Safont Armero, G.; Salazar Afanador, A.; Vergara Domínguez, L.; Gomez, E.; Villanueva, V. (2018). Probabilistic Distance for Mixtures of Independent Component Analyzers. IEEE Transactions on Neural Networks and Learning Systems. 29(4):1161-1173. https://doi.org/10.1109/TNNLS.2017.2663843

Probabilistic Distance for Mixtures of Independent Component Analyzers

Gonzalo Safont[(1)], Addisson Salazar[(1)], Luis Vergara[(1)], Enriqueta Gómez[(2)], Vicente Villanueva[(2)]
[(1)] Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, Spain
[(2)] Hospital Universitari i Politècnic La Fe, Valencia, Spain

***Abstract*— Independent component analysis (ICA) is a blind source separation technique where data are modeled as linear combinations of several independent non-Gaussian sources. The independence and linear restrictions are relaxed using several ICA mixture models (ICAMM) obtaining a two-layer artificial neural network structure. This allows for dependence between sources of different classes, and thus a myriad of multidimensional probability density functions (PDFs) can be accurate modeled. This paper proposes a new probabilistic distance (PDI) between the parameters learned for two ICA mixture models. The PDI is computed explicitly, unlike the popular Kullback-Leibler divergence (KLD) and other similar metrics, removing the need for numerical integration. Furthermore, the PDI is symmetric and bounded within 0 and 1, which enables its use as a posterior probability in fusion approaches. In this work, the PDI is employed for change detection by measuring the distance between two ICA mixture models learned in consecutive time windows. The changes might be associated with relevant states from a process under analysis that are explicitly reflected in the learned ICAMM parameters. The proposed distance was tested in two challenging applications using simulated and real data: (i) detecting flaws in materials using ultrasounds and (ii) detecting changes in electroencephalography signals from humans performing neuropsychological tests. The results demonstrate that the PDI outperforms the KLD in change-detection capabilities.**

***Index Terms*—independent component analysis, change detection, machine learning, probabilistic distance, EEG**

I. Introduction

Independent component analysis (ICA) is one of the most successful techniques for blind source separation. ICA considers the multidimensional mixed data, $\mathbf{x}$, as a linear mixture of statistically-independent sources, $\mathbf{s}$ ($\mathbf{x}=\mathbf{A}\cdot\mathbf{s}$, where $\mathbf{A}$ is called the mixing matrix). The sources are separated by the maximization of some indicator of their independence (e.g., mutual information or likelihood function) [1]. The restriction of independence of the sources of ICA is relaxed in ICA mixture models (ICAMM, first proposed in [2]), where sources of different classes need not be independent, but sources of the same class are still independent. In ICAMM, $\mathbf{x}_k=\mathbf{A}_k\cdot\mathbf{s}_k+\mathbf{b}_k$, where the sub-index $k=1...K$, denotes the k-th class of the mixture of K classes, and $\mathbf{b}_k$ is a bias term. Essentially, $\mathbf{b}_k$ determines the location of the cluster and $\mathbf{A}_k$ determines its shape. Both ICA and ICAMM have been successfully applied in a wide range of applications, from communications to bio-signal processing (see for instance [1, 3] and the references therein).

ICAMM can be interpreted in terms of a two-layer artificial neural network (ANN) whose output is an estimate of the probability density of the inputs, $p\left(\mathbf{x}|\boldsymbol{\theta}^{(i)}\right)$, where $\boldsymbol{\theta}^{(i)}$ is the set of parameters of the *i*-th probability density model. From this perspective, other ANN structures could be learned to estimate $p\left(\mathbf{x}|\boldsymbol{\theta}^{(i)}\right)$. However, this would require the definition of application-specific cost functions [4], and the distance between $p\left(\mathbf{x}|\boldsymbol{\theta}^{(1)}\right)$ and $p\left(\mathbf{x}|\boldsymbol{\theta}^{(2)}\right)$ cannot, in general, be estimated directly from the learned ANN parameters, (see Section III.A).

The degrees of freedom provided by ICAMM make it suitable to accurately model a myriad of multidimensional probability density functions, considering any kind of underlying distributions and arbitrary relationships among explicit and hidden variables. ICAMM can be used to attempt physical interpretation of their estimated parameters (sources, mixing matrices, and bias terms) or as a general-purpose data mining technique. Thus, a measure that allows two ICAMMs to be distinguished will be a powerful tool for many applications such as pattern recognition, clustering, and model selection [5]. An immediate use of this measure is change detection problems, where detection could be done by estimating whether or not the model changes by using this measure for successive models in time. ANNs have been used for change detection [6, 7] and outlier detection in previous works [8, 9].

This paper proposes a probabilistic distance (PDI) between the parameters learned for two ICA mixture models. To the best of our knowledge, there are no references to this kind of distance in the literature. There are three significant differences between

This work was supported by Spanish Administration and European Union under grant TEC2014-58438-R, and *Generalitat Valenciana* under grants PROMETEO II/2014/032 and GV/2014/034.

G. Safont, A. Salazar, and L. Vergara are with Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, Spain (e-mails: gonsaar@upvnet.upv.es, asalazar@dcom.upv.es, and lvergara@dcom.upv.es). E. Gómez and V. Villanueva are with Hospital Universitari i Politècnic La Fe, Valencia, Spain (e-mails: gomez_enrsiu@gva.es, vevillanuevah@yahoo.es).

the proposed distance and the KLD and other similar metrics: (i) PDI does not require numerical integration or simulated data; (ii) it is symmetric and bounded within 0 and 1, thus it can be used as a score or posterior probability for classification problems, facilitating late fusion with other scores [10]; and (iii) PDI is based on ICAMM, which is a very flexible technique that allows for meaningful subspace identification, discernible non-Gaussian source identification, and robustness with respect to noise and weak signals.

This work compares the PDI with the Kullback-Leibler divergence (KLD), one of the most classic distances between two probability densities. The KLD is quite popular in pattern recognition applications, and even more so in speech and image processing (e.g., [11]). Unlike the PDI, the KLD is theoretically not model-based. However, in practice, the probabilities compared with KLD have to be estimated using some sort of model or estimation method, i.e., assuming a probability density model or being limited by the estimation method (e.g., [12] estimates probability densities from user profiles using histograms).

The proposed PDI distance can be employed in several applications, since it enables to detect changes in the data generating process through differences in the multidimensional probability density function of data models. Taking advantage of this characteristic, the changes might be associated with relevant states from a process under analysis, which is usually called change detection [13, 14]. Depending on the application, the data generation of the process can be fixed (dataset) or continuous (datastream) on time. The latter might involve dynamic analysis of the evolution of unlabeled data generated from the process to adaptively changing the model, which is a current challenge in the framework of big data [15]. In essence, the proposed strategy is based on a two-step procedure. First, the parameters of the model are to be estimated from the observed data, and then the detection test is implemented. This is the idea behind well-known methods like the generalized likelihood ratio test (GLRT), which has been proven to yield good results in general, in spite of its strictly sub-optimal condition [16].

In order to demonstrate the performance of the proposed distance, two applications using simulated and real data were approached. The first one consists of detecting flaws in materials using non-destructive testing by ultrasounds and the second one is detecting changes in electroencephalographic (EEG) signals from humans performing neuropsychological tests. The results demonstrate that the PDI outperforms KLD and symmetric KLD in change detection capabilities.

A. Independent component analysis mixture model

Given its importance to the proposed method, the basics of ICA mixture models are reviewed here. Briefly, independent component analysis ([1]) is a blind source separation technique that assumes that the observations, $\mathbf{x}$, can be modeled as an instantaneous linear mixture of a set of independent sources, $\mathbf{s}$:

$$\mathbf{x} = \mathbf{A} \cdot \mathbf{s} \tag{1}$$

where the mixing matrix, $\mathbf{A}$, is of size $[L \times M]$, where L is the number of variables of each observation and M is the number of sources. We will assume for simplicity that there are as many observed variables as sources ($L = M$), and that the mixing matrix can be inverted to find the demixing matrix, $\mathbf{W} = \mathbf{A}^{-1}$. Individual sources can be extracted as $s_m = \mathbf{w}_m{}^T \cdot \mathbf{x}$, $m = 1...M$, and $\mathbf{w}_m{}^T$ is the m-th row of the demixing matrix.

One of the advantages of ICA is that the multivariate probability density function of the observations can be obtained as a product of one-dimensional densities:

$$p(\mathbf{x}) = \frac{1}{|\det \mathbf{A}|} p\left(\mathbf{A}^{-1}\mathbf{x}\right) = \det(\mathbf{W}) \cdot \prod_{m=1}^{M} p(s_m) \tag{2}$$

Thus, modeling the joint probability density $p(\mathbf{x})$ is decomposed into two parts: modeling the marginal densities $p(s_1) \cdot ... \cdot p(s_M)$, which implies M univariate models, and estimating the mixing matrix $\mathbf{A}$, which is responsible for the statistical dependence among the elements of $\mathbf{x}$. As long as the marginals are non-Gaussian, $\mathbf{x}$ will be multivariate non-Gaussian. There is a wide range of real multivariate non-Gaussian probability densities which can be modeled by adapting the marginal distributions and the mixing matrix, such as EEG data (see [17] and references therein).

The use of ICA mixture models was first proposed in [2], which considered a source model switching between Laplacian and bimodal densities. In ICAMM, it is assumed that the data are separated in mutually-exclusive classes, and each class is modeled using an ICA. The general expression of ICAMM requires some bias vectors to separate the components of the mixture, and thus (1) becomes

$$\mathbf{x} = \mathbf{A}_k \cdot \mathbf{s}_k + \mathbf{b}_k \tag{3}$$

where $\mathbf{x}$ belongs to class k, denoted by C_k; $\mathbf{A}_k$ and $\mathbf{s}_k$ are respectively the mixing matrix and the sources of the ICA model

of class *k*; and $\mathbf{b}_k$ is the corresponding bias vector. Essentially, $\mathbf{b}_k$ determines the location of the cluster and $\mathbf{A}_k$ determines its shape.

Given the extension to the mixture model, (2) becomes

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(C_k) \cdot \left|\det \mathbf{W}_k\right| \cdot p(\mathbf{s}_k) \tag{4}$$

where $P(C_k)$ is the prior probability of class *k* and $\mathbf{W}_k = \mathbf{A}_k^{-1}$.

