Sparse Separation: Principles and Tricks

Barak A. Pearlmutter* **Vamsi K. Potluru***

Abstract

Blind separation of linearly mixed white Gaussian sources is impossible, due to rotational symmetry. For this reason, all blind separation algorithms are based on some assumption concerning the fashion in which the situation departs from that insoluble case. Here we discuss the assumption of *sparseness* and try to put various algorithms that make the sparseness assumption in a common framework. The main objective of this paper is to give some rough intuitions, and to provide suitable hooks into the literature.

Introduction

What is blind separation?

Given observations from sensors such as microphones or MEG/EEG recordings, the process of extracting the underlying sources is called source separation. Doing so without strong additional information about the individual sources (aside from the weak assumption that the sources are independent and the mixing linear) or constraints on the mixing matrix is called *blind separation*.

Notationally, the $N \times T$ data matrix $\mathbf{X}$ has columns $\mathbf{x}(t)$ corresponding to the sensor readings at time t, so $x_i(t)$ is the reading of sensor i at time t. Blind separation (BSS) has a restricted case where the number of sources is the same as the number of sensors, noise is to be ignored, and the mixing process is instantaneous, called ICA. These ICA algorithms can be thought of as producing M components (estimated recovered sources) where component j has a spatial distribution, (the column vector) $\mathbf{a}_j$, and time course $s_j(t)$ (sometimes treated as a row vector) . ICA methods are decomposition algorithms, like PCA, in that they decompose the data into a sum of outer products

$$\mathbf{X} = \sum_{j=1}^{M} \mathbf{a}_j \, \mathbf{s}_j = \mathbf{AS} \tag{1}$$

where $\mathbf{S}$ has as its rows the time courses $\mathbf{s}_j$, and the matrix $\tilde{\mathbf{A}} = \mathbf{W}^{-1}$ has the vectors $\mathbf{a}_i$ as its columns. In blind separation the *unmixing matrix* $\mathbf{W}$ is estimated purely from information in the signals $\mathbf{X}$. It is generally preferable, for numeric reasons, to directly estimate $\mathbf{W}$ rather than to attempt to estimate $\mathbf{A}$ as an intermediate computation.

An M-dimensional Gaussian distribution can, by change of basis, be made spherical. This rotational symmetry makes it impossible to perform blind separation on white Gaussian sources. All BSS and ICA algorithms make some assumption regarding the fashion in which the sources differ from that impossible case. These can be thought of as cues for separation taken advantage of by various algorithms. It is important to note in passing that truly Gaussian signals are in practice extremely rare in the real world, except in sensor noise.

What is sparseness?

A signal is *sparse* when it is zero or nearly zero most of the time, or at least more of the time than one might expect from its variance. Another way of phrasing this is that its marginal distribution has a peak at zero larger than a Gaussian would, or has fatter tails than those of a Gaussian. Sparseness occurs often in the real world: sound comes from a small number of focal sources, rather than being generated by a plethora of acoustic sources evenly distributed in the environment; matter in the world is clumped; the people on the surface of the Earth are distributed in focal clumps. These properties form the basis of various inverse algorithms: methods by which properties of the world can be estimated from measurements.

Sites of neuronal activity are focal, as shown by FMRI studies. A typical algorithm that takes advantage of such sparseness is FOCUSS (Gorodnitsky and Rao, 1997), which estimates maximally sparse locations of activation and time courses to match EEG/MEG data to a forward model of electromagnetic propagation through the head.

Sparse separation

If $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$ where the s_j are independent and sparse, then the marginal density of $\mathbf{x}(t)$ cannot be spherically

*Hamilton Institute, NUI Maynooth, Co. Kildare, Ireland.

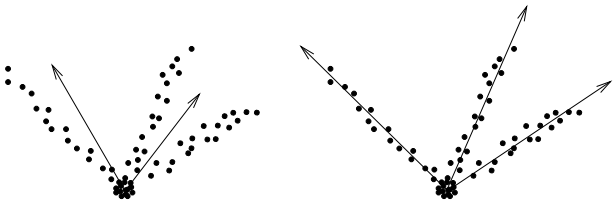

Figure 1: A mixture of $M = 3$ sources as scatter plot in $N = 2$ dimensional sensor space. Optimal complete basis (left) results in non-sparse data, while optimal overcomplete basis (right) allows data to be represented as a sparse subset of the coefficients.

symmetric. This can be exploited to find its preferred basis, *i.e.* to perform blind separation. Often sparseness is the sole cue used for separation.

