A unifying model for blind separation of independent sources

Aapo Hyvärinen
HIIT Basic Research Unit
Dept of Computer Science
University of Helsinki, Finland
aapo.hyvarinen@helsinki.fi
www.cs.helsinki.fi/aapo.hyvarinen/

17th November 2004

Abstract

Many algorithms have been proposed for blind separation of statistically independent sources. Most of the algorithms are based on one of the following properties: nongaussianity of the sources, their different autocorrelations, or their smoothly changing nonstationary variances. Each of the methods is able to separate sources if the respective assumptions are met. Here we propose a simple unifying model that is able to separate independent sources if any one of these three conditions is met. The model is a simple autoregressive model whose estimation can be performed by maximum likelihood estimation. We also propose a simple yet accurate approximation of the likelihood that gives a simple algorithm.

Keywords: Blind source separation, independent component analysis, nonstationary variance, autocorrelations

1 Introduction

Blind source separation (BSS) is typically performed in a setting where the observed signals are instantaneous noise-free linear superpositions of underlying hidden source signals. Let us denote the n source signals by $s_1(t),\dots,s_n(t)$, and the observed signals by $x_1(t),\dots,x_m(t)$, where t is the time index. Let a_{ij} denote the coefficients in the linear mixing between the source $s_j(t)$ and the observed signal $x_i(t)$. Further, let us collect the source signals in a vector $\mathbf{s}(t) = (s_1(t),\dots,s_n(t))^T$, and similarly we construct the observed signal vector $\mathbf{x}(t)$. Now the mixing can be expressed as the equation

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

where the matrix $\mathbf{A} = [a_{ij}]$ collects the mixing coefficients. No particular assumptions on the mixing coefficients are made. However, some weak structural assumptions are often made: for example, it is typically assumed that the mixing matrix is square, that is, the number of source signals equals the number of observed signals ($n = m$), which we will assume here as well. For technical simplicity, we shall also assume that all the signals have zero mean, but this is no restriction since is simply means that the signals have been centred [18]. The problem of blind source separation is now to estimate both the source signals $s_j(t)$ and the mixing matrix $\mathbf{A}$, based on observations of the $x_i(t)$ alone [19, 18].

In most methods, including the present one, the source signals are assumed statistically independent. Then, the model can be estimated if the source signals fulfill some additional assumptions, three of which are commonly used. First, if all the components (except perhaps one) have nongaussian distributions, the ensuing model is called independent component analysis (ICA) [7], and many techniques are available for estimation of the model [18]. Second, if the components have nonstationary, smoothly changing variances [20, 24, 12], the model can be estimated as well. Third, one can use temporal second-order correlations [25, 21, 3, 23] — an important difference from the two preceding principles is that in this case we also need to assume that the signals have *different* autocorrelation functions; mere existence of autocorrelations is not sufficient.

In this paper, we propose a simple model that unifies these three properties (nongaussianity, different autocorrelations, variance nonstationarity). Estimation of this model thus enables separation of sources that have any one of these properties. Moreover, the model uses all these properties simultaneously, which is likely to increase the performance of the separation method if the data has more than one of the properties. A theoretical treatment of the situation was given in [5]; our emphasis here is on developing a simple, concrete model and a practical algorithm.

First, we will review the properties used in previous methods in source separation, and their formulations in a maximum likelihood framework (Section 2). Then, we propose a new unifying model, formulate its likelihood, and propose an algorithm for source separation by maximum likelihood estimation (Section 3). Simulation results show that the model separates sources in cases where existing methods are not able to do so (Section 4), and finally we discuss related work and conclude the paper (Section 5).

2 Previous models

2.1 Nongaussianity

If the sources are assumed to be nongaussian and the time structure is ignored, we obtain the classic ICA model [7, 19] whose maximum likelihood estimation has been extensively investigated, see e.g. [18]. Denote by $\mathbf{W} = \mathbf{A}^{-1}$ the inverse of the mixing matrix. Denote the i-th row of $\mathbf{W}$ by $\mathbf{w}_i^T$. Assume that we have T observations $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ of the mixed data. Then the logarithm of the likelihood is given by

$$\log L(\mathbf{W}) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log p_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log |\det \mathbf{W}| \tag{2}$$

where p_i is the probability density function (pdf) of the i-th source (independent component).

