

Gaussian Processes for Source Separation in Overdetermined Bilinear Mixtures

Denis G. Fantinato¹(✉), Leonardo T. Duarte², Bertrand Rivet³,
Bahram Ehsandoust³, Romis Attux¹, and Christian Jutten³

¹ School of Electrical and Computer Engineering - University of Campinas,
Campinas/SP, Brazil,

² School of Applied Sciences - University of Campinas, Limeira/SP, Brazil,

³ GIPSA-Lab, Grenoble INP, CNRS, Grenoble, France,

{denisgf,attux}@dca.fee.unicamp.br,leonardo.duarte@fca.unicamp.br,
{bertrand.rivet,bahram.ehsandoust,christian.jutten}@gipsa-lab.grenoble-inp.fr

Abstract. In this work, we consider the nonlinear Blind Source Separation (BSS) problem in the context of overdetermined Bilinear Mixtures, in which a linear structure can be employed for performing separation. Based on the Gaussian Process (GP) framework, two approaches are proposed: the predictive distribution and the maximization of the marginal likelihood. In both cases, separation can be achieved by assuming that the sources are Gaussian and temporally correlated. The results with synthetic data are favorable to the proposal.

Keywords: Blind Source Separation, Bilinear Mixtures, Gaussian Process

1 Introduction

In the context of Blind Source Separation (BSS), it is desired to retrieve the original sources signals that were mixed together from a number of observations of these mixtures [1]. Classically, this problem is viewed from the perspective that the mixing process is linear and that the sources are mutually statistically independent. In this context, a number of methods based on Independent Component Analysis (ICA) were successfully employed in many practical applications [1]. However, in certain real world separation problems, e.g., chemical sensor arrays [2], the mixtures are evidently nonlinear. In such cases, the extension of the ICA methods towards a general nonlinear system is not straightforward [3]. In light of this, the studies on nonlinear BSS have been focused on classes of constrained mixing models in which the framework employed in the linear case can be effectively extended. A representative example of these constrained models is the Bilinear mixture equations [4], with actual applications like the show-through effect removal in scanned images [5] and the project of gas sensor arrays [6].

The bilinear model is of special interest for us due to three interesting features: (i) it can be formulated as a linear system with respect to the mixing coefficients; (ii) it is an initial step toward polynomial mixtures; and (iii) in the

overdetermined case (when the number of mixtures is greater than the number of sources), a linear structure may be enough for separation, under certain conditions on the number of sources and mixtures. In the literature, these features were exploited along with certain statistical properties about the sources, in which we can cite, for instance, circularity [7], finite alphabet [8], sparsity [9] and limited band [10]. In this work, however, we propose a different approach in which the sources are described as independent Gaussian Processes (GP) [11]. Although the GP framework is solely based on second-order statistics (SOS), this idea turns to be promising due to the linear properties found in the overdetermined bilinear model. Moreover, its formulation provides an attractive theoretical approach to perform source separation. The GP method encompassed here is twofold: we consider the GP predictive distribution (a semi-blind approach) and the maximization of the marginal likelihood, respectively.

2 Problem Statement

In the BSS problem, it is considered that a set of N sources are instantaneously mixed, giving rise to M observations according to the following relation:

$$\mathbf{x}(n) = \mathbf{f}(\mathbf{s}(n)) \quad (1)$$

where $\mathbf{s}(n) = [s_1(n) \cdots s_N(n)]^T$ is the source vector with N elements, $\mathbf{x}(n) = [x_1(n) \cdots x_M(n)]^T$ is the observation vector (of length M) at time instant n and $\mathbf{f}(\cdot)$ is a set of M functions, potentially nonlinear. By assuming the knowledge of only the mixtures $\mathbf{x}(n)$ and certain *a priori* information (e.g., statistical independence among sources), it is desired to recover the original sources $\mathbf{s}(n)$ by means of a separation process, up to scale and permutation factors [1].

Interestingly, for certain types of nonlinear functions $\mathbf{f}(\cdot)$, when the number of mixtures M is larger than that of sources N – which is referred to as the *overdetermined* case –, additional information can be used to systematically simplify the separation process. This attractive feature arises in the context of bilinear mixtures, which will be described in the following.