Mixture models emanate in a natural manner in the framework of classification/segmentation methods. If the data can be categorized into several mutually exclusive classes, where each class is characterized by a given multivariate probability density, the whole multivariate probability density of the data can be considered a mixture of the class-conditioned probability densities.

II. PROBABILISTIC DISTANCE FOR ICAMM (PDI)

A. Kullback-Leibler divergence in the detection context

The difference between two models can be estimated from the distance between their probability density functions (PDFs). One of the most common ways to measure this distance between PDFs is the Kullback-Leibler divergence [18]. Strictly speaking, the KLD is not a distance metric but an asymmetric measure of information gain or relative entropy. That is why the KLD has been used in coding problems in information theory. However, the KLD is also frequently used in detection problems (e.g., [19, 20] and references therein), sometimes in a symmetrized form.

Legitimation to use the KLD in the detection context can be obtained from optimum detection theory. Let us consider a data matrix of observed vectors $\mathbf{X} = [\mathbf{x}_1 ... \mathbf{x}_N]$. Based on these vectors, the optimum detector between two hypotheses H_1 and H_2 is given by the log-likelihood ratio (LLR) test:

$$LLR(\mathbf{X}) = ln \frac{p(\mathbf{X} \mid H_1)}{p(\mathbf{X} \mid H_2)} \underset{H_2}{\overset{H_1}{\gtrless}} ln\,\lambda \tag{5}$$

where $p(\mathbf{X} \mid H_i)$ is the probability density of $\mathbf{X}$ conditioned to hypothesis H_i, and λ is a threshold selected to satisfy some performance criterion. In general, computing $p(\mathbf{X} \mid H_i)$ is intractable due to the curse of dimensionality. In order to simplify this calculation, it is usually assumed that the observed vectors $\mathbf{x}_n$, $n = 1...N$, are statistically independent, so that $p(\mathbf{X} \mid H_i) = p(\mathbf{x}_1 \mid H_i) \cdot p(\mathbf{x}_2 \mid H_i) ... p(\mathbf{x}_N \mid H_i)$ and we can write the test in the form

$$\frac{1}{N} \sum_{n=1}^{N} ln \frac{p(\mathbf{x}_n \mid H_1)}{p(\mathbf{x}_n \mid H_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{1}{N} ln\,\lambda \tag{6}$$

where the left side of (6) is a sample mean of the individual LLR corresponding to every observation $\mathbf{x}_n$. Theoretically, (6) leads to the optimum detector. In practice, however, any model mismatch due to presence of noise, statistical dependence, or approximation errors in the estimation of $p(\mathbf{X} \mid H_i)$, may cause suboptimal performance. For example, if $p(\mathbf{x}_n \mid H_1) \approx 0$ for some $\mathbf{x}_n$, then $ln\,\Lambda(\mathbf{X}) \approx -\infty$ independently of the individual LLR yielded by the rest of the observations. Therefore, it is better to use a sample estimate of the expected value of the individual LLR relative to the distribution $p(\mathbf{x} \mid H_1)$, where the contribution to the LLR corresponding to $\mathbf{x}_n$ is made proportional to $p(\mathbf{x} \mid H_1)$:

$$\sum_{n=1}^{N} p(\mathbf{x}_n \mid H_1) ln \frac{p(\mathbf{x}_n \mid H_1)}{p(\mathbf{x}_n \mid H_2)} \underset{H_2}{\overset{H_1}{\gtrless}} ln\,\lambda \tag{7}$$

Notice that the left side of (7) is a numerical integration of the KLD between $p(\mathbf{x} \mid H_1)$ and $p(\mathbf{x} \mid H_2)$, where $\mathbf{x}_n$, $n = 1...N$, are realizations of the multivariate random vector $\mathbf{x}$. Hence, KLD seems to be an adequate reference for performance comparison with other options like the one proposed in this paper.

In this work, the KLD is used for detection as follows. It is assumed that each $\mathbf{x}_n$ corresponds to a multichannel sample at discrete time n. These data are divided in multiple epochs with some overlap and a different model $\boldsymbol{\theta}^{(i)}$ is fit using the data in each epoch (in our case, an ICAMM). For detection, we compare the data and the models from two consecutive epochs. Hypothesis H_1 is that the data are more likely to have been produced by $\boldsymbol{\theta}^{(1)}$, the model from the earlier epoch; conversely, hypothesis H_2 is that the data are more likely to have been produced by $\boldsymbol{\theta}^{(2)}$, the model from the later epoch. Thus, in the following we will identify $p\left(\mathbf{x}_n \mid H_i\right) \equiv p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(i)}\right)$.

As we will see in the experimental sections (IV and V), three variations of the KLD will be considered to compare with the proposed distance, namely

$$\begin{aligned} KLD_{12} &= \sum_{n=1}^{N} p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(1)}\right) ln \frac{p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(1)}\right)}{p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(2)}\right)} \\ KLD_{21} &= \sum_{n=1}^{N} p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(2)}\right) ln \frac{p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(2)}\right)}{p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(1)}\right)} \\ KLD_{sym} &= \frac{KLD_{12} + KLD_{21}}{2} \end{aligned} \tag{8}$$

The detection obtained from KLD_{12} or KLD_{21} is different from that obtained from KLD_{sym}. If KLD_{12} rises above the threshold, it can be said that model $\boldsymbol{\theta}^{(1)}$ is more likely to have produced the data than model $\boldsymbol{\theta}^{(2)}$, but there is no information on the likelihood of $\boldsymbol{\theta}^{(2)}$. On the other hand, if KLD_{sym} rises above the threshold, then it can be said that both models yield different PDFs, thus improving detection. The symmetrized KLD was first proposed in [18], but it remains in use until today (see for instance [19-21]).

To avoid overfitting, the computation of the values $p\left(\mathbf{x}_n \mid \boldsymbol{\theta}^{(i)}\right)$ required in (8) should use a different set of observations than the estimation of the ICAMM parameters $\boldsymbol{\theta}^{(i)}$. In practice, this would reduce the amount of data available for estimation, which could decrease performance. Another option is the generation of sample replicas once we have the ICAMM parameters. These problems can be avoided by using the distance proposed in the following section, as we obtain a closed form expression which allows direct computation of the proposed distance from the estimated ICAMM parameters.

B. Proposed distance

In this paper, we propose the use of a normalized quadratic distance (named "discordance" in [22]) between the PDFs of both models. In its general form, the discordance is:

$$\begin{aligned} D\left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right) &= \frac{\int \left(p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}\right) - p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(2)}\right)\right)^2 \cdot d\mathbf{x}}{\int \left(p^2\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}\right) + p^2\left(\mathbf{x} \mid \boldsymbol{\theta}^{(2)}\right)\right) \cdot d\mathbf{x}} = \\ &= 1 - \frac{2 \cdot \int p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}\right) \cdot p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(2)}\right) \cdot d\mathbf{x}}{\int p^2\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}\right) \cdot d\mathbf{x} + \int p^2\left(\mathbf{x} \mid \boldsymbol{\theta}^{(2)}\right) \cdot d\mathbf{x}} \end{aligned} \tag{9}$$

where $\boldsymbol{\theta}^{(i)}$ is the set of parameters of the $i-$th probability density (model). Unlike the Kullback-Leibler divergence, this distance is symmetric and bounded between 0 and 1 (with it being 0 if and only if $p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}\right) = p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(2)}\right)$). These limits allow this distance to be used as a sort of score or probability that both models are dissimilar. With this interpretation, the results of the probabilistic distance could be combined using soft fusion with the results of other methods, or those of a classifier (e.g., [23, 24]).

The distance in (9) cannot be calculated explicitly in its general form. We will demonstrate here, however, that a closed-form solution can be found if we assume that both models are ICA mixture models. In that case, the PDF of observation $\mathbf{x}$ under the i-th model is

$$p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(i)}\right) = \sum_{k=1}^{K^{(i)}} P\left(C_k^{(i)}\right) \cdot \left|\det \mathbf{W}_k^{(i)}\right| \cdot p\left(\mathbf{s}_k^{(i)} \mid \boldsymbol{\theta}^{(i)}\right) \tag{10}$$

where $\mathbf{s}_k^{(i)}$, $\mathbf{W}_k^{(i)}$, $\mathbf{b}_k^{(i)}$ are, respectively, the sources, demixing matrices and bias terms of the ICA for class $C_k^{(i)}$, the k-th class in model i, which has $K^{(i)}$ classes and $M^{(i)}$ sources; $P\left(C_k^{(i)}\right)$ is the prior probability that $\mathbf{x}$ belongs to class k in model i; and $\boldsymbol{\theta}^{(i)}$ is the set of ICAMM parameters for model i: $\mathbf{W}_k^{(i)}$, $\mathbf{b}_k^{(i)}$, $p\left(\mathbf{s}_k^{(i)}\right)$, $P\left(C_k^{(i)}\right)$, $k=1...K^{(i)}$. The density in (10) requires that the mixing matrices be square (i.e., both models have as many sources as variables are in $\mathbf{x}$) and thus $M^{(1)}=M^{(2)}=M$. Considering (10), the numerator in (9) can be expressed as

$$\begin{aligned}
&\int p\left(\mathbf{x}\mid\boldsymbol{\theta}^{(1)}\right)\cdot p\left(\mathbf{x}\mid\boldsymbol{\theta}^{(2)}\right)\cdot d\mathbf{x}= \\
&=\int\sum_{k=1}^{K^{(1)}} p\left(C_k^{(1)}\right)\cdot\left|\det\mathbf{W}_k^{(1)}\right|\cdot p\left(\mathbf{s}_k^{(1)}\mid\boldsymbol{\theta}^{(1)}\right)\cdot \\
&\cdot\sum_{l=1}^{K^{(2)}} p\left(C_l^{(2)}\right)\cdot\left|\det\mathbf{W}_l^{(2)}\right|\cdot p\left(\mathbf{s}_l^{(2)}\mid\boldsymbol{\theta}^{(2)}\right)\cdot d\mathbf{x}= \\
&=\sum_{k=1}^{K^{(1)}}\sum_{l=1}^{K^{(2)}} p\left(C_k^{(1)}\right)\cdot p\left(C_l^{(2)}\right)\cdot\left|\det\mathbf{W}_k^{(1)}\right|\cdot\left|\det\mathbf{W}_l^{(2)}\right|\cdot \\
&\cdot\int p\left(\mathbf{s}_k^{(1)}\mid\boldsymbol{\theta}^{(1)}\right)\cdot p\left(\mathbf{s}_l^{(2)}\mid\boldsymbol{\theta}^{(2)}\right)\cdot d\mathbf{x}
\end{aligned} \tag{11}$$

Due to the properties of ICA and ICAMM, the probability sources of the same class are independent. Thus, $p\left(\mathbf{s}_k^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)$ can be estimated as the product of the marginal probability densities of each source, $p\left(\mathbf{s}_k^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)=\prod_{m=1}^{M} p\left(s_{k,m}^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)$, where $p\left(s_{k,m}^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)$ is the marginal probability density of the m-th source of the k-th class of model i.

