Unsupervised linear component analysis for a class of probability mixture models

Marc Castella

Unsupervised linear component analysis for a class of probability mixture models

Marc Castella, *Senior Member, IEEE.*

***Abstract*—We deal with a model where a set of observations is obtained by a linear superposition of unknown components called sources. The problem consists in recovering the sources without knowing the linear transform. We extend the well-known Independent Component Analysis (ICA) methodology. Instead of assuming independent source components, we assume that the source vector is a probability mixture of two distributions. Only one distribution satisfies the ICA assumptions, while the other one is concentrated on a specific but unknown support. Sample points from the latter are clustered based on a data-driven distance in a fully unsupervised approach. A theoretical grounding is provided through a link with the Christoffel function. Simulation results validate our approach and illustrate that it is an extension of a formerly proposed method.**

***Index Terms*—Probability mixtures, Independent Component Analysis (ICA), Christoffel-Darboux kernel, unsupervised classification**

I. Introduction

In this paper, we consider multivariate data sets which come from probability mixture models. Such models can describe data sets generated by systems switching between different states, a situation which is likely to occur in many circumstances (see e.g. [1] for a different context). On top of the switching nature of the system, we consider that the recorded values stem from a linear transform of the signals of interest. Contrary to many machine learning and neural networks based methods, the context is here unsupervised or blind, which means that our method does not require any training set.

On one side, mixture models have been extensively studied for a long time. A classical point of view is to consider that the data generation is controlled by a hidden variable taking values in a finite set. Many probability models have been considered in this context, such as Markov models [1], Markov fields or graphical models [2]. Recovering the hidden process is then equivalent to a classification technique and this task is called unsupervised whenever the model parameters are unknown. Classification in presence of low probability events can also be seen as outlier detection [3]. Our paper differs from the latter by making no difference between inliers and outliers and by considering data with balanced proportions of the different classes.

On the other side, observations resulting from a transform of the unknown data is a commonly used model. Perturbation noise is generally added to the degradation and the transform is often assumed linear [4], [5], although nonlinearity is a more interesting but also much more challenging situation [6]. Contrary to many recovery methods which require precise knowledge of the transform, we deal with a blind context, assuming a linear but unknown transform, similarly to the well-known Independent Component Analysis (ICA) or related source separation techniques (see [7], [8], [9], [10], [11] or [12], [13], [14] for recent works with a similar linear model).

This work combines both previously described contexts and is a significant extension of [15], [16]. The novelties are:

1) we introduce a tool with strong theoretical foundation for unsupervised shape learning and detection of samples on a low dimensional nonlinear support which is not known and not parameterized.
2) we take advantage of an affine invariance property for combining this tool with a linear observation model.
3) compared with [15], [16]: (i) no model is required for the mixture components, (ii) no iterative procedure is required and (iii) multivariate data with dimension more than two can be dealt with.

The problem and model are described in Section II. An intuitive presentation is given in Section III before we provide a theoretical background in Section IV. The overall procedure is given in Section V. Simulations are given in Section VI and Section VII concludes our work.

II. Problem statement

A. Linear mixture model

Let us consider a data matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T] \in \mathbb{R}^{n \times T}$ corresponding to T observed samples of an n-dimensional signal. The column vectors $(\mathbf{x}_t)_{t=1}^T$ are assumed to come from a linear mixture, that is, there exists a fixed matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ which is assumed invertible and another matrix $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_T] \in \mathbb{R}^{n \times T}$ such that $\mathbf{X} = \mathbf{AS}$ or equivalently:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \qquad \forall t \in \{1, \dots, T\}. \tag{1}$$

Both matrices $\mathbf{S}$, the rows of which are called sources, and $\mathbf{A}$, which is called mixing matrix, are unknown. The objective is to recover the sources in $\mathbf{S}$ only from the recorded values in $\mathbf{X}$. This task is equivalent to estimating an inverse $\widehat{\mathbf{B}}$ of the mixing matrix. It is known that scaling and permutation ambiguities will necessarily remain in this blind context, which is similar to ICA [7], [8], [9], [10]. In the latter method, a usual assumption is the non gaussianity and independence of the components in each random vector $(\mathbf{s}_t)_{t=1}^T$. Here we consider on the contrary a model of dependent components.

