UNDERDETERMINED BLIND SOURCE SEPARATION USING A PROBABILISTIC SOURCE SPARSITY MODEL

Luis Vielva

Communications Engineering Department
Universidad de Cantabria
Spain
E-mail: luis@dicom.unican.es

Deniz Erdoğmuş, José C. Príncipe

Computational NeuroEngineering Laboratory
University of Florida
Gainesville, FL
E-mail: {deniz,principe}@cnel.ufl.edu

ABSTRACT

Blind source separation consists of recovering n source signals from m measurements that are an unknown function of the sources. In solving the underdetermined ($m < n$) linear problem three stages can be identified: to represent the signals in an appropriate domain, to estimate the mixing matrix, and to invert the linear problem to estimate the sources. As a consequence of having more degrees of freedom than constraints, the inverse problem has an infinite number of solutions. To choose the "best" solution, additional constraints have to be imposed on the basis of some performance criterion or previous knowledge. In this communication we present a method that choose the "best" demixing matrix in a sample by sample basis by using some previous knowledge of the statistics of the sources. The behaviour of the estimator is compared to the global pseudo inverse approach and with other local heuristic methods by means of Montecarlo simulations.

1. INTRODUCTION

The blind source separation problem consists of estimating n sources from m measurements that are an unknown function of the sources. The noise-free linear model for each sample is

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

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 the number of measurements is greater or equal than the number of sources ($m \geq n$) it is possible to separate statistically independent sources provided that at most one of them is gaussian [1, 2]. Once the mixing matrix is known, the sources can be readily obtained by the inverse matrix when $m = n$ or estimated by using the pseudo-inverse when $m > n$.

In the underdetermined case, when less measurements than sources are available ($m < n$) there is no unique inverse, which means that there exist an infinite number of source vectors that are solutions of the linear problem (1). A possible solution could be to use the Moore-Penrose pseudo inverse of the mixing matrix, this global method uses the same demixing matrix for the whole data. We could say that the "best" solution to the inverse problem is determined by the constraints that one imposes on $\mathbf{s}$ on the bases of some performance criterion or previous knowledge.

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 that the measurement vector is 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 $\mathbf{a}_j$. If the sources have a probability density function so that a high percentage of the samples are negligibly small —that is, if the sources are highly sparse— the measurements will tend to cluster around the directions imposed by the columns of the mixing matrix, as shown in figure 1, that allows to estimate $\mathbf{A}$ [4]. In many cases, even if the sources do not satisfy the sparsity premise, it is possible to apply a suitable linear transform —STFT, DCT, wavelet, ...— that does allow to represent the signals in a new space in which the coefficients are sparse [3, 4, 5].

In this communication we will assume that the signals are expressed in a suitable domain and that the mixing matrix has been appropiately estimated [6], thus we will focus on the third stage of the problem: estimating the sources from the measurements when the mixing matrix is known. In order to do that, we will introduce in section 2.2, equation (4), a probabilistic model that allows us to adjust the sparsity factor of the sources, i.e., the percentage of source coefficients in the representation domain that are negligible. With that probabilistic model in mind, we will derive heuristic separation approaches in section 2.1 and a Bayesian es-

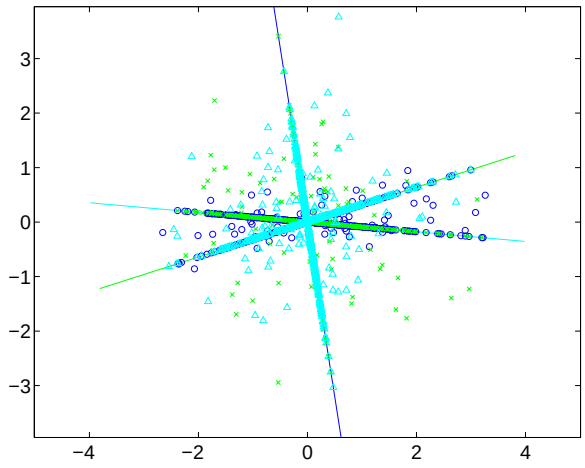

Fig. 1. Columns $\mathbf{a}_j$ of a mixing matrix $\mathbf{A} \in \mathbb{R}^{2\times 3}$ (solid lines) and measurements $\mathbf{x}$ (symbols) for sources with an 80% of zeros.

timator in section 2.2, the behaviour of both families of estimators will be compared in section 3.

