UNDERDETERMINED BLIND SOURCE SEPARATION IN A TIME-VARYING ENVIRONMENT

L. Vielva,[★] D. Erdoğmuş,[*] C. Pantaleón,[★] I. Santamaría,[★] J. Pereda,[★] J. C. Príncipe[*]

[★] Communications Engineering Department, Universidad de Cantabria, Spain
[*] Computational NeuroEngineering Laboratory, University of Florida, Gainesville, FL
E-mail: luis@dicom.unican.es, {deniz, principe}@cnel.ufl.edu

ABSTRACT

The problem of estimating n source signals from m measurements that are an unknown mixture of the sources is known as blind source separation. In the underdetermined —less measurements than sources— linear case, the solution process can be conveniently divided in three stages: represent the signals in a sparse domain, find the mixing matrix, and estimate the sources. In this paper we adhere to that approach and parametrize the performance of these stages as a function of the sparsity of the signals. To find the mixing matrix and track its variations in the dynamic case a non-parametric maximum-likelihood approach based on Parzen windowing is presented. To invert the underdetermined linear problem we present an estimator that chooses the "best" demixing matrix in a sample by sample basis by using some previous knowledge of the statistics of the sources. The results are validated by Montecarlo simulations.

1. INTRODUCTION

Suppose there exist n unknown random source signals that are combined in an unknown way to provide m random measurements. If you want to estimate the sources from the measurements you are faced with a blind source separation (BSS) problem. In the noise-free linear instantaneous case the observation vector can be written as a linear transformation on the source vector as

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

where $\mathbf{s} \in \mathbb{R}^n$ is the source random vector, $\mathbf{x} \in \mathbb{R}^m$ is the measurement random vector, and $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the unknown mixing matrix.

If only $\mathbf{x}$ is available, to estimate $\mathbf{s}$ can be conceptually devised as a two stage process: first estimate the mixing matrix from the measurements and then invert the linear problem (1). If the number of measurements equals the number of sources ($m = n$) the problem reduces to estimate the square matrix $\mathbf{A}$ [1, 2], since the linear problem is readily solved by the inverse matrix.

In the underdetermined case, when less measurements than sources are available ($m < n$), there is no unique inverse. In fact there exist an infinite number of source vectors that are solutions of the linear problem (1). We could say that the "best" solution is determined by the constraints that one imposes on $\mathbf{s}$ on the bases of some performance criterion or previous knowledge. We will show that to find a good inversion strategy, it is crucial to be able to represent the signals in a domain such that a high ratio of the coefficients are negligible [3, 4]. Consequently it is very convenient to parametrically model sources with different degrees of sparsity and have a framework to characterize the stages of the BSS taking the sparsity of the sources as a parameter. To that end we use the following probabilistic model for the distributions of the sources

$$p_{S_j}(s_j) = p_j\,\delta(s_j) + (1 - p_j) f_{S_j}(s_j), \quad j = 1, \ldots, n; \tag{2}$$

where p_j is the sparsity factor for source j, $\delta(\cdot)$ is the Dirac's delta, and $f_{S_j}(s_j)$ is the density when the corresponding source is active.

The organization of the paper is as follows. On the second section we show how to estimate the mixing matrix from the measurements both for the static case, in which the mixing matrix is constant, and for the time-varying case. On the third section we study the inversion procedure once the mixing matrix has been estimated and develop a maximum *a posteriori* (MAP) estimator. On the fourth section we present the conclusions extracted from the work and future lines.

2. ESTIMATION OF THE MIXING MATRIX

There have been different approaches to estimate the mixing matrix. Lin *et. al* use competitive learning in a feature extraction framework [5]. Bofill and Zibulevsky employ a potential function based clustering approach [3]. Wu uses eigenspread estimation to decide when only one source is active, and uses this information to find the columns of the mixing matrix [6]. In this paper we use a non-parametric maximum-likelihood approach, based on Parzen windowing. In this method probability distribution of sample directions is estimated and the peak points are shown to correspond to the directions defined by the column vectors of the mixing matrix.

