Proceedings of the
IASTED International Conference

IASTED

Signal and Image Processing

August, 13-16, 2001
Honolulu, Hawaii, USA

Editor: M.H. Hamza

A Publication of The International Association of Science and Technology for Development - IASTED

ISBN: 0-88986-297-4
ISSN: 1482-7921

ACTA Press

Anaheim * Calgary * Zurich

TABLE OF CONTENTS

Proceedings of the IASTED International Conference
SIGNAL AND IMAGE PROCESSING
August 13-16, 2001, Honolulu, Hawaii, USA

BLIND SOURCE SEPARATION FROM NOISY DATA USING BAYESIAN ESTIMATION AND GIBBS PRIORS

LUIGI BEDINI, EMANUELE SALERNO, ANNA TONAZZINI

Istituto di Elaborazione della Informazione
Area della Ricerca CNR di Pisa
Via G. Moruzzi, 1
I-56124 PISA (Italy)
<surname>@iei.pi.cnr.it

ABSTRACT

This paper deals with the blind separation and reconstruction of source signals from their mixtures with unknown coefficients, in the practical case where noise affects the mixtures themselves. We address the blind source separation problem within the ICA approach, i.e. assuming statistically independent source signals, and reformulate it in the framework of Bayesian estimation. In this way, the flexibility of the Bayesian formulation in accounting for available knowledge we may possess about the original signals can be exploited to describe time correlation of the single sources, through the use of suitable Gibbs priors. We propose a MAP estimation method based on simulated annealing to recover both the mixing matrix and the sources, and experimentally verify that a model for the source signals which accounts for time correlation is able to increase robustness of the estimates against noise in the data.

KEYWORDS

Blind Source Separation, Independent Component Analysis, Markov Random Fields, Bayesian Estimation.

1. INTRODUCTION

Blind source separation (BSS) is a recent emergent topic in signal processing, and even more recent in image processing and computer vision. It consists in reconstructing and separating a set of unknown signals from a set of mixtures, when no knowledge is assumed about the mixing coefficients. The most known application example of BSS is the so-called "cocktail-party" problem in speech recognition. Other applications include the removal of underlying artifact components of brain activity from EEG records, the search for hidden factors in parallel time series of financial data, and feature extraction or noise removal from natural images. In order to solve BSS, which is a severely ill-posed inverse problem, many techniques have been proposed so far, mainly based on the assumption of the statistical independence of the source signals through Independent Component Analysis (ICA) methods. Most of these methods were developed in the case of noiseless data, and differ in the manner they enforce independence. The Maximum Likelihood (ML) method [1] assumes a separable joint source distribution, the method of maximization of the mutual information [2] uses the entropy as a measure of independence, and the methods based on the minimization of contrast functions exploit high-order statistics to ensure independence [3]. The strict relationships among these methods have been investigated as well [4,5], and some fast and efficient algorithms have been proposed [6]. The requirement of independence can be verified in some practical applications, but in many cases there can be a clear evidence of correlation among the sources (e.g. in face recognition). Although some of the proposed algorithms have been experimentally shown to perform well even in the lack of independence, all of them perform poorly when some noise affects the data. Moreover, most of the methods do not account for a possible time correlation inside the single sources, and for different numbers of sources and data signals. Recently, some work has been done to partially overcome the limitations above [7]. In particular, the Fast ICA algorithm proposed in [6] has been extended to the noisy case [8,9], and an Independent Factor Analysis (IFA) method has been developed [10], which can account for noise as well. In [11,12,13] the Bayesian approach has been proposed as a solution to blind source separation, based on Maximum A Posteriori (MAP) estimation for both the sources and the mixing matrix. Moreover, in [13] it is shown that Bayesian estimation provides a unifying approach to source separation, within which the other approaches can be viewed as special cases.

