CLUSTERING APPROACH TO SQUARE AND NON-SQUARE BLIND SOURCE SEPARATION

Marc M. Van Hulle
K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg, Herestraat, B-3000 Leuven, BELGIUM
Tel.: + 32 16 34 59 61, Fax: + 32 16 34 59 93
E-mail: marc@neuro.kuleuven.ac.be
Web: simone.neuro.kuleuven.ac.be

Abstract. Recently, a number of heuristic techniques, mostly based on topographic maps, have been introduced in order to overcome (some) of the limitations of the Blind Source Separation (BSS) algorithms that are rooted in the theory of Independent Component Analysis. Here, we introduce a new heuristic that relies on the tendency of the mixture samples to *cluster* around the source directions in mixture space. We use linear mixtures of speech signals and consider BSS problems with not only square but also non-square mixing matrices (less mixtures than sources).

INTRODUCTION

The majority of the Blind Source Separation (BSS) algorithms are based on the theory of Independent Component Analysis (ICA) [1]. The idea is to find the statistically independent source signals $\mathbf{s}(t) = [s_i(t)]$, $i = 1, ..., m$, with t a time index, from the mixtures $\mathbf{v}(t) = A_{mm}\mathbf{s}(t)$, with A_{mm} the $m \times m$ mixing matrix, by optimizing a criterion such as maximum likelihood (ML). However, for reasons of the assumed linearity of the mixing, the ensuing increased algorithmic complexity to encompass non-linear mixtures, or the inability to address the case where there are fewer mixtures than sources, a number of heuristic techniques have been devised, mostly based on topographic maps. They meet one or several of the previous concerns and fall in two broad categories by their ability to separate either: 1) linear- or mildly non-linear mixtures of signals with either sub-Gaussian (*i.e.* approximatively uniform) [2, 3, 4, 5, 6], or 2) super-Gaussian (*i.e.* sharply peaked) source densities [7, 8]. The idea behind the second category, and to which we will restrict ourselves in this contribution, is as follows. The column vectors of the mixing matrix A_{mm} define, for the linear case, an independent component coordinate system the basis vectors of which are usually non-orthogonal. For sharply peaked source signals, such as speech signals, when only one signal s_i is

0-7803-5673-X/99/$10.00 © 1999 IEEE

active at a time, the mixed signal will be directed along a line parallel to A_{mmi}. The direction of this line can be retrieved by least-squares fitting [7] or by some sort of edge detection procedure *e.g.* the Hough transform [8]. Here, we propose a completely different approach. Basically, it relies on the tendency of the mixture samples to *cluster* around the source directions. We perform density-based clustering with a topographic map trained with the kernel-based Maximum Entropy learning Rule (kMER) [6].

CLUSTERING APPLIED TO BSS

Consider two sharply peaked source signals that are linearly mixed. As an example, we consider speech signals from the TIMIT database [9], namely *dr1/fcjf0/sa2* (TR_1) and *dr1/ftbr0/sa1* (TR_2), respectively 34509 and 55092 samples long and sampled at 16 *kHz* (Fig. 1A,B). We scale these signals so that they have a unit peak amplitude, and add zeros at the end of the first signal so that it also contains 55092 samples. The scaled signals are then linearly mixed using the following mixing matrix:

$$A_{22} = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}. \qquad (1)$$

The resulting scatter plot is shown in Fig. 1C. The high-density directions correspond to the coordinate axes of the source signals and thus to the directions defined by the column vectors of the mixing matrix.

In order to separate these speech signals from their mixtures, we adopt a three-stage approach. We first train a topographic map with kMER and use it for generating an estimate of the mixture distribution. We then use this estimate to perform a density-based clustering analysis and, finally, we perform BSS in both a parametric and a non-parametric manner.

Stage 1: Density estimation with kMER

Topographic map formation

Let A be a two-dimensional lattice of N formal neurons. To each neuron i corresponds a circular Receptive Field (RF) region with radius σ_i and center $\mathbf{w}_i = (w_{i1}, w_{i2})$. Both the centers and the radii are trained with our learning rule kMER of which the mathematical details, including a proof of convergence, are discussed elsewhere [6, 10]. In summary, the centers $\mathbf{w}_i$ are updated so that they become the medians of the neurons' RF regions, and the radii σ_i are updated so that the probability for neuron i to be active will be $\frac{1}{N}$, $\forall i$.

