MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Spectrogram Dimensionality Reduction with Independence Constraints

Kevin Wilson, Bhiksha Raj

TR2010-023 June 2010

Abstract

We propose an algorithm to find a low-dimensional decomposition of a spectrogram by formulating this as a regularized non-negative matrix factorization (NMF) problem with a regularization term chosen to encourage independence. This algorithm provides a better decomposition than standard NMF when the underlying sources are independent. It makes better use of additional observation streams than previous nonnegative ICA algorithms.

ICASSP 2010

SPECTROGRAM DIMENSIONALITY REDUCTION WITH INDEPENDENCE CONSTRAINTS

Kevin W. Wilson

Mitsubishi Electric Research Lab
Cambridge, MA, USA
wilson@merl.com

Bhiksha Raj

Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA, USA
bhiksha@cs.cmu.edu

ABSTRACT

We present an algorithm to find a low-dimensional decomposition of a spectrogram by formulating this as a regularized non-negative matrix factorization (NMF) problem with a regularization term chosen to encourage independence. This algorithm provides a better decomposition than standard NMF when the underlying sources are independent. It makes better use of additional observation streams than previous nonnegative ICA algorithms.

Index Terms— matrix decomposition

1. INTRODUCTION

This paper presents a new algorithm for finding a low-dimensional decomposition of a spectrogram in which each of the low-dimensional components evolves (nearly) independently. Such a representation can be useful when the spectrogram of interest was generated as a sum of independent sources. Because power- and magnitude-spectrograms are strictly nonnegative, this can be formulated as a problem of independent component analysis (ICA) of non-negative mixtures or alternatively as a problem of non-negative matrix factorization (NMF) with an appropriate independence constraint. This paper elucidates these relationships and presents a new method based on the NMF formulation of the problem.

Independent Component Analysis (ICA) aims to extract statistically independent components from observations through linear transformations. Given a collection of (column) vectors $\{\mathbf{v}\}$, which we can represent jointly as a matrix V, ICA attempts to estimate an "unmixing" matrix M such that the the rows of $H = MV$, *i.e.* the components of the vectors $\mathbf{h}$ that form the columns of H are statistically independent. If the vectors $\mathbf{v}$ were themselves obtained through a linear operation on independent sources, *i.e.* if $V = WX$, where X is a matrix composed of vectors $\mathbf{x}$ with statistically independent components, then M will be a scaling and permutation of the inverse of the original matrix W, *i.e.* $M \approx RW^{-1}$, where R is a scaling and permutation matrix, and $H \approx RX$. Alternately stated, the components of the observed vectors $\mathbf{v}$ are said to be *mixtures* of the original independent random variables represented in $\mathbf{x}$, where W is the *mixing matrix* that mixes the components of $\mathbf{x}$. ICA as it is normally performed [1, 2] aims to estimate an *unmixing* matrix M that can recover the original independent components (to within a permutation and scaling) from the mixed data in the observations.

The usual algorithms for ICA are agnostic to the polarity of the data. In other words, if M is a valid unmixing matrix, then ZM is also a solution, where $Z = diag(1, 1, \ldots, -1, \cdots,)$ is a diagonal matrix where some diagonal terms are 1 and the rest are -1. When both the original data X and their mixed observations V are known *a priori* to be strictly non-negative then the solutions obtained may not be satisfactory, since H is not guaranteed to be non-negative.

In [3] Plumbley presents a "non-negative ICA" algorithm that recovers non-negative independent components from mixtures of non-negative sources. The algorithm is based on a theorem that states that for any non-negative vector $\mathbf{x}$ with independent components drawn from well-grounded distributions (*i.e.* PDFs that extend to 0), multiplication by an orthogonal matrix Q results in a non-negative result $\mathbf{v}$ iff Q is a permutation matrix. This means that to derive the unmixing matrix M it is sufficient to derive an orthogonal matrix that decorrelates the rows of V while subject to the constraint that all components of MV are nonnegative. Plumbley's algorithm therefore proceeds accordingly: the observation matrix V is pre-whitened, followed by estimation of $\bar{Q}$ that results in non-negative H, the components of which are provably now independent.

In this paper we utilize a similar observation that actually leads to a reverse approach. We note that if a mixed non-negative matrix can be expressed as the product of two non-negative matrices such that the rows of the one of them are decorrelated, then the rows of that matrix are also independent. In other words, if we were to simultaneously estimate a non-negative *mixing* matrix W and a matrix of non-negative uncorrelated vectors H such that $V = WH$ the rows of H will also be independent. Additionally, there is no requirement for distributions to be grounded.