In this paper we address the blind source separation problem in the case where the noise is present on the data. We retain the independence constraint of the ICA approach, but reformulate the problem in a Bayesian framework, where both the mixing matrix and the sources are recovered via MAP estimation. The flexibility of the Bayesian formulation is exploited to account for available knowledge we may possess about the problem in general and about the sources in particular. Thus, besides enforcing the independence of the sources, we allow for time correlation inside each single source, and describe this correlation by using the Markov Random Field (MRF), or, equivalently, the Gibbs distribution formalism. We show how such a modeling of the source signals, which is however application-dependent, increases robustness of the estimates against noise in the data, both for the sources and the mixing matrix. For the joint maximization of the posterior probability with respect to all the unknowns, we propose a computational scheme where, within an overall simulated annealing algorithm, the coefficients of the mixing matrix are updated via the Metropolis algorithm and the sources are estimated through deterministic algorithms.

2. BSS THROUGH MAP ESTIMATION

According to the BSS formalism, the data generation model we consider is given by:

$$\mathbf{x}(t) = A\mathbf{s}(t) + \mathbf{n}(t) \qquad t=1,2,....T \tag{2.1}$$

where $\mathbf{x}(t)$ is the vector of the measurements, $\mathbf{s}(t)$ is the vector of the unknown sources, and $\mathbf{n}(t)$ is the noise or measurement error vector, at time t, and A is the unknown mixing matrix. We assume at the moment the same number N of measured and source signals, so that A is a NxN matrix. Considering the noise to be white and Gaussian with zero mean, the likelihood is given by:

$$P(\mathbf{x} \mid A,\mathbf{s}) = = \frac{1}{Z_\Sigma} \exp\left[-\frac{1}{2} \sum_t \left(A\mathbf{s}(t) - \mathbf{x}(t)\right)^T \Sigma^{-1}(t) \left(A\mathbf{s}(t) - \mathbf{x}(t)\right) \right] \tag{2.2}$$

where $\mathbf{x}=\{\mathbf{x}(1),..., \mathbf{x}(T)\}$, $\mathbf{s}=\{\mathbf{s}(1),..., \mathbf{s}(T)\}$, Σ is the covariance matrix of the noise, assumed to be time-dependent, and Z_Σ is the normalizing constant.

According to the Bayesian approach, a prior distribution can be assigned to both A and $\mathbf{s}$, and a posterior distribution can be derived in the form:

$$P(A,\mathbf{s} \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid A,\mathbf{s})P(A)P(\mathbf{s})}{P(\mathbf{x})} \tag{2.2}$$

This posterior distribution accounts for all information we have about the problem, and hopefully allows us to define a single solution, for instance as the global maximizer of the posterior itself (MAP estimate). To enforce independence of the sources, we assume a factorized prior for $\mathbf{s}$:

$$P(\mathbf{s}) = \prod_i P_i(s_i) \tag{2.3}$$

where the single P_i is the distribution of the i-th source, and has to be chosen on the basis of known properties of s_i. In particular, as it will be further developed later on, we adopt Gibbs distributions, related to Markov Random Field models, to describe global or local properties of regularity for the sources. This will be shown to increase the robustness of ICA against noise on the data.

By neglecting the term $P(\mathbf{x})$ which does not depend either on A or on $\mathbf{s}$, the solution to the BSS problem can then be defined via the following joint maximization:

$$(\hat{A},\hat{\mathbf{s}}) = \arg\max_{A,\mathbf{s}} P(\mathbf{x} \mid A,\mathbf{s})P(\mathbf{s})P(A) \tag{2.4}$$

where $P(\mathbf{s})$ takes on the factorized form of Eq. (2.3). As usual in blind signal/image estimation, this joint maximization is very hard, and needs to be reduced in some way. The most adopted strategy is to perform it by iteratively alternating steps of estimation with respect to A with steps of estimation with respect to $\mathbf{s}$. In formulas it is:

$$A^k = \arg\max_A P(\mathbf{x} \mid A,\mathbf{s}^k)P(A) \tag{2.5a}$$

$$\mathbf{s}^{k+1} = \arg\max_{\mathbf{s}} P(\mathbf{x} \mid A^k,\mathbf{s})P(\mathbf{s}) \tag{2.5b}$$