To simplify notation, we can divide the log-likelihood by T, and denote the average over the sample index t by an expectation operator $\hat{E}$, to obtain

$$\frac{1}{T} \log L(\mathbf{W}) = \sum_{i=1}^{n} \hat{E} \left\{ \log p_i(\mathbf{w}_i^T \mathbf{x}(t)) \right\} + \log |\det \mathbf{W}| \tag{3}$$

Another principle that gives an essentially equivalent objective function is minimization of mutual information [7, 18]. Nongaussianity is also often used in the form of cumulant-based methods such as maximization of kurtosis [8, 18]; using cumulants can be motivated as an approximation of the differential entropy used in mutual information [7, 18]. An intuitive interpretation of such estimation procedures as projection pursuit, i.e. finding the most nongaussian projections of the data, is also possible [18].

2.2 Nonstationary variances

The second statistical property that can be used for source separation is nonstationarity of the variance [20, 24, 12]: The variance of each independent source signal is assumed to change smoothly as a function of time.

Let us denote the probability density function of the signal $s_i(t)$ at time point t by $p_i(s,t)$. This depends on t because of the nonstationarity. Since we are only allowing nonstationarity of the variance, the pdf has the form

$$p_i(s_i,t) = \frac{1}{\sigma_i(t)} p_i\left(\frac{s_i}{\sigma_i(t)}\right) \tag{4}$$

where $p_i(s)$ is the underlying pdf in the hypothetical case where the nonstationarity is not present. Thus, the log-likehood of the model is given by

$$\frac{1}{T}\log L(\mathbf{W}) = \sum_{i=1}^{n} \hat{E}\left\{\log p_i\left(\frac{\mathbf{w}_i^T\mathbf{x}(t)}{\sigma_i(t)}\right) - \log\sigma_i(t)\right\} + \log|\det\mathbf{W}| \tag{5}$$

Here, it is assumed that the "nuisance" parameters $\sigma_i(t)$ are estimated separately, for a given $\mathbf{W}$, thus the σ_i are functions of $\mathbf{W}$. In fact, this is basically rather simple since for a given $\mathbf{W}$, we can take time windows centered around each time point t, and estimate the local variance for each estimate of the source signal inside that window. This is possible because of the crucial assumption that the variance changes smoothly.

The situation could be considerably simplified by assuming that the underlying densities p_i are gaussian [24]. However, this is a bit restrictive because then the marginal density over the whole data set has necessarily positive kurtosis, see e.g. [15]. In contrast, if the underlying density has a negative kurtosis, the overall density does not need to have positive kurtosis.

2.3 Different second-order autocorrelations

The third statistical property is the second-order autocorrelations of the signals, which have to be different (distinct) from each other [25, 21, 3]. We shall here formulate a very simple model of such signals to illustrate this principle and to prepare for the unifying model in the next section. A simple way of formulating a proper statistical model with autocorrelations of the sources is to express each source signal using a gaussian first-order autoregressive model:

$$s_i(t) = \alpha_i s_i(t-1) + n_i(t) \tag{6}$$

where $n_i(t)$ is a gaussian i.i.d. innovation process of zero mean. The variance of n_i can be computed to equal $E\{[\mathbf{w}_i^T\mathbf{x}(t) - \alpha_i\mathbf{w}_i^T\mathbf{x}(t-1)]^2\}$, which yields a simple moment estimator of the variance of the innovation when the sample average is used instead of the expectation. Since n_i is gaussian, the likelihood can then be calculated by plugging this variance estimate in a gaussian likelihood, which gives after some manipulations:

$$\frac{1}{T}\log L(\mathbf{W},\alpha_1,\ldots,\alpha_n) = -\sum_{i=1}^{n}\frac{1}{2}\log\hat{E}\{[\mathbf{w}_i^T\mathbf{x}(t) - \alpha_i\mathbf{w}_i^T\mathbf{x}(t-1)]^2\} + \log|\det\mathbf{W}| - n\log\sqrt{2\pi} - \frac{n}{2} \tag{7}$$