2. ESTIMATION OF THE SOURCES

The problem of estimating $\mathbf{s}$ from equation (1) when the mixing matrix $\mathbf{A}$ —which is assumed to be full rank— and $\mathbf{x}$ are known depends of the relation between m and n. If $m = n$ the problem is trivial, because the solution is given by $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$. In the overdetermined case ($m > n$), the pseudo inverse [7] $\mathbf{A}^+$ provides the solution $\mathbf{s} = \mathbf{A}^+\mathbf{x}$ that minimizes the L_2 norm of the residue, $||\mathbf{x} - \mathbf{A}\mathbf{s}||$. 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 norm L_p 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 [8]. As we will show next, if the signals admit a sparse representation, it is possible to design better inversion strategies.

2.1. Heuristic approaches

Let us suppose for a moment that for each source vector $\mathbf{s}$ —in the original domain could be a vector for each time sample, in the transformed domains, one for each coefficient— only the j-th component is not null. In this case, $\mathbf{x}$ will be collinear with the j-th column of the matrix $\mathbf{A}$ and the components of the source vector will be

$$s_k = \frac{\mathbf{a}_j^T \mathbf{x}}{\mathbf{a}_j^T \mathbf{a}_j} \delta_j^k, \quad k = 1, \ldots, n, \tag{2}$$

where the superscript T denotes transpose, and δ_j^k is the Kronecker delta. In a real situation, even with highly sparse sources, the signals will rarely be exactly zero, but at each sample there will be some probability of one of the sources being significatively bigger than the others. To estimate which one that component is, we could choose the one that maximizes the normalized projection on to the directions of each column of $\mathbf{A}$

$$j = \underset{k}{\text{argmax}} \frac{|\mathbf{a}_k^T \mathbf{x}|}{\mathbf{a}_k^T \mathbf{a}_k}, \quad k = 1, \ldots, n. \tag{3}$$

According to this, the following heuristic criterion —that we will call 1D— could be used to invert equation (1): apply (3) to calculate j for each sample and then use (2) to estimate the source vector. The performance of this method will depend on up to what point the premise of having only one significant source component at each sample is satisfied.

Another family of methods could be based on building at each time step a reduced square matrix $\mathbf{A}_r \in \mathbb{R}^{m\times m}$ using m vectors of $\mathbb{R}^m$, chosen between the n column vectors $\mathbf{a}_j$ according to some optimization criterion. The resulting source vector $\mathbf{s}$ will have $n - m$ zeros corresponding to the non-selected columns, and the other components will be given by $\mathbf{A}_r^{-1}\mathbf{x}$. There are many ways of selecting the appropiate columns of the reduced matrix. In [4] a method of this family is proposed for the $m = 2$ case. The criterion it uses is to divide $\mathbb{R}^2$ into the sectors defined by the column vectors $\mathbf{a}_j$ and to choose at each sample those two vectos that surround the measurement $\mathbf{x}$. Figure 1 provides geometric insight on this criterion. The lines represent the n column vectors $\mathbf{a}_j$ of $\mathbf{A} \in \mathbb{R}^{2\times 3}$ and the points represent the measurements. The sources have eighty percent of zeros and, in addition, satisify that no more that two sources are active at the same time. The symbols used for the points are chosen according to which source is imposed to be zero. In order for this method to work, the different symbols should cluster into different sectors; but as it can be seen, that is not always the case. Other possible criterion —that we will call m-D— that is valid for any m is to build the reduced matrix by using the m columns of $\mathbf{A}$ with the biggest projection of $\mathbf{x}$ on to its associated unitary vectors. Another criterion of the same family —that we will call m-DL_2— could be to select the m columns that yield minimum L_2 norm $\mathbf{s}$ which has at least $n - m$ zeros. In the section 3 we will compare these methods for a case with $m = 2$ measurements and $n = 3$ sources.