It can be observed that when, in agreement with the most general assessment of a BSS problem, we assume a uniform prior for A, step (2.5a) corresponds to the solution of a least squares problem of small size (the number N of sources and mixtures is generally low). When some prior is adopted for A, problem (2.5a) can loose concavity, and algorithms for non-convex optimization must be adopted, such as simulated annealing (SA). In this case, however, owing to the small number of variables, even SA results to be a reasonably cheap algorithm. Moreover, SA can be particularly suitable when P(A) enforces constraints on A that cannot be expressed in analytical form, such as bounds on the admitted values, etc. The algorithm to be adopted to solve the optimization problem (2.5b) mainly depends on the form adopted for the $P_i(s_i)$. On the other hand, we know from the ICA theory that the separation of the sources (at least in the noiseless case) is possible only when at most one of the sources is Gaussian. Thus, in general, problem (2.5b) will not be a quadratic one. However, when appropriate, the $P_i(s_i)$s can be chosen in such a way to ensure the concavity of the function to be optimized, so that a gradient ascent algorithm can be used. Otherwise, still relatively cheap algorithms for non-convex optimization can be used, such as Graduated Non-Convexity (GNC) [15], which was shown to be more efficient than SA. In the following Section we will show that the constraint we intend to enforce on the sources, i.e. correlation with respect to time, can be suitably expressed through MRF models whose related Gibbs distributions are suitable to be managed by GNC-like algorithms. On the basis of this choice, we will specialize the iterative scheme (2.5) and will propose a particular implementation of the scheme itself that exhibits a reduced computational complexity and for which experimental convergence can always be obtained.

3. GIBBS PRIORS AND DERIVATION OF THE ALGORITHM

MRF models have become very popular since the middle '80s, especially in connection to inverse, ill-posed problems of image processing, such as restoration, denoising, segmentation, optical flow estimation, and so on. Through MRF models it is indeed possible to describe local properties of the images, such as edges, in order to make space-variant the smoothness constraint which has to be enforced to regularize and stabilize the solutions of inverse problems in visual reconstruction. Furthermore, the local nature of these models allows us to define distributed and even parallel algorithms for the computation of the regularized solution. In this paper we propose to use MRF for modeling the properties of time correlation of the independent sources in a BSS and ICA context. We consider the one-dimensional case and apply MRF to signals sampled into T time instants.

Let us consider then the distribution of the i-th source $s_i(t)$ in our problem. According to the MRF formalism, it must have the following Gibbsian form:

$$P_i(s_i) = \frac{1}{Z_i}\exp[-U_i(s_i)] \qquad (3.1)$$

where Z_i is the normalizing constant and $U_i(s_i)$ is the prior energy in the form of a sum of potential functions over the set of cliques of interacting time instants. We consider the set of cliques constituted of a single instant t or two adjacent instants t and t+1. We then define $U_i(s_i)$ as:

$$U_i(s_i) = \sum_{t=1}^{T}\alpha_i f_i(s_i(t)) + \sum_{t=1}^{T-1}\lambda_i \phi_i(s_i(t) - s_i(t+1)) \qquad (3.2)$$