2.1 Bilinear Mixtures - The Overdetermined Case

The bilinear mixtures belong to a special class of nonlinear mixtures and can be represented by following expression [1, 5]:

$$\mathbf{x}_i(n) = \sum_{j=1}^N a_{ij} s_j(n) + \sum_{k \neq l} b_{ikl} s_k(n) s_l(n), \quad (2)$$

which can be viewed as a linear combination of the sources plus the cross product terms $s_k(n)s_l(n)$, for $k \neq l$. Interestingly, for this type of mixtures, if a given number of additional mixtures is available, it is possible to perform the suppression of the nonlinear terms, reducing the nonlinear problem to a linear one.

For the sake of simplicity, we assume henceforth the overdetermined case of $N=2$ sources and $M=3$ mixtures, which is a clarifying and still representative instance for practical scenarios [6]. In this case, by using a vector notation, the bilinear mixtures can be written as

$$\mathbf{x}(n) = \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{bmatrix} \begin{bmatrix} s_1(n) \\ s_2(n) \\ s_1(n)s_2(n) \end{bmatrix}, \quad (3)$$

note that we have used a simplified index notation for b , since there is only a single cross product term – i.e., $s_1(n)s_2(n)$. From a certain perspective, Eq. (3) can be viewed as a linear mixing problem with an additional statistically dependent source $s_1(n)s_2(n) = s_3(n)$ [9]. Regarding the separation task in the overdetermined case, we consider two approaches: (i) the one-stage and (ii) and the two-stage.

In the one-stage approach, a simple linear separating system is considered:

$$\begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} = \tilde{\mathbf{W}}\mathbf{x}(n), \quad (4)$$

where $\tilde{\mathbf{W}}$ is a non-square matrix of dimension 2×3 ($N \times M$). In this case, $\tilde{\mathbf{W}}$ is adjusted in a single stage.

For the two-stage approach, as shown in [9], a linear combination of the mixtures of the type

$$\begin{bmatrix} z_1(n) \\ z_2(n) \end{bmatrix} = \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} - \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} x_3(n), \quad (5)$$

is able to suppress the quadratic cross terms in Eq. (2) – i.e., the nonlinear part of the mixture is removed – if the values of α were properly adjusted. Theoretically, it can be shown that the optimum values for α are $\alpha_i = b_i/b_3$ [9] and, when this ideal case is achieved, $\mathbf{z}(n)$ is simply a linear mixture of the sources. We call this the first stage. For the second stage, since we face a linear BSS problem, we can write a 2-by-2 separating system as $\mathbf{y}(n) = \mathbf{W}\mathbf{z}(n)$, whose solution is straightforward to be reached – for example, via standard ICA methods [1]. Although only presented for the case of two sources and three mixtures, this idea can be generalized to N sources, but it will require at least $M = N(N+1)/2$ mixtures to cancel all nonlinear elements [10].

Interestingly, in both cases, a linear structure is able to perform separation. From the literature, for the one-stage approach, the non-square shape of the separating matrix avoids the direct application of certain classical ICA methods. However, it is still possible to use, for example, gradient-based methods for optimization [12]. On the other hand, the efforts in the two-stage approach are mainly aimed at solving the first stage (since the second stage is a well studied problem), usually being considered certain statistical properties about the sources [7–10].

These approaches are able to encompass a wide range of real world scenarios, but, in order to enlarge its scope, we propose, in this work, a different approach: sources that are described as Gaussian Processes. This idea will lead to two different methods in GP, depending on the approach, one- or two-stage.

3 Gaussian Processes in the Bilinear Mixtures

The main motivation for this work lies in Gaussian processes (GP), which can be defined as a collection of random variables (RV), any finite number of which have a joint Gaussian distribution [11]. GP is able to provide convenient methods, since it can be totally described by second-order statistics (SOS). In the context of bilinear mixtures, a GP can be constructed based on priors about the sources. More specifically, we assume that the sources are stationary, Gaussian distributed and mutually independent. Mathematically, this assumptions can be written, in the case of $N=2$ sources, as

$$\begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix} \sim \mathcal{N}\left(\mu, \begin{bmatrix} \sigma_{s_1}^2 & 0 \\ 0 & \sigma_{s_2}^2 \end{bmatrix}\right) \quad (6)$$

where $\mu = [\mu_1 \ \mu_2]^T$ is the column vector with the sources mean values and $\sigma_{s_i}^2$ the variance of the i -th source $s_i(n)$. For simplicity, we consider in this work that $\mu_i = 0$, for all i . In addition, sources are also assumed to be temporally colored (with different autocorrelation functions).