2.2. Bayesian estimation

If we let $\mathbf{A} = [\mathbf{a}_1|\cdots|\mathbf{a}_n]$, where $\mathbf{a}_j$ are the columns of $\mathbf{A}$, the measurement is

$$\mathbf{x} = \sum_{j=1}^{n} s_j \mathbf{a}_j = s_1\mathbf{a}_1 + \cdots + s_n\mathbf{a}_n.$$

If, at any given time, we knew that at most m components of the signal are non zero, the problem would not be underdetermined any more and we could invert it (provided that we know which are the non zero components). In order to estimate which sources are active at a given time, we introduce the following probabilistic source sparsity model for the distributions of the sources

$$p_{S_j}(s_j) = p_j\,\delta(s_j) + (1-p_j) f_{S_j}(s_j), \quad j = 1,\ldots,n; \quad (4)$$

where the parameter p_j, the sparsity factor, controls the percentage of sparsity of each source, $\delta(\cdot)$ is the Dirac's delta, and $f_{S_j}(s_j)$ are the distributions when the corresponding source is not silent. This densities allow us to parametrically model sources with different degrees of sparsity, and thus provide a framework to characterize the different inversion strategies as a function of the sparsity of the sources.

2.2.1. A priori probabilities

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 $C_{u,\ldots,w}$ that only and all of $s_u,\ldots,s_w$ are non-zero at the same time. According to (4), the *a priori* probabilities of these events (classes) are

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

2.2.2. Conditional probabilities

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 $C_{u,\ldots,w}$ had occurred, when number of active sources is less than m,

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

where

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

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. As we will show in the next section, we do not need to consider the case in which the number of active sources is greater than m.

2.2.3. MAP estimator

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 [10], and parametric estimation methods [11] are among the options. Once the *a posteriori* probabilities are known for all the events, the maximum *a posteriori* (MAP) estimator chooses the optimal source estimates corresponding to the event which maximizes $p(C_{u,\ldots,w}|\mathbf{x})$. If the selected event is $\bar{C}$, then the minimum norm solution provided by the pseudo-inverse is used.

3. NUMERICAL RESULTS

In order to compare the different methods, a problem with $m = 2$ measurements and $n = 3$ sources has been studied, assuming that the mixing matrix $\mathbf{A}$ is known. For the source distributions in (4), we chose Gaussian distributions with zero mean and unit variance for $f_{S_j}(s_j)$, and all p_j were assumed equal. In this case, it is not necessary to estimate $p(\mathbf{x})$ because $p(\mathbf{x}|C_{1,2,3})$ —the only event with a number of active sources greater than m— can be calculated analitically as a multivariate Gaussian with zero mean and variance $\mathbf{A}\mathbf{A}^T$ [12].

For illustrating the behaviour of the MAP estimator, we have performed a simulation using a mixing matrix with columns at 0, π, and $2\pi/3$ radians, and relative amplitudes

0.55, 1, and 0.85. We have evaluated the posterior probabilities $p(C_{u,\ldots,w}|\mathbf{x})$ on a grid, so that the clasification regions —corresponding to different events $C_{u,\ldots,w}$, and therefore to different inversion matrices— can be seen in figure 2. The columns of the mixing matrix are also shown. The color of the different regions correspond to the decissions of the MAP estimator. The black regions correspond to the event in which we estimate that all the sources are active, so that the pseudo inverse is applied. The regions with different gray scales correspond to events with two sources active at the same time —the source that is considered to be zero is the one associated to the mixing column most orthogonal to the region. The bright line on the horizontal axis corresponds to choosing the event C_1 as the most probable — the events C_2 and C_3, corresponding to points collinear to second and third mixing columns do not appear due to the discrete evaluation grid. It can be observed that the decision boundaries are highly non linear —as opposed to the linear decision boundaries of the heuristic criterion that divides the space into sectors according to the mixing columns. As can be observed in figure 1, most of the measurements are in a circle arround the origin with a radius similar to the length of the mixing columns of $\mathbf{A}$, where the non linearity of the regions is bigger. Some insight can also be obtained about the 1D heuristic criterion. Even for measurements collinear with mixing column $\mathbf{a}_u$, the most probable event is not always C_u, as can be observed in figure 2, where the bright horizontal line does not extend after a certain distance from the origin. The reason of this is that as the horizontal mixing column, $\mathbf{a}_1$, is the smallest of the three, then is not very probable to find far from the origin a measurement due only to $\mathbf{a}_1$.