where α_i and λ_i are positive weights, and f_i and ϕ_i are functions to be chosen according to our expectation about the probability law assigned to each sample of signal s_i, and about the degree of correlation we assign to couples of adjacent samples of the signal, respectively. From the ICA theory we know that, to ensure separability of the sources, these cannot be Gaussian, and, in many practical ICA algorithms, the probability distribution of the source samples is approximated via sub-Gaussian or super-Gaussian distributions, unless more precise knowledge about their nature is available. With respect to the choice of ϕ_i, this function (often called stabilizer) is devoted to describe the time correlation of the i-th source, in the form of regularity of the signal shape. This regularity is physically plausible in many real-word applications, and, as already said, it is an essential constraint to prevent the reconstructions from being unstable when the data are noisy. Nevertheless, the source signals can present some steep fronts which must be preserved as well. We thus refer to stabilizers for edge-preserving image recovering, and adapt them to the one-dimensional case. Some of the many proposed possess the characteristic of being convex [14], so that, if the f_i are also convex, the function to be maximized in step (2.5b) will result concave, and a gradient ascent can be used to perform the optimization. We consider instead the most general case of non-concavity, but adopt specific stabilizers that allows us to "correct" the mild non-concavity of the overall function by providing a sequence of approximations for it, according to the GNC strategy,.

Upon the above considerations, to further reduce the computational complexity of the iterative scheme (2.5) and ensure convergence, we propose a particular implementation of such a scheme, based on an overall simulated annealing for the estimation of A according to (2.5a), interrupted at each cycle, i.e. at each lowering of a temperature parameter τ after that a Markov chain is computed, to perform an update of the sources **s**, according to (2.5b). To this end, we first take the negative logarithm of the distribution $P(\mathbf{x}|A,\mathbf{s})P(\mathbf{s})P(A)$, thus obtaining the following energy function to be equivalently minimized in A and **s**:

$$E(\mathbf{s},A) = \frac{1}{2}\sum_t (A\mathbf{s}(t) - \mathbf{x}(t))^T \Sigma^{-1}(t)(A\mathbf{s}(t) - \mathbf{x}(t)) +$$
$$\sum_{t=1}^{T}\alpha_i f_i(s_i(t)) + \sum_{t=1}^{T-1}\lambda_i\phi_i(s_i(t) - s_i(t+1)) - \log P(A) \qquad (3.3)$$

where the constant terms coming from the partition functions have been neglected. A new distribution $P_\tau(\mathbf{s},A)$, to be used for simulated annealing, can thus be derived in the following way:

$$P_\tau(\mathbf{s},A) = \frac{1}{Z}\exp\left[-\frac{E(\mathbf{s},A)}{\tau}\right] \qquad (3.4)$$

The scheme of our algorithm is the following:

1. set k=0, $\mathbf{s}^{(k)}$, $A^{(k)}$, τ_k
2. set r=1, $A^{(r)} = A^{(k)}$

   for r=1, L

   compute $A^{(r+1)}$ according to $P_{\tau_k}(\mathbf{s}^{(k)}, A)$
3. set $A^{(k+1)} = A^{(L)}$
4. compute

   $$\mathbf{s}^{(k+1)} = \arg\min_{\mathbf{s}} E(\mathbf{s}, A^{(k+1)})$$
5. set k=k+1; go back to step 2 until a termination criterion is satisfied.

In the scheme above, the lowering to zero of the temperature ensures the convergence of the estimation of the mixing matrix A, in that distribution in Eq. (3.4) becomes a Dirac function when τ approaches zero. This ensures stabilization of the source estimates as well, since these are computed by minimizing the energy function with all fixed parameters.

4. EXPERIMENTAL RESULTS

The efficiency of the above algorithm with respect to the robustness of the estimates was tested on synthetic signals, belonging to the class of smooth or piecewise smooth signals. For generating these signals we did not refer to any specific probability law. In these conditions, the choice of a function f_i becomes difficult. We thus decided to enforce only generic constraints of time correlation for the sources, and dropped the first term from the prior energy of Eq. (3.2).

We performed a large set of numerical experiments, by letting the ideal matrix coefficients and the noise realization to be selected randomly. In order to further reduce the computational time, we found convenient to preprocess the data, in such a way to whiten them. To this purpose, we applied to the data **x** a matrix B such that $E[B\mathbf{x}\mathbf{x}^T B^T]=I$. B^{-1} was then assumed as starting point for estimating the mixing matrix. The starting point for the sources was instead always chosen randomly. For comparison purpose, in all cases we also computed the solutions of the Fast ICA algorithm.

