]p(\mathbf{x}_{-nm})d\mathbf{x}_{-nm} \\ = \int_{s_i} s_i\left[\int_{x_{-nm}} x_j p(s_i/x_{-nm})p(\mathbf{x}_{-nm})d\mathbf{x}_{-nm}\right]ds_i = \int_{s_i} s_i\left[\int_{x_{-nm}} x_j p(x_{-nm}/s_i)p(s_i)d\mathbf{x}_{-nm}\right]ds_i \\ = \int_{s_i} s_i p(s_i)\left[\int_{x_{-nm}} x_j p(x_{-nm}/s_i)d\mathbf{x}_{-nm}\right]ds_i = \int_{s_j} s_i p(s_i)\left[\int_{x_j} x_j p(x_j/s_i)dx_i\right]ds_i \\ = \int_{s_j}\int_{x_j} s_i x_j p(s_i, x_j)ds_i dx_j = E\left[s_i x_j\right] \\ E\left[\mathbf{f}\mathbf{q}^T\right] = -E\left[\mathbf{f}\mathbf{x}_{-nm}^T\mathbf{T}_{-nm}^T\right]\mathbf{W}^T = E\left[\mathbf{s}\mathbf{x}_{-nm}^T\mathbf{T}_{-nm}^T\right]\mathbf{W}^T = -E\left[\mathbf{W}\mathbf{x}\mathbf{x}_{-nm}^T\mathbf{T}_{-nm}^T\right]\mathbf{W}^T = -\mathbf{W}\mathbf{C}_{xx_{-nm}}\mathbf{T}_{-nm}^T\mathbf{W}^T \\ \quad = -\mathbf{W}\mathbf{C}_{xx}\mathbf{T}_{-nm}\mathbf{T}_{-nm}^T\mathbf{W}^T = -\left(\mathbf{W}^{-1}\right)^T(\mathbf{I} - \mathbf{T}_{nm}\mathbf{T}_{nm}^T)\mathbf{W}^T = \left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{nm}\mathbf{T}_{nm}^T\mathbf{W}^T - \mathbf{I} \\ (\mathbf{W}\mathbf{T}_{nm})^+ E\left[\mathbf{f}\mathbf{q}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T = (\mathbf{W}\mathbf{T}_{nm})^+\left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{nm} - (\mathbf{W}\mathbf{T}_{nm})^+\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \\ (\mathbf{W}\mathbf{T}_{nm})^+ E^T\left[\mathbf{f}\mathbf{q}^T\right]\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T = \mathbf{T}_{nm}^T\mathbf{W}^{-1}\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T - (\mathbf{W}\mathbf{T}_{nm})^+\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \end{array} \tag{A9}$$

Finally, considering (A3)–(A6) and (A8) we can write:

$$\begin{array}{l} \mathbf{C}_{e_{nm}e_{nm}} = \mathbf{T}_{nm}^T\mathbf{C}_{xx}\mathbf{T}_{nm} - (\mathbf{W}\mathbf{T}_{nm})^+\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T - \mathbf{T}_{nm}^T\mathbf{C}_{xx}\mathbf{T}_{nm} \\ +\mathbf{T}_{nm}^T\mathbf{W}^{-1}\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T + (\mathbf{W}\mathbf{T}_{nm})^+\left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{nm} - (\mathbf{W}\mathbf{T}_{nm})^+(\mathbf{I} - \mathbf{M}_{nm})\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \\ -(\mathbf{W}\mathbf{T}_{nm})^+\left(\mathbf{W}^{-1}\right)^T\mathbf{T}_{nm} + (\mathbf{W}\mathbf{T}_{nm})^+\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T - \mathbf{T}_{nm}^T\mathbf{W}^{-1}\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T + (\mathbf{W}\mathbf{T}_{nm})^+\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \\ = (\mathbf{W}\mathbf{T}_{nm})^+(\mathbf{M}_{nm})\left((\mathbf{W}\mathbf{T}_{nm})^+\right)^T \end{array} \tag{A10}$$

References

1. Baba, K.; Shibata, R.; Sibuya, M. Partial correlation and conditional correlation as measures of conditional independence. *Aust. N. Z. J. Stat.* **2004**, *46*, 657–664. [CrossRef]
2. Shuman, D.I.; Narang, S.K.; Frossard, P.; Ortega, A.; Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **2013**, *30*, 83–98. [CrossRef]

