A Semiparametric Approach to Source Separation using Independent Component Analysis

Ani Eloyan* and Sujit K Ghosh*

February 21, 2011

NC State University Department of Statistics Technical Report# 2635

Abstract

Data processing and source identification using lower dimensional hidden structure plays an essential role in many fields of applications, including image processing, neural networks, genome studies, signal processing and other areas where large datasets are often encountered. One of the common methods for source separation using lower dimensional structure involves the use of Independent Component Analysis, which is based on a linear representation of the observed data in terms of independent hidden sources. The problem thus involves the estimation of the linear mixing matrix and the densities of the independent hidden sources. However, the solution to the problem depends on the identifiability of the sources. This paper first presents a set of sufficient conditions to establish the identifiability of the sources and the mixing matrix using moment restrictions of the hidden source variables. Under such sufficient conditions a semi-parametric maximum likelihood estimate of the mixing matrix is obtained using a class of mixture distributions. The consistency of our proposed estimate is established under additional regularity conditions. The proposed method is illustrated and compared with existing methods using simulated and real data sets.

Keywords: Constrained EM-algorithm; Mixture Density Estimation; Source Identification

*Ani Eloyan is a graduate student and Sujit K Ghosh is a Professor, both in the Department of Statistics at North Carolina State University, Raleigh, NC 27695-8203. *Emails:* ani.eloyan@gmail.com and ghosh@stat.ncsu.edu

1 Introduction

The problem of finding a representation of multivariate random variables which maintains its essential distributional structure using a set of lower dimensional random variables has been of interest to researchers in statistics, signal processing and neural networks. Such representations of higher dimensional random vector using a lower dimensional vector provide a statistical framework to the identification and separation of the sources. Since the linear transformations of data are computationally and conceptually easier to implement, most of the methods are based on finding a linear transformation of the data. Some of the major approaches for solving this problem include principal component analysis (PCA), factor analysis (FA), projection pursuit (PP) and independent component analysis (ICA). A distinguishing feature of the ICA compared with other source separation methods is that the lower dimensional random variables are extracted as independent sources in contrast to uncorrelated random variables (e.g., as in PCA). Jutten and Herault (1991) were perhaps the first to state the problem and coin the name ICA. Some of the early approaches to ICA are based on estimating the mixing matrix of the linear transformation by the maximization of the mutual information or the negentropy function (see Comon (1994) for details). Other methods for estimating the mixing matrix are based on gradient algorithms or cumulant functions which are described in detail by Hyvarinen et al. (2001), Cardoso and Souloumiac (1993) and the references therein.

In general, separating the sources as independent components provides maximal separation of the sources from the observed signals and hence ICA has become a popular method among practitioners. In PCA, the goal is to reduce the dimension of the data by decorrelation. However, decorrelation may not be an adequate measure for source separation if the densities of underlying hidden sources are nongaussian. Embrechts et al. (2001) present an excellent overview of the limitations on using correlation as measure of dependence. Since in practice the data collected in signal processing are often nongaussian, the decorrelation approach does not usually result in adequate separation of the sources.

A general formulation of the source separation problem can be presented as follows. Given a random sample $X_1, \ldots, X_T$, where $X_i = (x_{i1}, \ldots, x_{in})^T$ are independent and identically distributed random vectors, can we find a unique transformation $g : \mathbb{R}^n \to \mathbb{R}^m$ for some $m \leq n$ and densities $f_1, \ldots, f_m$ such that

$$X_i \stackrel{d}{=} g(S)$$

where $S = (s_1, \ldots, s_m)^T$ is the vector of independent sources, in other words, $s_j \overset{indep}{\sim} f_j(\cdot)$ for $j = 1, \ldots, m$, $i = 1, \ldots, T$ and "$\stackrel{d}{=}$" denotes that the random quantities on either side of this equality have identical distribution. A special case emerges when the relationship is assumed to be linear. Then $g(S) = AS$ and the problem reduces to the estimation of the mixing matrix A and the probability densities $f_1, \ldots, f_m$.