B. Probability mixture model

We assume that the source samples $(\mathbf{s}_t)_{t=1}^T$ are drawn according to a probability mixture of two distributions $\mathbb{P}_0$ and

M. Castella (corresponding author) is with SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France. marc.castella@telecom-sudparis.eu

$\mathbb{P}_1$. Hence there exists $\eta \in [0,1]$ such that the probability distribution of $\mathbf{s}_t$ is given by:

$$\mathbb{P}(\mathbf{s}_t) = \eta \mathbb{P}_0(\mathbf{s}_t) + (1-\eta)\mathbb{P}_1(\mathbf{s}_t)\,, \quad \forall t \in \{1,\dots,T\}. \quad (2)$$

An equivalent model consists in introducing a binary hidden (or latent) random variable $r_t \in \{0,1\}$ such that $\mathbb{P}(r_t = 0) = \eta = 1 - \mathbb{P}(r_t = 1)$. The vector $\mathbf{s}_t$ can then be seen as the marginal of $(r_t, \mathbf{s}_t)$, where conditional distributions $\mathbb{P}(\mathbf{s}_t|r_t)$ are given by $\mathbb{P}_0$ and $\mathbb{P}_1$. Due to the invertible linear relation (1), the distributions of $\mathbf{x}_t$ and $\mathbf{s}_t$ are deduced one from another, up to a constant Jacobian term. Hence a strictly similar probability model holds for the observed data $\mathbf{X}$. Finally, we also consider that, for different $t \in \{1,\dots,T\}$, all variables are independent and identically distributed (i.i.d.). It follows that $\mathbb{P}(\mathbf{r}, \mathbf{X}) = \prod_{t=1}^{T} \mathbb{P}(r_t)\mathbb{P}(\mathbf{x}_t|r_t)$ where the conditional distributions are respectively $\mathbb{P}_0(\mathbf{x}_t)$ and $\mathbb{P}_1(\mathbf{x}_t)$. Although the same notation is used for both distributions of $\mathbf{s}_t$ and $\mathbf{x}_t$, there should be no confusion: both models are similar and only the latter will be involved in our method.

C. Unsupervised classification problem

Our methodology deals with situations where, due to the presence of $\mathbb{P}_1$, the classical assumption of ICA does not hold for $\mathbb{P}$. However, the distribution $\mathbb{P}_0$, which is the same for all $t \in \{1,\dots,T\}$, is assumed to be such that it satisfies the usual ICA requirements. In addition, $\mathbb{P}_1$ is assumed to be concentrated on a restricted nonlinear support, which is unknown. More precisely, the assumptions are:

- A1. $\mathbb{P}_0(\mathbf{s}_t)$ is such that the components of $\mathbf{s}_t$ are mutually independent and non Gaussian, except possibly one of them.
- A2. The distribution $\mathbb{P}_0$ is absolutely continuous with respect to the Lebesgue measure and the distribution $\mathbb{P}_1$ is singular with support on an algebraic set of Lebesgue measure zero[1].

To achieve reconstruction of $\mathbf{S}$, a possible intermediate goal is to learn the unknown support of $\mathbb{P}_1$ and classify the samples $(\mathbf{x}_t)_{t=1}^T$ according to whether $r_t = 0$ or 1. Removing the points with $r_t = 1$, the remaining samples are drawn from $\mathbb{P}_0$ and satisfy usual ICA assumption. Then, it is possible to identify the inverse of $\mathbf{A}$ by any classical algorithm such as in [17], [18], [7]. This basic idea has been introduced in [15], [16] and it has been shown that unsupervised classification can be successful with a very specific choice and knowledge of the corresponding model for $\mathbb{P}_1$. Our method here is more general.

III. A nonlinear distance based classifier

A. Intuitive justification

We will exploit the concentration of points of the distribution $\mathbb{P}_1$ in a specific region. Our method relies on quantifying how far a given sample deviate from it. For any vector $\mathbf{x} = (x_1,\dots,x_n)^\top \in \mathbb{R}^n$, a classical squared distance to the point cloud given by the samples in $\mathbf{X}$ is given by the quantity $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, where, writing $\mathbf{1}$ an all-one column vector of size T, $\boldsymbol{\mu} = \frac{1}{T}\mathbf{X}\mathbf{1}$ and $\boldsymbol{\Sigma} = \frac{1}{T}\mathbf{X}\mathbf{X}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ are the empirical mean and covariance matrices. Points equally far from the mean lie on an ellipsoid defined by the covariance matrix. Alternatively, one can include the constant 1 in the data feature space and introduce the extended covariance matrix