Based on the model description in Section 2 and on the prior given by Eq. (6), it is possible to assert some information about the mixtures $\mathbf{x}(n)$. From Eq. (3), we know that $s_3(n) = s_1(n)s_2(n)$ is the product of two Gaussian distributed RV, which results that $s_3(n)$ can be described by means of a double-sided chi-squared distribution; hence, the probability density function of the mixtures, $p(\mathbf{x})$, is not necessarily Gaussian – it will be approximately Gaussian for small values b in Eq. (3) or for a large number of sources, due to the central limit theorem [1], however, these cases will not be considered here and $p(\mathbf{x})$ will be assumed to be non-Gaussian.

In our approach, it turns to be interesting to express the estimated sources $\mathbf{y}(n)$ as a GP; in other words, it is desired that a conditional distribution of $\mathbf{y}(n)$ be Gaussian. We consider the two conditional posterior probabilities, $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{y}|\mathbf{z})$, which are related to the one- and two-stage approaches, respectively. As we intend to show, they will lead to two representative approaches in the GP formulation, one for each conditional probability: the predictive distribution and the marginal likelihood, as described in the following.

3.1 The Predictive Distribution

In the context of the one-stage approach, it is possible to assert a few comments on $p(\mathbf{y}|\mathbf{x})$. Using the Bayes' rule, $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})/p(\mathbf{x})$ and, since $p(\mathbf{x}|\mathbf{y})$ is not necessarily Gaussian and $p(\mathbf{x})$ is definitively not Gaussian, then

$p(\mathbf{y}|\mathbf{x})$ is not Gaussian as well. However, from a theoretical standpoint, by conditioning the posterior for a given source, e.g., $\mathbf{s}_i(n)$, we have now $p(\mathbf{y}|\mathbf{x},s_i) = p(\mathbf{x}|\mathbf{y},s_i)p(\mathbf{y}|s_i)/p(\mathbf{x}|s_i)$, which is Gaussian distributed: this result becomes more evident by verifying that $p(s_3|s_1)$ or $p(s_3|s_2)$ is Gaussian.

From a GP prediction standpoint, the distribution $p(\mathbf{y}|\mathbf{x},s_i)$ can be estimated from mixtures $\mathbf{x}(n)$ and a few known source values. This idea can be viewed as an interpolation problem, given certain reference values and the SOS defined by the GP [11]. Once obtained the predictive distribution $p(\mathbf{y}|\mathbf{x},s_i)$, the estimated sources $\mathbf{y}(n)$ can be obtained using Markov Chain Monte Carlo (MCMC) methods [12].

From the perspective of the bilinear mixture problem, this approach can contribute with an interesting theoretical understanding, but imposes some practical difficulties from the standpoint of blind separation, since the knowledge of certain reference values is difficult to obtain. In light of this, this method is classified as a semi-blind approach. Although this might seem restrictive, in certain cases it can be shown to be feasible, such as in the context of chemical sensor arrays, where the solution concentration measures are preceded by certain calibration points, which can be interpreted here as reference values [6]. Furthermore, the reference values can admit certain degree of error [11], which prompts the employment of this idea along with other methods that are able to provide ‘coarse’ estimates of sources.

In the GP formulation, the temporal information is crucial. Hence, we compose vectors $\mathbf{y}_i(\mathbf{l}) = [y_i(l_1) \dots y_i(l_L)]^T$, for $i = 1, \dots, N$, which is the i -th output at L time instants $\mathbf{l} = \{l_1, \dots, l_L\}$. Suppose now that the sources are known at P time instants $\mathbf{j} = \{j_1, \dots, j_P\}$, i.e., the vectors $\mathbf{s}_i(\mathbf{j}) = [s_i(j_1) \dots s_i(j_P)]^T$, for $i = 1, \dots, N$, are given (and they will be referred to as ‘target’ values). Also, the mixtures $\mathbf{x}(l_1), \dots, \mathbf{x}(l_L), \mathbf{x}(j_1), \dots, \mathbf{x}(j_P)$ are known (at the time instants \mathbf{l} and \mathbf{j}). Then, the goal is to predict the probability density function of $\mathbf{y}_i(\mathbf{l})$, the outputs at time instants \mathbf{l} .