These marginal densities are unknown in practice, and thus, they are usually learned using training data. To evaluate the PDF of the sources, $p\left(\mathbf{s}_k^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)$, a model of some sort is required. It has been shown that incorrect hypotheses on these densities (while following a parametric approach) can yield poor estimation performance, even complete failure, during source separation [25]. To tackle this issue, we used non-parametric kernel density estimation (KDE). This estimator is asymptotically unbiased and efficient, it has been shown to be a true density (i.e., it is non-negative and integrates to 1) and to converge to the true PDF under several conditions [26].

The choice of kernel for KDE has been shown to have little effect on the effectiveness of non-parametric estimation [26]. In this work, we have selected the Normal kernel due to its desirable properties, which enable to obtain closed analytical solutions, as we have done in equations (13) and (18). Bandwidth selection, which controls the functional, is contingent on the type of kernel. We have set the bandwidth to $h_{k,m}^{(i)}\approx 1.06\cdot\sigma_{k,m}^{(i)}\cdot\left(N_k^{(i)}\right)^{-1/5}$, the optimal value for Normal kernels (see [26, 27]).

Let us assume that the marginal densities, $p\left(s_{k,m}^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)$, are estimated by a KDE with a Normal kernel. In that case,

$$\begin{aligned}
&p\left(s_{k,m}^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)=\frac{1}{N_k^{(i)}}\cdot\left(2\pi\cdot\left(h_{k,m}^{(i)}\right)^2\right)^{-1/2}\cdot \\
&\cdot\sum_{n=1}^{N^{(i)}}\exp\left\{-\frac{1}{2}\frac{\left(s_{k,m}^{(i)}-\tau_{k,m}^{(i)}(n)\right)^2}{\left(h_{k,m}^{(i)}\right)^2}\right\}
\end{aligned} \tag{12}$$

where $h_{k,m}^{(i)}$ is the estimated bandwidth of the non-parametric KDE for source $s_{k,m}^{(i)}$, and $\tau_{k,m}^{(i)}(n)$ is the n-th training value available for that source and class, $n=1,...,N_k^{(i)}$. The parameters $\boldsymbol{\theta}^{(i)}$ are extended to include the bandwidths. The training values $\tau_{k,m}^{(i)}(n)$ are the sources extracted during the estimation of model i, as $\boldsymbol{\tau}_k^{(i)}(n)=\mathbf{W}_k^{(i)}\cdot\left(\boldsymbol{\xi}_k^{(i)}(n)-\mathbf{b}_k^{(i)}\right)$, where $\boldsymbol{\xi}_k^{(i)}(n)$ are the observations used to estimate the model.

Given KDE, the PDF of the sources can be found as the product of the probabilities in (12). For convenience, $p\left(\mathbf{s}_k^{(i)}\mid\boldsymbol{\theta}^{(i)}\right)$ can be expressed as a multidimensional KDE:

$$p\left(\mathbf{s}_k^{(i)} \mid \boldsymbol{\theta}^{(i)}\right) = (2\pi)^{-M/2} \cdot \left(N_k^{(i)}\right)^{-1} \cdot \left(\left|\det \mathbf{H}_k^{(i)}\right|\right)^{-1/2} \cdot \sum_{n=1}^{N_k^{(i)}} e^{-\frac{1}{2}\left(\mathbf{s}_k^{(i)} - \boldsymbol{\tau}_k^{(i)}(n)\right)^T \cdot \left(\mathbf{H}_k^{(i)}\right)^{-1} \cdot \left(\mathbf{s}_k^{(i)} - \boldsymbol{\tau}_k^{(i)}(n)\right)} \tag{13}$$

where $\mathbf{H}_k^{(i)} = \text{diag}\left(h_{k,1}^{(i)}, h_{k,2}^{(i)}, ..., h_{k,M^{(i)}}^{(i)}\right)^2$ is a diagonal matrix and $\boldsymbol{\tau}_k^{(i)}(n) = \left[\tau_{k,1}^{(i)}(n), \tau_{k,2}^{(i)}(n), ..., \tau_{k,M^{(i)}}^{(i)}(n)\right]^T$ are the sources extracted from the training data. Replacing (13) into (11) yields

$$\begin{aligned} &\int p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}\right) \cdot p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(2)}\right) \cdot d\mathbf{x} = \\ &= \sum_{k=1}^{K^{(1)}} \sum_{l=1}^{K^{(2)}} (2\pi)^{-M} \cdot a_k^{(1)} \cdot a_l^{(2)} \cdot \sum_{n=1}^{N_k^{(1)}} \sum_{n'=1}^{N_l^{(2)}} \int e^{-\frac{1}{2} u_{kl}(n,n')} \cdot d\mathbf{x} \end{aligned} \tag{14}$$

where $a_k^{(i)} = p\left(C_k^{(i)}\right) \cdot \left|\det \mathbf{W}_k^{(i)}\right| \cdot \left(N_k^{(i)}\right)^{-1} \cdot \left(\left|\det \mathbf{H}_k^{(i)}\right|\right)^{-1/2}$ and

$$\begin{aligned} u_{kl}(n,n') &= \left(\mathbf{s}_k^{(1)} - \boldsymbol{\tau}_k^{(1)}(n)\right)^T \cdot \left(\mathbf{H}_k^{(1)}\right)^{-1} \cdot \left(\mathbf{s}_k^{(1)} - \boldsymbol{\tau}_k^{(1)}(n)\right) + \\ &+ \left(\mathbf{s}_l^{(2)} - \boldsymbol{\tau}_l^{(2)}(n)\right)^T \cdot \left(\mathbf{H}_l^{(2)}\right)^{-1} \cdot \left(\mathbf{s}_l^{(2)} - \boldsymbol{\tau}_l^{(2)}(n')\right) \end{aligned} \tag{15}$$

The sources for the current observation can be estimated using the ICAMM parameters as $\mathbf{s}_k^{(i)} = \mathbf{W}_k^{(i)} \cdot \left(\mathbf{x} - \mathbf{b}_k^{(i)}\right)$. With some algebraic manipulations, $u_{kl}(n,n')$ can be split into one term that depends on $\mathbf{x}$ and several terms that do not:

$$\begin{aligned} u_{kl}(n,n') &= \left(\boldsymbol{\xi}_k^{(1)}(n)\right)^T \cdot \mathbf{D}_k^{(1)} \cdot \boldsymbol{\xi}_k^{(1)}(n) + \\ &\left(\boldsymbol{\xi}_l^{(2)}(n')\right)^T \cdot \mathbf{D}_l^{(2)} \cdot \boldsymbol{\xi}_l^{(2)}(n') - \left(\mathbf{f}_{k,l}^{(1,2)}\right)^T \cdot \left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1} \cdot \mathbf{f}_{k,l}^{(1,2)} + \\ &+ \left(\mathbf{x} - \left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1} \cdot \mathbf{f}_{k,l}^{(1,2)}\right)^T \cdot \mathbf{E}_{k,l}^{(1,2)} \cdot \left(\mathbf{x} - \left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1} \cdot \mathbf{f}_{k,l}^{(1,2)}\right) \end{aligned} \tag{16}$$

where $\mathbf{f}_{k,l}^{(1,2)} = \mathbf{D}_k^{(1)} \cdot \boldsymbol{\xi}_k^{(1)}(n) + \mathbf{D}_l^{(2)} \cdot \boldsymbol{\xi}_l^{(2)}(n')$ and $\mathbf{D}_k^{(i)} = \left(\mathbf{W}_k^{(i)}\right)^T \cdot \left(\mathbf{H}_k^{(i)}\right)^{-1} \cdot \mathbf{W}_k^{(i)}$. Only the last term in (16) depends on $\mathbf{x}$. This last term is part of the exponent of a multidimensional Normal distribution with covariance matrix $\left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1}$ and mean $\left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1} \cdot \mathbf{f}_{k,l}^{(1,2)}$. Thus, one can solve the integral in (14) as

$$\begin{aligned} &\int e^{-\frac{1}{2} u_{kl}(n,n')} \cdot d\mathbf{x} = e^{-\frac{1}{2}\left(\boldsymbol{\xi}_k^{(1)}(n)\right)^T \cdot \mathbf{D}_k^{(1)} \cdot \boldsymbol{\xi}_k^{(1)}(n) + \left(\boldsymbol{\xi}_l^{(2)}(n')\right)^T \cdot \mathbf{D}_l^{(2)} \cdot \boldsymbol{\xi}_l^{(2)}(n')} \cdot \\ &e^{\frac{1}{2}\left(\mathbf{f}_{k,l}^{(1,2)}\right)^T \cdot \left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1} \cdot \mathbf{f}_{k,l}^{(1,2)}} \cdot (2\pi)^{M/2} \cdot \left|\det\left(\mathbf{E}_{k,l}^{(1,2)}\right)\right|^{-1/2} \end{aligned} \tag{17}$$

The exponentials in (17) can be simplified by considering the definitions of $\left(\mathbf{f}_{k,l}^{(1,2)}\right)^T$, $\left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1}$, yielding:

$$\int p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(1)}\right)\cdot p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(2)}\right)\cdot d\mathbf{x} = \\ = \sum_{k=1}^{K^{(1)}}\sum_{l=1}^{K^{(2)}} a_k^{(1)}\cdot a_l^{(2)}\cdot(2\pi)^{-M/2}\cdot\left|\det\left(\mathbf{E}_{k,l}^{(1,2)}\right)\right|^{-1/2}\cdot \\ \cdot\sum_{n=1}^{N_k^{(1)}}\sum_{n'=1}^{N_l^{(2)}} e^{-\frac{1}{2}\left(\xi_l^{(2)}(n')-\xi_k^{(1)}(n)\right)^T\cdot\mathbf{D}_k^{(1)}\cdot\left(\mathbf{E}_{k,l}^{(1,2)}\right)^{-1}\cdot\mathbf{D}_l^{(2)}\cdot\left(\xi_l^{(2)}(n')-\xi_k^{(1)}(n)\right)} \tag{18}$$

This can be replaced back in (9) to obtain the distance indicator. The numerator of (9) can be straightforwardly calculated from (18), and each of the terms in the denominator can be obtained by replacing the adequate values in (18), since $\int p^2\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)\cdot d\mathbf{x} = \int p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)\cdot p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)\cdot d\mathbf{x}$.

