Markovian source separation in post-nonlinear mixtures

Anthony Larue[1], Christian Jutten[1] and Shahram Hosseini[2⋆]

[1] Institut National Polytechnique de Grenoble
Laboratoire des Images et des signaux (CNRS, UMR 5083)
F-38031 Grenoble Cedex, France
[2] Universite Paul Sabatier de Toulouse
Laboratoire d'Acoustique, Metrologie, Instrumentation
F-31062 Toulouse, France

Abstract. In linear mixtures, priors, like temporal coloration of the sources, can be used for designing simpler and better algorithms. Especially, modeling sources by Markov models is very efficient, and Markov source separation can be achieved by minimizing the conditional mutual information [1, 2]. This model allows to separate temporally colored Gaussian sources. In this paper, we extend this result for post-nonlinear mixtures (PNL) [3], and show that algorithms based on a Markov model of colored sources leads to better separation results than without prior, *i.e.* assuming iid sources. The paper contains theoretical developments, and experiments with auto-regressive (AR) source mixtures. PNL algorithms for Markov sources point out a performance improvement of about 7dB with respect to PNL algorithms for iid sources.

1 Introduction

First blind source separation methods, based on statistical independence of random variables and using higher (than 2) order statistics, does not take into account the temporal relation between successive source samples. However, early works [4, 5, 6, 7] show that it is possible to exploit source temporal correlation by considering simultaneously a few variance-covariance matrices, with various delays. In recent works [1, 2], for linear mixtures, we proposed Markov models of the sources for taking into account the temporal relation between samples. In this paper, we generalize the method to post-nonlinear mixtures (PNL). The paper is organized as follows: Section 2 provides the main theoretical foundations, Section 3 details two practical issues of the algorithm, Section 4 reports the experiments, before the conclusions in Section 5.

⋆ This work has been partly funded by the European project BLind Source Separation and applications (BLISS, IST-1999-14190).

2 Theoretical assessments

2.1 Mixing and separating models

Post-nonlinear (PNL) mixtures of n sources, represented by the figure 1, are characterized by a linear instantaneous mixtures, associated to a mixing matrix $\mathbf{A}$, followed by component-wise nonlinear distortions f_i. Considering a suited separating structure (Fig. 1, right side), it can be shown [3] that, under mild conditions[3], output independence leads to source separation, with the same indeterminacy than linear mixtures. The vectorial notation $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T$, also applied for e, x, z and y.

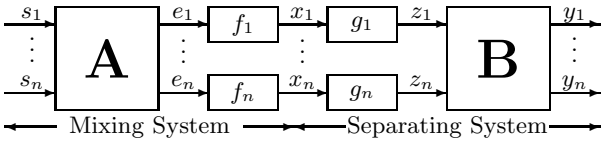

Fig. 1. The mixing-separating system for PNL mixtures.

Each source $s_i(t)$, $i = 1, \ldots, n$ is assumed to be temporally correlated (colored). It is modeled by a q-order Markov model, *i.e.* :

$$p_{s_i}(s_i(t)|s_i(t-1), \cdots, s_i(1)) = p_{s_i}(s_i(t)|s_i(t-1), \cdots, s_i(t-q)) \tag{1}$$

where p_{s_i} denotes the pdf of the random variable s_i.

2.2 Independence criteria

Since output independence leads to source separation, a possible approach for separating source is to consider a criterion measuring the independence of the output $\mathbf{y}$. Following [8, 2], one can use the conditional mutual information of $\mathbf{y}$, denoted by I :

$$I = \int p_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1), \cdots, \mathbf{y}(t-q)) \log \frac{p_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1), \cdots, \mathbf{y}(t-q))}{\prod_{i=1}^{n} p_{y_i}(y_i(t)|y_i(t-1), \cdots, y_i(t-q))} d\mathbf{y} \tag{2}$$

which is always nonnegative, and zero if and only if the variables $w_i(t) = y_i(t)|y_i(t-1), \cdots, y_i(t-q)$ are statistically independent for $i = 1, \cdots, n$, *i.e.* the signals $y_i(t)$, $i = 1, \cdots, n$ are independent Markovian process. Using the expectation operator $E[.]$, we can write:

$$I = E[\log p_{\mathbf{y}}(\mathbf{y}(t)|\mathbf{y}(t-1), \cdots, \mathbf{y}(t-q))] - \sum_{i=1}^{n} E[\log p_{y_i}(y_i(t)|y_i(t-1), \cdots, y_i(t-q))] \tag{3}$$