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} 1 \\ \mathbf{x}_t \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x}_t{}^\top \end{bmatrix}. \quad (3)$$

Using Schur complement as in [19], we obtain the same distance criterion up to a constant:

$$\begin{bmatrix} 1 & \mathbf{x}^\top \end{bmatrix} \widetilde{\boldsymbol{\Sigma}}^{-1} \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + 1\,. \quad (4)$$

The above notion, which is linked to an implicit Gaussian assumption, appears under the name of leverage-score or Mahalanobis distance [20]. In our context, we assume that the data drawn according to $\mathbb{P}_1$ is concentrated in the neighborhood of a lower dimensional subspace defined by nonlinear equations. A natural idea in a nonlinear context consists in further extending the data feature space by including additional monomials in a spirit similar to Taylor expansions or Volterra filters.

B. Proposed method

For any order $d \in \mathbb{N}$, denote by $[\mathbf{x}]_d$ a column vector containing a basis of all polynomials in $\mathbf{x}$ with maximal degree d. In practice, we included in $[\mathbf{x}]_d$ all monomials of degree less than or equal to d. To determine the points corresponding to $r_t = 1$, we propose to use this extended vector and compute for all $t \in \{1,\dots,T\}$ a score θ_t based on Equations (3) and (4). This score is then compared to a threshold value $\overline{\theta}$, the choice of which will be discussed later. The procedure for finding an estimate $\hat{r}_t$ of r_t hence consists of the steps given in Alg. 1.

Alg. 1 Classification method

Input: Data matrix $\mathbf{X} = (\mathbf{x}_t)_{t=1}^T$, threshold value $\overline{\theta}$.

1) Compute the extended empirical covariance matrix:

$$\widehat{\mathbf{M}}_d = \frac{1}{T}\sum_{t=1}^{T} [\mathbf{x}_t]_d [\mathbf{x}_t]_d{}^\top. \quad (5)$$

2) For $t = 1,\dots,T$, compute

$$\theta_t = [\mathbf{x}_t]^T \big(\widehat{\mathbf{M}}_d\big)^{-1} [\mathbf{x}_t]\,. \quad (6)$$

3) Set $\hat{r}_t = \begin{cases} 0 & \text{if } \theta_t > \overline{\theta}\,, \\ 1 & \text{if } \theta_t < \overline{\theta}\,. \end{cases}$

Output: Estimated classification $\hat{\mathbf{r}} = (\hat{r}_t)_{t=1}^T$.

IV. Connection with Christoffel-Darboux kernel

A theoretical justification of our method is provided by a link with the Christoffel function and the Christoffel-Darboux kernel. They have been known for long and are classical tools in interpolation and approximation with a close link to orthogonal polynomials. Their usefulness and relevance for data analysis tasks have been recently recognized [19], [21], [22]. A major asset of the Christoffel function is its ability to encode information about the shape of a distribution and, more importantly for us, it can detect the presence of a singular continuous component [23], [24].

[1]This means in practice that the support of $\mathbb{P}_1$ is the solution set of a finite number of polynomial equations.

A. Definitions and properties

Consider the probability distribution $\mathbb{P}(\mathbf{x})$ on the observed variables and assume that it is supported on a compact set $\mathbf{K} \subset \mathbb{R}^n$. The associated moment matrix is by definition:

$$\mathbf{M}_d^{\mathbb{P}} = \int_{\mathbf{K}} [\mathbf{x}]_d [\mathbf{x}]_d{}^\top \, \mathrm{d}\mathbb{P}(\mathbf{x}) \,, \tag{7}$$

where the integral is taken component-wise. Since any polynomial with degree less than d can be written $p(\mathbf{x}) = \mathbf{p}^\top [\mathbf{x}]_d$ with $\mathbf{p}$ the corresponding coefficients vector in the basis $[\mathbf{x}]_d$, one can see that

$$\mathbf{p}^\top \mathbf{M}_d^{\mathbb{P}} \mathbf{p} = \int_{\mathbf{K}} p(\mathbf{x})^2 \, \mathrm{d}\mathbb{P}(\mathbf{x}) \,.$$