We have named the proposed distance PDI (probabilistic distance between ICAMM).

As presented, the PDI measures the discordance between the probability density of the data, conditional to the ICAMM parameters ($\mathbf{W}_k, \mathbf{s}_k, \mathbf{b}_k$) of each model. Another approach would be to calculate the discordance between the PDF of the extracted sources instead, $p\left(\mathbf{s}_k^{(i)}\,|\,\boldsymbol{\theta}^{(i)}\right)$, or even a discordance that does not assume any model. These two alternatives to the proposed PDI, which we have denoted by PDI of the sources (PDI_S) and non-parametric discordance (NPD), respectively, are developed in the Appendices. Their results are compared with those of PDI in Sections IV and V of this work.

III. On the relation of PDI with other methods

A. *Relation between ICAMM and artificial neural networks*

Notice from equation (10) that ICAMM can be interpreted in terms of a two-layer artificial neural network (ANN) that, having the observed vector $\mathbf{x}$ at the input, produces an estimate of $p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)$ at the output (see Fig. 1 for $K = 2$ and $M = 2$). At the first layer, the input elements $\mathbf{x} = \left[x_1 \ldots x_M\right]^T$ are biased and linearly combined to yield the node elements $\mathbf{s}_k^{(i)} = \left[s_{k1}^{(i)} \ldots s_{kM}^{(i)}\right]^T$, $\mathbf{s}_k^{(i)} = \mathbf{W}_k^{(i)}\cdot\left(\mathbf{x}-\mathbf{b}_k^{(i)}\right)$, $k = 1 \ldots K^{(i)}$. Then, every node element $s_{km}^{(i)}$ is nonlinearly transformed by $p\left(s_{km}^{(i)}\,|\,\boldsymbol{\theta}^{(i)}\right)$, and the value $p\left(\mathbf{s}_k^{(i)}\,|\,\boldsymbol{\theta}^{(i)}\right) = \prod_{m=1}^{M} p\left(s_{k,m}^{(i)}\,|\,\boldsymbol{\theta}^{(i)}\right)$ is obtained. Finally, at the second layer, $p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)$ is computed as a weighted sum of $p\left(\mathbf{s}_k^{(i)}\,|\,\boldsymbol{\theta}^{(i)}\right)$, as indicated in (10). Once this ANN structure is set by learning all the weights and non-linear (activation) functions for $i = 1$ and for $i = 2$, the proposed distance can be computed following the above indicated procedure. From this perspective, other ANN structures that could be learned in an unsupervised manner, could be candidates to estimate $p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)$. However, this would require the definition of cost functions which are very application specific (see for example [4] and references therein) and the distance between $p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(1)}\right)$ and $p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(2)}\right)$ cannot, in general, be obtained directly from the estimated ANN parameters. In contrast, ICAMM has general applicability and the distance (PDI) can be computed directly from the learned models as indicated above, without relaying to KDE methods followed by KLD numerical integration.

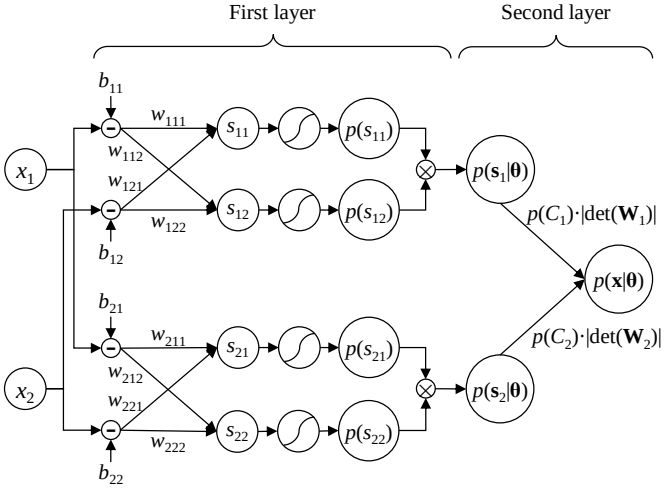

Fig. 1. Comparison between an ICA mixture model ($K = 2$ and $M = 2$) and an artificial neural network with two layers.

B. Comparison between PDI and adaptive ICA

PDI is a general-purpose distance that, basically, can be applied using two different procedures. In the first procedure, PDI is successively applied to compare a set of previous and a set of current ICAMM parameters for adaptively detecting of changes in monitoring a continuous process. The second procedure consists of comparing the model of the current subject with those in the database in applications such as Brain-Computer Interfaces (BCI) and biometric access control (subject identification/ authentication).

The applications of change detection approached in Sections IV and V employ the adaptive procedure. Note that the estimation of the ICAMM parameters is decoupled from the calculation of the PDI. Furthermore, the learned ICAMM parameters could also be used for other applications, such as classification or blind source separation.

There are some similarities between adaptive uses of PDI and adaptive ICA estimation methods such as adaptive eta ([28, 29]) or online recursive ICA [30]. Adaptive ICA methods obtain an instantaneous channel separation matrix $\mathbf{W}(t)$ that depends on the current time instant, t. The values $w_{ij}(t)$ of this mixture are calculated by optimizing some form of gradient with an adaptive learning rate [29] or forgetting factor [30]. This adaptive value could potentially be compared with a threshold in order to detect important changes in the ICA parameters and thus detecting changes in the data.

The main differences between the proposed change detection procedure based on PDI and adaptive ICA methods are the following: (i) the calculation of PDI is independent from the estimation of the ICA mixture models that are compared, while the adaptive learning rate (or forgetting factor) in adaptive ICA methods is linked with ICA parameter estimation; (ii) PDI is based on a more general modeling that involves mixture of several ICAs than adaptive ICA methods, which are based on only one ICA; (iii) the ICAMM parameters used by PDI, aside from source extraction, already include an analysis of the classification in categories of each observation of the data (see for instance [3]), while the ICA parameters obtained by adaptive ICA methods only include source extraction, without any further analysis; (iv) PDI compares ICAMM parameters that can be estimated using any ICA or ICAMM algorithm, which improves its flexibility in many applications (considering different kinds of learning, semi-blind source separation...), whereas in adaptive ICA methods the ICA parameters are always obtained using the methods in [28-30]; (v) as exposed previously, PDI could be immediately extended to other applications such as BCI or access control, while adaptive ICA methods are more limited in scope and, at any rate, their extension to these applications would require some modifications to the procedure.

IV. Experimental simulations

A. Stability analysis

An experiment with simulated data was performed in order to test the stability and accuracy of the proposed PDI distance. To this effect, several changes were introduced in an initial known ICA mixture model, and it was verified that the PDI increased as the models diverged further from the original known model. In order to compare these changes in the model, the PDI was compared with two commonly-used indicators of ICA extraction performance: the Amari index and the signal-to-distortion ratio (SDR) [31].

For each iteration of the experiment, $N = 2048$ samples were drawn from a randomly-initialized ICA mixture model with $M = 4$, $K = 2$, and sources following a K distribution [32]. This distribution is defined as

$$p(x) = \frac{2}{x}\left(\frac{L\nu x}{\mu}\right)^{\frac{L+\nu}{2}} \frac{1}{\Gamma(L)\Gamma(\nu)} K_{\nu-L}\left(2\sqrt{\frac{L\nu x}{\mu}}\right) \tag{19}$$

where μ is the mean and L, ν are shape parameters. In order to introduce some amount of differences in the model, some of the sources remained stationary with shape parameters $\mu = L = \nu = 1$, while other sources were set to $\mu = L = \nu = 1$ for the first 1024 samples and then ν was changed linearly from $\nu = 1$ to $\nu = 5$ during the last 1024 samples. The number of nonstationary sources was changed from 0 (all sources stationary) to 8 (all sources nonstationary). Nonstationary sources were distributed evenly between the classes of the model.

The simulated data were then split into epochs of length 512 samples with 90% overlap between epochs. A different ICA mixture model was fit to each epoch using the MIXCA algorithm ([38]), and PDI between models of consecutive epochs was calculated. To compare with PDI, the SDR between the extracted sources and the Amari index between the mixing matrices were also estimated. Since these two methods are used to compare single ICA, their results were averaged for all classes of the mixture. The above process was repeated for a total of 300 Monte Carlo experiments.

The results of the experiment are shown in Fig. 2. All three indicators remained more or less constant for the first half of the data, during which the data were stationary, and showed a decrease in performance during the second half of the data, during

which the data changed for some sources. In all cases, the number of affected sources directly affected the results of the indicators. The behavior of the PDI was more linear than the Amari index with respect to the number of affected sources. Furthermore, unlike both the SDR and the Amari index, the distance between models did not decrease during the latest epochs (i.e., for large values of the shape parameter, ν). Thus, the PDI proved its sensitivity with respect to changes in the compared models.

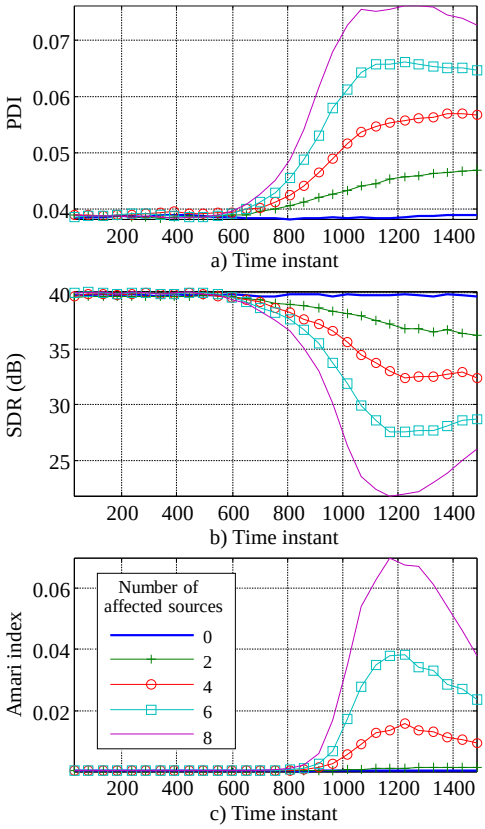

Fig. 2. Effect of changes in the data with respect to time and number of affected sources on several ICA performance indicators: a) the proposed distance; b) signal to distortion ratio (SDR); c) Amari index.