In matrix notation, a model for ICA can be written as,

$$X = AS + E, \tag{1}$$

where $X = (x_1, \ldots, x_n)^T$, $S = (s_1, \ldots, s_m)^T$, $A = (a_{ij})_{n \times m}$ and $E = (e_1, \ldots, e_n)^T$ is an $n \times 1$ vector of independent gaussian noise variables each with mean 0. Writing $B = (A \ I)$ and $Y = (S^T \ E^T)^T$ we can equivalently express (1) as

$$X = BY \tag{2}$$

Without any loss of generality we assume throughout the article that $E(X) = 0$ and $E(S) = 0$. Also, following the previous works on ICA, for the rest of the article we assume that $n = m$ (but see our remarks in Section 7). In many ICA algorithms such as the FastICA described by Hyvarinen and Oja (2000) it is further assumed that $E = 0$ and the model is called noise free.

In most of the early literature on likelihood based solutions to the ICA, the densities of the independent sources f_j, $j = 1, \ldots, m$, are prespecified parametrically and are chosen based on the fact that these densities should be nongaussian. Some of the commonly used choices are gamma or double exponential densities. Boscolo et al. (2004) proposed a pseudo-likelihood based method for ICA using model (1) where the densities of the sources are estimated nonparametrically by

using the kernel density estimate (Silverman, 1985). It was shown that simultaneously estimating the densities of the sources along with the mixing matrix improves the estimation compared to some of the parametric approaches. The performance of their method was evaluated by simulation studies in terms of maximizing the median signal-to-interference ratio (SIR) defined as $10 \log_{10} \sum_{i=1}^{n} s_{ij}^2/(\widehat{s}_{ij} - s_{ij})^2$, where s_{ij} is the original source value and $\widehat{s}_{ij}$ is the reconstructed value. A more recent nonparametric approach to the linear ICA model (1) proposed by Chen and Bickel (2006) is based on score functions. The score functions of the sources are estimated by using B-splines and the estimate of the unmixing matrix $W = A^{-1}$ is computed by a Newton-Raphson type optimization method. However, most of these previous methods for ICA are difficult to compare mutually, even based on simulated data sets, when the mixing matrix A and the source densities $f_1, \ldots, f_m$ are not uniquely identified.

One of the issues that is partly unresolved in the literature on ICA is the identifiability of the model given in (1). Comon (1994) describes the indeterminacies in the model succinctly as follows. If an information theoretic method is used for ICA and the original sources are 'as nongaussian as possible' then the model is identifiable up to matrix equivalence. Two square matrices A and B of the same dimension are called equivalent if each column of A is proportional to one of the columns of B and vice versa. In other words, there exist an $m \times m$ permutation matrix P, a diagonal matrix Λ with positive entries on its diagonal and a diagonal matrix D with diagonal entries equal to ± 1 such that

$$B = APD\Lambda.$$

Notice that $BS = AG$ if we choose $G = PD\Lambda S$ for any two equivalent matrices A and B which makes the representation (1) not identifiable if the goal is to estimate the matrix A and the densities of the independent components $s_1, \ldots, s_m$.

In most of the commonly used algorithms for ICA the fact that a model for ICA is not fully identifiable is often completely ignored (e.g., FastICA, JADE). Chen and Bickel (2006) proposed restricting the absolute median of the densities of independent sources be unity to partly resolve

the identifiability problem, but they correctly point out that there is still ambiguity due to sign changes and row permutations.

Boscolo et al. (2002) addressed the issue of the identifiability of ICA model where the extracted vector has more than one gaussian component as follows. Suppose that S is a vector of independent components of size $m \times 1$, k of which are gaussian random variables. Then it is proved that the $m-k$ nongaussian components can be extracted up to matrix equivalence from the linear mixture $X = AS$ if the matrix A is $m \times m$ and of full column rank and the mutual information is used for estimation of the unmixing matrix W.