[3] $\mathbf{A}$ is regular or full rank, with at least two non zero entries per row or per column, and f_i are invertible

Considering the separation structure (Fig. 1), where $\mathbf{y}(t) = \mathbf{B}\mathbf{z}(t)$ and $z_i(t) = g_i(\theta_i, x_i(t))$,[4] Eq. (3) becomes:

$$\begin{aligned} I = & E[\log p_{\mathbf{x}}(\mathbf{x}(t)|\mathbf{x}(t-1), \cdots, \mathbf{x}(t-q))] \\ & -E\left[\log\left|\prod_{i=1}^{n} \frac{\partial g_i(\theta_i, x_i(t))}{\partial x_i(t)}\right|\right] - \log|\det(\mathbf{B})| \\ & -\sum_{i=1}^{n} E[\log p_{y_i}(y_i(t)|y_i(t-1), \cdots, y_i(t-q))] \end{aligned} \tag{4}$$

The first term being independent of $\mathbf{B}$ and $\Theta = [\theta_1, \ldots, \theta_n]$, the separation structure can be estimated by minimizing:

$$\begin{aligned} J(\mathbf{B}, \Theta) = & -\sum_{i=1}^{n} E\left[\log\left|\frac{\partial g_i(\theta_i, x_i(t))}{\partial x_i(t)}\right|\right] - \log|\det(\mathbf{B})| \\ & -\sum_{i=1}^{n} E[\log p_{y_i}(y_i(t)|y_i(t-1), \cdots, y_i(t-q))] \end{aligned} \tag{5}$$

In practice, under the ergodicity conditions, the mathematical expectation (5) can be estimated by a time averaging, denoted $\widehat{J}(\mathbf{B}, \Theta)$, which requires the estimation of the conditional densities of the estimated sources. Asymptotically, extending the results for linear mixtures of Markovian sources [2], the equivalence of the mutual information minimization method with the Maximum Likelihood method still holds for PNL mixtures of Markovian sources.

2.3 Estimating equation

Estimation of $\mathbf{B}$ and Θ can be done by minimizing $J(\mathbf{B}, \Theta)$. Using a gradient method, one obtain two sets of estimating equations, which are the gradients of $J(\mathbf{B}, \Theta)$ with respect to $\mathbf{B}$ and with respect to $\theta_i,\ i = 1 \ldots n$, *i.e.*:

$$\frac{\partial J(\mathbf{B}, \Theta)}{\partial \mathbf{B}} = -\mathbf{B}^{-T} + E\left[\sum_{l=0}^{q} \psi_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1), \ldots, \mathbf{y}(t-q))\mathbf{z}^T(t-l)\right] \tag{6}$$

$$\begin{aligned} \frac{\partial J(\mathbf{B}, \Theta)}{\partial \theta_i} = & -E\left[\frac{\partial^2 g_i(\theta_i, x_i(t))}{\partial x_i(t)\partial \theta_i}\left(\frac{\partial g_i(\theta_i, x_i(t))}{\partial x_i(t)}\right)^{-1}\right] + \\ & E\left[\sum_{j=1}^{n} b_{ji} \sum_{l=0}^{q} \psi_{y_j}^{(l)}(y_j(t)|y_j(t-1), \ldots, y_j(t-q)) \frac{\partial g_i(\theta_i, x_i(t-l))}{\partial \theta_i}\right] \end{aligned} \tag{7}$$

where we define $q+1$ conditional score functions of a random variable w as $\psi_w^{(l)}(w_0|w_1, \ldots, w_q) = -\frac{\partial}{\partial w_l} \log p_w(w_0|w_1, \ldots, w_q), l = 0, \ldots, q$, and we denote

[4] $g_i(\theta_i, x_i(t))$ is a parametric model of $g_i(.)$, where θ_i can represent a set of parameters

$\psi_{\mathbf{y}}^{(l)}(\mathbf{y}(t)|\mathbf{y}(t-1),\ldots,\mathbf{y}(t-q))$ the n-th dimension vector whose i-th component is $\psi_{y_j}^{(l)}(y_j(t)|y_j(t-1),\ldots,y_j(t-q))$. One can remark that the gradients of the mutual information require first-order and second-order derivatives of the nonlinear mappings g_i's.