In Figures 1 and 2, we provide the results of two of the experiments performed, for the case of two sources and two data signals. In the experiment of Figure 1 we considered two step signals, of different scale, and added to their mixtures a 10% of noise. For these signals we found out convenient to assume as model a stabilizer with edge-preserving properties, for which we previously derived a family of approximations and a GNC-like algorithm [16]. The general form of this stabilizer is given by:

$$\phi(t) = \frac{|t|/\Delta}{1+|t|/\Delta} \tag{4.1}$$

where parameter Δ represents the threshold for the intensity gradient above which a steep front is likely to be present in the signal. We used two different values of the λ parameter for the two signals, due to their different scale. It is qualitatively evident the better fidelity of the separated signals to the original ones, with respect to the results of the Fast ICA algorithm, that are much more noisy. We also computed the root mean squared error (RMSE) between the ideal mixing matrix and the estimated one. We obtained for this experiment values of the RMSE around 0.046 for the Fast ICA and 0.03 for our method.

In the second experiment, shown in Figure 2, we considered as data a random mixture between a step signal and a sinusoidal signal, still added of a 10% of noise. For the first signal we adopted the same stabilizer as for the previous experiment, while for the sinusoidal signal we considered a simple parabola of the intensity gradient, which enforces global smoothness. Also in this case the reconstructions obtained with our method are much more robust against noise than those produced by the Fast ICA, and the RMSE between the ideal and the estimated mixing matrix is around 0.0157.

In all our experiments, we verified that the order in which we assign the prior to the signals is reflected in the order of the reconstructions, so that this approach does not present one of the typical ambiguities of BSS, i.e. possible permutation.

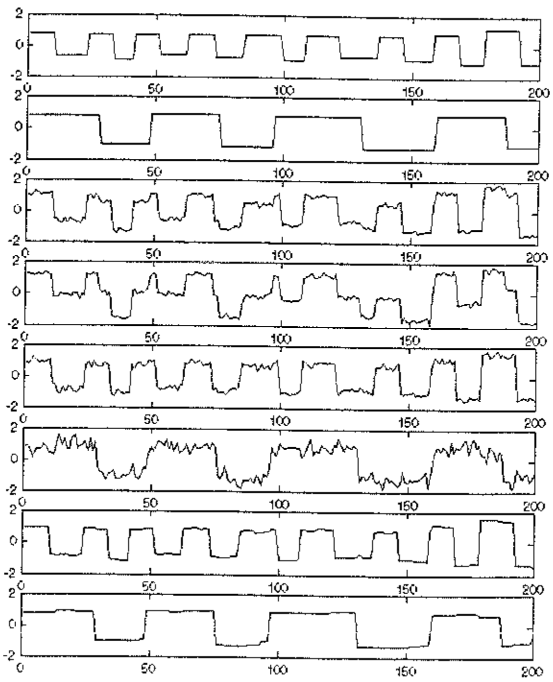

Fig. 1 – From top to bottom: original signals, mixed and noisy signals, reconstructions obtained by using the Fast ICA algorithm, reconstructions obtained by using the MAP estimation method.

5. CONCLUSIONS

We proposed a Bayesian formulation of ICA techniques for blind source separation, in the case where the data are noisy. We considered MRF models for the source signals which are suitable to describe both the independence of the sources themselves and the local time correlation for each single source. We proposed to implement alternating maximization for the joint MAP estimation of the mixing matrix and the sources by means of a simulated annealing scheme employing the Metropolis algorithm for the updating of the mixing matrix, and a deterministic algorithm for the updating of the sources, at each temperature. We experimentally verified that the introduction of *a priori* information about the time correlation of the sources can increase robustness of the estimates against noise in the data.