Therefore, $\mathbf{M}_d^{\mathbb{P}}$ is symmetric positive semi-definite. It is also positive definite under a non degeneracy condition which is satisfied for absolutely continuous measures [19], [22]. In our context, the presence of $\mathbb{P}_0$ with assumption A2 ensures that $\mathbf{M}_d^{\mathbb{P}}$ is non singular. Writing $\mathbf{y} = (y_1, \ldots, y_n)^\top$, the Christoffel-Darboux kernel associated to $\mathbb{P}$ can then be defined by

$$\kappa_d^{\mathbb{P}}(\mathbf{x}, \mathbf{y}) = [\mathbf{x}]_d{}^\top \left(\mathbf{M}_d^{\mathbb{P}}\right)^{-1} [\mathbf{y}]_d \,.$$

For any $\mathbf{z} = (z_1, \ldots, z_n)^\top$, another quantity of interest, referred to as the Christoffel function, is given by $C_d^{\mathbb{P}}(\mathbf{z}) = \frac{1}{\kappa_d^{\mathbb{P}}(\mathbf{z}, \mathbf{z})}$. It can be equivalently defined based on the following variational formula, where the minimization is with respect to polynomials with degree less than d and taking value 1 at $\mathbf{z}$ (see [19], [22] for details):

$$C_d^{\mathbb{P}}(\mathbf{z}) = \min_{p \in \mathbb{R}[\mathbf{x}]_d,\, p(\mathbf{z})=1} \int p(\mathbf{x})^2 \, \mathrm{d}\mathbb{P}(\mathbf{x}) \,. \tag{8}$$

From the above formula, one can understand that the shape of regions with high probability mass can be captured.

B. Empirical Christoffel function

In our practical setting, our method relies on the matrix $\widehat{\mathbf{M}}_d$ from Equation (5). Since the sample values $(\mathbf{x}_t)_{t=1}^T$ stored in $\mathbf{X}$ are i.i.d. and follow the distribution $\mathbb{P}$, $\widehat{\mathbf{M}}_d$ in (5) is the empirical estimate of the matrix $\mathbf{M}_d^{\mathbb{P}}$ from Equation (7). Importantly, as noted in [21], because $\mathbf{M}_d^{\mathbb{P}}$ is non singular, it holds for T large enough that $\widehat{\mathbf{M}}_d$ is almost surely invertible. As a consequence, the score θ_t computed in our method is the empirical estimate at $\mathbf{x}_t$ of $\mathbf{x} \mapsto \kappa_d^{\mathbb{P}}(\mathbf{x}, \mathbf{x})$, which is the inverse of the Christoffel function. Note that it has been shown in [21] that the empirical Christoffel function converges almost surely and uniformly in $\mathbf{x}$ to $C_d^{\mathbb{P}}(\mathbf{x})$ for large T.

C. Support and shape detection

1) Case of a singular support: Corresponding to the fact that $\mathbb{P}_1$ is concentrated on a specific set, we made Assumption A2 concerning the decomposition of $\mathbb{P}$ in (2) as a probability mixture. No additional assumption is made and in particular, nothing more is known about the support of $\mathbb{P}_1$. Contrary to [15], [16], no model is introduced for $\mathbb{P}_1$. The task of learning $\mathbf{r}$ is exclusively based on unsupervised identification of the support of $\mathbb{P}_1$.

For any measure such as $\mathbb{P}_1$ with singular support, the definition of the Christoffel-Darboux kernel requires attention because of the singularity of $\mathbf{M}_d^{\mathbb{P}_1}$. Fortunately the variational definition in (8) remains valid. From the latter, because any polynomial with $p(\mathbf{z}) = 1$ is positive in a small neighborhood of $\mathbf{z}$, we have $C_d^{\mathbb{P}_1}(\mathbf{z}) > 0$ for any $\mathbf{z}$ in the support of $\mathbb{P}_1$. More importantly for us, it has been proven that outside the support of $\mathbb{P}_1$, the inverse Christoffel function $\mathbf{x} \mapsto \kappa_d^{\mathbb{P}_1}(\mathbf{x}, \mathbf{x})$ on which our method is based grows (in d) at an exponentially fast rate and hence the Christoffel function goes to zero [21]. Knowing the Christoffel function associated to a measure therefore helps identifying its support.