B. Experiments on simulated ultrasound data

The experimental simulations corresponded to a case of flaw detection in materials using ultrasounds (US). When inspected with US, any sufficiently large internal inhomogeneity in the material produces a reflection that is captured by the recording device. These reflections induce a change in the PDF of the recorded signals that we attempted to detect with the PDI of locally-estimated ICAMM.

ICA and ICAMM have been widely used for flaw detection in industrial processes and non-destructive testing, such as: image texture and defect detection for solar panels [33] and defects in metallic parts [34]; detection of flaws in materials using sonic signals [35, 36]; and fabric defect detection (see several of the references in [37]). In most cases, it is assumed that the contribution of the defect is effectively independent from that of the main contribution (e.g., the infrared image of the solar panel or the backscattering of the sonic signal), and thus ICA is used to detect and extract this foreign component.

In this experiment, the data consisted of ultrasound targets buried into background noise. The targets were modeled using Gaussian-modulated tones with random initial phase, i.e., $x(t) = A \cdot \sin(2\pi f_c t + \phi_0) \cdot \exp\left[-\left(2(t-\tau)/T\right)^{\alpha}\right]$, where: A is the peak amplitude; f_c, τ and T are respectively the central frequency, time center, and duration of the tone; and α is an even number that determines the shape of the envelope of the pulse. We used $f_c = 20\,\text{kHz}$, $T = 1\,\text{ms}$, $\alpha = 4$, $\tau = 20$, 100 or 150 ms, ϕ_0 randomly drawn from a uniform distribution in the range $[0, 2\pi)$, and A calculated to obtain a peak signal-to-noise ratio (PSNR) of 0 dB. The background noise was modeled by a K distribution as in (19), which can describe the statistics of the envelope of the backscattered ultrasonic echo from a scattering medium [32]. The shape parameters were set to $\mu = L = 1$ and $\nu = 10$. Finally, it was assumed that the data were sampled at 50 kHz.

For each iteration of the experiment, we generated four channels filled with background noise during 200 ms, generating a total of 1000 samples. This noise was obtained by multiplying four independent K-distributed noise channels by a [4 × 4]

mixture matrix. Then, three ultrasound targets were buried into the noise at 20, 100 and 150 ms from the start of the simulation. The targets were obtained by mixing four Gaussian-modulated tones with random initial phases, using a different mixture matrix than the one used for the background noise. Thus, the areas of the simulation with background noise corresponded to a different model than the areas with both noise and US targets. An example of one channel of the generated data is shown in Fig. 3.a. Note that the targets (areas marked in red) are hardly distinguishable from the background noise, which shows the difficulty of the problem.

In complex applications, the tuning of the parameters for adaptive analysis is very difficult. Usually, a subset of the data is selected for supervised learning of the optimal parameters considering the objective function (i.e., maximize change detection). Thus, we estimated window analysis size (epoch size) and overlap between consecutive windows using Monte Carlo experiments. The obtained optimal parameters were epoch size = 256 samples (5-ms) and overlap = 50%.

After data generation, the simulated ultrasonic signals were split into epochs. A different ICA mixture model was fit to each epoch using the MIXCA algorithm ([38]) and the distance between the models of consecutive epochs was estimated. The models were one-class ICAMM with de-mixing matrices of size $M = 4$. It was hypothesized that the estimated ICAMM would change between epochs with no target and epochs with a target, and between epochs where the target was only partially present. Thus, the distance between models would mark the start and end of the ultrasound targets.

The distance between models was measured using PDI, PDI_S, NPD, and KLD from the previous model to the current one (KLD12) and vice versa (KLD21), and symmetrized KLD (SKLD). PDI, PDI_S and NPD were implemented according to the procedure explained in Section II and Appendices, and all variants of KLD were implemented using numerical integration with 1000 points. The NPD was implemented using numerical integration with 1000 points. The above process was repeated for a total of 1000 Monte Carlo experiments.

The distances obtained for one experiment are shown in Fig. 3. Note that the samples that we want to detect (see Fig. 3.a) correspond to the samples before and after each target, i.e., the transitions between noise and target, and vice versa. It can be seen that all distances consistently increased when the current epoch contained part of a target. In numerical terms, the correlation between the transitions and PDI, PDI_S, NPD, KLD12, KLD21, and SKLD were 0.999, 0.8649, 0.964, 0.714, 0.697, and 0.936, respectively. The methods based on the discordance (PDI, PDI_S, NPD), however, obtained detections that were more consistent in amplitude than all types of KLD. This was more noticeable for the proposed PDI distance, which consistently assigned large distances (>0.5) for targets and obtained a few false alarms. To better show this, the distance values have been compared with a threshold equal to the half value between the maximum and minimum distance (in the case of PDI, the threshold was set to 0.5). These thresholds are shown in Fig. 3.b through Fig. 3.g as horizontal dashed lines; values above the threshold were colored in red. Using these thresholds, PDI detected the six transitions (corresponding to the three targets), while the other methods missed one or more transitions. Note that one-sided KLDs tend to detect only the beginning (KLD21) or the end (KLD12) of each target, whereas the SKLD increases the detection in both cases.

$\text{KLD}\big(p(x), q(x)\big)$ can be considered as the information loss by approximating $p(x)$ using $q(x)$. The results obtained during simulation seem to indicate that there are several points where p can be approximated using q, yet q cannot be approximated using p. In the context of the change detection application, p is modeled by the previous ICAMM, and q is modeled by the current ICAMM. During data generation, the targets were over-imposed on the noise, i.e., the observations that contained target information also included noise information. Thus, a model estimated using data from the target would include more information than the model estimated using only noise. This would explained the observed behavior for KLD12 and KLD21, and predict the expected performance of the symmetrized KLD, which detects both cases. Measures based on the discordance (PDI, PDI_S, and NPD) are not affected by this effect (as shown in Fig. 3) because they are symmetric.

To further showcase the results obtained by the considered distances, the experiments were repeated three additional times with different values of PSNR: -7, -5, and -3 dB. These values were selected in order to simulate difficult detection cases, where targets are difficult to distinguish from the surrounding noise. For each value of PSNR, the average receiver operating characteristic (ROC) curve was calculated. The results are shown in Fig. 4. Note that to make the results more readable, the false positive rate (FPR) axis is in logarithmic scale. Experiments were performed with larger values of PSNR, but results were similar to those with 0 dB.

PDI obtained the best ROC curves for all PSNR, particularly for low PSNR, and the non-parametric discordance obtained the second best result, being similar to that of PDI for PSNR = 0 dB. The results obtained by PDI_S were the most affected by the PSNR value, yielding the worst result for -7 dB and a result similar to that of SKLD for 0 dB. The two one-sided versions of KLD obtained similar results in all cases, but symmetrized KLD yielded better results than both of them. The difference between PDI and the other methods were more important in the region of very low FPR. This region of FPR is particularly important in many real-world applications where operative working with relative high FPR is costly (e.g., fraud detection [39]). The ROC curve of the PDI rose rapidly with FPR, yielding over 90% TPR with 1% FPR even in the case with PSNR = -3 dB.

Considering computational costs, the average time per each Monte Carlo experiment of each method was: 6.44 ms for PDI, 6.50 ms for PDI_S, 31.86 ms for non-parametric discordance, and 91.46 ms for each KLD (182.92 s for symmetrized KLD). Thus, PDI and PDI_S were calculated 28 times faster than the symmetrized KLD and five times faster than non-parametric discordance.

However, the time for ICAMM calculation should also be taken into account. Adding the average time per ICAMM calculation (13.66 ms) to PDI, PDI_S, and NPD time calculation, we have that PDI and PDI_S were calculated 10 times faster than SKLD and twice as fast as NPD. This is because the result of KLD and NPD depend on the number of points of the numerical integration, whereas PDI is analytical and thus it only has to be computed once. Computational time was calculated on a computer running Windows 7 with four 3.3 GHz cores and 16 GB RAM.

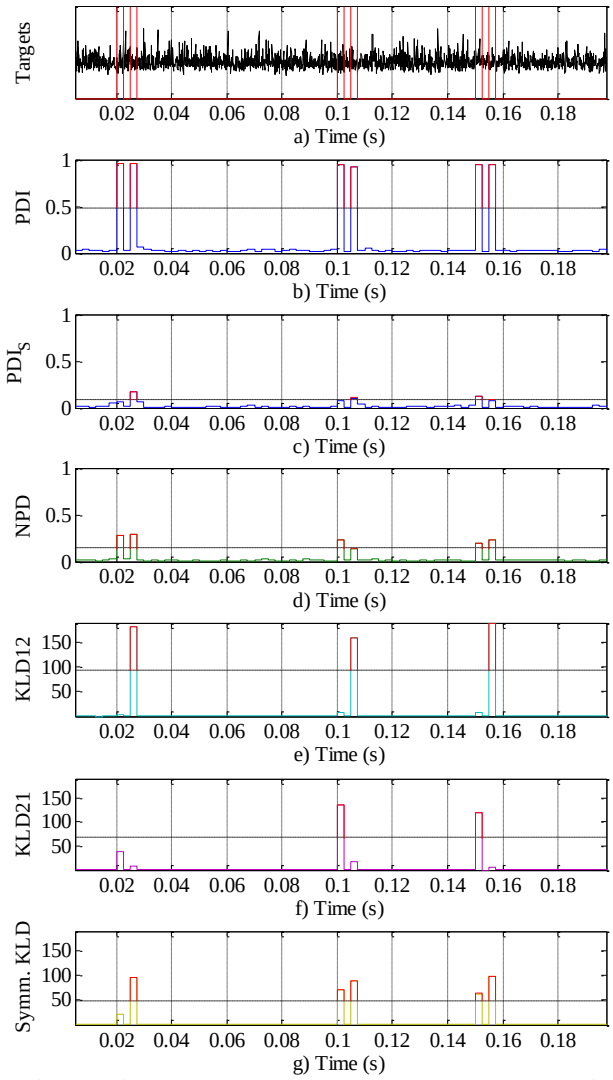

Fig. 3. Comparison between the transitions before and after each target with the distances between every pair of consecutive models for one iteration of the experiment: a) transitions before and after targets; b) PDI; c) PDI based on the sources; d) non-parametric discordance; e to f) KLD (both directions) and symmetrized KLD. Horizontal lines denote the change detection threshold for each distance.