Now, assume that the observed data is whitened and that $\mathbf{W}$ is constrained to be orthogonal. Assume further that the estimation of the autoregressive coefficients is decoupled from the estimation of the $\mathbf{w}_i$. In other words, the α_i are estimated for fixed $\mathbf{w}_i$ by maximization of the likelihood with respect to the α_i, which gives:

$$\hat{\alpha}_i = \hat{E}\{(\mathbf{w}_i^T\mathbf{x}(t))(\mathbf{w}_i^T\mathbf{x}(t-1))^T\} = \mathbf{w}_i^T\mathbf{C}_{-1}\mathbf{w}_i \tag{8}$$

where $\mathbf{C}_{-1} = \hat{E}\{\mathbf{x}(t)\mathbf{x}(t-1)^T\}$ is an autocovariance matrix. Then the log-likelihood can be expressed as a function of $\mathbf{W}$ alone. After tedious algebraic manipulations we have

$$\frac{1}{T}\log L(\mathbf{W}) = -\sum_{i=1}^{n}\frac{1}{2}\log[1 - (\mathbf{w}_i^T\mathbf{C}_{-1}\mathbf{w}_i)^2] - n\log\sqrt{2\pi} - \frac{n}{2} \tag{9}$$

This formulation of likelihood shows that these methods are not able to separate source signals that have equal autocovariances. The autocovariance of the sum $a_i s_i + a_j s_j$ of two independent signals is given by $a_i^2 \text{cov}(s_i(t), s_i(t-1)) + a_j^2 \text{cov}(s_j(t), s_j(t-1))$, and the same formula applies for the variance of the sum. If s_i and s_j have equal autocovariances, any mixing of s_i and s_j with an orthogonal mixing matrix will give two signals which have the same variances *and* the same autocovariances as the original signals. Thus, the likelihoods will be equal for any values of $\mathbf{W}$, and maximum likelihood estimation is not able to distinguish the two cases. The same applies even if we consider many time lags [3].

Some articles have also proposed combinations of nongaussianity and second-order autocorrelations [13, 22]. Our previous approach in [13] is close to the present one, but did not include nonstationary variances.

3 Unifying model

3.1 Definition of the model

Here we propose a signal model that incorporates the three properties commonly used for signal separation: nongaussianity, distinct autocorrelations, and a smoothly changing nonstationary variance. We model each source signal by an autoregressive model

$$s_i(t) = \sum_{\tau>0} \alpha_i^\tau s_i(t-\tau) + n_i(t) \tag{10}$$

where the zero-mean innovation term $n_i(t)$ can be nongaussian, and its variance can be nonstationary.

If $n_i(t)$ is nongaussian, the source signal $s_i(t)$ is nongaussian as well. If the autoregressive coefficients α_i^τ are not zero, the source signals have autocorrelations. Finally, if the variance of the innovation is nonstationary, the variance of the source signal is nonstationary. Thus, we are able to model all these three properties.

As is typically assumed in blind source separation, each observed signal is a linear combination of these source signals as in Eq. (1). For simplicity, it is assumed as above that the number of source signals equals the number of observed signals, let us denote it by n. Moreover, we must fix the scales of the source signal. As always in BSS, the scale of the signal is arbitrary, so we define the variance of each s_i to be equal to 1. The signals have necessarily zero mean because the innovations have zero mean.

3.2 Likelihood of the model

Assume first that we know the innovation signals $n_i(t)$. Then we could estimate the local standard deviation of the innovation $\sigma_i(t)$ by local averages. Denoting the estimator by $\hat{\sigma}_i(t)$, we can use:

$$\hat{\sigma}_i(t) = \sqrt{\sum_\gamma h(\gamma) n_i^2(t-\gamma)} \tag{11}$$

where $\mathbf{h} = (\ldots, h(-2), h(-1), h(0), h(1), h(2), \ldots)$ is some low-pass filter such that the sum of its weights (which are assumed nonnegative) is equal to 1. Note that the $\hat{\sigma}_i(t)$, when estimated for an estimate of $s_i(t)$ given by $\mathbf{w}_i^T \mathbf{x}$, are functions of $\mathbf{w}_i$, $\mathbf{h}$, and the coefficients α_i^τ. To emphasize this, we write $\hat{\sigma}_i(t, \mathbf{w}_i, \mathbf{h}, \alpha_i^\tau)$.