For mathematical simplicity, we define the following entities for the case of $N = 2$ sources: $\mathbf{y}(\mathbf{l}) = [\mathbf{y}_1^T(\mathbf{l}) \mathbf{y}_2^T(\mathbf{l})]^T$ is the column vector with all outputs at time instants \mathbf{l} ; similarly, $\mathbf{s}(\mathbf{j}) = [\mathbf{s}_1^T(\mathbf{j}) \mathbf{s}_2^T(\mathbf{j})]^T$ is the column vector with the targets; $\mathbf{X}(\mathbf{l}) = [\mathbf{x}(l_1) \dots \mathbf{x}(l_L)]$ and $\mathbf{X}(\mathbf{j}) = [\mathbf{x}(j_1) \dots \mathbf{x}(j_P)]$ are the mixtures matrices for time instants \mathbf{l} and \mathbf{j} , respectively. Hence, the conditional probability of $\mathbf{y}(\mathbf{l})$ can be denoted as $p(\mathbf{y}(\mathbf{l})|\mathbf{X}(\mathbf{l}),\mathbf{X}(\mathbf{j}),\mathbf{s}(\mathbf{j}))$.

Based on the priors and on the knowledge that $p(\mathbf{y}(\mathbf{l})|\mathbf{X}(\mathbf{l}),\mathbf{X}(\mathbf{j}),\mathbf{s}(\mathbf{j}))$ is Gaussian, we can write the following GP in the case of $N = 2$ sources:

$$\mathbf{y}(\mathbf{l})|\mathbf{X}(\mathbf{j}), \mathbf{s}(\mathbf{j}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(\mathbf{X}(\mathbf{l}),\mathbf{X}(\mathbf{l}))), \quad (7)$$

where $\mathbf{K}(\mathbf{X}(\mathbf{l}),\mathbf{X}(\mathbf{l}))$ is the covariance matrix of size $2L \times 2L$ in function of $\mathbf{X}(\mathbf{l})$. Eq. (7) means that the distribution for $\mathbf{y}(\mathbf{l})$ is jointly Gaussian with zero mean and covariance matrix $\mathbf{K}(\mathbf{X}(\mathbf{l}),\mathbf{X}(\mathbf{l}))$. However, $\mathbf{K}(\mathbf{X}(\mathbf{l}),\mathbf{X}(\mathbf{l}))$ must be defined so that it is able to encompass the temporal structure of the sources and the mutual independence information. Hence, in this work, we propose the use of a

block-diagonal covariance matrix of the type:

$$\mathbf{K}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{l})) = \begin{bmatrix} \mathbf{K}_{y_1}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{l})) & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{y_2}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{l})) \end{bmatrix} \quad (8)$$

where $\mathbf{K}_{y_i}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{l}'))$ is a covariance submatrix of dimension $L \times L$, whose element of the l -th row and l' -th column is given by the squared-exponential (SE) function [11]:

$$\mathbf{K}_{y_i}(\mathbf{X}(l), \mathbf{X}(l')) = \gamma_i \exp\left(\frac{-1}{2}(\mathbf{x}(l) - \mathbf{x}(l'))^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}(l) - \mathbf{x}(l'))\right), \quad (9)$$

with $\boldsymbol{\Sigma}_i = \sigma_i^2 \mathbf{I}_M$, being σ_i^2 the estimated variance of source i , γ_i a scale factor, and \mathbf{I}_M the identity matrix of order M (here $M=3$).