3. Sandryhaila, A.; Moura, J.M.F. Discrete signal processing on graphs. *IEEE Trans. Signal Process.* **2013**, *61*, 1644–1656. [CrossRef]
4. Ortega, A.; Frossard, P.; Kovacevic, J.; Moura, J.M.F.; Vandergheynst, P. Graph Signal Processing: Overview, challenges and applications. *Proc. IEEE* **2018**, *106*, 808–828. [CrossRef]
5. Zhang, C.; Florencio, D.; Chou, P.A. *Graph Signal Processing—A Probabilistic Framework*; Tech. Rep. MSR-TR-2015-31; Microsoft Research Lab: Redmond, WA, USA, 2015.
6. Pávez, E.; Ortega, A. Generalized precision matrix estimation for graph signal processing. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 6350–6354.
7. Mazumder, R.; Hastie, T. The graphical lasso: New insights and alternatives. *Electron. J. Stat.* **2012**, *6*, 2125–2149. [CrossRef] [PubMed]
8. Hsieh, C.J.; Sustik, M.A.; Dhillon, I.S.; Ravikumar, P. Sparse inverse covariance matrix estimation using quadratic approximation. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 2330–2338.
9. Chen, X.; Xu, M.; Wu, W.B. Covariance and precision matrix estimation for high-dimensional time series. *Ann. Stat.* **2013**, *41*, 2994–3021. [CrossRef]
10. Öllerer, V.; Croux, C. Robust high-dimensional precision matrix estimation. In *Modern Multivariate and Robust Methods*; Nordhausen, K., Taskinen, S., Eds.; Springer: New York, NY, USA, 2015; pp. 329–354.
11. Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432–441. [CrossRef]
12. Peng, J.; Wang, P.; Zhou, N.; Zhu, J. Partial correlation estimation by joint sparse regression model. *J. Am. Stat. Assoc.* **2009**, *104*, 735–746. [CrossRef]
13. Belda, J.; Vergara, L.; Salazar, A.; Safont, G. Estimating the Laplacian matrix of Gaussian mixtures for signal processing on graphs. *Signal Process.* **2018**, *148*, 241–249. [CrossRef]
14. Salazar, A.; Vergara, L. *Independent Component Analysis (ICA): Algorithms, Applications and Ambiguities*; Nova Science Publishers: New York, NY, USA, 2018.
15. Common, P.; Jutten, C. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*; Academic Press: Cambridge, MA, USA, 2010.
16. Hyvarinen, A. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
17. Lee, T.W. *Independent Component Analysis: Theory and Applications*; Kluwer: Norwell, MA, USA, 1998.
18. Chai, R.; Naik, G.R.; Nguyen, T.N.; Ling, S.H.; Tran, Y.; Craig, A.; Nguyen, H.T. Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 715–724. [CrossRef] [PubMed]
19. Liu, H.; Liu, S.; Huang, T.; Zhang, Z.; Hu, Y.; Zhang, T. Infrared spectrum blind deconvolution algorithm via learned dictionaries and sparse representation. *Appl. Opt.* **2016**, *55*, 2813–2818. [CrossRef]
20. Naik, G.R.; Selvan, S.E.; Nguyen, H.T. Single-Channel EMG Classification with Ensemble-Empirical-Mode-Decomposition-Based ICA for Diagnosing Neuromuscular Disorders. *IEEE Trans. Neural Syst. Rehab. Eng.* **2016**, *24*, 734–743. [CrossRef] [PubMed]
21. Guo, Y.; Huang, S.; Li, Y.; Naik, G.R. Edge effect elimination in single-mixture blind source separation. *Circuits Syst. Signal Process.* **2013**, *32*, 2317–2334. [CrossRef]
22. Yuejie, Ch. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE J. Select. Top. Signal Process.* **2016**, *10*, 782–794.
23. Pendharkara, G.; Naik, G.R.; Nguyen, H.T. Using blind source separation on accelerometry data to analyze and distinguish the toe walking gait from normal gait in ITW children. *Biomed. Signal Process. Control* **2014**, *13*, 41–49. [CrossRef]
24. Guo, Y.; Naik, G.R.; Nguyen, H.T. Single channel blind source separation based local mean decomposition for biomedical applications. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, 3–7 July 2013; pp. 6812–6815.
25. Liming, W.; Chi, Y. Blind Deconvolution from Multiple Sparse Inputs. *IEEE Signal Process. Lett.* **2016**, *23*, 1384–1388.
26. Salazar, A.; Vergara, L.; Serrano, A.; Igual, J. A General Procedure for Learning Mixtures of Independent Component Analyzers. *Pattern Recognit.* **2010**, *43*, 69–85. [CrossRef]