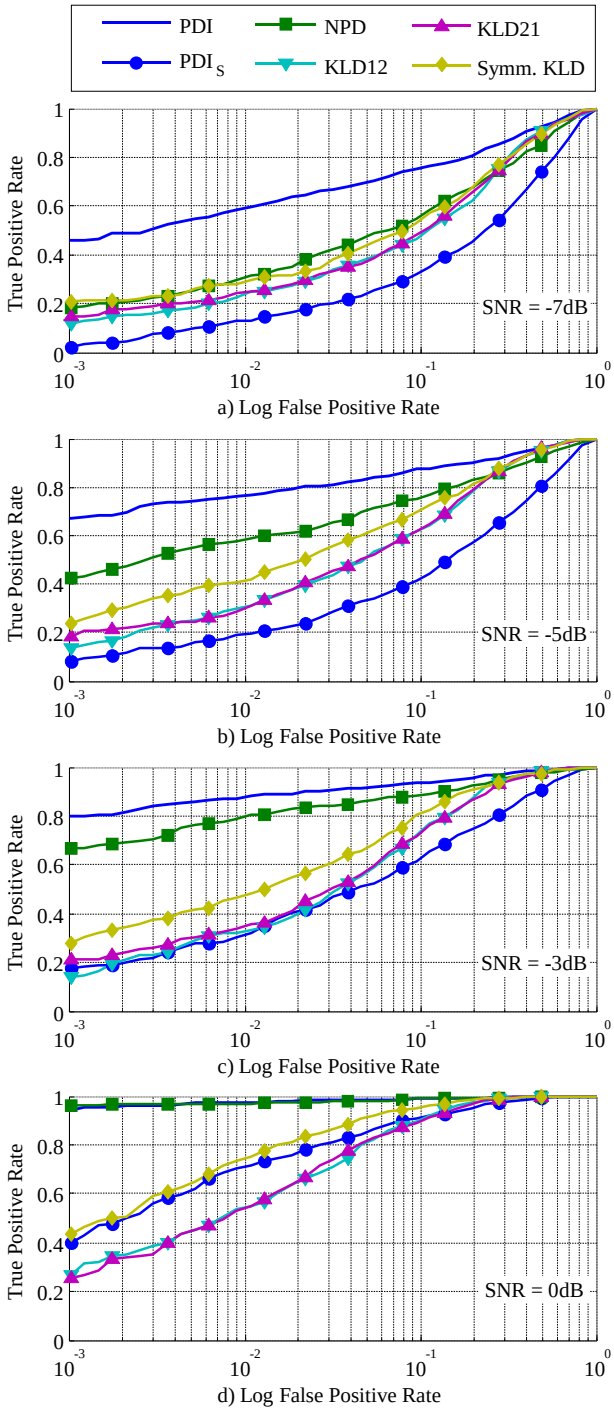

Fig. 4. Average ROC curves obtained for simulated data with different values of PSNR: a) PSNR = -7 dB; b) PSNR = -5 dB; c) PSNR = -3 dB; d) PSNR = 0 dB.

V. Application on Real Data

The proposed probabilistic distance was devised for a real-life application: automatic staging of EEG data from six epileptic patients performing neuropsychological tests. This is an essential area of clinical neurophysiology, since the evaluation of the learning and memory functions of the patient is a critical part of their neuropsychological assessment. Information cannot be processed if the brain is unable to store a certain amount of it in short-term (working) memory or to recall past experiences, events and strategies from long-term memory. Conversely, information stored in short- or long-term memory is useless without the means to properly access and activate it. The biosignals and the tests were synchronized using a graphic user interface designed by the authors, and the data were captured by the Neurology and Neurophysiology Units at Hospital Universitari i Politècnic La Fe, Valencia (Spain).

The evaluated neuropsychological tests were the following: the visual paired associates and figural memory subtests from the

Wechsler Memory Scale (WMS-R, [40]), the digit span subtest from the Wechsler Adult Intelligence Scale (WAIS, [40]), and the TAVEC ([41]). The latter is commonly considered as the Spanish version of the California Verbal Learning Test ([40]). The first two tests are based on visual stimuli, whereas the last two tests rely on auditory stimuli. All the tests can be split in two stages, stimulus display (D) and subject response (R) that are repeated in several trials changing the stimuli.

The use of an ICAMM-based distance for this application is motivated by many successful works in the literature that are based on ICA and ICAMM for processing EEG signals, see for instance [17]. We hypothesized that the learned ICAMM would change between the two stages (D and R) of the test. If the models correspond to different classes or to a transition, they should be more different than if the models belong to the same class. Thus, PDI could be used to search for increases in distance, which would mark the commuting between stages.

The EEG was composed of $M = 19$ EEG channels that were sampled at 500 Hz. These data were split into 4-seconds epochs with 90% overlap, and a different ICA mixture model with $K = 2$ classes was fit to each epoch using MIXCA. These models had de-mixing matrices of size [19 × 19]. Then, the distance between the models of every pair of consecutive epochs was estimated as explained in Section II.B.

Due to epoch size and overlap between epochs, the model does not instantly transit between classes, but there is a transition zone instead (see Fig. 5.a). This zone is related with uncertainty of the change detection algorithm, not with uncertainty of source separation. In this application we were interested in detecting changes using all the ICAMM parameters jointly (mixing matrices, sources, and bias terms) instead of an accurate estimation of the sources. The resolution of the algorithm depends on epoch size and overlap. For lasting changes, this transition zone is not critical. For sudden changes, however, the resolution becomes an important part of the method: the larger the overlap, the lower the chance to lose detection. Thus, the ideal behavior is choosing the highest possible overlap between epochs. In practice, overlap is a compromise between resolution and other limitations (e.g., time costs).

The results of the experiment of the TAVEC test for one of the subjects are shown in Fig. 5. The results of discordance-based tests are noisier than the results for KLD, but PDI and PDI_S obtained the best result. The behavior of PDI followed the changes in test stage more closely than KLD, which obtained more local peaks. Furthermore, KLD has no upper bound, whereas PDI is always constrained within [0, 1]. Thus, the values of PDI can be compared with a fixed threshold (e.g., 0.5, as shown in Fig. 5.b) in order to detect changes between stages. Conversely, KLD would have to be compared with some empirically-determined threshold value. For instance, the thresholds shown in Fig. 5.d through Fig. 5.f were half of the maximum KLD. In numerical terms, the correlation between the real changes in the stage of the test and the considered methods were: 0.870, 0.775, 0.008, 0.754, and 0.756 for PDI, PDI_S, NPD, KLD12, KLD21, and SKLD, respectively.

The detection obtained by NPD was no longer valid owing to two reasons: (i) the higher dimension of the real-data experiment, which might mean that the multivariate KDE would require a larger amount of data in order to adapt to the data; and (ii) there were greater similarities between the PDF of the data of consecutive epochs. All versions of the KLD obtained a worse result than the PDI in part for lack of a definite bound, since each possible detection obtained a different value of KLD. This makes setting the threshold a difficult prospect. Conversely, the good results of the PDI are explained by the robustness of the method, the values limited within [0, 1], and the suitability of the ICAMM to EEG data. Finally, the PDI obtained a few false alarms between 780 and 900 seconds. These alarms were likely due to the subject being distracted from the task at hand, since they were struggling to produce more words for several seconds, to no avail. However, these false alarms could be removed by a slight increase of the detection threshold.

Detection results for all tests from the same subject in Fig. 5 are summarized in Fig. 6, which shows the ROC curves for each test. Results were similar for the other subjects. The results are consistent with those in Fig. 5, with PDI and PDI_S yielding the best results, followed by SKLD, then the one-sided versions of the KLD, and finally the NPD yielded the worst result. Again, this would suggest that the multidimensional KDE used in NPD was not enough to properly capture the differences between epochs, or that the epochs are very similar.

As seen in Fig. 6, detection for visual tests (Fig. 6.c and Fig. 6.d) was slightly better than detection for auditory tests (Fig. 6.a and .b). Using the thresholds shown in Fig. 5, the average detection probabilities for auditory tests were: 94.43%, 94.16%, 2.03%, 30.65%, 33.67%, and 36.69% for PDI, PDI_S, NPD, KLD12, KLD21, and SKLD, respectively.

In comparison, the average results for visual tests were: 99.00%, 98.73%, 3.81%, 35.31%, 41.31%, and 47.98% for PDI, PDI_S, NPD, KLD12, KLD21, and SKLD, respectively. These differences could be due to the patients being more engaged by the visual tests, which would produce more pronounced changes in their mental state between stages.

With regard to computational costs, the average time per epoch of each method was: 1.53 s, 1.52 s, 0.83 s, 1.69 s, 1.69 s, and 3.37 s for PDI, PDI_S, NPD, KLD12, KLD21, and SKLD, respectively. The times for KLD and PDI, however, do not include the 6.40 s spent on average on ICAMM estimation. Thus, PDI was on average 1.23 times faster than SKLD (2.2 times without considering ICAMM), but took on average 9.60 times as long as NPD (1.86 times without considering ICAMM). Again, computational time was calculated on a computer running Windows 7 with four 3.3 GHz cores and 16 GB RAM. The reduction in the computational burden ratio between PDI and KLD compared with simulation results of Section IV was due to the increased size of the model ($M = 19$ for real data versus $M = 4$ for simulated data) and the higher epoch size (2000 versus 250 samples). Although these results are not real time, they could be brought down to real time by reducing epoch size or by

exploiting parallel computing techniques that reduce the estimation stage of the methods [42].