Note that $V = WH$ is identical to the decomposition used in non-negative matrix factorization (NMF). NMF [4] decomposes a non-negative matrix into the product of two non-negative matrices through a set of simple multiplicative rules that optimize one of several objective functions. NMF by itself does not guarantee any statistical relationships between the terms it computes; however such relationships can be imposed through regularization terms [5].

We recast the problem of deriving independent non-negative components from their non-negative mixtures as a soft-constrained NMF problem, where the constraint is the requirement that the rows of H be maximally uncorrelated. This constraint, applied using a decorrelating update mechanism proposed by Parra et. al. [6] as a regularization term within NMF yields a set of simple multiplicative update rules. We refer to our algorithm as *NMFICA*.

Simulations show that NMFICA is able to estimate mixing matrices accurately, and results in estimates of unmixed independent components that are comparable (in terms of SNR) to or better than those obtained by other ICA algorithms, particularly in the presence of noise. In particular, when the mixing matrix W is not square (and has more rows than the number of independent sources) we achieve superior results to other ICA techniques.

2. THE NMFICA ALGORITHM

We wish to solve the following problem. Given a nonnegative matrix V, we wish to factorize it as WH, where both W and H are nonnegative matrices. Further, we seek H such that its rows are independent. More formally, we seek W and H such that

$$\begin{aligned} V &= WH \\ W_{ab} &\geq 0 \ \forall a, b \\ H_{bc} &\geq 0 \ \forall b, c \\ V_{ac} &\geq 0 \ \forall a, c \\ P(H_{ic}H_{jc}) &= P(H_{ic})P(H_{jc}) \ \forall i, j, c \end{aligned} \tag{1}$$

where W_{ab} and H_{bc} are components of W and H respectively. The fifth condition in Equation 1 expresses independence of the rows of H. (We will show below that it is sufficient to decorrelate the columns of H to achieve this independence.) Since all terms here are non-negative, we recognize the above as a problem of regularized non-negative matrix factorization.

Theorem 1 from Oja and Plumbley [7] states that independent, nonnegative basis functions can be recovered by decorrelating observations and then finding an orthogonal transformation that leaves them nonnegative. [7] does not constrain the mixing matrix to be nonnegative, so they must require that probability distributions of the sources be "well-grounded," i.e. have support down to zero, in order to recover independent components. In our formulation, we require that the mixing matrix is nonnegative, and as a result we do not require any "well-groundedness" of the source distributions. We outline our argument below.

Lemma 1: Let U be a non-negative random variable whose components are independent. Let Y be a non-negative matrix such that $H = YU$ is also non-negative. The rows of H will be uncorrelated if and only if they are also independent, in which case Y will a permutation-and-scaling matrix. The intuition behind this lemma is that the mixing matrices that maintain uncorrelatedness can be expressed as compositions of isometric transformations and single-coordinate scalings; however, the only isometries that can be expressed with a nonnegative mixing matrix Y are permutations. (For example, rotations that do not result in complete permutations require negative values in Y.)

Lemma 2: If a non-negative matrix V has been obtained from a non-negative matrix U with independent rows as $V = ZU$, it can also be expressed as $V = WH$, where Z and W are also non-negative and the columns of H have the same dimensionality as those of U, then the rows of H can be expressed as the permutation of the rows of U and are also independent. This follows as a consequence of non-negativity and Lemma 1.

It follows from the above that a non-negative random variable V that has been obtained through a linear combination of independent variables can be expressed as the product of a non-negative matrix W and a nonnegative vector H such that the components of H are uncorrelated, H will be a permutation (and scaling) of the original independent variables that compose V. Any linear transformation that is uncorrelated but not independent will necessarily take on negative values in W and/or H.

The consequence of the above fact is that in order to derive the independent components from a non-negative matrix V, it is sufficient to decompose it as $V = WH$, where both W and H are non-negative, and the components of H are uncorrelated.

Following the approach suggested in [5] we seek a solution to Equation 1 using the following regularized NMF objective function that must be minimized with respect to W and H:

$$D(W, H) = \frac{1}{2}||V - WH||_F^2 + \alpha J(H) \tag{2}$$

Here $||V - WH||_F^2$ measures the reconstruction error, the Frobenius norm of the difference between the mixed matrix V and its decomposition WH. For perfect decomposition this term would be 0. $J(H)$ represents a regularization term that expresses some property that we require from H. α is a scalar weight given to this regularization term in the optimization process.