2) Threshold value: The previous elements justify to consider points with large values of the inverse Christoffel function as an estimation for points outside the singular component $\mathbb{P}_1$ of $\mathbb{P}$: this is precisely what is done in our method by computing θ_t for each $t \in \{1, \ldots, T\}$ and comparing it to the threshold $\overline{\theta}$. More precisely, based on [21], $\overline{\theta}$ should be proportional to the binomial coefficient $\binom{n+d}{n} = \frac{(n+d)!}{n!d!}$. This is confirmed in [3] and we precisely choose $\overline{\theta} = \eta\binom{n+d}{n}$ in our experiments.

V. Application to linear component analysis

We come back to our initial objective of obtaining a linear decomposition similar to the ICA model in (1). Therefore, it is natural to ask how the tools introduced previously behave under linear transformation of the data.

A. Affine invariance

Let us consider an invertible matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$. Given the probability distribution $\mathbb{P}$ on $\mathbf{x}$, this matrix induces a probability distribution denoted $\mathbb{P}_{\mathbf{B}}$ for the corresponding variable $\mathbf{z} = \mathbf{B}\mathbf{x}$. It has been proven (see e.g. [19], [21]) that the Christoffel-Darboux function satisfies an invariance property by any invertible affine transform and in particular:

$$\kappa_d^{\mathbb{P}}(\mathbf{x}, \mathbf{x}) = \kappa_d^{\mathbb{P}_{\mathbf{B}}}(\mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x})$$

As a consequence, we have the following proposition:

Proposition 1: For a given realization of $(\mathbf{s}_t)_{t=1}^T$ and for any $(\mathbf{x}_t)_{t=1}^T$ such that (1) holds with invertible $\mathbf{A}$, the scores $(\theta_t)_{t=1}^T$ defined in (5)-(6) take values independent of $\mathbf{A}$.

This property is of high importance in our method: contrary to [15], [16], we will not use any iterative procedure, but the values of $(\theta_t)_{t=1}^T$ will be computed only once based on $(\mathbf{x}_t)_{t=1}^T$. These values are identical to the values that would have been computed based on $(\mathbf{s}_t)_{t=1}^T$.

B. Method for linear decomposition

In our context of linear mixture, thanks to the above affine invariance, the Christoffel-Darboux function is a particularly well suited tool for classifying points from components $\mathbb{P}_0$ or $\mathbb{P}_1$. Relying on an existing ICA algorithm denoted `ICA`$(.)$ that returns the inverse of the mixing matrix such as CoM2 [7], FastICA [18], JADE [17], [25], the proposed global procedure also performs a linear decomposition as in (1). We sum it up in Alg. 2.

Alg. 2 Linear decomposition

Input: Data matrix $\mathbf{X} = (\mathbf{x}_t)_{t=1}^T$, algorithm $\texttt{ICA}(.)$

- Perform steps 1) to 3) as in Section III-B, Alg. 1.
- Define $\widehat{\mathbf{X}}_0$ the submatrix of $\mathbf{X}$ with columns indexed by $\{t = 1, \ldots, T \,|\, \hat{r}_t = 0\}$.
- Perform $\widehat{\mathbf{B}} = \texttt{ICA}(\widehat{\mathbf{X}}_0)$.

Output: Estimated linear components $\widehat{\mathbf{S}} = \widehat{\mathbf{B}}\mathbf{X}$.

VI. Simulations

A. Experimental setup

We have tested our method on synthetic data where the sources in $\mathbf{S}$ were randomly drawn according to (2). The distribution $\mathbb{P}_0$ satisfied (A1-A2) with components uniformly distributed, centered and unit variance. The different choices for $\mathbb{P}_1$ are detailed in the next section. The matrix $\mathbf{A}$ has been systematically randomly drawn with i.i.d. Gaussian entries and our ICA algorithm was CoM2 [7]. To quantify the success of our method, we eliminated the inherent ambiguities of ICA and considered the average mean square error (MSE) on the components of the recovered $(\mathbf{s}_t)_{t=1}^T$. In addition the probability of correctly estimating $(r_t)_{t=1}^T$, has been computed, given by $\Upsilon = \frac{\#\{t \,|\, \hat{r}_t = r_t\}}{T}$ where $\#$ is the cardinality of the set. For comparison, we considered the result of ICA applied directly on the data $\mathbf{X}$, hence ignoring the probability model assumed in Section II-B. We also considered the ideal supervised case with known true values of $(r_t)_{t=1}^T$, keeping the samples with $r_t = 0$ as an input for the ICA algorithm. All presented results are mean values over 1000 Monte-Carlo realizations, after discarding the bottom/top 1% values.