We now use kMER for developing a *lin-polar* map of the mixture coordinate frame with amplitude $R = \sqrt{v_1^2 + v_2^2}$ and phase angle:

$$\theta = \min(\arctan \frac{v_2}{v_1}, \pi + \arctan \frac{v_2}{v_1}). \qquad (2)$$

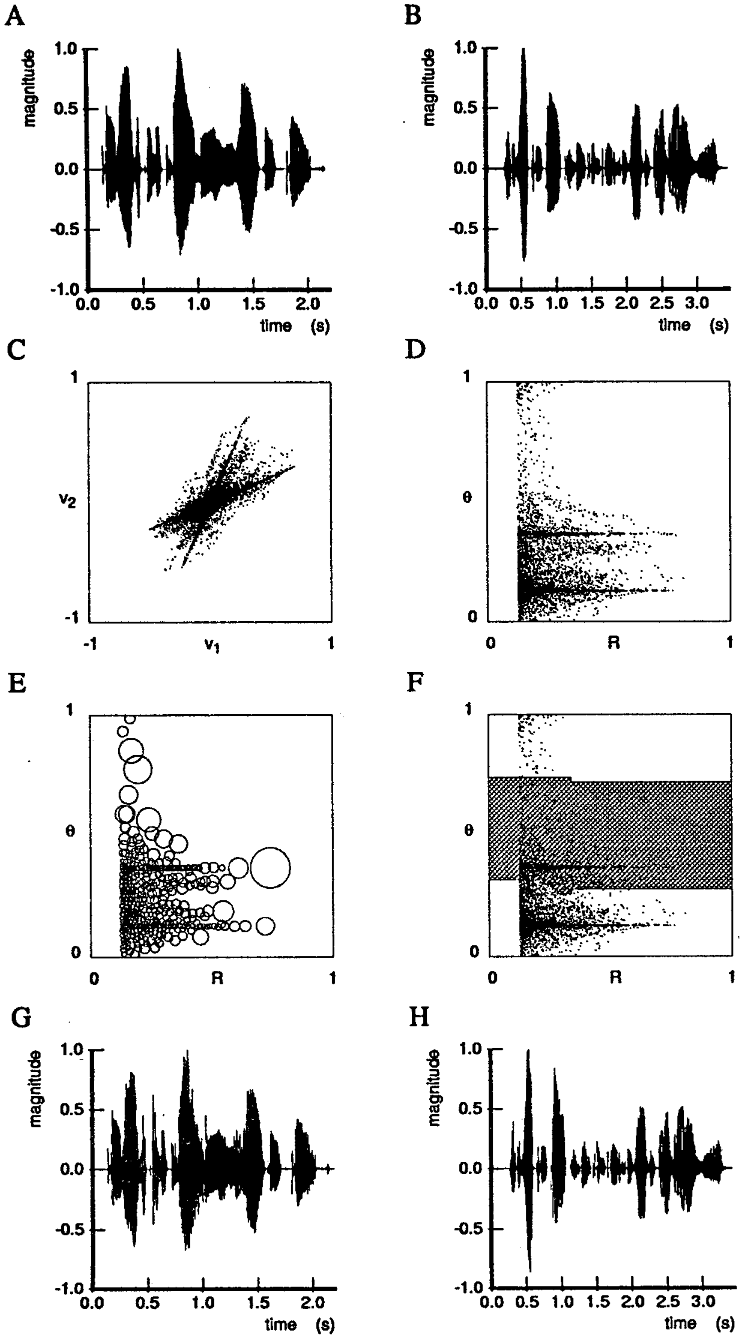

Figure 1: Continues on next page.

Figure 1: Clustering approach to Blind Source Separation. (A,B) Two example speech signals, taken from the TIMIT database, that are subject to linear mixing. (C,D) Scatter plot of the signal mixtures (v_1, v_2) in the original Cartesian coordinate frame (C) and in the lin-polar frame (R, θ) for $R \geq 0.125$ (D). No signal sharpening or whitening has been performed prior or after the mixing process. (E) RF regions of a 24 × 24 lattice trained with kMER on the data shown in (D). (F) Result of density-based clustering with the SKIZ technique, starting from the variable kernel density estimate corresponding to (E). The "influence zones" of the two clusters, of which one is shaded, are shown together with the training data. Note the effect of the toroidal extension along the θ-axis. (G,H) The speech signal estimates obtained after demixing with the first strategy (CL). The signals are scaled for unit peak amplitude.

