

Conditions for Non-Negative Independent Component Analysis

Mark Plumbley

Abstract

We consider the noiseless linear independent component analysis problem, in the case where the hidden sources \mathbf{s} are non-negative. We assume that the random variables s_i s are *well-grounded* in that they have a non-vanishing pdf in the (positive) neighbourhood of zero. For an orthonormal rotation $\mathbf{y} = \mathbf{W}\mathbf{x}$ of pre-whitened observations $\mathbf{x} = \mathbf{Q}\mathbf{A}\mathbf{s}$, under certain reasonable conditions we show that \mathbf{y} is a permutation of the \mathbf{s} (apart from a scaling factor) if and only if \mathbf{y} is non-negative with probability 1. We suggest that this may enable the construction of practical learning algorithms, particularly for sparse non-negative sources.

Index Terms

Independent component analysis, Non-negative matrix factorization, Sparse coding, ICA.

I. INTRODUCTION

SEVERAL authors have suggested that decomposition of a observation into non-negative factors, or *Non-Negative Matrix Factorization*, is able to produce useful and meaningful representations in real-world problems [1]. In this letter we show that, for m observations generated by a nonsingular linear combination of $n \leq m$ non-negative independent sources, the combination of pre-whitening (removal of covariance) with finding a rotation to give non-negative outputs is both necessary and sufficient to finding the set of underlying independent components, provided that the sources have non-vanishing pdf down to zero.

II. PROBLEM STATEMENT

Suppose we have an n -dimensional random vector $\mathbf{s} = [s_1, \dots, s_n]^T$ where the real-valued component random variables (*sources*) s_i are independent, and have bounded non-zero variance.

We call a source *non-negative* if $\Pr(s < 0) = 0$, and *well-grounded* if it has a non-vanishing pdf in any (positive) neighbourhood of zero, such that for any $\delta > 0$ we have $\Pr(s < \delta) > 0$. We also suppose are given some (noiseless) observations $\mathbf{z} = \mathbf{A}\mathbf{s}$ where \mathbf{A} is a bounded $m \times n$ matrix with $m \geq n$ and full rank n .

We shall assume, without loss of generality, that the sources have unit variance $\sigma_{s_i}^2 = E((s_i - \bar{s}_i)^2) = 1$. If this were not the case, we could simply scale the sources to $\mathbf{s}' = \mathbf{G}\mathbf{s}$ giving $\mathbf{z} = \mathbf{A}'\mathbf{s}'$ with $\mathbf{A}' = \mathbf{A}\mathbf{G}$, where $\mathbf{G} = \text{diag}(g_1, \dots, g_n)$ and $g_i = +\sigma_{s_i}^{-1}$ with $0 < g_i < \infty$ since the sources have bounded non-zero variance. This means that any set of sources can only be identified subject to a (strictly positive) scaling factor.

We would like to find a linear transformation $\mathbf{y} = \mathbf{V}\mathbf{z}$ that unmixes the sources such that $\mathbf{y} = \mathbf{s}$. In fact, it will not be possible to identify the original scaling of the sources, or the specific order of the source components, so it is sufficient for us to find a $\mathbf{y} = \mathbf{V}\mathbf{z}$ such that $\mathbf{y} = \mathbf{P}\mathbf{s}$ where \mathbf{P} is a permutation matrix.

Audio & Music Lab, Department of Electronic Engineering, King's College London, Strand, London WC2R 2LS, UK.
Email: mark.plumbley@kcl.ac.uk.

Rather than deal with the original observations \mathbf{z} , our task will be simplified if we perform a *pre-whitening* step [2], generating $\mathbf{x} = \mathbf{Q}\mathbf{z}$ where the $n \times m$ matrix \mathbf{Q} is chosen to make the covariance $\mathbf{C}_\mathbf{x} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T] = \mathbf{I}_n$ where $\bar{\mathbf{x}}$ is the mean of \mathbf{x} . Note that, in contrast to many other pre-whitening methods, we do not subtract the mean of the data during pre-whitening, since this would lose any information about non-negativity of the sources.