In this paper we derive the conditions for the uniqueness of the linear representation (1) by imposing a set of minimal moment constraints on the distributions of the independent sources. We then use a newly proposed semi-parametric density estimation method based on a suitable class of mixture densities that allows to conserve the moment restrictions needed for identifiability and establish consistency of our proposed estimator of W. Next, we present an iterative method for computing the proposed estimate of the unmixing matrix W and the source densities $f_1, \ldots, f_m$ simultaneously. Finally, we present empirical analysis based on simulated data and compare the performance of our method to three existing competitive methods for which software are available and an illustrative example based on a real dataset.

2 Parameter Identifiability of the ICA

Suppose that a vector of observed values $X = (x_1, \ldots, x_m)^T$ is known to be a mixture of some underlying independent sources $S = (s_1, \ldots, s_m)^T$ as given in (1). The problem is the estimation of the matrix A and the densities of the underlying sources $s_1, \ldots, s_m$. The statistical estimation of the mixing matrix A (or its inverse W) and the source densities $f_1, \ldots, f_m$ remains an ill-posed problem until the 'true parameters,' the mixing matrix and the source densities are uniquely defined in the statistical model given by (1). In this Section we derive a set of sufficient conditions under which the ICA model has a solution and it is unique.

To begin with, we re-state a characterization result due to Kagan et al. (1973, p. 315) showing the existence of the solution and its uniqueness up to matrix equivalence.

Theorem 2.1. *(Kagan et al. (1973)). Suppose X can be expressed as in (2) where the matrix B is such that the columns corresponding to the nongaussian components of Y are linearly independent. Then X can be expressed as in (1)*

$$X = AS + E,$$

where E has a multivariate normal distribution, S is a vector of nongaussian independent components and it is independent of E and the columns of A are linearly independent. The decomposition is unique in that if X has another representation given by $X = BG + \widetilde{E}$ then E and $\widetilde{E}$ have the same multivariate normal distribution, S and G have the same distribution except for change of scale and location. The matrices A and B are equivalent.

Hence by the above theorem not only A is statistically identifiable but also the distributions of the independent sources S are uniquely identifiable. In addition, one set of sufficient conditions for existence and uniqueness of the solution for ICA model up to matrix equivalence is that the independent components are assumed to be nongaussian and that A is of full column rank. In other words if X has two representations given by $X = AS + E$ and $X = BG + \widetilde{E}$ then the vectors E and $\widetilde{E}$ have the same multivariate normal density, A and B are equivalent and S and G belong to the same location and scale family, i.e., if the joint density of the source vector S is given by $f_s(\cdot)$ then there exist a vector $C \in \mathbb{R}^m$ and a scalar $b > 0$ such that the joint density of G can be expressed as $f_g(Y) = f_s\{(Y - C)/b\}$. Hence, there exist matrices P, D and Λ, such that $B = APD\Lambda$ and $G = \Lambda^{-1}D^{-1}P^TS$, where P is a permutation matrix, Λ is a diagonal matrix with positive diagonal elements and D is a diagonal matrix with diagonal values ± 1. Based on the above result and taking into account the three sources of nonidentifiability we obtain the following result to resolve the identifiability of the ICA model.

Theorem 2.2. *Suppose the mixing matrix A in noisy ICA model (1) is of full column rank and the independent sources are all nongaussian. Further, suppose that all of the third moments of*

the sources exist and they satisfy the following conditions,

$$E(s_j) = 0, \;\; E(s_j^2) = 1, \; \textit{for } j = 1, \ldots, m \textit{ and} \\ 0 < E(s_1^3) < E(s_2^3) < \ldots < E(s_m^3) \tag{3}$$

then the ICA model is fully identifiable, in the sense if X has two representations given by $X = AS + E = BG + \widetilde{E}$, where A and B are each of full column rank and both S and G satisfy the conditions in (3), then $A = B$ and $S \stackrel{d}{=} G$.