We now compute the likelihood with respect to the innovation processes. We do not consider the prior distribution of the $\sigma_i(t)$. This can be justified on the grounds of simplicity; the likelihood with respect to the n_i seems to be quite enough for succesful separation (according to the simulations below).

To be able to compute the innovations, we need estimates of the autoregressive coefficients, $\hat{\alpha}_i^\tau$. Note that these estimates, just like $\sigma_i(t)$, depend on $\mathbf{w}_i$ because $\mathbf{w}_i$ defines the estimate of the source signal from which the estimates are obtained. To emphasize this we write $\hat{\alpha}_i^\tau(\mathbf{w}_i)$ in the following. The actual method for estimating the autoregressive coefficients is not our main interest here.

Following Section 2.2, the pdf of each innovation can be expressed as

$$p_i(n_i,t) = \frac{1}{\sigma_i(t)} p_i\left(\frac{n_i}{\sigma_i(t)}\right) \tag{12}$$

Like in Eq. (5), we then express the log-likelihood by considering these marginal distributions. Replacing $n_i(t)$ by its estimate $\mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))$ and $\sigma_i(t)$ by its estimate in (11), we obtain

$$\frac{1}{T}\log L(\mathbf{W}) = \sum_{i=1}^{n} \hat{E}\left\{G_i\left(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\right) - \log\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)\right\} + \log|\det\mathbf{W}| \tag{13}$$

where $\hat{E}$ denotes the expectation over t (sample average). The function G_i is the logarithm of the probability density function of the “underlying” innovation process, i.e. the pdf of $n_i(t)$ in the hypothetical case where it is stationary and its variance is equal to one.

3.3 Estimation algorithm

Now we develop an estimation algorithm based on maximization of the likelihood in Eq. (13).

First, it must be noted that the ensuing algorithm easily becomes unstable. More precisely, the estimates of G_i must be very accurate to prevent $\mathbf{W}$ from going to zero or infinity. Therefore, we use a classic trick of stabilizing BSS algorithms: we prewhiten the data and constrain $\mathbf{W}$ to be orthogonal [18]. After this stabilization, the log-likelihood is simplified to:

$$\frac{1}{T}\log L(\mathbf{W}) = \sum_{i=1}^{n} \hat{E}\left\{G_i\left(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\right) - \log\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)\right\} \tag{14}$$

As derived in the Appendix, the gradient of (14) can be approximated as follows:

$$\nabla_{\mathbf{w}_i}\frac{1}{T}\log L(\mathbf{W}) \approx \hat{E}\left\{\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\left(\mathbf{x}(t) - \sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau)\right) g_i\left(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\right)\right\} \tag{15}$$

and $g_i(.)$ is the derivative of $G_i(.)$ and $\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)$ is defined as

$$\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau) = \sqrt{\sum_\gamma h(\gamma)[\mathbf{w}_i^T(\mathbf{x}(t-\gamma) - \sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\gamma-\tau))]^2} = \sqrt{\sum_\gamma h(\gamma)\hat{n}_i(t-\gamma)^2} \tag{16}$$

This approximation of the gradient is quite well justified, because it is asymptotically exact in the case where the data is actually generated according the autoregressive model, the G_i are equal to the log-likelihoods of the underlying stationary (normalized) innovation proceeses, the local standard deviations and the innovations are estimated exactly, and the innovations have a symmetric distribution.

To improve the convergence, it is quite useful to perform a projection of the gradient on the tangent surface of the set of orthogonal matrices [9]. This means replacing the gradient $\nabla_{\mathbf{W}}$ with respect to by:

$$\nabla_{\mathbf{W}}^{ort} = \nabla_{\mathbf{W}} - \mathbf{W}\nabla_{\mathbf{W}}^T\mathbf{W} \tag{17}$$