$J(H)$ must be chosen to express some function of H that, when minimized, will also minimize the correlation between the rows of H. We choose the Frobenius norm of the empirical correlation matrix of H.

$$J(H) = ||C(H)||_F^2 \tag{3}$$

$$C(H) = P_H^{-1/2} H H^\mathsf{T} P_H^{-1/2} \tag{4}$$

where $C(H)$ is the energy-normalized correlation matrix of H [6]. P_H is a diagonal matrix of the energies (sums of squares) of the rows of H. The diagonal elements of the $C(H)$ are always equal to 1; as a result minimizing its Frobenius norm will force the off-diagonal elements toward zero.

We use the general form of the NMF update with regularization on H from [5]:

$$\begin{aligned} W_{ab} &\leftarrow W_{ab} \frac{[VH^\mathsf{T}]_{ab}}{[WHH^\mathsf{T}]_{ab}} \\ H_{bc} &\leftarrow H_{bc} \frac{\left[[W^\mathsf{T}V]_{bc} - \alpha\varphi(H_{bc})\right]_\varepsilon}{[W^\mathsf{T}WH]_{bc} + \varepsilon} \end{aligned} \tag{5}$$

where ε is a small positive constant and $[\,]_\varepsilon$ indicates that any values within the brackets less than ε should be replaced with ε to prevent violations of the nonnegativity constraint. $\varphi(H)$ is the gradient of $J(H)$ with respect to H.

$$\varphi(H_{bc}) = \frac{\partial J(H)}{\partial H_{bc}} \tag{6}$$

$$= \sum_i \sum_j C_{ij} \frac{\partial C_{ij}}{\partial H_{bc}} \tag{7}$$

It is the straightforward to show that $\partial C_{ij}/\partial H_{bc}$ has the form:

$$\frac{\partial C_{ij}}{\partial H_{bc}} = \frac{B_{ij}(\partial A_{ij}/\partial H_{bc}) - A_{ij}(\partial B_{ij}/\partial H_{bc})}{B_{ij}^2} \tag{8}$$

where we define intermediate variables A and B as follows for notational convenience:

$$A = HH^\mathsf{T} \tag{9}$$

$$B = NN^\mathsf{T} \tag{10}$$

$$N_b = ||H_b|| \tag{11}$$

$$\partial A_{ij}/\partial H_{bc} = 1_b H_c^\mathsf{T} + H_c 1_b^\mathsf{T} \tag{12}$$

$$\partial B_{ij}/\partial H_{bc} = H_{bc}(U 1_b 1_b^\mathsf{T} + 1_b 1_b^\mathsf{T} U^\mathsf{T}) \tag{13}$$

$$U = N(N^{-1})^\mathsf{T} \tag{14}$$

where 1_b is an indicator vector that is zero everywhere except for having the b^{th} element equal to one. N is a vector whose elements

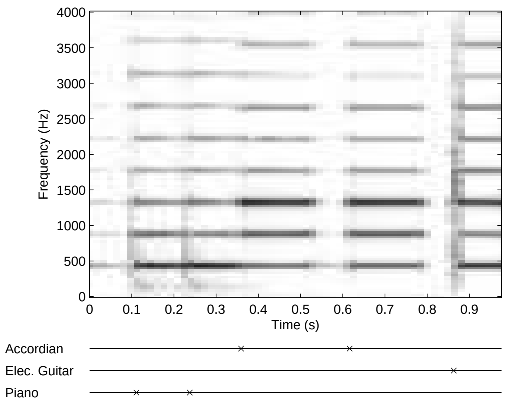

Fig. 1. Spectrogram of a single-channel signal in which three different instruments (piano, electric guitar, and accordion) play the same note (440 Hz) at random times.

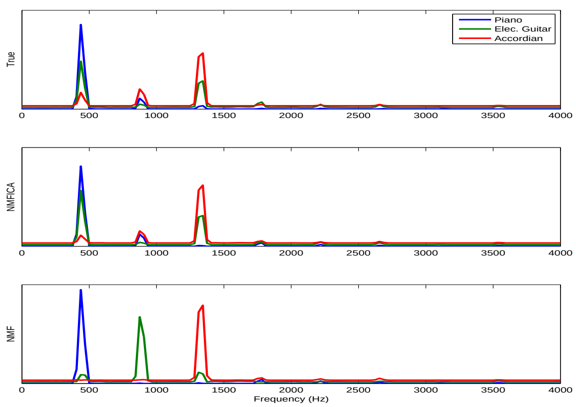

Fig. 2. Top: true average spectra of the three instruments from Figure 1. Middle: spectral basis functions found by NMFICA. Bottom: spectral basis functions found by standard NMF.

are the norms of the rows of H, and U is an outer product of N with its element-wise inverse.