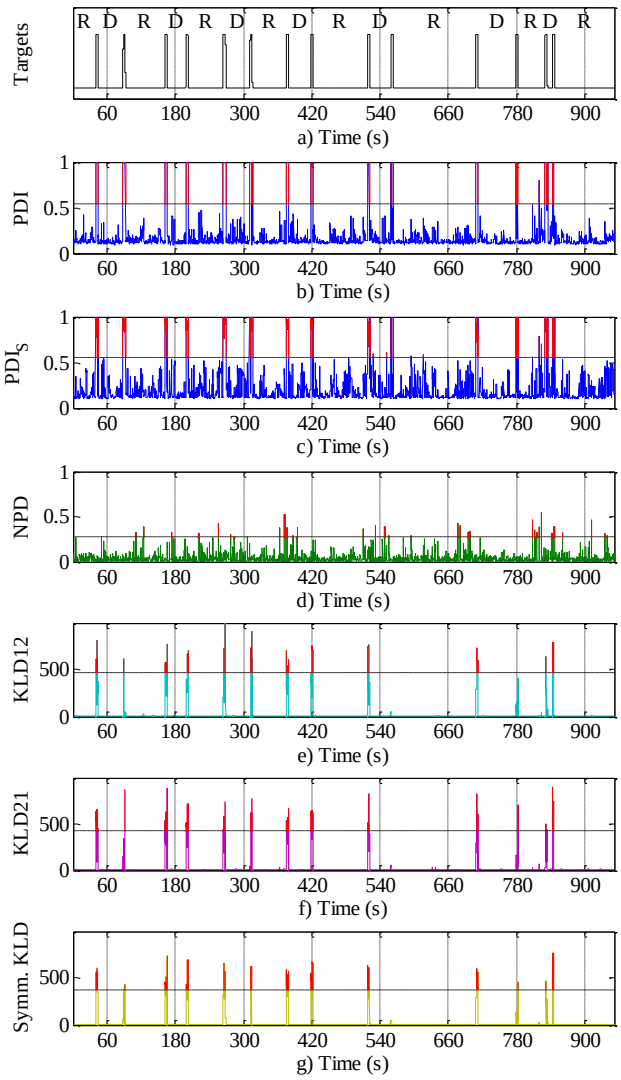

Fig. 5. Comparison between the transitions before and after each stage of the TAVEC test and the distances between every pair of consecutive models: a) transitions between stages of the TAVEC memory test; b) PDI; c) PDI_S; d) non-parametric discordance; e to g) KLD (both directions) and symmetrized KLD. The horizontal lines denote the change detection threshold for each distance.

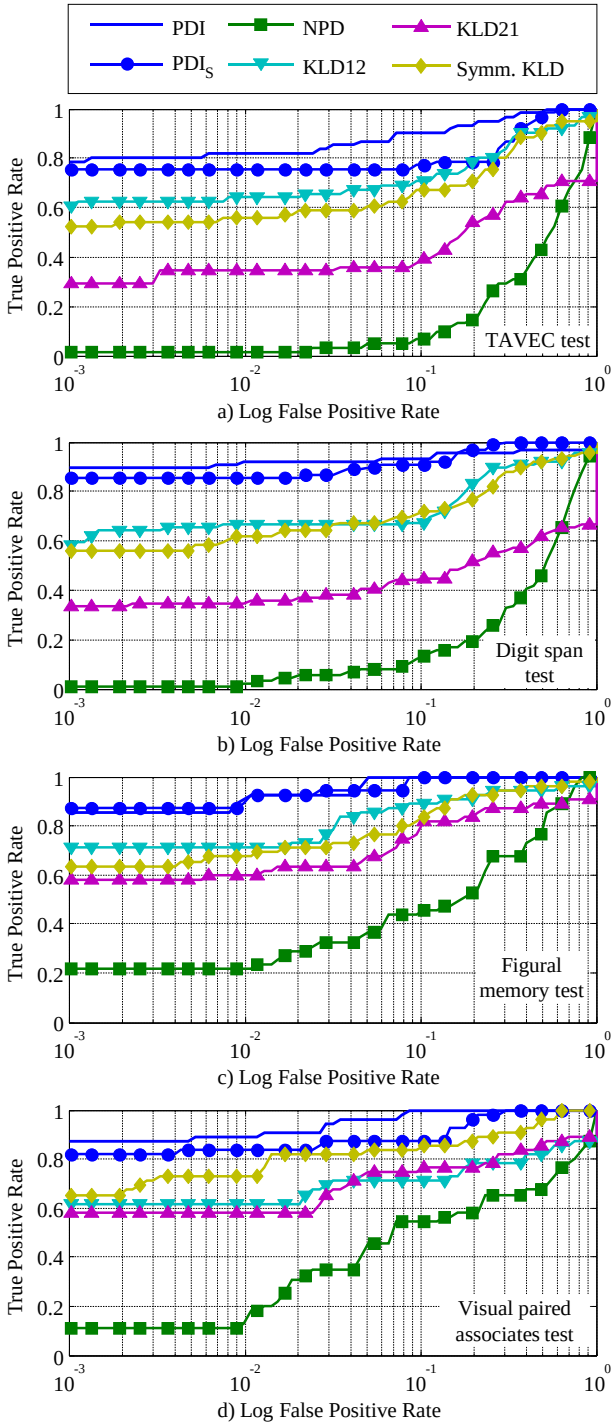

Fig. 6. ROC curves obtained for the experiment on real data for four neuropsychological tests: a) TAVEC; b) digit span; c) figural memory; d) visual paired associates.

VI. Discussion and Conclusion

A novel probabilistic distance for mixtures of independent component analyzers (ICAMM), which we have named PDI, has been proposed. Unlike similar metrics such as the Kullback-Leibler divergence (KLD), the proposed distance does not require numerical integration and also it is symmetric and bounded within 0 and 1. An experiment with simulated data compared the stability of the PDI with that of two commonly-used indicators, the signal-to-distortion ratio and Amari index. Results showed that the proposed distance was similarly sensitive to changes in the data, and the results were just as stable as those of the other methods while considering more information about the model.

PDI outperformed KLD in two challenging change-detection applications: detecting flaws in materials using ultrasounds and detecting changes in electroencephalography signals from epileptic patients performing neuropsychological tests. PDI followed

more accurately the actual changes in the underlying data distributions, even in cases of low signal to noise ratio. ROC analyses demonstrated the efficiency of PDI for detecting small changes in the underlying data generation models at very low false positive rates (FPR). Furthermore, since the PDI is bounded within [0, 1], a detection threshold was easy to set up.

The definition of PDI as a probabilistic-driven model distance entails relevant advantages compared with classic distances. After the change-detection step, the estimated parameters of the models can be used for subsequent processing steps. For instance, in the EEG application studied herein, the parameters of ICAMM might be used to analyze some brain properties like level of activation or define brain functional regions for the different stages of the neuropsychological tests. Besides, the estimated non-Gaussian sources would facilitate further analyses for clinical research.

The capability of PDI for accurate detection of changes in the stages of an EEG-based neuropsychological test suggests its application to other EEG staging problems such as the study of stages during sleep. It could also improve the performance of real-life applications where a more precise temporal location of each stage could in turn yield more precise spatial locations during subsequent studies of the extracted EEG sources. For instance, this is the case with focal epileptic seizures, where precise location can yield more information about the focus of the seizure [43]. Future works will attempt to apply the PDI on data from subjects in real time, using parallel processing techniques to improve estimation times.

The normalized range of values of PDI allows for this distance to be directly used as a score in the fusion with other detectors in complex problems where only one detector cannot give appropriate performance. This could leave to further work in fields such as credit card fraud detection, multimodal classifier fusion, or data mining. Another possible line of work is the creation of a combined method that includes both PDI and ICAMM estimation. By embedding PDI in the ICAMM estimation procedure, the method could adapt to new data (e.g., by building new classes) while performing estimation and source extraction.

APPENDICES

A. Probabilistic distance between sources

Section II.B introduced the PDI as the discordance between the probability density of the data, conditional to the ICAMM parameters ($\mathbf{A}_k, \mathbf{s}_k, \mathbf{b}_k$) of each model. Another approach would be to calculate the discordance between the PDF of the extracted sources instead, $p\left(\mathbf{s}_k^{(i)} \mid \boldsymbol{\theta}^{(i)}\right)$. There is, however, some uncertainty in this comparison: since there are multiple classes in each model, one would need to know which class from model 1 corresponds to each class in model 2. This uncertainty is further compounded by the fact that model 1 and model 2 need not have the same number of classes, so there could potentially be multiple matches. In order to consider all the possible combinations of sources for a detection problem, we define the following discordance between sources:

$$D'\left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right) = 1 - \frac{2 \cdot \sum_{k=1}^{K^{(1)}} \sum_{l=1}^{K^{(2)}} d\left(\mathbf{s}_k^{(1)}, \mathbf{s}_l^{(2)} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right)}{\sum_{k=1}^{K^{(1)}} \sum_{l=1}^{K^{(1)}} d\left(\mathbf{s}_k^{(1)}, \mathbf{s}_l^{(1)} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(1)}\right) + \sum_{k=1}^{K^{(2)}} \sum_{l=1}^{K^{(2)}} d\left(\mathbf{s}_k^{(2)}, \mathbf{s}_l^{(2)} \mid \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(2)}\right)} \tag{A.1.a}$$

$$d\left(\mathbf{s}_k^{(i)}, \mathbf{s}_l^{(j)} \mid \boldsymbol{\theta}^{(\mathrm{i})}, \boldsymbol{\theta}^{(\mathrm{j})}\right) = \int p\left(\mathbf{s}_k^{(i)} \mid \boldsymbol{\theta}^{(i)}\right) \cdot p\left(\mathbf{s}_l^{(j)} \mid \boldsymbol{\theta}^{(j)}\right) \cdot d\mathbf{s} \tag{A.1.b}$$

which is similar to the PDI, but considering only the probability density of the sources. Using the same development outlined in Section II.B, it can be seen that

$$\begin{aligned} & d\left(\mathbf{s}_k^{(i)}, \mathbf{s}_l^{(j)} \mid \boldsymbol{\theta}^{(\mathrm{i})}, \boldsymbol{\theta}^{(\mathrm{j})}\right) = (2\pi)^{-M/2} \cdot \\ & \cdot \left(N^{(1)}\right)^{-M} \cdot \left(\left|\det \mathbf{H}_k^{(1)}\right|\right)^{-1/2} \cdot \left(N^{(2)}\right)^{-M} \cdot \left(\left|\det \mathbf{H}_l^{(2)}\right|\right)^{-1/2} \\ & \cdot \left|\det\left(\left(\mathbf{H}_k^{(1)}\right)^{-1} + \left(\mathbf{H}_l^{(2)}\right)^{-1}\right)\right|^{-1/2} \cdot \\ & \cdot \sum_{n=1}^{N^{(1)}} \sum_{n'=1}^{N^{(2)}} e^{-\frac{1}{2}\left(\tau_l^{(2)}(n') - \tau_k^{(1)}(n)\right)^T \cdot \left(\mathbf{H}_k^{(1)} + \mathbf{H}_l^{(2)}\right)^{-1} \cdot \left(\tau_l^{(2)}(n') - \tau_k^{(1)}(n)\right)} \end{aligned} \tag{A.2}$$

As per the PDI, the solution to (A.2) can be plugged into (A.1) to obtain the final discordance between sources. This distance has also been considered in the following, denoted by $\mathrm{PDI_S}$.