Hence, from the GP classical results [11], the predictive distribution can be obtained by

$$p(\mathbf{y}(\mathbf{l}) | \mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{j}), \mathbf{s}(\mathbf{j})) \sim \mathcal{N}(\mathbf{K}(\mathbf{X}(\mathbf{j}), \mathbf{X}(\mathbf{l})) \boldsymbol{\Gamma}^{-1} \mathbf{s}(\mathbf{j}), \mathbf{K}(\mathbf{X}(\mathbf{j}), \mathbf{X}(\mathbf{j})) - \mathbf{K}(\mathbf{X}(\mathbf{j}), \mathbf{X}(\mathbf{l})) \boldsymbol{\Gamma}^{-1} \mathbf{K}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{j}))). \quad (10)$$

where $\boldsymbol{\Gamma} = \mathbf{K}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{l})) + \boldsymbol{\Phi}$, with $\boldsymbol{\Phi} = [\epsilon_1^2 \mathbf{I}_L \ \mathbf{0}; \ \mathbf{0} \ \epsilon_2^2 \mathbf{I}_L]$, a diagonal matrix which is able to consider the degree of error (or uncertainty) in the target samples by means of noise variances ϵ_i^2 associated to each source. In that sense, the targets $\mathbf{s}(\mathbf{j})$ can admit certain level of error, and the accuracy of the predictive distribution will depend on the estimation of the error variances ϵ_i^2 . Eq. (10) shows that $p(\mathbf{y}(\mathbf{l}) | \mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{j}), \mathbf{s}(\mathbf{j}))$ is Gaussian distributed with mean $\mathbf{K}(\mathbf{X}(\mathbf{j}), \mathbf{X}(\mathbf{l})) \boldsymbol{\Gamma}^{-1} \mathbf{s}(\mathbf{j})$ and covariance matrix $\mathbf{K}(\mathbf{X}(\mathbf{j}), \mathbf{X}(\mathbf{j})) - \mathbf{K}(\mathbf{X}(\mathbf{j}), \mathbf{X}(\mathbf{l})) \boldsymbol{\Gamma}^{-1} \mathbf{K}(\mathbf{X}(\mathbf{l}), \mathbf{X}(\mathbf{j}))$.

The complexity of the method exponentially increases with the number of considered time instants (for both target, \mathbf{j} , and predicted samples, \mathbf{l}). In view of this, for implementation purposes, a Cholesky decomposition can be used to simplify the inversion of $\boldsymbol{\Gamma}$ and to generate the predicted samples of $\mathbf{y}(\mathbf{l})$ [11]. The variables $\tilde{\boldsymbol{\theta}} = \{\gamma_1, \gamma_2, \sigma_1, \sigma_2, \epsilon_1, \epsilon_2\}$ are called hyperparameters of the GP and can be adjusted by maximizing the marginal likelihood given the targets [11]. Once the predicted source samples are obtained, the separation matrix $\tilde{\mathbf{W}}$ can be directly estimated via supervised approaches [1], if necessary.

3.2 Maximization of the Marginal Likelihood

For the two-stage case, using Bayes' rule, $p(\mathbf{y} | \mathbf{z}) = p(\mathbf{z} | \mathbf{y}) p(\mathbf{y}) / p(\mathbf{z})$, $p(\mathbf{z})$ – and, consequently, $p(\mathbf{y})$ – can be Gaussian, but not necessarily (note that the output $\mathbf{z}(n)$ will be associated with a Gaussian distribution if the values of $\boldsymbol{\alpha}$ are properly adjusted). Nonetheless, when $p(\mathbf{z})$ is Gaussian, $p(\mathbf{y} | \mathbf{z})$ will also be. In this case, the marginal likelihood $p(\mathbf{y} | \mathbf{z})$ can be maximized with respect to the hyperparameters of the system, forcing it towards the Gaussian distribution.

Using the previously defined notation, for a given set of observations $\mathbf{X}(\mathbf{l})$ and for given values of $\boldsymbol{\alpha}$ and \mathbf{W} , it is possible to obtain $\mathbf{Z}(\mathbf{l})$ and, in the sequence,

$\mathbf{y}(\mathbf{l})$ – see Eq. (5). Thus, we wish that $p(\mathbf{y}(\mathbf{l})|\mathbf{Z}(\mathbf{l}))$ be described according to a GP, i.e.,

$$\mathbf{y}(\mathbf{l}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(\mathbf{Z}(\mathbf{l}), \mathbf{Z}(\mathbf{l}))), \quad (11)$$

being $\mathbf{K}(\mathbf{Z}(\mathbf{l}), \mathbf{Z}(\mathbf{l}))$ a block-diagonal covariance matrix, as defined in Eq. (8) with inputs $\mathbf{Z}(\mathbf{l})$ instead of $\mathbf{X}(\mathbf{l})$.