Given that $\mathbf{C}_\mathbf{s} = \mathbf{I}_n$ it is clear that the covariance of \mathbf{z} given by $\mathbf{C}_\mathbf{z} = E[(\mathbf{z} - \bar{\mathbf{z}})(\mathbf{z} - \bar{\mathbf{z}})^T] = \mathbf{A}\mathbf{C}_\mathbf{s}\mathbf{A}^T = \mathbf{A}\mathbf{A}^T$ has exactly n non-zero bounded eigenvalues since \mathbf{A} is bounded and has rank n . To calculate a suitable \mathbf{Q} , let the compact eigenvector-eigenvalue decomposition of $\mathbf{C}_\mathbf{z}$ be given by $\mathbf{C}_\mathbf{z} = \mathbf{U}_z\Lambda_z\mathbf{U}_z^T$ where \mathbf{U}_z is a matrix of n orthogonal m -dimensional column vectors such that $\mathbf{U}_z^T\mathbf{U}_z = \mathbf{I}_n$ and $\Lambda_z = \text{diag}(\lambda_{z_1}, \dots, \lambda_{z_n})$ where the λ_{z_i} are bounded and non-zero. If we set $\mathbf{Q} = \mathbf{R}\Lambda_z^{-1/2}\mathbf{U}_z^T$ where \mathbf{R} is any $n \times n$ orthonormal rotation matrix with $\mathbf{R}\mathbf{R}^T = \mathbf{R}^T\mathbf{R} = \mathbf{I}_n$ (for example, we may choose $\mathbf{R} = \mathbf{I}_n$), then $\mathbf{C}_\mathbf{x} = \mathbf{Q}\mathbf{C}_\mathbf{z}\mathbf{Q}^T = \mathbf{I}_n$ as required. Note that the matrix product $\mathbf{Q}\mathbf{A}$ is square, and is orthonormal since $(\mathbf{Q}\mathbf{A})(\mathbf{Q}\mathbf{A})^T = \mathbf{I}_n$.

If we write $\mathbf{y} = \mathbf{W}\mathbf{x}$ where \mathbf{W} is an $n \times n$ orthonormal rotation matrix with $\mathbf{W}\mathbf{W}^T = \mathbf{W}^T\mathbf{W} = \mathbf{I}_n$ then we have $\mathbf{y} = \mathbf{W}\mathbf{Q}\mathbf{A}\mathbf{s} = \mathbf{U}\mathbf{s}$ where $\mathbf{U} = [u_{ij}]$ must be square and orthonormal with $\mathbf{U}\mathbf{U}^T = \mathbf{I}_n$, since both \mathbf{W} and $(\mathbf{Q}\mathbf{A})$ are square and orthonormal. Our task is therefore to find an orthonormal rotation \mathbf{W} , for which \mathbf{y} is a permutation of \mathbf{s} , i.e. that \mathbf{U} is a permutation matrix [2], from which we can also construct $\mathbf{V} = \mathbf{W}\mathbf{Q}$ if required.

III. MAIN THEOREM

First we shall write down a few useful results. In the following, let c be any positive real constant.

Suppose p and q are real-valued independent random variables. Then we have $\Pr(p < q) \geq \Pr(p < q, p < c, c \leq q) = \Pr(p < c, c \leq q)$ so

$$\Pr(p < q) \geq \Pr(p < c)\Pr(c \leq q). \quad (1)$$

Suppose also that $\mathcal{P} = \{p_i | p_i > 0\}$ is a non-empty set of strictly positive independent random variables with $i \in \mathcal{I} \neq \emptyset$. If $p_i < c/|\mathcal{I}|$ for all $i \in \mathcal{I}$, where $|\mathcal{I}|$ is the number of elements in \mathcal{I} , then $\sum_{i \in \mathcal{I}} p_i < c$. Therefore since p_i are independent, we have

$$\Pr\left(\sum_{i \in \mathcal{I}} p_i < c\right) \geq \prod_{i \in \mathcal{I}} \Pr(p_i < c) \quad (2)$$

For the case where $\mathcal{P} = \mathcal{I} = \emptyset$, the empty sum will be zero ($< c$) and the empty product will be 1, so the inequality (2) also holds with $1 = 1$.