The highly influential Infomax algorithm of Bell and Sejnowski (1995) can be viewed as making the sparseness assumption in a subtle fashion, by performing a maximum likelihood fit (Pearlmutter and Parra, 1996; Cardoso, 1997) against a simple parametric model involving a linear mixture of fat-tailed distributions. From this perspective there are two interesting tricks to the BS-Infomax algorithm. First, the model is parameterized not by the mixing matrix, as is natural for a forward model, but rather by its inverse. Although the inverse contains the same information, it makes for a much simpler update rule and enormously superior numeric conditioning. The second trick is to take a naive stochastic gradient descent algorithm and modify it by multiplying the gradient by a particular positive-definite matrix, $\mathbf{W}^T\mathbf{W}$, where $\mathbf{W}$ is the current estimate of the unmixing matrix. While Amari et al. (1996) provides a theoretical derivation of this based on information geometry, a naive perspective would be that the multiplier chosen fortuitously happens to eliminate a matrix inversion, making each step of the algorithm much faster, and also makes the convergence rate independent of the condition number of the unmixing matrix.

As shown schematically in Figures 1 and 2, when the sources are sufficiently sparse it is possible to use clustering to directly estimate the rows of the mixing matrix $\mathbf{A}$, which correspond to the coordinates of the cluster centers. This is due to the observation that with sufficiently sparse [1] This is due to the observation that with sufficiently sparse sources at most one source will generally be active at a time. When only one source is active the attenuation between that source and the sensors corresponds to the sensor reading vector, up to a constant factor.

Often sources, considered in the time domain, are only

[1] Actually line directions, since the clusters are lines through the origin. One can use conventional clustering algorithms by projecting onto unit sphere, or use a special-purpose clustering algorithm which characterizes each cluster by a line through the origin instead of a center point.

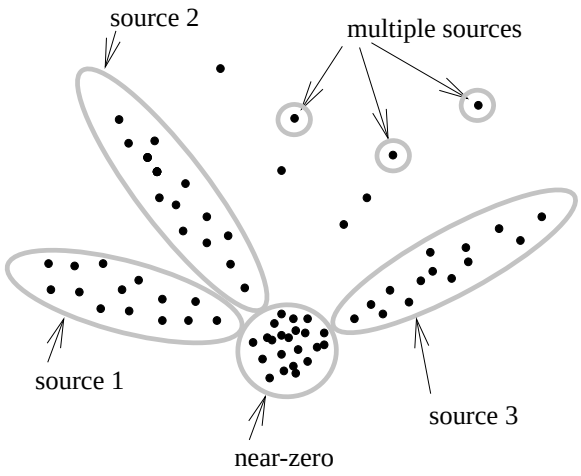

Figure 2: Diagrammatic scatter plot of data, in sensor space, showing how sufficiently sparse sources allow direct estimation of the mixing matrix using clustering, and even direct assignment of coefficients to sources even when $N < M$. Coefficients can be classed into islands depending on whether they result from no sources (near-zero), a single source, or multiple sources. When few coefficients come from multiple sources correct separation can be accomplished by simply reconstructing sources from the coefficients assigned to them.

slightly sparse, *i.e.* have slightly fat tails, but (as shown in Figure 3, doing a transform to another domain, *e.g.* a short-time FFT or a wavelet transform, makes each source extremely sparse in coefficients. This insight makes algorithms of this class practical, but limits their applicability to domains where this assumption is met. In particular, there must be sufficiently little sensor noise so as to not bring the background fluctuations up to the significant coefficients. Chen et al. (1999) noted the importance of sparsity in a transform domain constructed an algorithm which finds a transform in which the data becomes sparse. In their work sparsity of the data in the transformed domain is explicitly modeled by minimizing the L_1 norm of the coefficients. This gives an analysis of the data in the overcomplete basis.

