

Application of Geometric Dependency Analysis to the Separation of Convolved Mixtures

Samer Abdallah¹ and Mark Plumbley¹

Centre for Digital Music, Queen Mary, University of London
{samer.abdallah, mark.plumbley}@elec.qmul.ac.uk
<http://www.elec.qmul.ac.uk/digitalmusic>

Abstract. We investigate a generalisation of the structure of frequency domain ICA as applied to the separation of convolved mixtures, and show how a geometric representation of residual dependency can be used both as an aid to visualisation and intuition, and as tool for clustering components into independent subspaces, thus providing a solution to the source separation problem.

1 Introduction

Geometric dependency analysis (GDA) was introduced in [1, ch. 8] as way to represent geometrically the residual dependencies in a distributed representation such as those generated by ICA, both as an aid to visualisation and as a basis for further processing. In this paper, we investigate how ICA and GDA, when applied to two-channel audio data, can yield a solution to the problem of separating convolutively mixed sources. The approach is conceptually quite simple in that it involves very few assumptions about the problem domain: the fact that there are two microphones is not explicitly modelled; neither is the assumption that the sources are mixed convolutively in the time domain. Instead, after training, the ICA weight matrix implicitly represents these aspects of the system. The final separation of the sources is based on clustering of components in a low-dimensional geometric space, which could in principle be done in an unsupervised manner, though in the present system it was done manually.

The description below will be in terms of a 2-by-2 (2 microphones, 2 sources) system, but can be generalised naturally to an m -by- m system.

2 An overview of frequency domain source separation

We begin with an overview of a typical frequency-domain approach to the separation of convolved mixtures, (see, e.g., [2] for more details) emphasizing how the entire system can be understood as a composition of constrained sparse matrices followed by a partition of the resulting components into two subspaces.

First, the signals from the two microphones are buffered into frames of length L samples, which we will denote by the vectors \mathbf{x}_1 and \mathbf{x}_2 , both in \mathbb{R}^L . The next step, motivated by the duality between convolution in the time domain

and multiplication in the frequency domain, is to compute the discrete Fourier transform to each frame. This is usually done using a complex-valued fast Fourier transform (FFT); however, each complex Fourier coefficient then represents a 2-dimensional subspace of \mathbb{R}^L . Since we aim to understand the overall process in terms of an analysis of subspaces, we choose to work instead with a real-valued Fourier transform: each coefficient then represents a 1-dimensional subspace, and higher dimensional subspaces must be formed explicitly by grouping components. In this case, the Fourier transform can be represented as an $L \times L$ orthogonal matrix \mathbf{F} , where the rows of \mathbf{F} form an orthonormal basis of \mathbb{R}^L . Assuming L is even, these basis vectors are sinusoids covering $L/2 + 1$ different frequencies: the zero and Nyquist frequencies are represented by one basis vector each, while the other frequencies each inhabit 2-D subspaces spanned by two basis vectors in quadrature phase.

To do frequency domain ICA, the corresponding per-frequency subspaces from both microphones are brought together to form $L/2 + 1$ low-dimensional ICA problems, each of which is solved independently. These two steps can be represented as the product of a permutation matrix (to interleave the Fourier coefficients from the two channels), and a constrained *block diagonal* ICA weight matrix \mathbf{V} , where the first and last blocks are 2×2 and the rest are 4×4 . The entire process so far can therefore be written as

$$\mathbf{s} = \mathbf{V}\mathbf{P} \begin{pmatrix} \mathbf{F}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{F}^T \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad (1)$$

where \mathbf{P} represents the permutation $[1, 2, \dots, 2L] \mapsto [1, L, 2, L + 1, \dots, 2L]$, and \mathbf{V} is of the form

$$\mathbf{W} = \begin{pmatrix} \mathbf{V}_0^{(2 \times 2)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_1^{(4 \times 4)} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{V}_L^{(2 \times 2)} \end{pmatrix} \quad (2)$$

Note also that the Fourier matrix can also be written as the product of $O(\log L)$ sparse matrices (hence the FFT algorithm). Indeed, it is the proliferation of sparse matrices that makes the computation rather tractable even for large frames.

Finally, the $2L$ components of \mathbf{s} are partitioned into two groups (one for each source) containing representatives from each of the $L/2 + 1$ ICA sub-problems. The partition defines two orthogonal subspaces; we consider the problem solved if each subspace represents activity from only one source, in which case either source can be reconstructed in signal domain by setting the components in the other subspace to zero and inverting the system.