Equation 5, with $\varphi(H)$ as defined by Equations 6-14, forms the substance of the update rule for NMFICA.

Figures 1 and 2 show the unmixing achieved by NMFICA for a simple audio example. The spectrogram in Figure 1 shows a segment of the input signal, in which a synthesized piano, electric guitar, and accordion each play the same note (440 Hz) repeatedly and at random times. (The actual input to NMF and NMFICA in this example is ten seconds long; only one second is shown.) As shown in Figure 2, NMFICA finds three spectral basis functions that each almost perfectly match the average spectra of an individual instrument. Standard NMF finds three basis functions that roughly represent the fundamental and first two harmonics. These basis functions can be combined to represent each instrument, but individual basis functions do not capture individual instruments.

3. RELATION TO OTHER METHODS

In contrast to NMFICA, NMF by itself is not guaranteed to recover independent signals; any such recovery is purely incidental. The decorrelating regularization term is critical for independence.

Decorrelation in itself does not work as an independence criterion for ICA, and most ICA algorithms actually attempt to manipulate higher-order moments of the data either directly or indirectly to achieve independence in the unmixed outcome. Under some conditions, however, decorrelation can directly result in independence. Fancourt and Parra [6] have previously employed decorrelation as an independence criterion for nonstationary (although not non-negative) signals where they seek a solution that decorrelates the reconstructed sources at multiple points in time. In fact, we have drawn the principle of minimization of the Frobenius norm of the correlation matrix as an objective for decorrelation from their work. Oja and Plumbley [7] also use decorrelation (without enforcing higher order independence) as part of their nonnegative ICA algorithm. They prove that this decorrelation criterion is sufficient for use as an independence criterion for nonnegative ICA as long as the source PDFs are "well-grounded". The key contrast between our work and these prior approaches is that we aim to estimate the mixing matrix, whereas prior methods have invariably attempted to estimate the unmixing matrix. We will show in our results that our approach yields better results in the noisy, overdetermined case.

Additional distinctions exist with respect to prior algorithms for ICA of non-negative data. In contrast to Plumbley's approach which only ensures that H is non-negative, and requires it to be "well-grounded", our approach ensures that both W and H are nonnegative and does not require H to be grounded. For some applications, this may be an important distinction. On the other hand, Plumbley's approach leads to a convex problem, whereas our algorithm, like all NMF formulations, is guaranteed only to find a local minimum. The two constraints of decorrelation and nonnegativity will only be achieved when a perfect decomposition $V = WH$ is found. For locally optimal solutions the decorrelation may not be complete, from which it follows that H will not be truly independent.

Nevertheless, we believe that our algorithm is useful because it can deal with non-square mixing matrices. We are not aware of an extension to Plumbley's approach to non-square matrices. We will show in the following section that in noisy conditions with non-square mixing matrices, our algorithm can outperform other approaches.

4. RESULTS

We test our algorithm on a simple synthetic problem against three related algorithms.

We generate synthetic observations Y by $Y = [MX + Z]_\varepsilon$, where the elements of M and X are independently chosen from a uniform distribution on $[0, 1]$ and where Z is IID Gaussian noise. We take as a baseline a problem with 500 samples of a 3-dimensional source, i.e. X is 3×500, and we vary the observation dimensionality and noise level. For NMFICA and for standard NMF, we initialize the entries of W and H with uniform random values from [0, 1].

We compare against three other methods: unregularized NMF, i.e. Equation 5 with $\alpha = 0$, FastICA [1], a popular ICA implementation (that does not include a nonnegativity constraint), and Oja and Plumbley's nonnegative ICA algorithm from [7].

Figure 3 shows results for a square (3×3) mixing matrix, and Figure 4 shows results for a 6×3 mixing matrix, resulting in twice as many observations as sources. Both figures show output SIR as input SNR is varied. We define "input SNR" to be the power ratio between MX and Z, the ratio of the mixed source power to the additive noise power. "Output SIR" refers to the signal-to-interferer ratio (SIR), the ratio between the recovered source power and the residual power

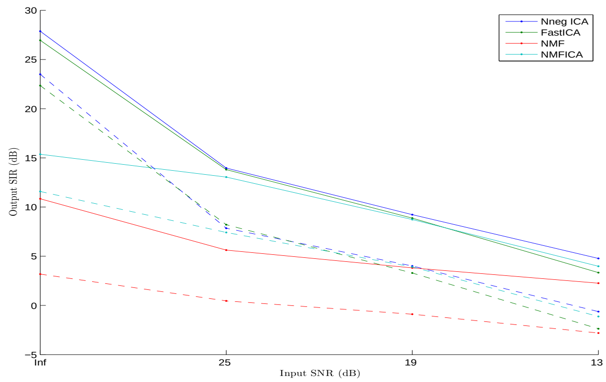

Fig. 3. Output SIR vs. input SNR for a square mixing matrix. ($n_o = n_s = 3$). Solid lines show average performance. Dashed lines show the worst SNR out of the three recovered sources.