B. Simulation results

1) Comparison with [15]: The dependent sources given by Example 1 in [15] satisfy conditions (A1-A2) and our method is indeed successful in separating them with a computational complexity that is reduced compared to the iterative procedure in [15] (see Table I). As shown next, our method goes beyond this very specific case where $n = 2$.

Runtime	Method	$\eta = 0.2$	$\eta = 0.4$	$\eta = 0.6$
1.8 ms	Ignore $\mathbb{P}_1$	0.5849	0.5838	0.2493
33.6 ms	Method in [15]	0.5848	0.1953	0.0072
5.1 ms	Proposed (order $d = 6$)	**0.1232**	**0.0763**	**0.0389**
1.2 ms	Known $\mathbf{r}$	0.0011	0.0006	0.0003

TABLE I
MSE results on sources from [15, Example 1] (T=2000 samples) and comparison of runtime with $\eta = 0.6$.

2) Case of 3 sources: As an extension, we considered $n = 3$ sources where $\mathbb{P}_1$ is given as follows: s_1 uniform on $[-\beta, \beta]$, s_2 uniform on $[-\gamma, \gamma]$ and $s_3 = \frac{1}{\beta^2} s_1^3$. We took $\beta = 3/2$ and a typical realization of such sources is given on the top row of Figure 1, for $\gamma = 0$ (left) and $\gamma = 1$ (right): to show the success of our unsupervised classification, the points classified with $\hat{r}_t = 1$ are plotted in green and the shape of the hypersurface appears clearly. Correspondingly, the values of the MSE (middle) and of Υ are plotted depending on η. Our method has good performance and shows better results with data concentrated on a low dimensional subset.

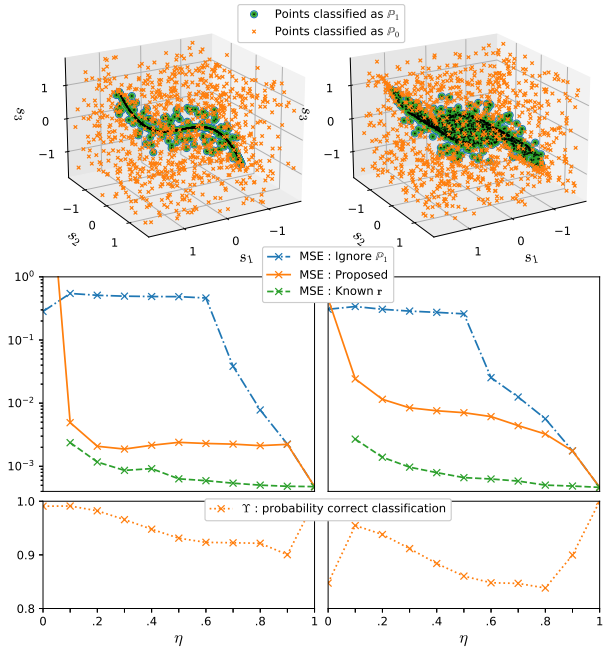


Fig. 1. Sources from Sec. VI-B2: typical classification and corresponding linear component analysis results ($T = 2000$ samples, order $d = 6$).

3) Simultaneously vanishing sources: For a more concrete example, we considered a scenario where two sources are simultaneously switched off. This might occur in different applications. We took $n = 5$ random independent uniformly distributed sources when $r_t = 0$, whereas we set components 4 and 5 of $\mathbf{s}_t$ to zero when $r_t = 1$. Table II shows the success of our method. The best results were obtained experimentally with the order $d = 6$, which seems a compromise between good modeling capabilities and numerical stability.