Our aim in this paper is to investigate what happens if the three matrices in (1) are replaced by a single unconstrained ICA weight matrix, and how an analysis of residual dependency can be used to partition the resulting components into two subspaces. Although this clearly requires more computation and much more training data to fit the larger number of parameters, the system is

conceptually simpler, and shows how sensible processing strategies for dealing with stereo signals can emerge in an unsupervised way.

3 Unconstrained ICA of buffered stereo data

The data for the unconstrained ICA system consists of the same packed stereo frames $\mathbf{x} \equiv (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^N$ (with $N = 2L$) as described in the previous section. For these experiments, we used recordings from two microphones placed in a normally reverberant room along with two loudspeakers playing (38 s) short extracts from two different radio programmes. A female presenter was speaking in one programme, a male in the other. The signals were sampled at 16 kHz.

A natural gradient maximum-likelihood ICA algorithm [3] was used to estimate an $N \times N$ weight matrix \mathbf{W} , yielding the estimated independent components $\mathbf{s} = \mathbf{W}\mathbf{x}$ for each frame. The weight updates were of the form

$$\mathbf{W} \mapsto \mathbf{W} + \eta \langle \mathbf{I} - \boldsymbol{\varphi}(\mathbf{s})\mathbf{s}^T \rangle \mathbf{W}, \quad (3)$$

where $\langle \cdot \rangle$ denotes an average taken over the training data (or a smaller batch randomly sampled from the whole), η is the learning rate, and the function $\boldsymbol{\varphi} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is the gradient of the negative log-prior on the components: $\boldsymbol{\varphi}(\mathbf{s}) = -\nabla_{\mathbf{s}} \log p(\mathbf{s})$. A generalised exponential factorial prior was used:

$$p(\mathbf{s}) = \prod_{i=1}^N p(s_i), \quad p(s_i) = \frac{\exp -|s_i|^{\alpha_i}}{2\Gamma(1 + \alpha_i^{-1})}. \quad (4)$$

During training, the exponents α_i were periodically re-estimated from the data using a maximum-likelihood gradient method [4]. Some of the resulting stereo basis vectors (the columns of $\mathbf{A} = \mathbf{W}^{-1}$) are illustrated in fig. 1(a). If we take each component as an ‘atom’ of sound (in the case of perfect separation, it will come from just one of the sources), its stereo basis vector encodes how that atomic sound is received at the two microphones. Thus, the basis matrix contains information about the relative path delay and frequency-dependent response between each sources and microphone. Comparing the unconstrained ICA basis with the equivalent basis vectors for the frequency domain ICA system—some of which are illustrated in fig. 1(b)—it is clear that both systems exploit phase and amplitude differences to distinguish the sources, but only the unconstrained ICA system is able to exploit *time delays*, since, unlike the frequency domain system, it includes temporally localised basis vectors.

The goodness of fit of both ICA models was assessed in terms of the average log-probability of a frame:

$$L = \langle \log p(\mathbf{x}) \rangle = \langle \log \det \mathbf{W} + \log p_{\mathbf{s}}(\mathbf{W}\mathbf{x}) \rangle. \quad (5)$$

This is related to the average coding cost-per-vector: the higher the score the lower the cost. The unconstrained ICA system achieved a score 1516.2, whereas the frequency domain ICA achieved 1347.9. In terms of coding cost, this is a difference of 168.3 nats (242.9 bits) per frame.

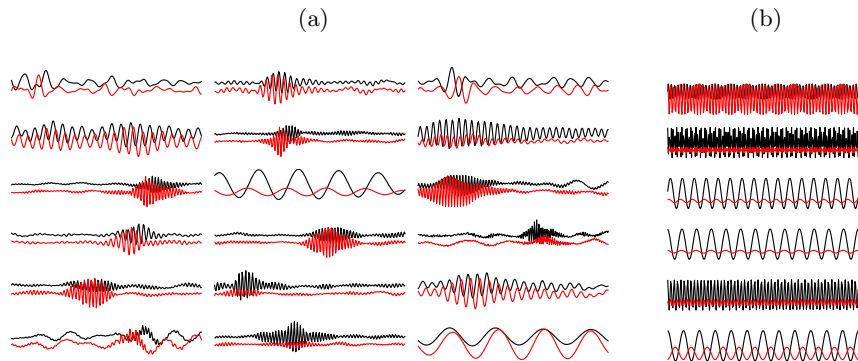


Fig. 1. (a) Some of the stereo basis vectors obtained by ICA of data recorded from two microphones with two sources present in the room. (The right-channel signal is offset below the left-channel.) The frame size for each channel was 256, so the ICA system is 512-dimensional. (b) Equivalent basis vectors implicit in frequency domain ICA of the same system, obtained by inverting the three matrices in (1).