The details of the proof are given in Appendix A. Notice that the conditions $E(s_j^3) > 0$ for $j = 1, \ldots, m$ can be replaced by $E(s_j^3) < 0$ for all $j = 1, \ldots, m$ and the third moments are ordered. Here we chose to distinguish the source densities according to their skewness measure. However, other statistical measures of the shape of the source densities can be used for indentifying the densities of the original sources.

In some applications, we often have subject matter knowledge that the original sources are positive valued random variables. Since the third moment of a positive valued random variable is necessarily positive, the following result can be obtained immediately from Theorem 2.2.

Corollary 2.3. *Suppose $s_1, \ldots, s_m$ in the ICA model (1) are positive valued random variables with $E(s_j) = 1$, for $j = 1, \ldots, m$. Then the model is fully identifiable if* $\mathrm{var}(s_1) < \ldots < \mathrm{var}(s_m)$.

Next we show that the sufficient conditions stated in Theorem 2.2 are minimal if we assume that skewness of the source variables are distinct.

Theorem 2.4. *The conditions (3) given in Theorem 2.2 are minimal if the sources $s_1, ..., s_m$ in the ICA model (1) are assumed to have third order moments and further assuming that the skewness measures of the densities are distinct.*

The proof of the theorem is given in Appendix B. The above result also facilitates the implementation of an algorithm where the independent sources with mean zero can be transformed to satisfy the conditions (3). Section 4 provides more details on source density estimation.

3 A Semiparametric ICA Model

By Theorem 2.2 a set of sufficient conditions requiring the existence of the third moments of densities $f_1, \ldots, f_m$ in ICA model makes the mixing matrix A identifiable. Eloyan and Ghosh (2011) developed a flexible class of models based on a mixture of densities (for instance gaussian kernels) to estimate a univariate density subject to moment constraints. We extend their method to the multivariate case for estimating the densities of the independent components of S.

Following the work of Eloyan and Ghosh (2011), we propose to estimate each of the source densities f_j by the following mixture of densities

$$f_j(s) = \sum_{k=1}^{N_j} \theta_{jk} \phi \left(\frac{s - \mu_{jk}}{\sigma_{N_j}} \right) \frac{1}{\sigma_{N_j}}, \tag{4}$$

where $\mu_{j1} < \mu_{j2} < \ldots < \mu_{jN_j}$ is a suitable sequence of known numbers (knots) and $\sigma_{N_j} > 0$ is chosen as a function of μ_{jk}'s and N_j, $\phi(\cdot)$ is a kernel density function satisfying a set of regularity conditions. Given the μ_{jk}, σ_{N_j} and N_j, the weights θ_{jk} are estimated subject to a set of restrictions implying that $f_j, j = 1, \ldots, m$ satisfy a set of sufficient conditions for identifiability (e.g., as in (3) in Theorem 2.2). In particular, in order to satisfy the set of three conditions given in (3), we estimate the θ_{jk}'s subject to the following necessary conditions:

$$\begin{aligned} &(i) \quad \sum_{k=1}^{N_j} \theta_{jk} \mu_{jk} = 0, \ \sum_{k=1}^{N_j} \theta_{jk} \mu_{jk}^2 = 1 - \sigma_{N_j}^2 \text{ and } \sum_{k=1}^{N_j} \theta_{jk} \mu_{jk}^3 > 0 \\ &(ii) \quad \sum_{k=1}^{N_j} \theta_{jk} = 1 \text{ and } \theta_{jk} \geq 0. \end{aligned} \tag{5}$$

Clearly the first set of restrictions (i) correspond to the three conditions in (3) and the last condition (ii) in (5) is needed to assure that the mixture density in (4) is a legitimate probability density function. A constrained EM algorithm (see Eloyan and Ghosh (2011)) can be used to estimate not only the θ_{jk}'s subject to restrictions given in (5), but also N_j's. Next, we describe a method to simultaneously estimate the matrix W and the weights θ_{jk} for a given sequence of the knots μ_{jk} and σ_{N_j}. Notice that the number of components N_j, $j = 1, \ldots, m$ are not fixed but rather estimated making our method fully automatic not requiring any tuning parameter selection. The densities can then be reordered according to their third order moments as in (3).