Such a map is more suited for density-based clustering than a Cartesian map: the mixture samples now display the tendency to cluster along lines parallel to the R-axis and hence, the clusters will be better defined when the mixture samples have smaller amplitudes. In order to avoid the region with the highest density in mixture space, *i.e.* where the source signal amplitudes are close to their mean values, we only consider a mixture sample when its amplitude $R \geq R_{min}$, with $R_{min} > 0$. Here, we will take $R_{min} = 0.125$ so that $M = 7859$ lin-polar transformed mixture samples remain. For the sake of convenience, we scale the range of the polar axis from $[0, \pi]$ to $[0, 1]$ (Fig. 1D). Finally, we re-consider the $N = 24 \times 24$ lattice, with the same Gaussian neighborhood function, cooling scheme, number of epochs, and radius initialization procedure as in [6]. The weights are randomly initialized in the unit square $[0, 1]^2$. Contrary to others, we do not pre-process the audio signals to make them even more sharply tuned [8], and we do not perform any whitening of the lattice inputs prior to training [5, 7]. The resulting map of the RF regions is shown in Fig. 1E.

Density estimation

Once the lattice has converged, we determine the density estimate of the mixture distribution as follows. We allocate at each neuron weight $\mathbf{w}_i$ a radially-symmetric, equi-volume kernel with center $\mathbf{w}_i$ and radius σ_i. Since the radii are adapted to the local sample density, and since the kernels have an equal volume, we obtain a *variable* kernel density estimate [11]:

$$\widehat{p_{\rho_s}}(\mathbf{v}) = \sum_{i=1}^{N} \frac{\exp(-\frac{\|\mathbf{v}-\mathbf{w}_i\|^2}{2(\rho_s \sigma_i)^2})}{Z_i}, \tag{3}$$

when Gaussian kernels are used, and with ρ_s a factor with which the radii σ_i are scaled, and Z_i a proper normalizing factor. The "optimal" degree of smoothness ρ_s is determined with the technique described in [10]. The result is $\rho_s = 1.3$, when optimizing in steps of 0.1.

Stage 2: Density-based clustering

In order to determine how many clusters there are, and to identify the areas spanned by them (*i.e.* the "influence zones"), we apply the SKeleton by Influence Zones (SKIZ) technique [12]. Basically, this technique considers L equidistant levels at which cross-sections with $\widehat{p_{\rho_s}}$ are taken. From the connected regions obtained in this way, the "influence zones" are identified. For the simulations reported on, we have used $L = 16$ levels. The result for Fig. 1E is shown in Fig. 1F.

Stage 3: Blind source separation

We compare three strategies. First, we consider the mixture samples, which belong to a given influence zone, as samples originating from the same source channel. In other words, we perform a classification of the mixture samples (case labeled CL). The source signals are then reconstructed by supplementing the R-value of the mixing signal with the corresponding sign (by definition of our phase angle θ). Note that, as a result of classification, there is only one "winning" channel at any given time. For source signals with sharply peaked densities, such as speech signals, this is a reasonable assumption since these signals hover around their mean values most of the time.

Second, we classify the weights of the converged lattice, and determine the median along the θ-axis for each weight class: the two medians then serve as regression lines for the corresponding weight classes (case labeled CL_REG_W). (Note that medians are less sensitive to outliers than means and, in this way, a better definition of the regression lines is obtained in practice.) From the resulting θ's, the column vectors of the mixing matrix A_{22} can be determined, up to a scale factor, the inverse of the mixing matrix calculated, A_{22}^{-1}, and the source signals reconstructed.

Finally, we repeat the previous case but now starting from a classification of the mixture samples, instead of the lattice weights (case labeled CL_REG_S).

RESULTS

The source signal estimates obtained with the first strategy (CL case) are shown in Fig. 1G,H; the estimates for the other two strategies cannot be distinguished from the originals by visual inspection. In order to quantify all these results, we first scale the source signal estimates so that their variances equal those of the corresponding source signals s_1 and s_2. We then determine the mean squared error (MSE) between the originals and the estimates, as well as the signal-to-noise-ratio (SNR) (in dB). Furthermore, we also consider two other speech signals for testing purposes, $dr5/fbmh0/sx146$ (TE_1) and $dr1/ftbr0/sa2$ (TE_2), which are, respectively, 41268 and 44852 samples long. The results are summarized in Table 1. Note the high precision achieved with the CL_REG_S and CL_REG_W strategies (sometimes bounded by the single precision representation used in the algorithm). The estimate obtained for the mixing matrix eq. (1) is, *e.g.* for the CL_REG_S case:

$$\widehat{A_{22}} = \begin{pmatrix} 0.6962 & 0.3000 \\ 0.3038 & 0.7000 \end{pmatrix}. \tag{4}$$