Equation (1) can be interpreted from a geometrical point of view as the projection of the source vectors $\mathbf{s}$ from $\mathbb{R}^n$ into the vector space $\mathbb{R}^m$ of the measurement vectors $\mathbf{x}$. If we denote by $\mathbf{a}_j$ the j-th column of the mixing matrix $\mathbf{A}$, (1) can be rewritten as $\mathbf{x} = \sum_{j=1}^{n} s_j \mathbf{a}_j$, that explicitly shows the measurement vector as a linear combination of the columns of the mixing matrix. According to this interpretation, if at a given time only the j-th source is non zero, the meassurement vector will be collinear with $\mathbf{a}_j$. Scatter-plots of $\mathbf{x}$ for a case with two measurements and three sources are shown in figures 1a and 1c for sparsity factors of 0.1 and 0.8 respectively. Notice that when the histogram of the angle of samples is considered, as shown in figures 1b and 1d, the three directions that correspond to the columns of the mixing matrix are clearly identified. However, the resolution of the histogram-based estimation of the column vectors is limited by the bin length that is assumed in evaluating the histogram. To overcome this problem,

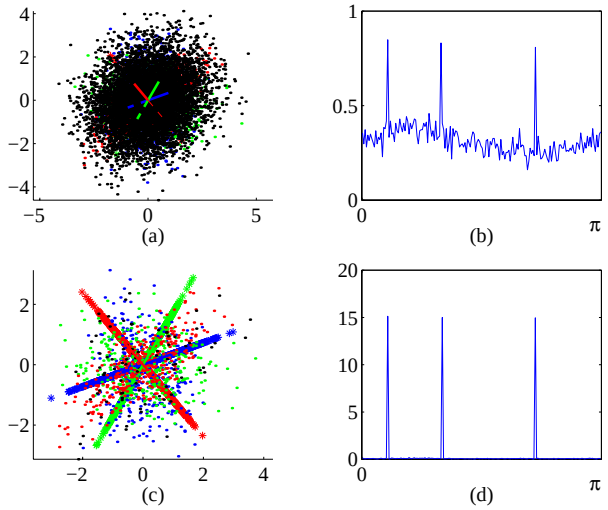

Fig. 1. Scatter plot of measurements and histogram of angles for sparsity factors 0.1 —(a) and (b)— and 0.8 —(c) and (d).

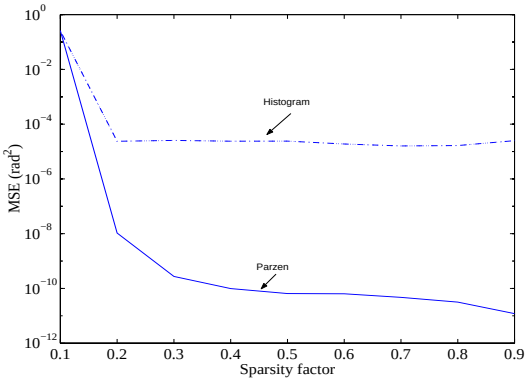

Fig. 2. MSE in angle estimation vs sparsity factor for Parzen window (using optimal kernel size) and histogram.

we propose the use of Parzen windowing [7] to estimate the probability density of the angle, and then pick the n largest peaks of this distribution as the estimates for the directions of the n columns of $\mathbf{A}$.

The Parzen window density estimation for the angle given the samples and a kernel function $\kappa_\sigma(\cdot)$ is given by

$$p(\theta) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(\theta - \theta_i), \tag{3}$$

where the samples of the angle are evaluated, from $\theta_i = \arctan(x_2/x_1)$. The zero measurements are simply omitted, as they have no well-defined angle. Once this density estimation is obtained, we utilize the steepest ascent algorithm to find the angles corresponding to the peaks of the density function.

2.1. Static case

In the static case, the mixing matrix is assumed to be constant; therefore, all the measurement samples can be used in the density estimation for the angle in a batch-learning scheme. Since we are looking for the largest n peaks of the estimated density, and we are going to use steepest ascent, our initial estimates must be in the domain of attraction of those solutions that we seek. Accordingly, the direction estimates obtained from the histogram method are used as initial conditions to the steepest ascent algorithm. For example, by using 180 bins in the interval $[0, \pi]$ we can obtain initial estimates that are closer than one degree to the solutions. Note that it is sufficient to consider the angles in this interval only, since a sign ambiguity is acceptable in BSS. Furthermore, we can assume that the columns of $\mathbf{A}$ are unit length, since this corresponds to an ambiguity in the scaling factor, which is also acceptable in BSS.

Once the initial estimates are obtained from the histogram method, the following gradient expression of the "cost function" in (3) is used to refine the estimates until convergence to the maximum is achieved

$$\nabla_\theta p(\theta) = \frac{1}{N}\sum_{i=1}^{N} \nabla\kappa_\sigma(\theta - \theta_i).$$

This procedure is repeated for the n initial conditions provided by the histogram. Since Parzen windowing provides a continuous estimate of the density function of interest, in theory it is possible to achieve a very high resolution, provided that the kernel size is chosen sufficiently small so that there are no artifact peaks; whereas in the histogram, in order to increase resolution, more bins are necessary.