By denoting $\boldsymbol{\theta} = \{\gamma_1, \gamma_2, \sigma_1, \sigma_2, \boldsymbol{\alpha}, \mathbf{W}\}$ the vector of all hyperparameters, it is possible to write the log likelihood

$$\mathcal{L} = \frac{-1}{2} \mathbf{y}^T(\mathbf{l}) \mathbf{K}^{-1}(\mathbf{Z}(\mathbf{l}), \mathbf{Z}(\mathbf{l})) \mathbf{y}(\mathbf{l}) - \frac{1}{2} \log |\mathbf{K}(\mathbf{Z}(\mathbf{l}), \mathbf{Z}(\mathbf{l}))| - \frac{(L+1)}{2} \log 2\pi, \quad (12)$$

being $\mathbf{K}(\mathbf{Z}(\mathbf{l}), \mathbf{Z}(\mathbf{l}))$ a function of $\mathbf{Z}(\mathbf{l})$ and $\boldsymbol{\theta}$.

Hence, by maximizing \mathcal{L} with respect to the hyperparameters $\boldsymbol{\theta}$, we hope to obtain a conditional probability $p(\mathbf{y}(\mathbf{l})|\mathbf{Z}(\mathbf{l}))$ that is Gaussian and, in this case, the optimal parameters $\boldsymbol{\alpha}$ and \mathbf{W} are the solution of the separation problem. Note that, in this case, the two-stages are solved simultaneously. As in the predictive case, the complexity grows exponentially with the number of samples and large data sets should be avoided.

The effectiveness of this approach comes from the fact that different time delays are being compared in the GP by construction. In fact, since we are considering only SOS, the temporal structure is essential for separation. This idea is encompassed by the block diagonal covariance function $\mathbf{K}(\mathbf{Z}(\mathbf{l}), \mathbf{Z}(\mathbf{l}))$ which allows temporal correlation between samples of the same output, but applies decorrelation (in different time delays) for different outputs. This is also valid for the predictive case. In the sequence, we compare the performance of the two proposed GP approaches.

4 Simulation Results

In order to test the two proposed methods, we consider a simulation scenario for the 2 source and 3 mixture case. The sources are assumed to be two colored Gaussian distributed, obtained from *i.i.d.* Gaussian sampled, that are temporally colored by the finite impulse response (FIR) filters with impulse response $h_1(z) = 1+0.6z^{-1}-0.3z^{-2}$ and $h_2(z) = 1-0.8z^{-1}$, each source separately. We considered the following mixing matrix,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{bmatrix} = \begin{bmatrix} -0.8049 & 0.0938 & -0.0292 \\ -0.4430 & 0.9150 & 0.6006 \\ -1.4434 & -0.4997 & 0.5180 \end{bmatrix}, \quad (13)$$

whose mixtures were obtained using Eq. (3).

The predictive distribution is the first approach to be tested. From a set of 100 samples of the mixtures $\mathbf{x}(n)$, we consider the cases in which $P = 10$ and $P = 5$ targets $\hat{\mathbf{s}}(j)$ are available (with and without noise). For this purpose, within the time window of 100 samples, P time instants are randomly picked,

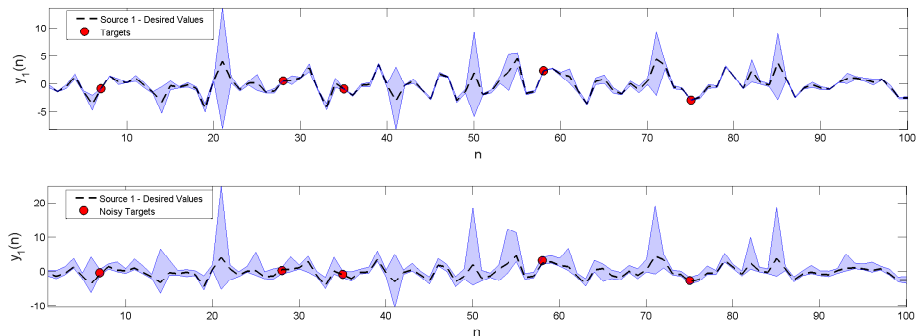


Fig. 1. Predictive - Noise (top) and Noiseless (bottom) Targets. The shaded area denotes the prediction uncertainty.