Suppose also that $\mathcal{Q} = \{q_j | q_j \geq 0\}$ is a non-empty set of non-negative independent random variables with $j \in \mathcal{J} \neq \emptyset$. If $q_{j'} \geq c$ for some $j' \in \mathcal{J}$, then $\sum_{j \in \mathcal{J}} q_j \geq c$, and hence

$$\Pr\left(\sum_{j \in \mathcal{J}} q_j \geq c\right) \geq \Pr(q_{j'} \geq c) \quad (3)$$

for any $j' \in \mathcal{J}$.

Lemma 1: Let $\mathbf{U} = [u_{ij}]$ be an $n \times n$ orthonormal matrix such that $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}_n$. Then all elements of \mathbf{U} are non-negative iff \mathbf{U} is a permutation matrix, i.e. a matrix for which, given a sequence $\{j_i | 0 \leq i \leq n\}$ of n distinct integers $0 \leq j_i \leq n$, we have $u_{ij} = 1$ if $j = j_i$ and $u_{ij} = 0$ otherwise.

Proof: Suppose that $u_{ij} \geq 0$ for all i, j . Since $\mathbf{U}^T \mathbf{U} = \mathbf{I}_n$, we have $\sum_i u_{ij} u_{ik} = 0$ for all $j \neq k$, and the positivity of the elements means that $u_{ij} \neq 0$ implies $u_{ik} = 0$ for all $k \neq j$. Similarly from $\mathbf{U}^T \mathbf{U} = \mathbf{I}_n$ we have $\sum_j u_{ij} u_{kj} = 0$, so $u_{ij} \neq 0$ implies $u_{kj} = 0$ for all $k \neq i$. Thus only one element can be non-zero in every row or column. We also have that $\sum_i u_{ij}^2 = 1$ for all j , i.e. that the sum of the squares of every element in each column j is unity, so the non-zero elements of \mathbf{U} (which must be positive) are +1. Hence \mathbf{U} must be a permutation matrix.

The converse is clearly true, since the elements of a permutation matrix take values 0 or 1 only, which are non-negative. \blacksquare

We are now in a position to state the main theorem.

Theorem 1: Let \mathbf{s} be an n -dimensional random vector of real-valued, non-negative and well-grounded independent sources, each with unit variance, and let $\mathbf{y} = \mathbf{U}\mathbf{s}$ be an orthonormal rotation of \mathbf{s} , where $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}_n$. Then \mathbf{U} is a permutation matrix if and only if \mathbf{y} is non-negative with probability 1.

Proof: In the case where \mathbf{U} is a permutation matrix, \mathbf{y} is simply a permutation of the non-negative source vector \mathbf{s} , so is itself non-negative.

To prove the other direction, we suppose that \mathbf{y} is non-negative with probability 1, which is true iff $\Pr(y_i < 0) = 0$ for all $0 \leq i \leq n$.