The choice of the kernel size in Parzen windowing is crucial to an accurate estimation of the correct directions of the columns. For that end, we have performed a Montecarlo simulation and chosen the kernel size as the one that minimizes the MSE, defined as

$$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{j=1}^{n}(\theta_j - \hat{\theta}_j)^2,$$

in the estimation of the actual directions (in radians squared); where M is the number of simulations used in the Montecarlo method.

It is also interesting to investigate the behavior of the MSE of estimation as a function of the sparsity rate. One would expect to observe an increasing performance in estimation as the sparsity increases, since there will be more and more samples that are perfectly aligned with the columns compared to the outliers that are generated as a result of more than one active source.

Figure 2 shows that, just as expected, when the sparsity rate increases the MSE decreases. Another important conclusion drawn from this plot is that refining the estimates with the Parzen windowing method remarkably improves upon the initial estimates given by the histogram. Note that, since a steepest ascent algorithm with a fixed step size is utilized in training, the actual maxima are not exactly obtained. A portion of the MSE in the Parzen window method is thus due to this slight misadjustment from the optimal values. In contrast, a greater portion of the MSE in the histogram estimates is due to the finite bin length. Assuming a uniform distribution in each bin of length one degree, the associated variance would be on the order of 10^{-5} rad^2. We observe that the results in figure 2 agree with this expectation.

2.2. Dynamic case

In the dynamic case, we assume that the mixing matrix is time varying, although the mixing procedure is still linear and instantaneous. This situation may occur, for example, when there are fixed microphones in a room and the speakers are moving around, so that the attenuation experienced by the speech signal until it reaches the microphone varies with time [8].

In order to track the columns of a changing matrix, the training algorithm presented in the static case must be slightly modified. Since the algorithm will try to track the directions of the columns blindly, a sufficiently accurate initial estimate is crucial. The static Parzen windowing method can be used to achieve this initialization assuming that the columns are rotating "slowly". In order to initialize the angle estimations, a number of samples must be collected first. The number of samples required for this initialization procedure can be in the order of hundreds or smaller. Once the initialization is achieved, a modified version of the steepest ascent algorithm is applied to adapt the angle estimates on-line on a sample-by-sample basis. We use a forgetting factor approach and estimate the new density using a linear combination of the estimate from the previous sample and the kernel evaluated at the current sample. In this case, the cost function, i.e. the density estimate, at time instant k reads as

$$p_k(\theta) = \alpha p_{k-1}(\theta) + (1-\alpha)\kappa_\sigma(\theta - \theta_k),$$

where α is the forgetting factor. This formulation of the density estimation also gives rise to a recursive algorithm for the evaluation of the gradient. The gradient expression to be used in the update at time instant k, in terms of the gradient from the previous samples and the kernel function, now becomes

$$\nabla_\theta^k p = \alpha \nabla_\theta^{k-1} p + (1-\alpha)\nabla\kappa_\sigma(\theta - \theta_k),$$

and is evaluated at the current estimate of the angle θ. In the update phase, only one of the n angles is updated, and that is determined by comparing the difference between the angle of the current sample and the estimates of the angles from the previous update.

A number of simulations have been carried on to evaluate the performance of this tracking algorithm. It has been determined that the tracking ability of the algorithm is limited by the first and second derivatives of the angles of the columns with respect to time. Using the value 0.9 for the forgetting factor, $3 \cdot 10^{-3}$ rad for the size of the Gaussian kernel, and a step size of 10^{-7}, the algorithm was able to track signals with first order derivatives on the order of 10^{-5} rad/sample [8].

3. ESTIMATION OF THE SOURCES

In the underdetermined case ($m < n$) the problem (1) has an infinite number of solutions, so it is necessary to impose some aditional criterion to select one solution vector $\mathbf{s}$. One possible criterion of general applicability could be to impose some L_p norm of the solution to be a minimum. Specifically, the solution provided by the pseudo inverse is the one that minimices the L_2 norm of the solution $||\mathbf{s}||$, and with no additional knowledge of the statistics of the sources could be the canonical option to choose. As we will show next, if the signals admit a sparse representation, it is possible to design better inversion strategies.