The quality of our source estimates (CL case) seems to be superior to that obtained by Lin and co-workers, albeit that the estimates they showed were for a mildly non-linear but still whitened mixing (compare their Fig. 4 in [7] to our Fig. 1 – unfortunately they did not show the results for their linear mixing case or quantify the performance). When a regression analysis on the weights is performed, Lin and co-workers report for their linear case a mismatch of less than 1 % in the orthogonality of the source signal estimates, *i.e.* when the estimates are expressed as linear combinations of the original source signals, whereas we find 2.5 % on average (CL_REG_W strategy). This discrepancy in performance is, in our opinion, due to the fact that Lin and co-workers have whitened their mixture signals: whitening orthogonalizes the independent coordinate axes.

Table 1: Performance for a square BSS problem expressed in terms of the mean squared error (MSE) and signal-to-noise-ratio (SNR) in dB (between brackets). The first two signals (TR_1, TR_2) were used for training; the last two for testing (TE_1, TE_2).

MSE	CL	CL_REG_W	CL_REG_S
TR_1	2.16×10^{-3} (7.52)	5.64×10^{-8} (53.3)	7.41×10^{-13} (102.2)
TR_2	1.07×10^{-3} (9.41)	2.75×10^{-5} (25.3)	1.10×10^{-6} (39.3)
TE_1	4.60×10^{-3} (5.25)	1.12×10^{-7} (51.4)	4.10×10^{-12} (95.7)
TE_2	2.68×10^{-3} (8.39)	3.49×10^{-5} (27.3)	1.39×10^{-6} (41.2)

BSS FROM FEWER MIXTURES

By virtue of our clustering approach, we can also tackle the "non-square" BSS problem where the number of mixtures is less than the number of sources. Additional sources introduce additional density directions in mixture space. As an example, we consider the case of 3 sources and 2 mixtures. The CL strategy stays the same but the other two have to be adapted since the mixing matrix is no longer square. We suggest the following heuristic: we first project the mixture sample onto each of the three source directions that we have identified. We then select the source direction which produces the largest projection result (*i.e.* the smallest angular difference in the lin-polar map) and assign the projection result to the corresponding source channel. For the residue of that projection, we invert the remaining 2×2 mixing matrix and determine the other two source signal estimates. We use the previous

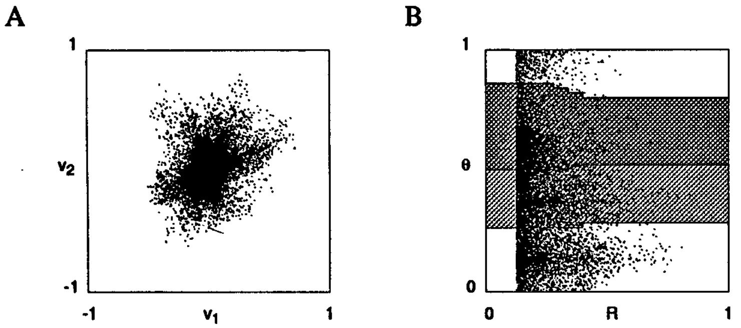

Figure 2: Blind Source Separation from fewer mixtures. Scatter plot produced by a mixture of three signals in the Cartesian coordinate frame (A) and the lin-polar frame (R, θ) for $R \geq 0.125$ (B). In (B), the influence zones of the three clusters are indicated with different gray levels.

two source signals TR_1 and TR_2 and select a third one form the TIMIT database TR_3 ($dr5/fbmh0/si1136$; 71476 samples) for training our lattice. We use the following mixing matrix:

$$A_{23} = \begin{pmatrix} 0.7 & 0.3 & -0.3 \\ 0.3 & 0.7 & 0.7 \end{pmatrix}, \tag{5}$$

and again consider 55092 samples for mixing (Fig. 2A). This results in 13519 lin-polar transformed mixture samples (Fig. 2B). For testing, we use the TIMIT signals $dr1/fcjf0/sa1$ (46797 samples), $dr1/ftbr0/sa2$ (44852 samples), and $dr5/fbmh0/sx146$ (41268 samples) (labeled further as TE_1, TE_2, TE_3). Again, no signal sharpening or whitening has been performed. The clustering result is shown in Fig. 2B; the BSS results are summarized in Table 2. The estimate of the mixing matrix eq. (5), obtained with the CL_REG_S strategy, is:

$$\widehat{A_{23}} = \begin{pmatrix} 0.7133 & 0.2930 & -0.3000 \\ 0.2867 & 0.7070 & 0.7000 \end{pmatrix}. \tag{6}$$

Table 2: Performance for a non-square 2 × 3 BSS problem.