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 (4) 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

$$\text{SNR} = -20 \log \frac{||\hat{\mathbf{s}} - \mathbf{s}||}{||\mathbf{s}||} \quad \text{(dB)}.$$

Figure 3 shows the results. The pseudo inverse solution does not depend on the sparsity factor, since it is fixed once the mixing matrix is known. When the sparsity factor is low, the performance of the pseudo inverse is better than all the heuristic methods, but for a sparsity factor arround 70%, the heuristic methods start to outperform the pseudo inverse. For all the cases, the best performance is obtained with the MAP estimator. It can be noticed that when the number of zeros is very small, the pseudo inverse acts as a lower bound for the performance of the MAP estimator, since there is no additional knowledge about the clustering of the sources.

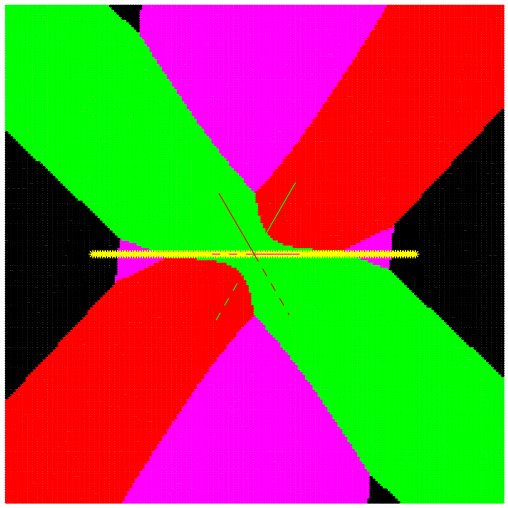

Fig. 2. Classification regions of the MAP estimator and columns of the mixing matrix. The black region corresponds to the event $C_{1,2,3}$, the gray regions correspond to events $C_{u,v}$, the bright horizontal line corresponds to the event C_1.

4. CONCLUSIONS

In this communication, we have studied the problem of separating the sources in the instantaneous underdetermined linear mixing problem. The canonical solution given by the pseudo inverse does not provide a sufficient performance. Since the pseudo inverse is a matrix that depends only on the mixing matrix, and not on the individual samples of measurements, its performance is constant over changing sparsity of the sources. When the sources are sparse enough, heuristic approaches can be formulated that outperform the pseudo inverse solution.

We have considered the case where source densities are parametrized by a sparsity factor, and presented an MAP estimator. By using this additional knowledge of 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. Like the heuristic approaches, the MAP estimator chooses the "best" inversion matrix on a sample by sample basis.

As a final conclusion, in order to achieve good performance, the underdetermined separation problem requires a highly sparse representation of the sources. When the original sources do not satisfy the sparsity condition, as is the

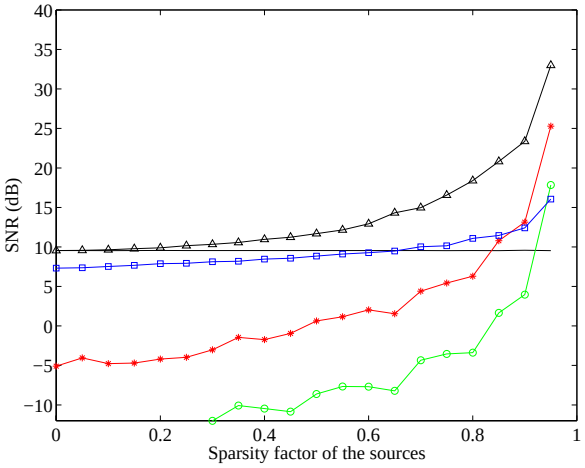

Fig. 3. SNR of the source separation by the heuristic methods 1D ($*$), m-D ($\circ$), and m-DL_2 ($\square$), the MAP estimator ($\triangle$), and the pseudo inverse (solid line).

case with speech signal in the time domain, a suitable linear transformation should be applied beforehand. At the moment, the authors are working on sparse representations for underdetermined speech separation.

Acknowledgments: This work was partially supported by the NSF grant ECS-9900394.

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. In press.

[4] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *submitted to Signal Processing*, 2000.

[5] 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.

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

[7] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, pp. 77–85. SIAM, 1997.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, pp. 270–274. Johns Hopkins University Press, 3rd ed., 1996.

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

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

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

[12] L. L. Scharf, *Statistical Signal Processing*, pp. 59–60. Addison Wesley, 1991.