Let us also tentatively suppose that there is at least one strictly negative element $u_{i'j'} < 0$ of \mathbf{U} . Let us write $y_{i'} = \sum_j u_{i'j} s_j = \sum_{j \in \mathcal{J}_{i'}^+} u_{i'j} s_j - \sum_{j \in \mathcal{J}_{i'}^-} (-u_{i'j}) s_j$ where the sets $\mathcal{J}_{i'}^+ = \{j | u_{i'j} > 0\}$ and $\mathcal{J}_{i'}^- = \{j | u_{i'j} < 0\}$ are mutually exclusive ($\mathcal{J}_{i'}^+ \cap \mathcal{J}_{i'}^- = \emptyset$), and the zero-valued $u_{i'j}$ s are omitted. Note that $j' \in \mathcal{J}_{i'}^-$ since $u_{i'j'} < 0$. Then $y_{i'} < 0 \iff \sum_{j \in \mathcal{J}_{i'}^+} u_{i'j} s_j < \sum_{j \in \mathcal{J}_{i'}^-} (-u_{i'j}) s_j$ and therefore

$$\Pr(y_{i'} < 0) = \Pr \left(\sum_{j \in \mathcal{J}_{i'}^+} u_{i'j} s_j < \sum_{j \in \mathcal{J}_{i'}^-} (-u_{i'j}) s_j \right). \quad (4)$$

Let us choose a positive constant $c_{i'} = -u_{i'j'} \bar{s}_{j'} > 0$, where $\Pr(s_{j'} \geq \bar{s}_{j'}) > 0$ since $s_{j'}$ has non-zero variance. Then applying the results (1), (2) and (3) above to equation (4), while noting that the s_j s are independent and each appear at most once in the RHS of (4), in the case where $\mathcal{J}_{i'}^+ \neq \emptyset$ we get

$$\Pr(y_{i'} \leq 0) \geq \Pr \left(\sum_{j \in \mathcal{J}_{i'}^+} u_{i'j} s_j < c_{i'} \right) \Pr \left(c_{i'} \leq \sum_{j \in \mathcal{J}_{i'}^-} (-u_{i'j}) s_j \right) \quad (5)$$

$$\geq \prod_{j \in \mathcal{J}_{i'}^+} \Pr(u_{i'j} s_j < (c_{i'} / |\mathcal{J}_{i'}^+|)) \Pr(c_{i'} \leq (-u_{i'j'}) s_{j'}) \quad (6)$$

$$= \Pr(s_{j'} \geq \bar{s}_{j'}) \prod_{j \in \mathcal{J}_{i'}^+} \Pr(s_j < (c_{i'} / (|\mathcal{J}_{i'}^+| u_{i'j}))). \quad (7)$$

Now since the random variables are all *well-grounded*, all the probabilities on the RHS are strictly positive, as is the resulting product, implying that $\Pr(y_{i'} \leq 0) > 0$. It is simple to verify that this also holds in the case of $\mathcal{J}_{i'}^+ = \emptyset$. But we know that $\Pr(y_{i'} \leq 0) = 0$, therefore our tentative supposition that at least one element of \mathbf{U} is negative is false. Thus all elements of \mathbf{U} must be non-negative, and from Lemma 1, \mathbf{U} must be a permutation matrix. \blacksquare

Many practical algorithms to perform ICA and related tasks are expressed in terms of minimization of a cost function or ‘error function’. The following corollary follows directly.

Corollary 1: For the same conditions as Theorem 1, suppose that $J(\mathbf{U})$ is a cost function such that $J(\mathbf{U}) = 0$ if and only if $\mathbf{y} = \mathbf{U}\mathbf{s}$ is non-negative with probability 1. Then $J(\mathbf{U}) = 0$ if and only if \mathbf{U} is a permutation matrix, i.e. that \mathbf{y} is a permutation of the sources \mathbf{s} .

Of course, in practical algorithms the cost function will actually be determined by \mathbf{W} . For example, suppose we have a rectified output $\mathbf{y}^+ = [y_1^+, \dots, y_n^+]^T$ where $y_i^+ = \max(0, y_i)$, and we construct a re-estimate of \mathbf{x} given by $\hat{\mathbf{x}} = \mathbf{W}^T \mathbf{y}^+$, noting that $\mathbf{W}\hat{\mathbf{x}} = \mathbf{W}\mathbf{W}^T \mathbf{y}^+ = \mathbf{y}^+$ and $\mathbf{y}^+ = \mathbf{y}$ if and only if $y_i \geq 0$ for all i . Then it is easy to verify that the cost functions $J_0(\mathbf{W}) = \Pr(\hat{\mathbf{x}} \neq \mathbf{x})$ and $J_2(\mathbf{W}) = E(|\mathbf{x} - \hat{\mathbf{x}}|^2)$ are both suitable cost functions for Corollary 1.