In the $N = M$ case, once the mixing matrix is found it can be inverted and the problem is solved. However when $N < M$ it is not optimal to linearly map from sensors back to sources even when A is known exactly. Instead some nonlinear process must be used, and there is unavoidable error. This can be easily seen by noting that the posterior, according to Bayes' rule, broadens from a single delta function into a range of possible reconstructions.

The ability to directly estimate the mixing matrix by

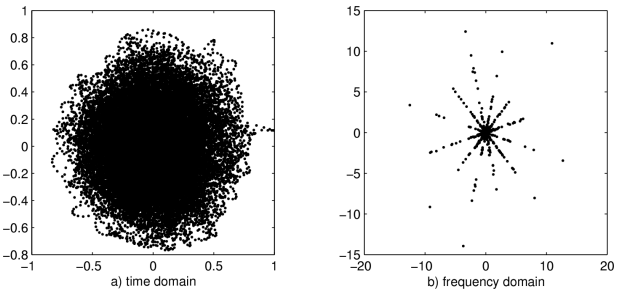

Figure 3: Scatter plot of the readings of two sensors receiving a synthetic instantaneous mixture of six voices. In the time domain (left) the signals are not sparse, but the coefficients resulting from a short-time FFT (right) are sparse, allowing visual identification of the six sources. Taken with permission from Zibulevsky et al. (2001).

clustering in some domain in which the sources become extremely sparse was exploited by Zibulevsky et al. (2001); Zibulevsky and Pearlmutter (2001), which to our knowledge exhibited the first practical blind algorithm for the case of $N < M$. Strong sparseness can be considered minimization of the L_0 norm, since that norm measures the number of non-zero coefficients and minimizing this number maximizes sparseness. It is known in the optimization community that the L_0 norm is often well approximated by the L_1 norm. Here this results in a linear programming formulation of the problem of partial assignment of coefficients to sources.

The algorithm of Roweis (2001) is similar, but can work with $N = 1$, *i.e.* with only a single sensor. There the assumption is unique assignment (each coefficient to one source) and strong source models. The source models are rich enough to give strong correlational structure to the coefficients, *i.e.* to say which sets of coefficients belong to the same source. In this case, HMMs of the sources allowed separation of two people speaking using one microphone. Note that such strong source models mean the algorithm was far from blind.

DUET

The DUET algorithm (Rickard and Dietrich, 2000) combines these two ideas, in that it uses clustering in a two-dimensional space to find "islands" belonging to each source (Figure 4), and then does hard assignment of coefficients to sources. With DUET the two-dimensional space is a time-frequency transform space, necessitated by a more complex acoustic mixing process that includes delays. The model in DUET is that of time delayed and attenuated sources captured at $N = 2$ sensors through an

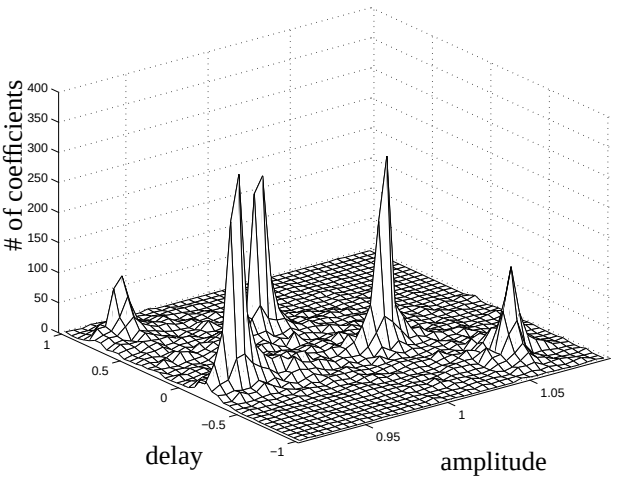

Figure 4: DUET algorithm assigns the coefficients in each island to separate sources. (From Rickard and Dietrich (2000, Figure 1), with permission.)