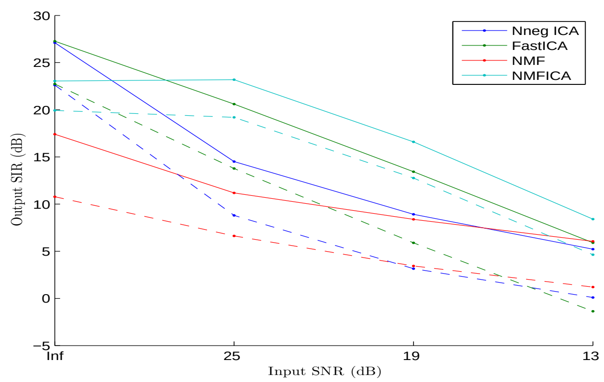

Fig. 4. Output SIR vs. input SNR for a mixing matrix with twice as many observations as sources. ($n_o = 6, n_s = 3$). Solid lines show average performance. Dashed lines show the worst SNR out of the three recovered sources.

from other source channels remaining in the reconstruction. In each figure, solid lines represent mean SIR and dashed lines represent the minimum (worst) SIR of the three recovered sources. This minimum SIR is important because we often want reasonable reconstructions of all sources rather than very good reconstructions of some sources and very poor reconstructions of others. Each point in the figures is an average value over 100 realizations of the problem.

In general, we do not expect unregularized NMF ("NMF" in the results tables) to perform particularly well because it incorporates no independence constraint. It can be expected to achieve low reconstruction error, i.e. $WH \approx Y$, but the rows of H will not necessarily be a permutation of the rows of Y. Our results in general show that source reconstruction by NMF is poor.

Figure 3 shows that for the square mixing matrix and infinite SNR (no noise), FastICA and nonnegative ICA do much better than NMFICA. However, when even a small amount of noise is added, the performance of FastICA and nonnegative ICA become comparable to that of NMFICA.

For the non-square (extra observations) case in Figure 4, NMFICA performs best in all but the no-noise case. We are unaware of an extension of Oja and Plumbley's nonnegative ICA algorithm to handle non-square matrices, so instead we use only the first 3 observations as a square mixing problem. For this reason, nonnegative ICA performance is nearly the same in the two figures. Other than the fact that the number of sources was specified, FastICA was used with its default parameters. In this non-square case, the finding of the independent components is a form of dimensionality reduction.

We believe this scenario in which there are many noisy measurements of a relatively small number of independent sources is an important one, for example when a spectrogram with hundreds of frequency bins can be described using a relatively small number of basis functions. In our previous work, we have encountered this scenario while applying NMF-based techniques to speech denoising [8, 9]. In future work, we hope to combine NMFICA with the regularization techniques in [8, 9] and apply it to denoising and source separation of speech and other nonstationary signals.

5. CONCLUSION

We have presented an NMF-based algorithm for independent component analysis of non-negative data such as power- or magnitude-spectrograms. In contrast to previous methods, we estimate a mixing matrix, rather than an unmixing matrix. Experiments show that we are able to achieve unmixing comparable to other methods of ICA for square mixing matrices, and significantly better when the mixing matrix is not square, particularly in noise.

6. REFERENCES

[1] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999. [Online]. Available: citeseer.ist.psu.edu/hyv99fast.html

[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995. [Online]. Available: citeseer.ist.psu.edu/bell95informationmaximization.html

[3] M. Plumbley, "Conditions for non-negative independent component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 177–180, 2002.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems*, 2000, pp. 556–562. [Online]. Available: citeseer.ist.psu.edu/lee01algorithms.html

[5] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2006, pp. 621–625.

[6] C. Fancourt and L. Parra, "A comparison of decorrelation criteria for the blind source separation of non-stationary signals," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002, pp. 165–168.

[7] E. Oja and M. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Computation*, vol. 16, no. 9, pp. 1811–1825, 2004.

[8] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[9] K. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Interspeech*, 2008.