Method	$\eta = 0.2$	$\eta = 0.4$	$\eta = 0.6$	$\eta = 0.8$
Ignore $\mathbb{P}_1$	0.2239	0.1994	0.1988	0.0056
Proposed (order $d = 2$)	0.0647	0.0140	0.0089	0.0056
Proposed (order $d = 4$)	0.0103	0.0067	0.0049	0.0038
Proposed (order $d = 6$)	**0.0058**	**0.0044**	**0.0038**	**0.0028**
Proposed (order $d = 8$)	0.0042	0.0042	0.0048	0.0035
Known $\mathbf{r}$	0.0034	0.0018	0.0013	0.0011

TABLE II
MSE results on sources from Sec. VI-B3 (T=2000 samples).

VII. Conclusion

We have considered an extension of ICA for data switching between two probability distributions, only one of which satisfies the ICA assumptions. For such a case, we proposed an intuitively simple and theoretically grounded method for performing a linear decomposition in a blind context. An unsupervised selection of data samples in accordance with the ICA assumptions is performed by identifying points clustered in a restricted region through the use of the Christoffel function. Due to its affine invariance, this tool is particularly well suited for this linear superposition context. Simulations show the interest and good performance of the approach for different examples and models, including cases where a previously proposed approach is not applicable.

REFERENCES

[1] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[3] K. Ducharlet, L. Travé-Massuyès, J.-B. Lasserre, M.-V. Le Lann, and Y. Miloudi, "Leveraging the Christoffel-Darboux Kernel for Online Outlier Detection," 2022, hal-03562614. [Online]. Available: https://hal.laas.fr/hal-03562614

[4] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable Metric Forward-Backward Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function," *Journal of Optimization Theory and Applications*, vol. 162, no. 1, pp. 107–132, 2014.

[5] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems," *Inverse Problems*, vol. 23, no. 4, pp. 1495–1518, 2007.

[6] A. Marmin, M. Castella, J.-C. Pesquet, and L. Duval, "Sparse signal reconstruction for nonlinear models via piecewise rational optimization," *Signal Processing*, vol. 179, pp. 107 835:1–107 835:13, Feb 2021.

[7] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[8] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.

[9] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[11] C. Simon, P. Loubaton, and C. Jutten, "Separation of a class of convolutive mixtures: a contrast function approach," *Signal Processing*, vol. 81, no. 4, pp. 883–887, 2001.

[12] S. Cruces, "On the Minimum Perimeter Criterion for Bounded Component Analysis," in *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5.

[13] A. Weiss and A. Yeredor, "Exact algebraic blind source separation using side information," in *28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1941–1945.

[14] M. Saleh, A. Karfoul, A. Kachenoura, L. Albera, and L. Senhadji, "Independent Component Analysis Based on Non-Polynomial Approximation of Negentropy: Application To MRS Source Separation," *52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 2179–2183, 2018.

[15] M. Castella, S. Rafi, P. Comon, and W. Pieczynski, "Separation of instantaneous mixtures of a particular set of dependent sources using classical ICA methods," *EURASIP J. Adv. Signal Process.*, no. 62, 2013.

[16] S. Rafi, M. Castella, and W. Pieczynski, "An extension of the ICA model using latent variables," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 3712–3715.

[17] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings F Radar and Signal Processing*, vol. 140, no. 6, p. 362, 1993.

[18] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[19] J.-B. Lasserre and E. Pauwels, "Sorting out typicality with the inverse moment matrix SOS polynomial," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 29, 2016.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., 2006.

[21] J.-B. Lasserre and E. Pauwels, "The empirical Christoffel function with applications in data analysis," *Adv Comput Math*, vol. 45, pp. 1439–1468, 2019.

[22] J.-B. Lasserre, E. Pauwels, and M. Putinar, *The Christoffel-Darboux Kernel for Data Analysis*, ser. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2022.

[23] E. Pauwels, M. Putinar, and J.-B. Lasserre, "Data analysis from empirical moments and the Christoffel function," *Foundations of Computational Mathematics*, vol. 21, pp. 246–273, 2021.

[24] M. Korda, M. Putinar, and I. Mezić, "Data-driven spectral analysis of the Koopman operator," *Applied and Computational Harmonic Analysis*, vol. 48, no. 2, pp. 599–629, 2020.

[25] J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, pp. 157–192, 1999.

APPENDIX

We gather in this Appendix some supplemental material and results that may be useful.