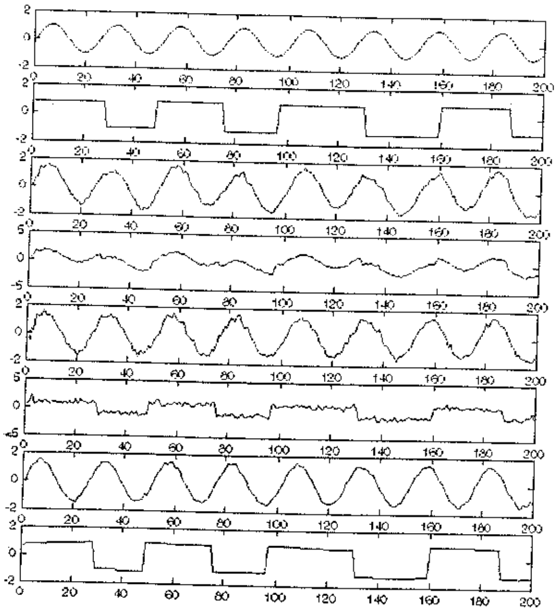

Fig. 2 – From top to bottom: original signals, mixed and noisy signals, reconstructions obtained by using the Fast ICA algorithm, reconstructions obtained by using the MAP estimation method.

REFERENCES

1. A.J. Bell and T.J. Sejnowski, An information maximization approach to blind separation and blind deconvolution, Neural Computation, 7(6), 1995, 1129-1159.
2. T. Lee, M. Lewicki and T. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources, Neural Computation, 11, 1999, 409-433.
3. J.-F. Cardoso, High-order contrasts for independent component analysis, Neural Computation, 11, 1999, 157-192.
4. J.-F. Cardoso, Infomax and maximum likelihood for source separation, IEEE Letters on Signal Processing, 4, 1997, 112-114.
5. A. Hyvärinen and E. Oja, Independent Component Analysis: Algorithms and Applications, Neural Networks, 13(4-5) , 2000, 411-430.
6. A. Hyvärinen, Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, IEEE Trans. on Neural Networks, 10(3), 1999, 626-634.
7. T. Lee, M. Lewicki, M. Girolami and T. Sejnowski, Blind source separation of more sources than mixtures using overcomplete representation, IEEE Signal Processing Letters, 6(4), 1999, 87-90.
8. A. Hyvärinen, Independent Component Analysis in the Presence of Gaussian Noise by Maximizing Joint Likelihood, Neurocomputing, 22, 1998, 49-67.
9. A. Hyvärinen, P. Hoyer and E. Oja, Image Denoising by Sparse Code Shrinkage, in S. Haykin and B. Kosko (Eds), *Intelligent Signal Processing* (IEEE Press, 2000).
10. H. Attias, Independent Factor Analysis, Neural Computation, 11, 1999, 803-851.
11. K. Knuth, Bayesian source separation and localization, Proc. SPIE'98, Bayesian Inference for Inverse Problems, San Diego, USA, 1998, 147-158.
12. S.E. Lee and S.J. Press, Robustness of Bayesian factor analysis estimates, Communications in statistics - Theory and Methods, 27(8), 1998.
13. A. Mohammad-Djafari, A Bayesian approach to source separation, 19th Int. Workshop on Maximum Entropy and Bayesian Methods, Boise, Idaho, USA, 1999.
14. S.Z. Li, Close-form solution and parameter selection for convex minimization-based edge-preserving smoothing, IEEE Trans. on PAMI, 20(9), 1998, 916-932.
15. L. Bedini, I. Gerace, E. Salerno and A. Tonazzini, Models and algorithms for edge-preserving image reconstruction, in P.W. Hawkes (Ed.), *Advances in Imaging and Electron Physics*, 97 (San Diego: Academic Press, 1996) 86-189.
16. L. Bedini, I. Gerace and A. Tonazzini, A GNC algorithm for constrained image reconstruction with continuous-valued line processes, Pattern Recognition Letters, 15(9),1994, 907-918.