The key geometric intuition to develop heuristic inversion procedures is to think of $\mathbf{x}$ as a linear combination of the vectors defined by the columns of the mixing matrix. Consider the case with $m = 2$ measurements and $n = 3$ sources that is shown in figure 1c. If at a given time all the sources are zero, $\mathbf{x}$ will be placed at the origin. If only one source is active, $\mathbf{x}$ will be collinear with the corresponding column vector of the mixing matrix (those are the three solid lines in figure 1c). There are three different combinations with two sources active at the same time, and since any two non-collinear vectors are a base of the plane, any $\mathbf{x}$ can be due to any two sources active at the same time. Of course, if the three sources are all active, $\mathbf{x}$ can be placed at any point on the plane. However, if the sources are sparse enough, the events corresponding to higher number of sources active at the same time will be less and less probable. According to this, it is possible to derive very simple heuristic approaches to invert the linear problem (1). The simplest of these approaches, that we call 1D, consists of considering that at most one source is active at a given time: a measurement vector is supposed to be due to the column that maximizes the scalar product with $\mathbf{x}$. Another heuristic approach, that we call 2DL_2, considers that $\mathbf{x}$ is a linear combination of two source vectors; from the three combinations, the one that minimices the norm of the solution is choosen. In figure 3, the performance of these heuristic approaches is compared with the pseudo inverse; as predicted, the heuristic approaches improve as the sparsity of the sources increases. For a sparsity factor around 0.7, both of them outperform the pseudo inverse. However, as we show next, by using the sparse probability model (2) it is possible to obtain even better estimators.

3.1. Bayesian estimation

If, at any given time, we knew that at most m given components of the signal are non zero, the problem (1) would not be underdetermined any more and we could invert it.

Let us denote by C_0 the event that all the components of the source vector are zero at a given time, by C_u the event that only component s_u is non-zero, by $C_{u,v}$ the event that s_u and s_v are the only non-zero components, and in general by $C_{u,\dots,w}$ that only and all of $s_u, \dots, s_w$ are non-zero at the same time. According to (2), the *a priori* probabilities of these events are

$$p(C_{u,\dots,v}) = \prod_{j=u,\dots,w}(1-p_j)\prod_{j\neq u,\dots,w} p_j.$$

Next we will consider the conditional densities of the observations given the events. When all the sources are silent, $P(\mathbf{x}|C_0) = \delta(\mathbf{x})$. When only source s_u is active,

$$p(\mathbf{x}|C_u) = \frac{1}{|a_{iu}|} f_{S_u}\left(\frac{x_i}{a_{iu}}\right),$$

where x_i is the ith component of the measurement $\mathbf{x}$ corresponding to a non zero matrix component a_{iu}. In general, given that the event $C_{u,\dots,w}$ had occurred, when number of active sources is less than m,

$$p(\mathbf{x}|C_{u,\dots,w}) = \frac{1}{|\det \mathbf{A}_{u,\dots,w}|}\prod_{j=u,\dots,w} f_{S_j}(\hat{s}_j),$$

where

$$\mathbf{A}_{u,\dots,w} = \begin{bmatrix} a_{ku} & \dots & a_{kw} \\ \vdots & & \vdots \\ a_{lu} & \dots & a_{lw} \end{bmatrix}, \quad \begin{bmatrix} \hat{s}_u \\ \vdots \\ \hat{s}_w \end{bmatrix} = \mathbf{A}_{u,\dots,w}^{-1} \begin{bmatrix} x_k \\ \vdots \\ x_l \end{bmatrix},$$

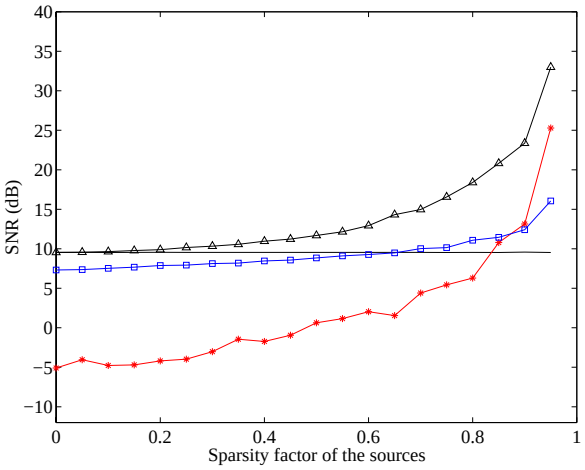

Fig. 3. SNR of the source separation by 1D ($*$), m-DL_2 ($\square$), MAP estimator ($\triangle$), and pseudo inverse (solid line).

and the rows $k, \ldots, l$ have been choosen from $\mathbf{A}$ so that $\mathbf{A}_{u,\ldots,w}$ is invertible. When the number of assumed active sources equals m, the previous equation still applies by using the complete columns of $\mathbf{A}$ corresponding to the active sources.