In matrix notation the noise free ICA model (see Comon (1994), Chen and Bickel (2006), Hyvarinen et al. (2001), etc.) is given by

$$X = SA,$$

where X is the $T \times m$ matrix of observed values (signals), S is the $T \times m$ matrix of underlying (hidden) sources and A is the $m \times m$ unknown mixing matrix. In addition, suppose that the mixing matrix is nonsingular and define its inverse as $W = A^{-1}$. As discussed in Section 1, the densities of the independent components are given by $s_{ij} \sim f_j$, for each $j = 1, \ldots, m$ and hence the column S_j is a sample of T independent and identically distributed variates from the density f_j.

Suppose $\widehat{W}^{(0)}$ is an initial estimate for W found by some preliminary but fast estimation method (e.g. singular value decomposition). Next, for each hidden component s_j, let $N_j^{(0)}$ be the initial number of components in the mixture used for estimating the density of the source variable s_j, $\{\mu_{j1}^{(0)}, \ldots, \mu_{jN_j^{(0)}}^{(0)}\}$ be a starting set of known means and $\sigma_{N_j}^{(0)}$ a known common variance for the components of the mixture density.

Our goal is to estimate the true unknown unmixing matrix W_0 and the densities of the sources f_j simultaneously using an iterative method. For the iteration step $M \in \{1, 2, \ldots\}$, we obtain a pseudo-sample of the matrix of independent sources as $\widehat{S}^{(M)} = X\widehat{W}^{(M-1)}$. For each $j = 1, \ldots, m$ to estimate the density of the hidden source s_j using the pseudo-sample $\widehat{s}_{1j}^{(M)}, \ldots, \widehat{s}_{Tj}^{(M)}$ let $N_j^{(M)} = N_j^{(M-1)} + 1$ and suppose the set of means $\{\mu_{j1}^{(M)}, \ldots, \mu_{jN_j^{(M)}}^{(M)}\}$ and variance $\sigma_{N_j}^{(M)}$ are chosen so that the sieve structure of the sets of means is conserved. In other words if we define $\mathcal{M}_{N_j^{(M)}} = \{\mu_{j1}^{(M)}, \ldots, \mu_{jN_j^{(M)}}^{(M)}\}$, then $\mathcal{M}_{N_j^{(M-1)}} \subset \mathcal{M}_{N_j^{(M)}}$. Notice that by this construction the standard deviations σ_{N_j} are fixed, however the choice of the bandwidth is controlled by estimation of N. Further details for constructing the means that conserve the sieve structure can be found in Eloyan and Ghosh (2011). Next, the constrained EM-algorithm as described in Eloyan and Ghosh (2011) can be used to compute the weights of the mixture density $(\widehat{\theta}_{j1}^{(M)}, \ldots, \widehat{\theta}_{jN_j^{(M)}}^{(M)})$ that minimize the Kullback-Leibler discrepancy (KLD) between the

true and the estimated densities of s_j defined as

$$KLD(\widehat{f}, f) = \int f(s) \log \frac{f(s)}{\widehat{f}(s)} ds.$$

Hence for each $j = 1, \ldots, m$ the density estimate is constructed as follows

$$\widehat{f}_j^{(M)}(s) = \sum_{k=1}^{N_j^{(M)}} \widehat{\theta}_{jk}^{(M)} \phi \left(\frac{s - \mu_{jk}^{(M)}}{\sigma_{N_j}^{(M)}} \right) \frac{1}{\sigma_{N_j}^{(M)}}, \tag{6}$$

where $\phi(\cdot)$ is the density of a gaussian random variable with mean zero and variance one. Other popular Kernels with mean zero and variance unity can also be used in (6). By construction the densities of the independent sources are nongaussian.