MSE	CL	CL_REG_W	CL_REG_S
TR_1	6.00×10^{-3} (3.08)	2.49×10^{-3} (6.90)	2.39×10^{-3} (7.07)
TR_2	4.24×10^{-3} (3.42)	4.58×10^{-3} (3.08)	4.11×10^{-3} (3.55)
TR_3	2.93×10^{-3} (4.91)	2.27×10^{-3} (6.02)	2.26×10^{-3} (6.04)
TE_1	6.55×10^{-3} (1.99)	3.68×10^{-3} (4.49)	3.58×10^{-3} (4.61)
TE_2	6.66×10^{-3} (4.44)	7.17×10^{-3} (4.12)	6.31×10^{-3} (4.68)
TE_3	4.44×10^{-3} (5.40)	3.44×10^{-3} (6.52)	3.39×10^{-3} (6.57)

CONCLUSION

We have introduced a new heuristic approach to the Blind Source Separation (BSS) of linear mixtures of signals with sharply peaked source densities, such as speech signals. Our technique relies on the tendency of the mixture samples to *cluster* around the source directions in mixture space. We have performed a density-based clustering analysis in order to determine how many source channels there are, and how to separate the source signals from the observed mixtures. We have quantified the BSS performance in the case of square as well as non-square mixing matrices applied to speech signals taken from the TIMIT database [9].

Acknowledgments

The author wishes to thank T. Gautama for his help in developing the "non-square" BSS example. M.M.V.H. is a research associate of the Fund for Scientific Research – Flanders (Belgium) and is supported by research grants received from the Fund for Scientific Research (G.0185.96), the National Lottery (Belgium) (9.0185.96), the Flemish Ministry of Education (GOA 95/99-06), and the Flemish Ministry for Science and Technology (VIS/98/012).

REFERENCES

[1] P. Comon, "Independent component analysis – a new concept?," Signal Processing, vol. 36(3), pp. 287-314, 1994.

[2] T. Kohonen, K. Raivo; O. Simula, O. Ventä, and J. Henriksson, "Combining linear equalization and self-organizing adaptation in dynamic discrete-signal detection," Proc. IJCNN'96 (San Diego, CA, June 17-21), pp. I223-I228, 1996.

[3] E. Moreau, and O. Macchi, "High-order contrasts for self-adaptive source separation, Int'l J. of Adaptive Control and Signal Processing, vol. 10, pp. 19-46, 1996.

[4] P. Pajunen, A. Hyvärinen, and J. Karhunen, "Nonlinear Blind Source Separation by Self-Organizing Maps," Progress in Neural Information Processing (Proceedings of the International Conference on Neural Information Processing, ICONIP'96), S.-I. Amari, L. Xu, L.-W. Chan, I. King, and K.-S. Leung (Eds.), Vol. 2, pp. 1207–1210, 1996.

[5] M. Herrmann, and H.H. Yang, "Perspectives and limitations of self-organizing maps in blind separation of source signals," Progress in Neural Information Processing (Proceedings of the International Conference on Neural Information Processing, ICONIP'96), S.-I. Amari, L. Xu, L.-W. Chan, I. King, and K.-S. Leung (Eds.), Vol. 2, pp. 1211-1216, 1996.

[6] M.M. Van Hulle, "Kernel-based equiprobabilistic topographic map formation," Neural Computation, vol. 10, pp. 1847-1871, 1998.

[7] J.K. Lin, D.G. Grier, J.D. Cowan, "Faithful representations of separable distributions," Neural Computation, vol. 9, pp. 1305-1320, 1997.

[8] J.K. Lin, D.G. Grier, J.D. Cowan, "Feature extraction approach to blind source separation," Proc. IEEE NNSP97 (Amelia Island Plantation, Florida), pp. 398-405, 1997.

[9] TIMIT, "DARPA TIMIT acoustic-phonetic speech corpus," NIST Speech Disc 1-1.1 (CD-ROM), 1990.

[10] M.M. Van Hulle, "Clustering with kernel-based equiprobabilistic topographic maps," Proc. IEEE NNSP98 (Cambridge, UK), pp. 204-213, 1998.

[11] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall: London, 1986.

[12] J. Serra, Image analysis and mathematical morphology, New York: Academic Press, 1982.