for which the sources are assumed to be known. Hence, the objective is to correctly predict the system output $\mathbf{y}(\mathbf{l})$, given $\mathbf{s}(\mathbf{j})$, $\mathbf{X}(\mathbf{j})$ and $\mathbf{X}(\mathbf{l})$. Additionally, a perturbation can be considered with additive white Gaussian noise (to simulate targets with certain degree of error) with variance $\sigma^2 = 1e-1$, resulting a SNR level of 14.9 dB. 1000 independent experiments were considered, from which each realization encompassed a new set of mixtures and targets. We start by adjusting the hyperparameters $\tilde{\theta}$, which were empirically chosen to be $\gamma_1=10$, $\gamma_2=30$, $\sigma_1=4.5$, $\sigma_2=5.48$, $\epsilon_1=\epsilon_2=0$. For illustrative purposes, we display in Fig. 1 one realization of the predicted distribution for $P = 5$ and for noiseless (top) and noisy (bottom) targets. The dashed line represents the desired output and the red circle the known targets. The shaded area comprises the region where a realization of $p(y|\mathbf{x}, \mathbf{s})$ could fall (more precisely, the region denotes the predicted mean value ± 3 times the predicted standard deviation for each time instant). Hence, a large shaded area means the prediction is less accurate. It is possible to note for the noiseless targets that, with exception of certain small regions, the predicted region is small and falls really close to the desired output. Although we do not show here, for $P = 10$ noiseless targets, the prediction is very accurate. Notwithstanding, when the provided targets present certain level of error, the accuracy is reduced for all time instants; as shown in Fig. 1 (bottom), the shaded area is increased with respect to the noiseless case. In the noisy case, we have chosen $\epsilon_1^2 = \epsilon_2^2 = \sigma^2 = 1e-1$.

In order to evaluate the quality of the predicted samples, we measured the *signal-to-interference ratio* (SIR), defined as $\text{SIR} = 10 \log (E[y_i(n)^2]/E[(s_i(n) - y_i(n))^2])$. Since we are interested in the mean performance, it was considered only the predicted mean for computing the SIR (and averaged over all realizations). The results are displayed in Tab. 1. As previously discussed, in the noiseless case ($\sigma^2 = 0$), we can see high values of SIR for both values of P (above 50 dB), which means that the sources are recovered with small error. However, the reduction of the number of targets causes the performance to slightly decrease.

Table 1. Mean SIR [dB] for the maximum likelihood and the predictive approaches

	Predictive				Maximum Likelihood
	$\sigma^2 = 0$		$\sigma^2 = 1e-1$		
	$P = 10$	$P = 5$	$P = 10$	$P = 5$	
Source 1	73.41	64.18	38.08	24.78	13.70
Source 2	74.30	65.35	40.59	27.81	16.26

In the case of noisy targets, the SIR is reduced even more, to approximately 40 dB for $P = 10$ and to 25 dB for $P = 5$. In that sense, it is clear that the higher the number of targets and its quality, the higher is the SIR of the prediction.

For the second approach, we consider the maximization of the log likelihood, in which we wish to adjust the hyperparameters θ to its optimal values. The adaptation can be performed according to several optimization methods [12]. However, in this work, we adopt the metaheuristic called *Differential Evolution* (DE) [13] for optimization, which is an efficient technique to explore the search space and avoid convergence to local optima (for more details, please refer to [13]). The chosen DE parameters are $N_P=300$ (population size), $F=0.7$, $CR=0.7$ and 100 iterations.

In this case, we consider a set of 45 samples of $\mathbf{x}(n)$ and performed the adaptation of θ via the DE method. This procedure was repeated 10 times, for different realizations of $\mathbf{x}(n)$. Even with a reduced number of samples, each optimization leads to good results, with mean value of $\alpha_1=0.0456$ and $\alpha_2=1.1498$, which are close to the ideal values $\alpha_1=0.0564$ and $\alpha_2=1.1595$, respectively. The mean SIR values can be seen in Tab. 1. Although this SIR value is not quite impressive, the performance is good if the reduced number of samples is taken into account.