IV. DISCUSSION

Fig. 1 shows an illustration of the separation process. From this figure it is intuitively clear that there is only one possible rotation for which neither of the outputs y_i are ever negative. However, one can appreciate that this depends on the well-groundedness of the sources \mathbf{s} : if their pdf did not extend down to zero, there would be some ‘slack’, allowing the output \mathbf{y} to rotate without going negative.

A number of other interesting points and insights emerge. Firstly, in this system there is no requirement for the (unwhitened) inputs \mathbf{z} or the underlying generator matrix \mathbf{A} to be non-negative, in contrast to e.g. non-negative matrix factorization [1]. This indicates that the non-negative ICA approach may be applicable to problems where a non-negative source causes a negative effect in an observation, such as the presence of a shadow in an image, reducing the pixel intensity.

In addition, apart from the rectification nonlinearity, this approach only requires second order statistics. It does not need to make assumptions as to the shape of the source pdfs, such as their kurtosis, except for the requirement for the source pdfs to be non-negative and well-grounded at zero. It would be interesting to explore whether these requirements could be relaxed in favour of a bounded-minimum requirement.

We note that the proof relies on the sources being what we call *well-grounded*, i.e. that they have a non-vanishing pdf down to zero. There is much current interest in the analysis of observations generated by *sparse* sources [3], [4], [5]. Sparse sources have large concentration of probability around zero, representing a high probability of being ‘off’. We would expect that a learning algorithm developed from the principles outlined in this letter would work particularly well for sources with this type of distribution.

In a previous paper [6] we experimentally investigated learning algorithms based on inhibitory connections to remove the covariance in \mathbf{y}^+ . The current proof is slightly different, in that the pre-whitening ensures that the covariance in \mathbf{y} (rather than \mathbf{y}^+) is removed, and therefore Theorem 1 is not a convergence proof for those algorithms. However, a number of authors have studied learning algorithms involving the orthonormal matrices \mathbf{W} that we have used here, but using other types of cost functions or contrast functions. Fiori [7] discusses a framework for such learning algorithms based on the concept of a *Stiefel manifold*: a set $\mathcal{I}_1^{n \times n} \subset \mathbb{R}^{n \times n}$ of matrices \mathbf{W}^T such that $\mathbf{W}\mathbf{W}^T = \mathbf{I}_n$, together with an inhomogeneous function $C(\mathbf{W}) \neq C(\mathbf{R}\mathbf{W})$ for an orthonormal matrix $\mathbf{R} \neq \mathbf{I}_n$. Corollary 1 suggest that both $J_0(\mathbf{W})$ and $J_2(\mathbf{W})$ would be candidates for such an inhomogeneous function, and that it might be possible to use these to construct a learning algorithm to find the separated basis.

Finally, we note that this is a *noise-free* approach. Any significant noise, for example on the original observations \mathbf{z} , would be likely to disrupt the pre-whitening step (since apparent correlations would be reduced), but more seriously it would mean that there may not be any rotation \mathbf{W} for which \mathbf{y} is entirely non-negative, and hence perfect reconstruction would not be possible. Nevertheless,

it would be interesting to investigate the effect of noise on any learning algorithms developed. We might conjecture, for example, that a learning algorithm based on $J_2(\mathbf{W})$ would be more robust than one based on $J_0(\mathbf{W})$, since the former might be expected to rely more heavily on larger scale inputs.

V. CONCLUSIONS

We have considered the task of identifying independent sources from a noiseless nonsingular linear mixture, in the case where the sources are known to be non-negative.