The likelihood function of the unmixing matrix $W = ((w_{lj}))$ is given by

$$l(W, F) = \prod_{i=1}^{T} \prod_{j=1}^{m} f_j \left(\sum_{l=1}^{m} x_{il} w_{lj} \right) |\det(W)|^T,$$

where $F = (f_1, \ldots, f_m)$ is the vector of densities of hidden sources. By using the estimates given in (6) and writing $\widehat{F} = (\widehat{f}_1, \ldots, \widehat{f}_m)^T$ the loglikelihood function of the unmixing matrix is given by

$$L(W, \widehat{F}) = \sum_{i=1}^{T} \sum_{j=1}^{m} \log \left\{ \sum_{k=1}^{N_j^{(M)}} \widehat{\theta}_{jk}^{(M)} \phi \left(\frac{\sum_{l=1}^{m} x_{il} w_{lj} - \mu_{jk}^{(M)}}{\sigma_{N_j}^{(M)}} \right) \frac{1}{\sigma_{N_j}^{(M)}} \right\} + T \log |\det W|. \tag{7}$$

Notice that by the choice of the estimating densities of original sources the gradient vector $\nabla L(W, F)$ and hessian matrix $\nabla^2 L(W, \widehat{F})$ of the loglikelihood above can be computed analytically and are given in Appendix D. Hence by using a hill-climbing version of the Newton-Raphson algorithm (see Section 4 for more computational details) an update of the unmixing matrix can be computed as follows

$$\widehat{W}^{(M+1)} = \widehat{W}^{(M)} - \nabla^2 L(\widehat{W}^{(M)}, \widehat{F})^{-1} \nabla L(\widehat{W}^{(M)}, \widehat{F}).$$

Let $\mathcal{F}_N = \{f : f(x) = \sum_{k=1}^{N} \theta_k \phi[(x - \mu_k)/\sigma]/\sigma, x \in \mathbb{R}\}$ be a set of finite mixture densities with N mixture components. Let $\mathcal{F}$ denote a class of densities satisfying the following regularity

conditions, for any $f \in \mathcal{F}$,

(*i*) $0 \leq f(x) \leq L$ for some $L > 0$ and for $x \in \text{supp}(f) = \{x \in \mathbb{R} : f(x) > 0\}$.

(*ii*) $|\int_S f(x) \log f(x) dx| < \infty$ and $|\int_S f(x)[-\log \phi\{(x-\mu)/\sigma\}]dx| < \infty$ for any $\mu \in \text{supp}(f)$ and $\sigma > 0$.

We assume that the true source densities f_j satisfy the identifiability conditions (3) given in Theorem 2.2. The following result provides a set of regularity conditions under which we establish the consistency of the estimator of W obtained by maximizing the log-likelihood function given by (7) and simultaneously estimating the densities $\widehat{f}_1, \ldots, \widehat{f}_m$.

Theorem 3.1. *Suppose in the ICA model (1) the following conditions hold.*

1. *The densities of the hidden sources $f_j \in \mathcal{F}$, for $j = 1, \ldots, m$ and are nongaussian.*

2. *There exists a sequence of known quantities $\mathcal{M}_{N_j} = \{\mu_{j,1} < \ldots < \mu_{j,N_j} : \mu_{j,k} \in supp(f_j),\ k = 1, \ldots, N_j, j = 1, \ldots, m\}$, such that $\mathcal{M}_{N_j} \subset \mathcal{M}_{N_j+1}$ and $\max_{1 \leq k < N_j} (\mu_{j,k+1} - \mu_{j,k}) = o(1)$ as $N_j \to \infty$.*

3. *There exists a sequence of known quantities σ_{N_j} satisfying $\sigma_{N_j} = o(1)$ as $N_j \to \infty$.*

4. *The estimated densities $\widehat{f}_j \in \mathcal{F}_N$ and satisfy the constraints (3).*

5. *The true mixing matrix A is nonsingular.*

If $KLD(f_j, \widehat{f}_j) \to 0$ as $N_j \to \infty$ and $\widehat{W} = \text{argmax} L(W, \widehat{f})$ then

$$\widehat{W} \to W_0 \text{ almost surely, as } T \to \infty,$$

where W_0 is the true value of the inverse of the mixing matrix, i. e., $W_0 = A_0^{-1}$.