NOTATION CLARIFICATION (EXAMPLE))

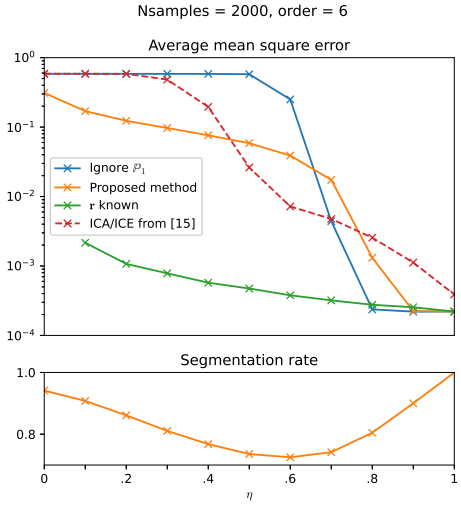
In this paper, the notation $[\mathbf{x}]_d$ denotes a vector containing a basis of polynomials in $\mathbf{x}$ with maximal degree d. For example, if the monomial basis is considered, this yields with $n = 2$ and $d = 2$, $d = 3$ respectively:

$$[\mathbf{x}]_2 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1x_2 \\ x_2^2 \end{bmatrix} \quad [\mathbf{x}]_3 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1x_2 \\ x_2^2 \\ x_1^3 \\ x_1^2x_2 \\ x_1x_2^2 \\ x_2^3 \end{bmatrix}$$

It is known that the vector space of polynomials of degree at most d in $\mathbf{x} = (x_1, \ldots, x_n)$ has dimension $\binom{n+d}{n} = \frac{(n+d)!}{n!d!}$, which is therefore the size of $[\mathbf{x}]_d$.

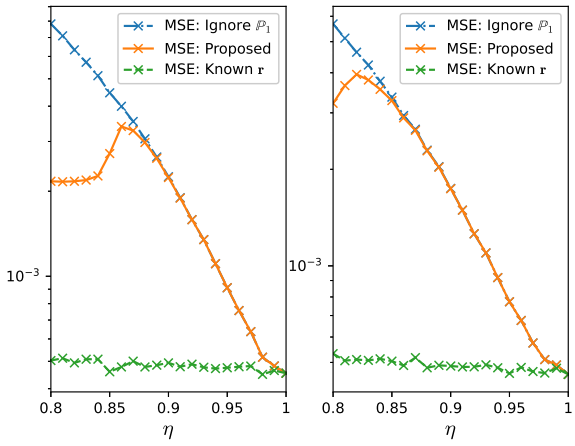
ADDITIONAL DETAILS ON SECTION VI-B1

The plot of the complete simulation results corresponding to Table I is given below. For small and big values of η (that is $\eta < 0.5$ or $\eta \geq 0.8$), one can see that the proposed method performs better than the method in [15], although the generated data is more favorable to the method in [15]. Note also that for values of $\eta \geq 0.8$, the perturbation introduced by $\mathbb{P}_1$ is small enough and better or equivalent results are obtained by just applying ICA on the perturbed data.



ADDITIONAL DETAILS ON SECTION VI-B2

Below are results showing more precisely how our method behaves for η between 0.8 and 1 for the corresponding sources of Figure 1. One can see that above a threshold value of η (approximately 0.85), our method yields results corresponding to the classification $\hat{r}_t = 0$ for all t.



COMMENT ON THE CHOICE OF d

As mentioned above, the size of $[\mathbf{x}]_d$ is $\binom{n+d}{d}$. It is also the size of the square matrices $\mathbf{M}_d$ and $\widehat{\mathbf{M}}_d$. The corresponding values are given below depending on n and d:

	$d=1$	$d=2$	$d=4$	$d=6$	$d=8$
$n=2$	3	6	15	28	45
$n=3$	4	10	35	84	165
$n=5$	6	21	126	462	1287
$n=8$	9	45	495	3003	12870

The complexity of our method given in Alg. 1 is linked to the inversion of $\widehat{\mathbf{M}}_d$ and hence to its size $\binom{n+d}{d}$ as given above. Note that when the inversion of $\widehat{\mathbf{M}}_d$ is too costly, it is possible to solve a quadratic optimization problem for each evaluation of θ_t in Equation (6) (see [22]).

Finally, let us mention that choosing the hyperparameter d requires a compromise: greater values of d indeed provide better modeling capabilities. However, we observed badly conditioned moment matrices for values above $d \geq 8$. This phenomenon seems more limiting than computational load difficulties.