By applying Bayes rule, we can calculate the *a posteriori* probabilities of the defined events given the measurements as

$$p(C_{u,\ldots,w}|\mathbf{x}) \propto p(\mathbf{x}|C_{u,\ldots,w})p(C_{u,\ldots,w}),$$

where we evaluate the *a posteriori* probabilities for all the events with a number of active sources less than or equal to m. The rest of the events, corresponding to the cases where the number of active sources is greater than m, are combined into one single event, $\bar{C}$, so that

$$p(\bar{C}|\mathbf{x}) = p(\mathbf{x}) - \sum p(C_{u,\ldots,w}|\mathbf{x}).$$

For estimating $p(\mathbf{x})$ a number of methods are readily available. Polynomial expansion approaches [9], kernel-based methods [7], and parametric estimation methods [10] are among the options. Once the *a posteriori* probabilities are known for all the events, the MAP estimator chooses the optimal source estimates corresponding to the event which maximizes the *a posteriori* probability. If the selected event is $\bar{C}$, then the minimum norm solution provided by the pseudo-inverse is used.

In order to compare the behaviour of the pseudo inverse, the different heuristic methods, and the MAP estimator, a Montecarlo simulation has been performed. We have generated 10000 source vectors $\mathbf{s}$ according to (2) with Gaussian $f_{S_j}(s_j)$. For each value of the sparsity factor we have randomly generated 500 mixing matrices with uniform distribution on the angles and uniform distribution on the magnitude of the column vectors. As a measure of the error of the estimation $\hat{\mathbf{s}}$, we have used the signal to noise ratio, that is shown in figure 3 as a function of the sparsity factor.

4. CONCLUSIONS

In this paper we have shown that the underdetermined blind source separation problem can be conveniently separated into three stages: representation of the signals in an sparse domain, estimation of the mixing matrix, and inversion of the underdetermined mixing model. Adhering to this framework, we have parametrized the source densities by a sparsity factor, so we have focused on the last two stages, considering the sparsity of the sources as a parameter. For the second stage, we have developed a nonparametric algorithm that has allowed us to estimate the mixing in the static case and to track its variations in a dynamic environment. For the third stage, we have developed an MAP estimator that chooses the "best" inversion matrix on a sample by sample basis. By using additional knowledge on the sources, the MAP estimator is shown to improve performance over both the pseudo inverse —that acts as a lower bound when there is no sparsity on the sources— and the heuristic approaches. As a final conclusion, we have shown that while to find a sparse representation of the signals merely facilitates the task of finding the mixing matrix, it is crucial for the last stage of inverting the linear problem to estimate the sources. When the original sources do not satisfy the sparsity condition, as is the case with speech signal in the time domain, a suitable linear transformation (as short-time Fourier tranform or wavelet transform) could be applied beforehand.

Acknowledgments: This work was partially supported by the NSF grant ECS-9900394 and by the European Commission and the Spanish Government under contracts 1FD97-1863-C02-01 and 1FD97-1066-C02-01.

5. REFERENCES

[1] A. Hyvärinen, "Survey on independent component analysis," in *Neural Computer Surveys*, no. 2, pp. 94–128, 1999.

[2] J. Cardoso, *Proceedings of the IEEE, special issue on blind identification and estimation*, ch. Blind signal separation: statistical principles. IEEE, 1988.

[3] M. Zibulevsky, B. Pearlmutter, P. Bofill, and P. Kisilev, *Independent Components Analysis: Principles and Practice*, ch. Blind source separation by sparse decomposition in a signal dictionary. Cambridge University Press, 2000.

[4] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, no. 37, pp. 33311–3325, 1997.

[5] J. K. Lin, D. G. Grier, and J. D. Cowan, "Faithful representation of separable distributions," *Neural Computation*, vol. 9, pp. 1303–1318, 1997.

[6] H.-C. Wu, *Blind Source Separation using Information Measures in the Time and Frequency Domains*. PhD thesis, CNEL, University of Florida, 1999.

[7] E. Parzen, *Time Series Analysis Papers*, ch. On Estimation of a Probability Density Function and Mode. Holden-Day, 1967.

[8] D. Erdoğmuş, L. Vielva, and J. Príncipe, "Nonparametric estimation and tracking of the mixing matrix for underdetermined blind source separation," in *ICA and BSS*, (San Diego, CAL, USA), 2001.

[9] J. H. Friedman, "Exploratory projection pursuit," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 249–266, 1987.

[10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd ed., 1991.