We also need to compute the nonlinear functions g_i used in the algorithm. These are the derivatives of the log-densities G_i. In ICA, it is well-known that the exact form of the non-quadratic function used to probe higher-order statistics is not very important [6, 18]. We may therefore optimistically assume that the exact form of the function G is not very important here either, as long as it is qualitatively similar enough. Thus, function g_i can be chosen as in ordinary ICA, but according to the probability distribution of the estimate of the innovation process normalized to unit variance. If the innovation is supergaussian, $g(u) = -\text{sign}(u)$ is suitable. This could also be approximated by a smoother function $g(u) = -\tanh(u)$ [2, 18]. For subgaussian innovations, one could use $g(u) = -u + \tanh(u)$ [10], or $g(u) = -u^3$, for example. For almost gaussian innovations, $g(u) = -u$ could be used, but it is probably not necessary to consider this as a special case; the same function as for supergaussian innovations seems to work in simulations. (We have here omitted some multiplicative constants which are considered immaterial in ICA estimation.)

Finally, the autoregressive coefficients have to be estimated. The statistically optimal way to accomplish this would be to maximize the likelihood, but a computationally simpler method can be found by using classical least-squares methods. In fact, a large number of methods have been developed for estimating the coefficients in this case. In the case of an first-order autoregressive model, the very simple formula used in Equation (8) could be used.

Thus, the estimation consists of the following steps:

0. Remove the mean from the data and whiten it. Denote the preprocessed data by $\mathbf{x}(t)$. Choose a (random) initial value for the matrix $\mathbf{W}$.

1. Compute estimates of the source signals as $\hat{s}_i(t) = \mathbf{w}_i^T \mathbf{x}(t)$.

2. Compute estimates for the autoregressive coefficients $\hat{\alpha}_i^\tau$, for example, by a classical least-squares method.

3. Compute estimates of the innovations as $\hat{n}_i(t) = \hat{s}_i(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau \hat{s}_i(t-\tau)$. Compute estimates of the local standard deviations $\hat{\sigma}_i(t)$ according to Eq. (16).

4. Choose a nonlinearity g_i for each source based on the distributions of the normalized innovations $\hat{n}_i(t)/\hat{\sigma}_i(t)$. For example, if the kurtosis of $\hat{n}_i$ is positive, take $g_i(u) = -\text{sign}(u)$, otherwise take $g_i(u) = -u + \tanh(u)$.

5. (a) Compute the gradient with respect to $\mathbf{W}$ as given (separately for each row) in Eq. (15).

   (b) Compute the projected gradient by (17).

   (c) Do a gradient step
   $$\mathbf{W} \leftarrow \mathbf{W} + \mu \nabla_{\mathbf{W}}^{ort} \tag{18}$$
   where μ is a small step size constant.

   (d) Orthogonalize $\mathbf{W}$ by
   $$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W} \tag{19}$$

The five steps 1–5 are repeated until $\mathbf{W}$ has converged. This is the version with symmetric orthogonalization, but a deflationary version (one-by-one estimation) can be readily used as well [18, 13].

4 Simulations

To validate the algorithm, we performed blind source separation experiments with artificial data.

In each trial, we created six source signals. First, we created four signals using a first-order autoregressive model with constant variances of the innovations (i.e. constant $\sigma_i(t)$), with 5000 time points. Of these four, signals #1 and #2 were created with supergaussian innovations, and the signals #3 and #4 with gaussian innovations. All these innovations

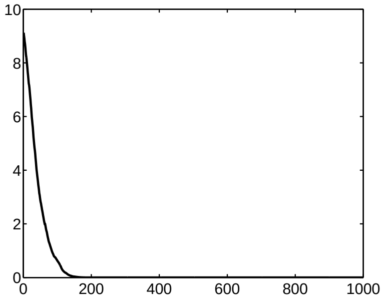

Figure 1: Convergence of our unified BSS algorithm for artificially generated data. Vertical axis: error, horizontal axis: iteration count. The error index shown is the squared distance of the separating matrix times the whitened mixing matrix from the nearest (signed) permutation matrix. The median was taken over 100 runs with different random matrices and initial conditions. To see if there were cases were the algorithm did not converge, we computed the number of cases where the error was was larger than 0.01 after 1000 iterations. There were 3, which means the estimation was succesful in 97% of the iterations.

had constant unit variance. The signals #1 and #3 had identical autoregressive coefficients (0.33), and therefore identical autocovariances; the signals #2 and #4 had identical coefficients (0.75) as well. Finally, we created signals the #5 and #6 so that they had smoothly changing variances as follows: we created two gaussian signals with the same autoregressive method as above (except that the coefficient was 0.9), and then completely randomizing the signs of the signals by multiplying the signals by two binary i.i.d. signals that took the values ± 1 with equal probabilities [12]. These six signals were then mixed as in ICA, using random mixing matrices.