4 Geometric dependency analysis

The aim of GDA is to represent each element of a distributed representation as a point in a metric space such that the distance between each pair is inversely related to the mutual information between the corresponding components. Truly independent components are pushed infinitely far apart, while dependent components form clusters or manifolds. As argued in [5], such residual dependencies can be useful in interpreting the representation and organising the next stage of processing. The method described in [1, ch. 8] involves estimating the mutual information between each pair of components s_i, s_j in terms of a nonlinear correlation coefficient in the range $[-1, 1]$,

$$\rho_f(S_i, S_j) = \text{corr}[f_1(S_i), f_2(S_j)] = \frac{\text{cov}[f_1(S_i), f_2(S_j)]}{\sqrt{\text{var } f_1(S_i) \text{var } f_2(S_j)}}, \quad (6)$$

where S_i and S_j denote the random variables whose realisations are s_i and s_j , f_1 and f_2 are rectifying (that is, even) nonlinear functions, and $f(S)$ is shorthand for the random variable obtained by applying the function f to realisations of S . From this, a matrix of pair-wise distances is defined:

$$D_{ij} = \sqrt{-2 \log |\rho_f(S_i, S_j)|}. \quad (7)$$

Finally, multidimensional scaling [6] is used to generate a spatial configuration of N points r_i in an E -dimensional metric space \mathcal{M} such that their pair-wise distances according to a predetermined metric $d: \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}^+$ approximate the correlation distances, that is $d(r_i, r_j) \approx D_{ij}$ for all pairs i, j .

Before presenting the results of GDA on the ICA systems described in previous sections, we describe a refinement to the nonlinear correlation that should in principle give a more accurate measure of dependence.

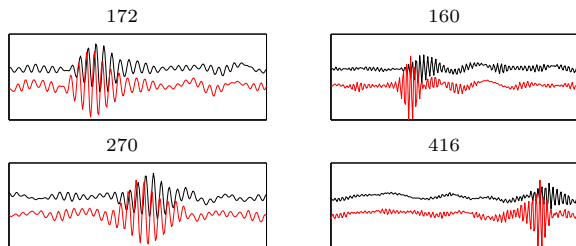


Fig. 2. Two pairs of basis vectors whose \mathcal{F} -correlation is increased by inclusion of lags. The original stereo data was buffered using a hop size of 32, and the function space \mathcal{F} included lags up to 4, i.e. only the coefficients a_0 to a_4 in (10) were allowed to vary from zero. The \mathcal{F} -correlation of the pair on the left (172 and 270) increased from 0.42 to 0.83, while the correlation of the pair (160,416) increased from 0.51 to 0.87.

4.1 Estimating residual dependency using the \mathcal{F} -correlation

The \mathcal{F} -correlation [7] can be thought of as a generalisation nonlinear correlation: instead of two fixed functions f_1 and f_2 , we allow f_1 and f_2 to range freely over a function space \mathcal{F} and define $\rho_{\mathcal{F}}$ as the maximal correlation so obtained:

$$\rho_{\mathcal{F}} = \sup_{f_1, f_2 \in \mathcal{F}} \text{corr}[f_1(S_1), f_2(S_2)]. \quad (8)$$

If \mathcal{F} is a linear space, then the computation of the \mathcal{F} -correlation is equivalent to canonical correlation analysis (CCA) and can be solved as a generalised eigenvalue problem. The spectrum of canonical correlations can then be used to compute the \mathcal{F} -correlation and the so-called *generalised variance*, both of which can be used as measures of statistical dependence.