An outline of the proof of the theorem is presented in Appendix C.

4 An Iterative Method to Compute the MLE of W

We first describe a method to find a quick and good starting value for W. The $m \times m$ covariance matrix of X defined as $C_x = \text{cov}(X) = \sum_{i=1}^{T}(X_i - \bar{X})(X_i - \bar{X})^T/(T-1)$ can be factorized as

$$C_x = A^T C_s A,$$

where $C_s = \text{cov}(S) = I$ by constraints (3). Hence $C_x = A^T A$. By spectral decomposition we obtain $C_x = Q\Lambda Q^T$, where Λ is the diagonal matrix of eigenvalues of C_x and Q is an orthogonal matrix of corresponding eigenvectors satisfying $Q^T Q = I$. This implies that a good choice for A can be obtained as $A = Q\Lambda^{1/2}Q^T$. We choose a starting value for W as

$$\widehat{W}^{(0)} = A^{-1}.$$

By using the above starting value of W given by $\widehat{W}^{(0)}$, let $S^{(0)} = XW^{(0)}$, $N_j^{(0)} = 1 + 2[\max(S_j^{(0)}) - \min(S_j^{(0)})]/3$, $\mu_{j1}^{(0)} = \min(S_j^{(0)})$ and $\mu_{jN_j^{(0)}}^{(0)} = \max(S_j^{(0)})$. Finally let $\sigma_{N_j}^{(0)2} = 2(\mu_{j1}^{(0)} - \mu_{jN_j^{(0)}}^{(0)})/[3(N_j^{(0)} - 1)]$ and construct $\mathcal{M}_{N_j^{(0)}} = \{\mu_{j1}^{(0)} < \ldots < \mu_{jN_j^{(0)}}^{(0)}\}$ (see Eilers and Marx (1996) and Komarek et al. (2005)). An iterative algorithm for finding the estimate of the unmixing matrix W is given as follows.

For iteration step $M \in \{1, 2, \ldots\}$,

1. Let $S^{(M)} = XW^{(M)}$.
2. For each $j = 1, \ldots, m$ set $N_j^{(M)} = N_j^{(M-1)} + 1$ and construct the set of means that satisfy $\mathcal{M}_{N_j^{(M)}} \supseteq \mathcal{M}_{N_j^{(M-1)}}$.
3. By using constrained EM algorithm obtain the estimate $(\widehat{\theta}_{j1}^{(M)}, \ldots, \widehat{\theta}_{jN_j^{(M)}}^{(M)})$. Notice that the EM algorithm described by Eloyan and Ghosh (2011) estimates the density subject to constraints on the mean(=0) and variance(=1) of the random variable.
4. Sort the densities $f_1, \ldots, f_m$ in ascending order of their third order moments.
5. Compute the gradient $\nabla L(\widehat{W}^{(M)}, \widehat{F})$ and hessian matrix $\nabla^2 L(\widehat{W}^{(M)}, \widehat{F})$ (see Appendix F for exact analytical expressions).

6. Update the unmixing matrix by setting

$$\widehat{W}^{(M+1)} = \widehat{W}^{(M)} - \nabla^2 L(\widehat{W}^{(M)}, \widehat{F})^{-1} \nabla L(\widehat{W}^{(M)}, \widehat{F}).$$

7. If $L(\widehat{W}^{(M+1)}, \widehat{F}) < L(\widehat{W}^{(M)}, \widehat{F})$, set $\widehat{W}^{(M+1)} = \widehat{W}^{(M)}$ and repeat steps 5-6. In other words increase the number of mixture components and implement steps 5-6 until a new value of $\widehat{W}$ is obtained. Otherwise return to step 1.