5 Conclusions

In this work, we proposed two GP formulations to solve the overdetermined bilinear mixing problem concerning the one- and two-stage approaches. Our propositions are both based on the prior that the sources are mutually independent Gaussian distributed, however, they differ in the application of the GP method. We have shown that, for the GP formulation consistency, it was required in the former case the conditioning of the output distribution to the knowledge of certain reference samples, named targets, what prompted us to adopt a predictive based GP approach. In the second case, the two-stage approach allowed the adjustment of the hyperparameters via the maximization of the marginal likelihood. As shown in the simulation results, the predictive approach tends to provide better results, depending on the quality and number of the targets, however, the maximization of the marginal likelihood is also able to perform the

separation of the sources. Although these GP methods present certain algorithmic complexity, the computational burden can be reduced by using a relatively small number of samples.

For future works, we consider a deeper analysis of the requirements for separation – e.g., a spectral density analysis of the temporal structure – and the extension of these theoretical analysis to noisy scenarios and to other classes of mixing systems, like the Linear Quadratic.

Acknowledgements. This work was partly supported by FAPESP (2013/14185-2, 2015/23424-6), CNPq and ERC project 2012-ERC-AdG-320684 CHESSE.

References

1. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press (2010)
2. Duarte, L. T., Jutten, C., Moussaoui, S.: A Bayesian Nonlinear Source Separation Method for Smart Ion-Selective Electrode Arrays. *IEEE Sensors Journal* 9(12), 1763–1771 (2009)
3. Hosseini, S., Jutten, C.: On the Separability of Nonlinear Mixtures of Temporally Correlated Sources. *IEEE Signal Processing Letters* 10(2), 43–46 (2003)
4. Hosseini, S., Deville, Y.: Blind Separation of Linear-Quadratic Mixtures of Real Sources Using a Recurrent Structure. *International Work-Conference on Artificial Neural Networks*, 241–248 (2003)
5. Merrikh-Bayat, F., Babaie-Zadeh, M., Jutten, C.: Linear-Quadratic Blind Source Separating Structure for Removing Show-Through in Scanned Documents. *International Journal on Document Analysis and Recognition (IJ DAR)* 14(4), 319–333 (2011)
6. Duarte, L. T. and Jutten, C.: Design of Smart Ion-Selective Electrode Arrays Based on Source Separation Through Nonlinear Independent Component Analysis. *Oil & Gas Science and Technology* 69(2), 293–306 (2014)
7. Abed-Meraim, K., Belouchiani, A., Hua, Y.: Blind Identification of a Linear-Quadratic Mixture of Independent Components Based on Joint Diagonalization Procedure. *IEEE ICASSP* 5, 2718–2721 (1996)
8. Castella, M., Inversion of Polynomial Systems and Separation of Nonlinear Mixtures of Finite-Alphabet Sources. *IEEE Transactions on Signal Processing* 56(8), 3905–3917 (2008)
9. Duarte, L.T., Ando, R.A., Attux, R.R.F., Deville, Y., Jutten, C.: Separation of Sparse Signals in Overdetermined Linear-Quadratic Mixtures. *International Conference on Latent Variable Analysis and Signal Separation*, 239–246 (2012)
10. Ando, R.A., Duarte, L.T., Attux, R.R.F.: Blind Source Separation for Overdetermined Linear Quadratic Mixtures of Bandlimited Signals. *IEEE International Telecommunications Symposium (ITS)* (2014)
11. Rasmussen, C.: *Gaussian Processes for Machine Learning*. Citeseer (2006)
12. Deville, Y., Duarte, L. T.: An Overview of Blind Source Separation Methods for Linear-Quadratic and Post-Nonlinear Mixtures. *International Conference on Latent Variable Analysis and Signal Separation*, 155–167 (2015)
13. Price, K., Storn, R., Lampinen, J.: *Differential Evolution: A Practical Approach to Global Optimization*. Springer (2005)