In this application, we used a function space spanned by a basis of *lagging* functions l_{τ} : if S represents the sequence of values of one component as successive frames are processed, then the lagged component $l_{\tau}(S)$ is the same sequence delayed by τ frames. In addition, we used as our rectification nonlinearity a form of *generalised energy* derived from the generalised exponential prior (4):

$$\mathcal{E}(s_i) = |s_i|^{\alpha_i}. \quad (9)$$

A typical element of this function space is

$$f : S_i \mapsto f(S_i), \quad f(S_i) = a_0 l_0(\mathcal{E}(S_i)) + a_1 l_1(\mathcal{E}(S_i)) + a_2 l_2(\mathcal{E}(S_i)) \dots \quad (10)$$

where the a_{τ} are weighting coefficients. Note that, strictly speaking, because of the dependence of the generalised energy on the index of the component i , the functions f_1 and f_2 lie in two of N *separate* spaces \mathcal{F}_i depending on the indices of the components S_i and S_j .

The motivation behind using a space of lagging functions is to capture temporal dependencies where activity in one component implies activity in another a certain time later. Fig. 2 shows two pairs of basis vectors whose \mathcal{F} -correlation is significantly increased by the inclusion of lags, which is unsurprising since they appear to be shifted versions of the same stereo waveform.

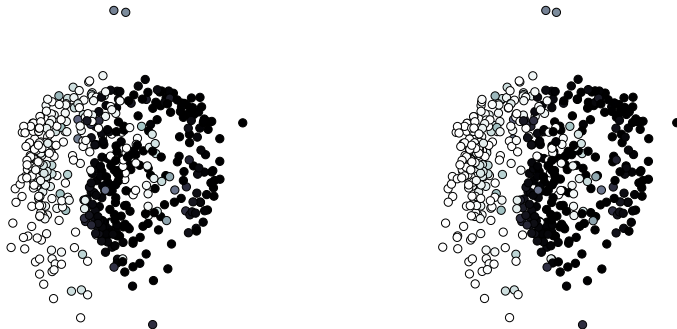


Fig. 3. MDS results obtained with a hop size of 32, lags 0 to 6, $q = 2$, using Kruskal’s stress function (see [6]) in a 3-D Euclidean space. The shading indicates which side of a manually chosen separating plane each point lies. The points form two crescent-shaped clusters lying side-by-side with a gap in between them. (The two images are a stereo pair with the left-eye image on the left.)

4.2 Multidimensional scaling results

Given a particular weight matrix \mathbf{W} , there are many variables involved in determining the final spatial configuration discovered by MDS: the hop size between frames and the lags used to compute the \mathcal{F} -correlation, the dimension E of the MDS embedding space, and stress function used in the MDS algorithm. Due to lack of space we can only illustrate one example here. The points in fig. 3 form two crescent-shaped clusters lying side by side, with, roughly, high frequency basis vectors are at one end and low frequencies are at the other. An inspection of the basis vectors in each cluster shows that those basis vectors which are readily interpretable as belonging to one or other source (such as those in fig. 2) are consistently segregated. In the next section we assess how well the original sources can be reconstructed on the basis of this partition.

5 Source reconstruction and evaluation

The reconstruction process involves setting some of the components s_i to zero and inverting the ICA system. Letting $\mathbf{y} \equiv (\mathbf{y}_1, \mathbf{y}_2)$ denote a reconstructed stereo frame, this can be expressed as $\mathbf{y} = \mathbf{W}^{-1}\mathbf{H}\mathbf{s} = \mathbf{W}^{-1}\mathbf{H}\mathbf{W}\mathbf{x}$, where \mathbf{H} is a diagonal matrix. The process can be thought of as ‘ICA domain filtering’ in direct analogy with frequency domain filtering, the difference being that we use ICA instead of a Fourier transform before applying a diagonal operator and transforming back. In this application, we aim to place ones and zeros on the diagonal of \mathbf{H} in order to select only those components which belong to a particular source. Given a partitioning of the components into two groups, we can therefore define two complementary ICA domain filters $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ to reconstruct the sources.

The partition was determined by manually positioning a separating plane between the two clusters of points found by MDS. Each source was reconstructed as a stereo signal (using a Hanning window to recombine overlapping frames) and compared with those obtained from a frequency domain algorithm [2] using the same frame size of 256 samples from each channel. At a sampling rate of 16 kHz, these frames are only 16 ms long—much shorter than the room’s impulse response—which places limits on the potential performance of both systems. Hence, neither was able to separate the sources perfectly, but the reconstructions were of similar quality.