B. Non-parametric discordance

The PDI distances developed in Section II.B and Appendix A assume an ICA mixture model. In order to test the performance of this parametric model, the PDI was compared with a non-parametric version of the discordance in (9). This is done by estimating the PDF of the data, $p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)$, using multivariate non-parametric kernel density estimation [26]. Assuming a multivariate Normal kernel,

$$p\left(\mathbf{x}\,|\,\boldsymbol{\theta}^{(i)}\right)=\left(2\pi\right)^{-M/2}\cdot\left(N^{(i)}\right)^{-1}\cdot\left(\left|\det\mathbf{H}^{(i)}\right|\right)^{-1/2}\cdot \sum_{n=1}^{N^{(i)}}\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\xi}^{(i)}\left(n\right)\right)^{T}\cdot\left(\mathbf{H}^{(i)}\right)^{-1}\cdot\left(\mathbf{x}-\boldsymbol{\xi}^{(i)}\left(n\right)\right)\right\} \tag{A.3}$$

where M is the number of variables of the data; $N^{(i)}$ is the number of data available for model i; $\boldsymbol{\xi}^{(i)}\left(n\right)$ are the data available for model i, $n=1...N^{(i)}$; and $\mathbf{H}^{(i)}$ is the bandwidth matrix for model i. Here, $\boldsymbol{\theta}^{(i)}$ is the set of values of the bandwidth matrix.

In this work, we used a diagonal matrix, $\mathbf{H}^{(i)}=\mathrm{diag}\left(h_1^{(i)},h_2^{(i)},...,h_M^{(i)}\right)^2$, whose elements were calculated as $h_m^{(i)}=\hat{\sigma}_m^{(i)}\cdot\left(\frac{4}{M+2}\right)^{1/(M+4)}\cdot\left(N^{(i)}\right)^{1/(M+4)}$ [44].

We have denoted this distance as NPD.

ACKNOWLEDGMENT

This work was supported by Spanish Administration and European Union under grant TEC2014-58438-R, and *Generalitat Valenciana* under grants PROMETEO II/2014/032 and GV/2014/034.

REFERENCES

[1] P. Common and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. USA: Academic Press, 2010.

[2] T-W. Lee, M.S. Lewicki and T.J. Sejnowski, "ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22 (10), pp.1078-1089, 2000.

[3] A. Salazar, *On Statistical Pattern Recognition in Independent Component Analysis Mixture Modelling*. Berlin, Germany: Springer, 2013.

[4] A. Ghosh, B.N. Subudhi, and L. Bruzzone, "Integration of Gibbs Markov random field and Hopfield-type neural networks for unsupervised change detection in remotely sensed multitemporal images," *IEEE Transactions on Image Processing*, vol. 22 (8), pp.3087-3096, 2013.

[5] R.O. Duda, P.E., Hart, and D.G. Stork, Pattern Classification, 2nd ed. NY, USA: Wiley, 2001.

[6] C. Alippi and M. Roveri, "Just-in-time adaptive classifiers - Part I: Detecting nonstationary changes," *IEEE Transactions on Neural Networks*, vol. 19 (7), pp. 1145-1153, 2008.

[7] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change Detection in Synthetic Aperture Radar Images Based on Deep Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27 (1), pp. 125-138, 2016.

[8] Q. Cai, H. He, and H. Man, "Spatial outlier detection based on iterative self-organizing learning model," *Neurocomputing*, vol. 117, pp. 161-172, 2013.

[9] H. Ferdowsi, S. Jagannathan, and M. Zawodniok, "An online outlier identification and removal scheme for improving fault detection performance," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25 (5), pp. 908-919, 2014.

[10] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, pp. 345-379, 2010.

[11] N. Singh and R.K. Agrawal, "Combination of Kullback–Leibler divergence and Manhattan distance measures to detect salient objects," *Signal, Image and Video Processing*, vol. 9 (2), pp. 427-435, 2015.

[12] J. Parra-Arnau, D. Rebollo-Monedero, and J- Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Generation Computer Systems*, vol. 33, pp. 53-63, 2014.

[13] A. Polunchenko and A. Tartakovsky, "State-of-the-art in sequential change-point detection," *Methodology and Computing in Applied Probability*, vol. 14, no. 3, pp. 649–684, 2012.

[14] C. Alippi, G. Boracchi, and M. Roveri, "Hierarchical Change-Detection Tests," *IEEE Transactions on Neural Networks and Learning Systems*, available online (doi: 10.1109/TNNLS.2015.2512714), 2016.

[15] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 44:1-44:37, 2014.

[16] L.L. Scharf, *Statistical signal processing*. Addisson-Wesley, Reading, MA (USA), 1991.

[17] T. Jung and T.-W. Lee, "Applications of Independent Component Analysis to Electroencephalography," in *Statistical and Process Models for Cognitive Neuroscience and Aging*. USA: Psychology Press, 2012.

[18] S. Kullback and R.A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79-86, 1951.

[19] J. Harmouche, C. Delpha, and D. Diallo: "Incipient fault detection and diagnosis based on Kullback–Leibler divergence using Principal Component Analysis: Part I," *Signal Processing*, vol. 94, pp 278-287, 2014.

[20] W. Wang, B. Zhang, D. Wang, Y. Jiang, S. Qin, and L. Xue, "Anomaly detection based on probability density function with Kullback-Leibler divergence," *Signal Processing*, vol. 126, pp. 12-17, 2016.

[21] A. Heuser, O. Rioul, S. Guilley, and J.L. Danger, "Information Theoretic Comparison of Side-Channel Distinguishers: Inter-class Distance, Confusion, and Success," In *Trusted Computing for Embedded Systems* (pp. 187-225), Springer International Publishing, 2015.

[22] S. Ray, "Distance-Based Model Selection with Application to the Analysis of Gene Expression Data," PhD. Thesis, Dept. of Statistics, Pennsylvania State University, 2003.
[23] L. Vergara, A. Soriano, G. Safont, A. Salazar, "On the fusion of non-independent detectors," *Digital Signal Processing*, vol. 50, pp. 24-33, 2016.
[24] A. Soriano, L. Vergara, B. Ahmed, A. Salazar, "Fusion of Scores in a Detection Context Based on Alpha Integration," *Neural Computation*, vol. 27, pp 1983-2010, 2015.
[25] J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86 (10), pp. 2009-2025, 1998.
[26] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1985.
[27] A. Sáez, C. Serrano, and B. Acha, "Model-Based Classification Methods of Global Patterns in Dermoscopic Images," *IEEE Transactions on Medical Imaging*, vol. 33 (5), pp. 1137-1147, 2014.
[28] A. Cichocki, S. Amari, M. Adachi, and W. Kasprzak, "Self-adaptive neural networks for blind separation of sources," in *Proceedings of the Third IEEE International Symposium on Circuits and Systems*, pp. 157-160, Rodos, Greece, 1996.
[29] N. Murata, K.-R. Mueller, A. Ziehe, and S. Amari, "Adaptive on-line learning in changing environments," in *Proceedings of the 11th Annual Conference on Neural Information Processing Systems (NIPS'97)*, pp. 599-605, Denver, CO, USA, 1997.
[30] S.H. Hsu, T. Mullen, T.P. Jung, and G. Cauwenberghs, "Real-time Adaptive EEG Source Separation using Online Recursive Independent Component Analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, available online (doi: 10.1109/TNSRE.2015.2508759), 2015.
[31] K. Nordhausen, E. Ollila, and H. Oja, "On the performance indices of ICA and blind source separation," in *Proceedings of the IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications*, pp. 486-490, San Francisco (USA), 2011.
[32] T. Eltoft, "Modeling the Amplitude Statistics of Ultrasonic Images," *IEEE Transactions on Medical Imaging*, vol. 25 (2), 2006.
[33] T. Du-Ming, W. Shih-Chieh, and W.Y. Chiu, "Defect Detection in Solar Modules Using ICA Basis Images," *IEEE Transactions on Industrial Informatics*, vol. 9, pp. 122-131, 2013.
[34] L. Cheng, B. Gao, G.Y. Tian, W.L. Woo, and G. Berthiau, "Impact damage detection and identification using eddy current pulsed thermography through integration of PCA and ICA," *IEEE Sensors Journal*, vol. 14 (5), pp. 1655-1663, 2014.
[35] A. Salazar, L. Vergara, J. Igual, and J. Gosalbez "Blind source separation for classification and detection of flaws in impact-echo testing," *Mechanical Systems and Signal Processing*, vol. 19 (6), pp. 1312-1325, 2005.
[36] A. Salazar, L. Vergara, and R. Llinares, "Learning material defect patterns by separating mixtures of independent component analyzers from NDT sonic signals," *Mechanical Systems and Signal Processing*, vol. 24(6), pp. 1870-1886, 2010.
[37] A. Kumar, "Computer-vision-based fabric defect detection: a survey," *IEEE Transactions on Industrial Electronics*, vol. 55, pp. 348-363, 2008.
[38] A. Salazar, L. Vergara, A. Serrano and J. Igual, "A General Procedure for Learning Mixtures of Independent Component Analyzers," *Pattern Recognition*, vol. 43, pp. 69-85, 2010.
[39] A. Salazar, G. Safont, and L. Vergara, "Surrogate techniques for testing fraud detection algorithms in credit card operations," in *Proceedings of the 48th Annual IEEE International Carnahan Conference on Security Technology (ICCST 2014)*, pp. 1-6, Rome, Italy, 2014.
[40] E. Strauss, *A Compendium of Neuropsychological Tests*. UK: Oxford University Press, 2006.
[41] M. Benedet and M. Alejandre, *Test de Aprendizaje Verbal España-Complutense*. Spain: TEA Ediciones, 1998.
[42] G. Safont, A. Salazar, L. Vergara, and A. Vidal, "Study on the computational cost of EEG dynamic modeling methods," in *Proceedings of the SAI Computing Conference*, pp. 108-112, London (UK), 2016.
[43] A.E. Hramov, A.A. Koronovskii, V.A. Makarov, A.N. Pavlov, and E. Sitnikova, *Wavelets in Neuroscience*. Springer, 2014.
[44] W.K. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models*. Springer, NY, 2012.