27. Safont, G.; Salazar, A.; Vergara, L.; Gomez, E.; Villanueva, V. Probabilistic distance for mixtures of independent component analyzers. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 1161–1173. [CrossRef]
28. Salazar, A.; Igual, J.; Safont, G.; Vergara, L.; Vidal, A. Image applications of agglomerative clustering using mixtures of non-Gaussian distributions. In Proceedings of the 2015 International Conference on Computational Science and Computational Intelligence, Las Vegas, NV, USA, 7–9 December 2015; pp. 459–463.
29. Safont, G.; Salazar, A.; Rodriguez, A.; Vergara, L. On recovering missing ground penetrating radar traces by statistical interpolation methods. *Remote Sens.* **2014**, *6*, 7546–7565. [CrossRef]
30. Salazar, A.; Safont, G.; Soriano, A.; Vergara, L. Automatic credit card fraud detection based on non-linear signal processing. In Proceedings of the IEEE International Carnahan Conference on Security Technology, Boston, MA, USA, 15–18 October 2012; pp. 207–212.
31. Salazar, A.; Igual, J.; Vergara, L.; Serrano, A. Learning hierarchies from ICA mixtures. In Proceedings of the IEEE International Joint Conference on Artificial Neural Networks, Orlando, FL, USA, 12–17 August 2007; pp. 2271–2276.
32. Vergara, L.; Bernabeu, P. Simple approach to nonlinear prediction. *Electron. Lett.* **2001**, *37*, 928–936. [CrossRef]
33. Celebi, E. General formula for conditional mean using higher-order statistics. *Electron. Lett.* **1997**, *33*, 2097–2099. [CrossRef]
34. Lee, T.W.; Girolami, M.; Sejnowski, T.J. Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources. *Neural Comput.* **1999**, *11*, 409–433. [CrossRef]
35. Cardoso, J.F. Blind beamforming for non-Gaussian signals. *IEE Proc. F-Radar Signal Process.* **1993**, *140*, 362–370. [CrossRef]
36. Hyvärinen, A.; Oja, E. A fast fixed-point algorithm for Independent Component Analysis. *Neural Comput.* **1997**, *9*, 1483–1492. [CrossRef]
37. Salazar, A.; Vergara, L.; Miralles, R. On including sequential dependence in ICA mixture models. *Signal Process.* **2010**, *90*, 2314–2318. [CrossRef]
38. Lang, E.W.; Tomé, A.; Keck, I.R.; Górriz-Sáez, J.; Puntonet, C. Brain connectivity analysis: A short survey. *Comput. Intell. Neurosci.* **2012**, *2012*. [CrossRef]
39. Fiedler, M. Algebraic connectivity of graphs. *Czecoslovak Math. J.* **1973**, *23*, 298–305.
40. Merris, R. Laplacian matrices of a graph: A survey. *Linear Algebra Appl.* **1994**, *197*, 143–176. [CrossRef]
41. Lake, B.; Tenenbaum, J. Discovering structure by learning sparse graph. In Proceedings of the 32nd Annual Meeting of the Cognitive Science Society CogSci 2010, Portland, OR, USA, 11–14 August 2010; pp. 778–783.
42. Dong, X.; Thanou, D.; Frossard, P.; Vandergheynst, P. Learning Laplacian matrix in smooth graph signal representations. *IEEE Trans. Signal Process.* **2016**, *64*, 6160–6173. [CrossRef]
43. Moragues, J.; Vergara, L.; Gosálbez, J. Generalized matched subspace filter for nonindependent noise based on ICA. *IEEE Trans. Signal Process.* **2011**, *59*, 3430–3434. [CrossRef]
44. Egilmez, H.E.; Pavez, E.; Ortega, A. Graph learning from data under Laplacian and structural constraints. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 825–841. [CrossRef]