Ordinary ICA methods based on nongaussianity would be able to separate only signals #1 and #2. Methods based on second-order correlations would not be able to separate any of the signals, since there was no signal with a unique autocorrelation. Methods based on nonstationary variances would be able to separate only signals #5 and #6. Methods combining nongaussianity with autocorrelations [13, 22] would be able to separate the first four signals only. Thus, to our knowledge, all existing source separation algorithm would fail with this data.

We ran our algorithm on 100 data sets generated as described above. The step size μ was taken equal to 0.1, and the nonlinearity was fixed a priori as $g(u) = -\text{sign}(u)$. The length of the filter h was 6, and its coefficients were all equal $(1/6)$. Symmetric orthogonalization was used.

Figure 1 shows the convergence of our algorithm. The algorithm correctly estimated the independent components, in around 200 iterations. Note that a single generic nonlinearity that corresponds to supergaussian innovations was able to separate both gaussian and supergaussian signals, which indicates that the method is robust with respect to the choice of nonlinearity in the much same way as ICA.

5 Discussion

A different line of research is considering the case where the source signals are *not* independent. We refer the reader to the existing literature [4, 14, 15, 1, 11, 26, 16, 17]. Future research may find a unifying framework where some dependencies are taken into account in addition to the three properties considered here. A step in that direction was taken in [17].

We have not proven the exact consistency conditions for our method. This is actually a rather complicated question because it depends on how precisely we model the pdf's of the innovations, the linear autoregressive model, and the variance dependencies over time. Assuming that all these are exactly known, we conjecture that our method can separate sources if at least one of the classic conditions discussed in Section 2 is true for each source. In particular, if there is more than one source that is gaussian and stationary, the autocorrelation structures of those sources (but only those) must be different (distinct) as typical with methods based on second-order autocorrelations only.

To conclude, we have proposed a very simple model for signal separation that combines the three basic properties used in blind separation of independent sources. This is based on an autoregressive model of the sources, where the innovations are nongaussian and have nonstationary variances. It is possible to formulate the likelihood of the model in closed form. The gradient of the likelihood can be simplified to yield a relatively simple algorithm. Simulations show that the algorithm is able to separate sources when other existing methods fail.

Appendix: Approximation of the gradient of the log-likelihood

The gradient of $\log L$ in (14) with respect to $\mathbf{w}_i$ can be obtained straight-forwardly as

$$\begin{aligned}
\nabla_{\mathbf{w}_i} \log L(\mathbf{W}) = \hat{E}\Bigg\{ & \frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \Big(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau)\Big) g_i\Big(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\Big)\Bigg\} \\
+ \hat{E}\Bigg\{ & \frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \Big(\sum_{\tau>0} (\nabla_{\mathbf{w}_i}\hat{\alpha}_i^\tau(\mathbf{w}_i))\mathbf{w}_i^T \mathbf{x}(t-\tau)\Big) g_i\Big(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\Big)\Bigg\} \\
- \hat{E}\Bigg\{ & \frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \nabla_{\mathbf{w}_i}\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau) \\
+ & \frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)^2} \nabla_{\mathbf{w}_i}\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)\mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau)) g_i\Big(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\Big)\Bigg\}
\end{aligned} \tag{20}$$

We shall now claim that the second and third terms are negligible with respect to the first one, near the convergence points. To accomplish this simplification, we assume that all the nuisance parameters in the model are estimated exactly, and only the main parameter of interest, $\mathbf{W}$ is not known. That is, the autoregressive coefficients α_i^τ, the local variances σ_i^2, and the log-densities G_i are known exactly. Also we need to assume, the data is generated according to the model and that the innovations are estimated exactly as well.