3 Algorithm

In this section, we focus on two points for practically implementing the algorithm. The first one concerns the estimation of conditional score functions. The second one is a trick for computing a good initialization point of the algorithm, which leads to enhanced speed of convergence. The algorithm is as follows :

1. initialization of the separating matrix $\mathbf{B}$ and the nonlinear parameters Θ
2. estimation of the conditional score functions
3. computation of the gradients (6) and (7)
4. updating of $\mathbf{B}$ and Θ according to a gradient descent
5. computation of the linearized observations z_i and the estimated sources y_i
6. normalization step

We iterate from 2 to 6 until convergence. The normalization step is required for taking into account scale indeterminacies in $\mathbf{B}$ and in g_i's estimations.

3.1 Estimating the conditional score functions

For estimating the conditional score functions, we can firstly estimate the conditional densities and compute then the conditional score functions by computing the gradient of their logarithms. For a q-order Markovian source, the estimation of the conditional densities may be done using the estimation of the joint pdf of $q+1$ successive samples of each source by a kernel method, which is very time consuming and requires a lot of data. It must be also noticed that the distribution of the data in $(q+1)$-th dimensional space is sparse (curse of dimensionality) and not symmetric because of the temporal correlation between the samples. Thus, one should either use non symmetrical kernels or apply a pre-whitening transformation on data.

Recently, Pham [8] has proposed another algorithm for computing the conditional score functions. The method starts with a pre-whitening stage for obtaining non correlated temporal data. Pham suggests also that the time pre-whitening can allow to reduce the dimension of the used kernels because a great part of the dependence between the variables is cancelled. The influence of the pre-whitening on the estimation of the score functions is computed and will be later compensated using an additive term. Afterwards, the joint entropies of whitened data are estimated using a discrete Riemann sum and the third order cardinal spline kernels. The conditional entropies, defined as

$$H(y_i(t)|y_i(t-1),\cdots,y_i(t-q)) = -E[\log p_{y_i}(y_i(t)|y_i(t-1),\cdots,y_i(t-q))] \quad (8)$$

are computed by estimating the joint entropies:

$$H(y_i(t)|y_i(t-1),\cdots,y_i(t-q)) = H(y_i(t),y_i(t-1),\cdots,y_i(t-q)) - H(y_i(t-1),\cdots,y_i(t-q)) \quad (9)$$

The estimator $\hat{H}(y_i(t)|y_i(t-1),\cdots,y_i(t-q))$ is a function of the observations $y_i(1),\cdots,y_i(N)$, where N is the sample number. The l-th component of the conditional score function in a sample point $y_i(n)$ is computed as:

$$\hat{\psi}_{y_i}^{(l)}(y_i(t)|y_i(t-1),\cdots,y_i(t-q))|_{t=n} = N\frac{\partial \hat{H}(y_i(t)|y_i(t-1),\cdots,y_i(t-q))}{\partial y_i(n-l+1)} \quad (10)$$

The method is very powerful and provides a quite good estimation of the conditional score functions.

3.2 Initializing the nonlinear function

The convergence speed can been enhanced by choosing a relevant starting point, especially for the parameters of the functions g_i. As presented in [9, 10], the idea is based on two remarks: (i) each mixture of sources, e_i, is a random variable close to Gaussian, and (ii) due to the nonlinear distortions, the random variable, $x_i = f_i(e_i)$ is farther to the Gaussian than e_i. Consequently, the nonlinear transform $\hat{g}_i = \Phi^{-1} \circ F_{x_i}$, where Φ is the cumulative density function of the Gaussian and F_{x_i} is the cumulative density function of x_i, transforms x_i to a Gaussian random variable z_i. If x_i is exactly a Gaussian random variable, then $\hat{g}_i = f_i^{-1}$; if it is approximately Gaussian, it is a rough estimation of f_i^{-1}. Thus, we estimate the initial parameter θ_i by minimization of the mean square error between $\hat{g}_i$ and $g_i(\theta_i, .)$ Since the Gaussian assumption of x_i is not completely fulfilled, we used the above idea for computing a good starting point of the algorithm.

4 Experiments

The aim of this section is to check if a Markov model of the sources is able to improve the performance of the algorithm. We will consider two kinds of colored sources, modeled both by first order auto-regressive (AR) filters, whose input is an iid random signal with either a Gaussian or a uniform distribution. We restrict the study to post-nonlinear mixtures of 2 sources.

We compared two algorithms, the first one with 1-st order Markov model and the second one with order 0, *i.e.* without modeling source temporal correlation.