anechoic mixing process,

$$x_1(t) = \sum_{j=1}^{M} s_j(t) + \text{noise} \quad (2)$$

$$x_2(t) = \sum_{j=1}^{M} a_j \, s_j(t - \delta_j) + \text{noise} \quad (3)$$

The mixing parameters here correspond to attenuation and delay factors, given by the vectors $\mathbf{a}$ and δ, respectively. It is assumed that the distance between the sensors is small enough to make the narrowband assumption as referred to in the array signal processing literature, *i.e.* the delays must be less than some maximum tolerable delay. Transforming the signals into a time-frequency domain and assuming sparsity (referred to in Rickard and Dietrich (2000) as the W-disjoint orthogonal property), DUET finds clusters of the sources as a histogram mapping of the attenuations vs delays. These in turn can be used to find the mixing parameters and the sources. It can be observed that there is no restriction on the number of sources, which can be estimated from the number of distinct cluster centres. The only requirement is that they satisfy the W-disjoint orthogonal property. For speech, the DUET algorithm was able to separate five speakers using just two sensors.

DUET was, to our knowledge, the first practical blind algorithm that could operating in real time on real acoustic data with more sources than sensors. Its primary limitation is that the algorithm requires delays to manifest themselves as phase shifts in a time-frequency transform domain. In concrete terms, this means the two microphones must be placed within a wavelength of each other, *i.e.* for

voice a maximum distance of a few centimeters.

Learning overcomplete representations

Lewicki and Sejnowski (2000) represent data in an overcomplete basis using an underlying sparse source assumption. The sparsity is modelled by using maximum likelihood estimation with a peaked prior distribution such as a Laplacian, which corresponds to an L_1 norm in the same way that a Gaussian corresponds to an L_2 norm. (Standard approaches like Fourier and wavelets do not need to do this, as they are not over-complete and the representation of the signal is unique.)

With a Laplacian prior and no noise, the problem reduces to linear programming. The data vectors are assumed to be independent (for simplicity) and the data likelihood is calculated. This turns out to be intractable in the overcomplete case, and here it is approximated by fitting a Gaussian distribution around the posterior mode of the sources. Since the data is modelled by an overcomplete basis, we have *a priori* information about the sparsity. This makes for a better representation to extract the underlying structure of the data, as shown in Figure 1.

Support Vector Machines

SVMs, a kind of classifier used in machine learning, are sparse in another sense: the SVM training algorithm uses only a small but carefully selected subset of the training samples to define the classification boundaries. Hochreiter and Mozer (2001) show a surprising link between the sparse separation algorithms of Zibulevsky and Pearlmutter (2001) and SVM training. As sparse methods invade more domains, we expect the sparseness prior to raise its head often in machine learning, in various guises.

Conclusion

The sparseness assumption is surprisingly powerful, not just for denoising and wavelets but also for inverse problems and source separation. It has been conjectured that all useful statistical structure in signals can be represented as sparseness in an appropriate domain. Although we cast no light here on that strong conjecture, it is certainly the case that the sparseness assumption has led to a large number of important and practical algorithms. As this path continues, it seems likely that sparse algorithms will soon enjoy wide deployment in a variety of applications.

Acknowledgements

Supported by US NSF CAREER 97-02-311 and Science Foundation Ireland.

References

Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 8*. MIT Press.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Cardoso, J.-F. (1997). Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Gorodnitsky, I. F. and Rao, B. D. (1997). Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616.

Hochreiter, S. and Mozer, M. C. (2001). Monaural separation and classification of mixed signals: A support-vector regression perspective. In ee, T.-W., Jung, T.-P., Makeig, S., and Sejnowski, T. J., editors, *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, CA.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–365.

Pearlmutter, B. A. and Parra, L. C. (1996). A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing*, pages 151–157, Hong Kong. Springer-Verlag.

Rickard, S. and Dietrich, F. (2000). DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, pages 311–314, Pocono Manor, PA.

Roweis, S. T. (2001). One microphone source separation. In *Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press.

Zibulevsky, M. and Pearlmutter, B. A. (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882.

Zibulevsky, M., Pearlmutter, B. A., Bofill, P., and Kisilev, P. (2001). Blind source separation by sparse decomposition in a signal dictionary. In Roberts, S. J. and Everson, R. M., editors, *Independent Components Analysis: Principles and Practice*, pages 181–208. Cambridge University Press.