To make a quantitative evaluation, a different data set was used: the two sources were recorded in situ as before, but separately rather than simultaneously. The two (fully reverberant) stereo recordings were mixed artificially and analysed using previously fitted ICA model. The match between the separately recorded sources and their reconstructions was then measured by a matrix of correlation coefficients:

$$J_{kl} = \frac{\left[\sum_{ij} X_{ij}^k Y_{ij}^l \right]^2}{\sum_{ij} [X_{ij}^k]^2 \sum_{i',j'} [Y_{i'j'}^l]^2}, \quad (11)$$

where X_{ij}^k is the j th sample from the i th channel (left or right) of the k th source, and Y_{ij}^l is similarly defined for the l th reconstruction. Perfect separation and reconstruction is achieved when $J_{kl} = \delta_{kl}$. The correlation matrices obtained for the two systems are tabulated below:

	recon. 1	recon. 2		recon. 1	recon. 2
source 1	0.357	0.218	source 1	0.400	0.189
source 2	0.089	0.421	source 2	0.230	0.295
(a) Frequency domain ICA			(b) Unconstrained ICA		

The unconstrained system does not quite achieve the contrast between sources in each reconstruction that the frequency domain method does; this may be because of sub-optimal convergence of the ICA model, either because of local minima, or because of over-fitting due to insufficient training data, though this is yet to be confirmed experimentally. Note that the frequency domain system was trained using a fixed-point algorithm, which generally gives better convergence than the natural gradient optimiser used in the unconstrained system.

6 Discussion and conclusions

In this paper, we have shown that GDA is capable of generating meaningful, readily interpretable structures when applied stereo audio data, in this case, providing, on the basis of very few assumptions, a solution to the speaker separation problem comparable to that produced by a more specialised algorithm. The approach essentially boils down to finding independent subspaces in high

dimensional data, which in this case, happens to be segments of a stereo signal. Hence, it can be compared with other systems that analyse the clustering of dependent components [5, 8].

One benefit of GDA is that it may reveal, as complex geometric forms, dependency structures that do not emerge clearly in other subspace analysis methods. A disadvantage of our approach in this particular application is that the unconstrained ICA phase of the process does not scale to long frames, because of the amount of training data required to adequately adapt the N^2 elements of \mathbf{W} . The weight matrix contains specific information about the physical configuration of the sources and microphones, as well general information about the statistical structure of speech. Thus, one approach might be to factorise the system along the lines of (1), but using a general speech-trained ICA transform [1, ch. 5] on each channel instead of a Fourier transform, followed by sparse matrix ICA in adaptive local ‘modules’ defined by residual dependency, as described in [9].

Acknowledgements The authors would like to thank Nikolaos Mitianoudis for helpful discussions, for the test data used in these experiments, and for the frequency domain ICA results quoted in section 5. This work was supported by EPSRC grant GR/R54620.

References

1. Abdallah, S.A.: Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models. PhD thesis, Department of Electronic Engineering, King’s College London (2002)
2. Mitianoudis, N., Davies, M.E.: New fixed-point solutions for convolved mixtures. In: 3rd Intl. Conf. on Independent Component Analysis and Source Separation (ICA2001), San Diego, California (2001)
3. Cardoso, J.F., Laheld, B.: Equivariant adaptive source separation. *IEEE Trans. on Signal Processing* **44** (1996) 3017–30
4. Everson, R., Roberts, S.: Independent component analysis: A flexible nonlinearity and decorrelating manifold approach. *Neural Computation* **11** (1999) 1957–1983
5. Hyvärinen, A., Hoyer, P., Inki, M.: Topographic independent component analysis. *Neural Computation* **13** (2001) 1527–1558
6. Cox, T., Cox, M.A.A.: *Multidimensional Scaling*. Chapman Hall/CRC, London (2001)
7. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. Technical Report UCB/CSD-01-1166, Division of Computer Science, University of California, Berkeley (2001)
8. Bach, F.R., Jordan, M.I.: Finding clusters in independent component analysis. In: 4th Intl. Symp. on Independent Component Analysis and Signal Separation (ICA2003), Nara, Japan (2003)
9. Matsuda, Y., Yamaguchi, K.: Linear multilayer ICA integrating small local modules. In: 4th Intl. Symp. on Independent Component Analysis and Signal Separation (ICA2003), Nara, Japan (2003)