Each experiment is repeated about 16 times, with random choice of the AR coefficients, of the mixing matrix and of the nonlinear parameters :

- AR coefficients, ρ_i, $i = 1, 2$, are chosen so that $0.2 <| \rho_i |< 0.9$ and $| \rho_1 - \rho_2 | > 0.2$, since the source spectra must be different.

– The main diagonal entries a_{ii} of mixing matrix $\mathbf{A}$ are enforced to 1, while the other are chosen in the range $0.2 <\mid a_{ij} \mid< 1$. This choice allows to avoid mixing matrices close to diagonal matrices, which provide post-nonlinear observations x_i which would be still independent.

– Each nonlinear function f_i, $i = 1, 2$, defined by the relation (11) (see below) with parameter β_i, is chosen so that $0.1 < \beta_i < 5$.

4.1 Simple nonlinear functions

In this first set of experiments, we use three nonlinear invertible functions, f_i, depending on one parameter β, and their inverses, g_i, too. Since we can compute the theoretical parameter of g_i, this experiment allows to measure the parametric error in the estimation of the nonlinear function.

Example 1. The main advantage of this nonlinear function is to have a very simple inverse, expression of which is linear with respect to the parameter θ:

$$f(\beta, e) = \frac{sign(e)}{2\beta}(-1 + \sqrt{1 + 4\beta|e|}) \quad \Rightarrow \quad g(\theta, x) = x + \theta x|x| \qquad (11)$$

Examples 2 *and* 3. We can defined two others saturating non linear functions :

$$f(\beta, e) = sign(e) \times (\tanh(|e|))^{(1/\beta)} \Rightarrow \; g(\theta, x) = sign(x) \times \tanh^{-1}(|x|^{\theta}) \qquad (12)$$

$$f(\beta, e) = \frac{\beta e}{\beta + |e|} \qquad \Rightarrow \qquad g(\theta, x) = \frac{\theta x}{\theta - |x|} \qquad (13)$$

The concavity of the function (12) can be varied with the parameter, but, contrary to the function (13) the magnitude of the saturation is not adjustable.

The figure 2 shows the shapes of the three functions defined by (11),(12) and (13) for three values of β: 0.1 in solid line, 1 in dashed line and 5 in dotted line.

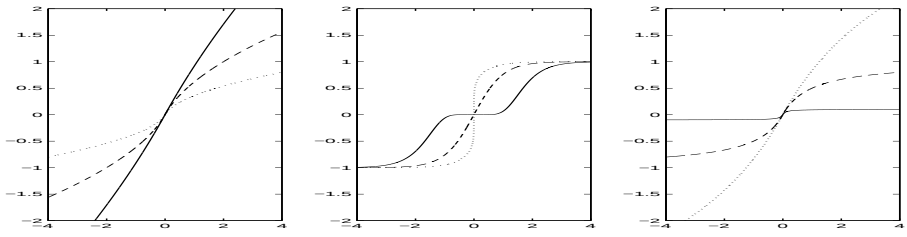

Fig. 2. Examples of simple nonlinear mappings. Left: non-linearity (11); middle: non-linearity (12); right: non-linearity (13).

Eq. (7) depends on the parametric models $g_i(\theta_i, x)$. In the experiments, we will use either specific models or polynomials. However, the derivatives with respect to parameters θ_i are evident and will not be given in this paper.

4.2 Polynomial approximation of inverse function

We also used a more general and more flexible model for estimating the nonlinear functions $g_i(\theta_i, .)$, based on a polynomial expression of $g_i(\theta_i, x) = \sum_{m=0}^{P} \theta_{im} x^m$. This expression being linear with respect to the parameter θ_{ik}, the gradient (7) becomes :

$$\frac{\partial J(\mathbf{B}, \Theta)}{\partial \theta_{ik}} = -E\left[\frac{k x_i(t)^{k-1}}{\sum_{m=0}^{P} m\theta_{im} x_i(t)^{m-1}}\right] + E\left[\sum_{j=1}^{n} b_{ji} \sum_{l=0}^{q} \psi_{y_j}^{(l)}(y_j(t)|y_j(t-1), \ldots, y_j(t-q)) x_i(t-l)^k\right] \tag{14}$$

The polynomial model is well suited to the inversion of saturating nonlinear transformation. For more general mappings, we could extend the model by adding rational powers in the polynomial.