Given certain reasonable conditions, including that the sources have non-vanishing pdf in the positive neighbourhood of zero, we showed that with pre-whitened sources \mathbf{x} multiplied by a square orthonormal rotation matrix \mathbf{W} to give the outputs $\mathbf{y} = \mathbf{W}\mathbf{x}$, the underlying sources will be identified if and only if all components of \mathbf{y} are non-negative with probability 1, or equivalently, the reconstruction error from a rectified output \mathbf{y}^+ is reduced to zero.

We suggest that this non-negative ICA approach may lead to practical learning algorithms, and these may be particularly suitable for the separation of sparse sources.

VI. ACKNOWLEDGEMENTS

The author would like to thank Samer Abdallah and Lai-Wan Chan for useful discussions and suggestions.

REFERENCES

- [1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 21 October 1999.
- [2] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
- [3] D. J. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, pp. 559–601, 1994.
- [4] D. Charles and C. Fyfe, "Modelling multiple-cause structure using rectification constraints," *Network: Computation in Neural Systems*, vol. 9, pp. 167–182, 1998.
- [5] P. O. Hoyer and A. Hyvärinen, "A non-negative sparse coding network learns contour coding and integration from natural images," Submitted manuscript. Available online at <http://www.cis.hut.fi/aapo/pub.html>, 2001.
- [6] M. D. Plumbley, "Adaptive lateral inhibition for non-negative ICA," 2001, Submitted to the *International Conference on Independent Component Analysis and Blind Signal Separation, ICA2001*.
- [7] S. Fiori, "A theory for learning by weight flow on Stiefel-Grassman manifold," *Neural Computation*, vol. 13, pp. 1625–1647, 2001.

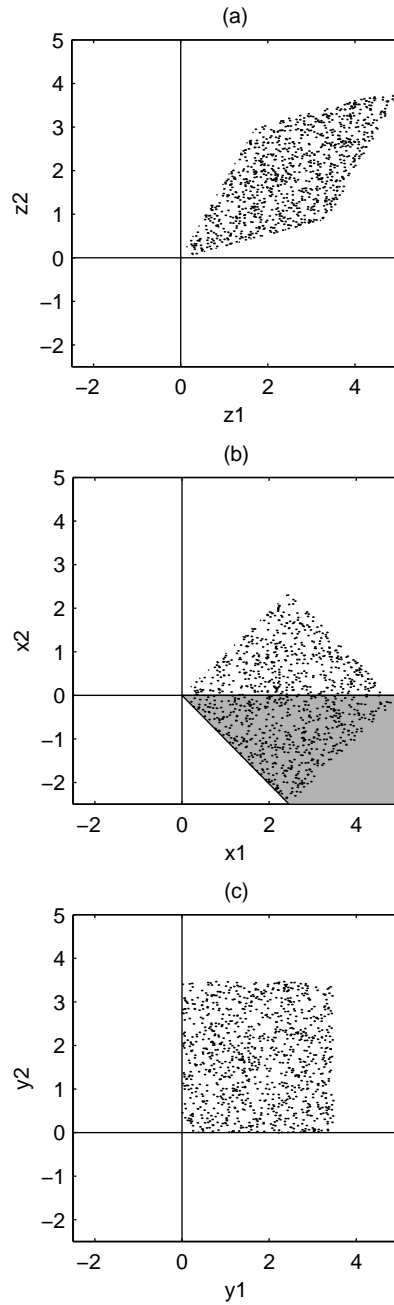


Fig. 1. Visualization of the separation process, for a simple artificial example, showing (a) the unwhitened observations \mathbf{z} , (b) the pre-whitened observations \mathbf{x} , and (c) the final non-negative outputs \mathbf{y} . The shading in (b) shows the region where $x_2 < 0$, which must be removed by the rotation $\mathbf{y} = \mathbf{W}\mathbf{x}$.