First we apply the following well-known lemma, proven, for example, in [13]:

Lemma 1 *For any random variable x with a smooth density p_x and satisfying $E\{x\} = 0$, we have*

$$E\{x \frac{p_x'(x)}{p_x(x)}\} = -1 \tag{21}$$

Applying this lemma for the normalized innovations, we have

$$\hat{E}\left\{ \frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau)) g_i(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)} \mathbf{w}_i^T(\mathbf{x}(t) - \sum_{\tau>0} \hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))) \right\} \rightarrow -1, \tag{22}$$

asymptotically when the sample size goes to infinity.

The normalized innovation is independent of any transformation of $\sigma_i(t)$. By assumption, the local variances are estimated perfectly, which means that

$$\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau) = \sqrt{\sum_j (\mathbf{w}_i^T\mathbf{a}_j)^2\sigma_i^2(t)} \tag{23}$$

whose gradient with respect to $\mathbf{w}_i$ is independent of normalized innovations as well. We can thus calculate

$$\begin{aligned}
&\hat{E}\{\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)^2}\nabla_{\mathbf{w}_i}\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)\mathbf{w}_i^T(\mathbf{x}(t)-\sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))g_i(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\mathbf{w}_i^T(\mathbf{x}(t)-\sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))\} \\
&\approx \hat{E}\{\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\nabla_{\mathbf{w}_i}\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)\} \\
&\times \hat{E}\left\{\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\mathbf{w}_i^T(\mathbf{x}(t)-\sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau))g_i\left(\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\mathbf{w}_i^T(\mathbf{x}(t)-\sum_{\tau>0}\hat{\alpha}_i^\tau(\mathbf{w}_i)\mathbf{x}(t-\tau)\right)\right\} \\
&\longrightarrow \hat{E}\{\frac{1}{\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)}\nabla_{\mathbf{w}_i}\hat{\sigma}_i(t,\mathbf{w}_i,\mathbf{h},\alpha_i^\tau)\}\times(-1)
\end{aligned} \tag{24}$$

which implies that the third term is approximately zero asymptotically.

Second, the quantity $\sum_{\tau>0}(\nabla_{\mathbf{w}_i}\hat{\alpha}_i^\tau(\mathbf{w}_i))\mathbf{w}_i^T\mathbf{x}(t-\tau)$ depends only on the past values of $\mathbf{w}_i^T\mathbf{x}(t)$. Therefore, it is approximately independent from the estimate of the normalized innovation at time t; the approximation is exact if the innovations are estimated exactly. Moreover, the term has zero mean. Thus the second term in (20) approximatively vanishes as well. This gives the approximation in (15).

References

[1] F. R. Bach and M. I. Jordan. Beyond independent components: trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.

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

[3] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

[4] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, 1998.

[5] J.-F. Cardoso. The three easy routes to independent component analysis: contrasts and geometry. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, 2001.

[6] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[7] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[8] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[9] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[10] M. Girolami. An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, 10(8):2103–2114, 1998.

[11] J. Hurri and A. Hyvärinen. Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. *Network: Computation in Neural Systems*, 14(3):527–551, 2003.

[12] A. Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474, 2001.

[13] A. Hyvärinen. Complexity pursuit: Separating interesting components from time-series. *Neural Computation*, 13(4):883–898, 2001.

[14] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[15] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[16] A. Hyvärinen and J. Hurri. Blind separation of sources that have spatiotemporal variance dependencies. *Signal Processing*, 84(2):247–254, 2004.

[17] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *J. of the Optical Society of America A*, 20(7):1237–1252, 2003.

[18] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[19] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[20] K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

[21] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3636, 1994.

[22] K.-R. Müller, P. Philips, and A. Ziehe. $JADE_{TD}$: Combining higher-order statistics and temporal information for blind source separation (with noise). In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 87–92, Aussois, France, 1999.

[23] D.-T. Pham. Blind separation of instantaneous mixtures of sources via the Gaussian mutual information criterion. *Signal Processing*, 81:855–870, 2001.

[24] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49(9):1837–1848, 2001.

[25] L. Tong, R.-W. Liu, V.C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38:499–509, 1991.

[26] H. Valpola, M. Harva, and J. Karhunen. Hierarchical models of variance sources. *Signal Processing*, 84:267–282, 2004.