4.3 Results

The following tables give the mean residual cross-talk, as well as max and min between brackets, expressed in dB for 16 random configurations. Since s_i and y_i are unit power signals, the residual cross-talk is $E[(y_i - s_i)^2]$.

With simple functions. We compare two algorithms based on minimization of the mutual information (MIM):

- the PNL algorithm developed in this paper, denoted MIM Markov 1, for $q = 1$ (order 1, Markov source),
- the PNL algorithm, denoted MIM iid, which does not take into account the source time structure (in fact, it correspond to MIM Markov 0, and is a special case of the algorithm developed in this paper).

Sources	MIM Markov 1	MIM iid
AR Gaussian	(-20.7) -16.2 (-14.1)	(-18.4) -8.9 (-5.0)
AR uniform	(-23.1) -16.1 (-13.6)	(-23.7) -14.5 (-12.3)

We remark that the Markov model of the source improves the performance: about 7 dB, on the average for Gaussian input, and 1.6 dB for uniform input. The improvement is then very sensitive, but much less important than for linear mixtures [2] (-34.5 dB for Markov model, and -24.2 dB for iid algorithm for uniformly distributed innovation process). This is mainly due to the nonlinear part of PNL: a small error in the nonlinear parameter estimation can imply a poor estimation of the separating matrix $\mathbf{B}$. We also remark that, like in linear mixtures, time correlation modeling (here with Markov models) allows to separate Gaussian sources.

With polynomials. The functions g_i's are now modeled by 7-degree polynomials. We again compare the two algorithms, MIM Markov 1 and MIM iid, according to the notations of the previous paragraph.

Sources	MIM Markov 1	MIM iid
AR Gaussian	(-16.7) -14.0 (-11.8)	(-12.2) -7.8 (-5.1)
AR uniform	(-17.0) -14.3 (-11.9)	(-15.3) -11.1 (-5.4)

5 Conclusions

In this paper, we presented an algorithm modeling the temporal relation between successive source samples with a Markovian model, in post nonlinear mixtures. For various parametric model of the nonlinear mappings g_i's, Markov model of the sources provides a performance improvement for separating first order autoregressive sources. The computing time increases as 3^{q+1}, where q is the Markov model order: it is mainly due to the estimation of conditional score functions. Further works include (i) the comparaison of our algorithm with TDSEP [7] (using second order statistics) (ii) the relevance of suitability between the Markov order, q, and AR source order.

References

1. Hosseini, Sh., Jutten, C. and Pham, D. T.: Blind Separation of Temporally Correlated Sources Using A Quasi-Maximum Likelihood Approach. Proceedings ICA'01, San Diego (CA, USA) (2001) 586–590
2. Hosseini, Sh., Jutten, C. and Pham, D. T.: Markovian Source Separation. IEEE Trans. on Signal Processing **51** (2003) 3009–3019
3. Taleb, A., Jutten, C.: Source separation in post nonlinear mixtures. IEEE Trans. on Signal Processing **47** (1999) 2807–2820
4. Tong, L., Soon, V., Huang, Y. and Liu, R.:AMUSE: a new blind identification algorithm. Proceedings ISCAS'90, New Orleans (USA), 1990
5. Molgedey, L., Schuster, H. G.: Separation of a mixture of independent signals using time delayed correlation. Physical Review Letters **72** (1994) 3634–3636
6. Belouchrani, A., Abed Meraim, K., Cardoso, J.-F. and Moulines, E.: A blind source separation technique based on second order statistics. IEEE Trans. on Signal Processing **45** (1997) 434-444
7. Ziehe, A. and Müller, K.-R.: TDSEP: an efficient algorithm for blind separation using time structure. Proceedings of ICANN'98, Skvde (Sweden) (1998) 675–680
8. Pham, D. T.: Fast algorithm for estimating mutual information, entropies and score functions. Proceedings ICA'03, Nara (Japan) (2003) 17–22
9. Ziehe, A., Kawanabe, M., Harmeling, S., Müller, K.-R.: Blind separation of post-nonlinear mixtures using gaussianizing transformations and temporal decorrelation. Proceedings ICA'03, Nara (Japan) (2003) 269-274
10. Solé, J., Babaie-Zadeh, M., Jutten, C., Pham, D. T.: Improving algorithm speed in PNL mixture separation and Wiener system inversion. Proceedings ICA'03, Nara (Japan) (2003) 639-644

This article was processed using the LaTeX macro package with LLNCS style