Repeat the steps 1-7 above until convergence. In our numerical illustrations, we have used the stopping rule as $\max_{1 \le l,j \le m} |\widehat{w}_{lj}^{(M+1)} - \widehat{w}_{lj}^{(M)}| < \epsilon$ with $\epsilon = 10^{-3}$.

5 Simulation Study

We illustrate the proposed estimation method by evaluating the performance of the algorithm under different scenarios. First we set the number of underlying sources to $m = 2$ and the number of observations $T = 1000$. The true value of the mixing matrix is set as

$$A = \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & -0.5 \end{pmatrix}.$$

We generate the hidden source variables using three sets of nongaussian distributions:

Case I: $m = 2$

(a) Shifted and scaled Gamma densities

$f_1(s) = \frac{4^4}{2\Gamma(4)} \left(\frac{1}{2}s + 1\right)^3 e^{-4(s/2+1)}$,

$f_2(s) = \frac{1}{4} \left(\sqrt{2}s + 4\right) e^{-(\sqrt{2}s+4)/2}$.

(b) Shifted and scaled Weibull densities

$f_1(s) = 3 \left\{ s\sqrt{\Gamma(5/3) - \Gamma(4/3)^2} + \Gamma(4/3) \right\}^2 e^{-\{s\sqrt{\Gamma(5/3)-\Gamma(4/3)^2}+\Gamma(4/3)\}^3}$,

$f_2(s) = e^{-(s+1)^2}$.

(c) Shifted and scaled Gamma and a mixture of normals

$f_1(s) = 3\phi(5x + 4) + 2\phi(5x - 6)$,

$f_2(s) = \frac{4^4}{2\Gamma(4)} \left(\frac{1}{2}s + 1\right)^3 e^{-4(s/2+1)}$,

where $\phi(\cdot)$ denotes the density function of a normal random variable with mean 0 and variance 1. Notice that the above densities satisfy the conditions (3) required for the identifiability of ICA, in other words the means are equal to 0, the variances are equal to 1 and the skewness of all of the above densities are positive and increasing for each case. The computations are performed using the R software. For comparison, the following three algorithms were also used: (i) `fastICA` (FICA) proposed by Hyvarinen and Oja (2000), (ii) `PearsonICA` (PICA) proposed by Karvanen and Koivunen (2002) and (iii) `JADE` proposed by Cardoso and Souloumiac (1993) (the corresponding R packages are available online at `http://cran.r-project.org/`). In the rest of this Chapter we will use Mixture ICA (MICA) to denote our proposed method.

The performance of the method is evaluated by a commonly used error criterion in signal processing literature called the Amari error (Amari, 1998). For a given known $m \times m$ mixing matrix A and estimated unmixing matrix $\widehat{W}$ the Amari error is defined as

$$AE(A, \widehat{W}) = \frac{1}{2m} \sum_{i=1}^{m} \left(\sum_{j=1}^{m} \frac{|p_{ij}|}{max_k |p_{ik}|} - 1 \right) + \frac{1}{2m} \sum_{j=1}^{m} \left(\sum_{i=1}^{m} \frac{|p_{ij}|}{max_k |p_{kj}|} - 1 \right),$$

where $P = A\widehat{W}$. Notice that for any two matrices A and $\widehat{W}$, the Amari error satisfies the inequality $0 \leq AE \leq m - 1$ and is equal to zero if and only if the true mixing matrix A and the estimated mixing matrix $\widehat{A} = \widehat{W}^{-1}$ are equivalent in the sense as defined in Section 2. However, the Amari error is not invariant to a constant multiplier, in other words $AE(A, \widehat{W}) \neq AE(A, \widehat{W}\Lambda)$, where Λ is a diagonal matrix with positive elements on the diagonal. Hence before computing this error we rescale the columns of the matrices A and $\widehat{W}$ to have Euclidean norm unity so that the estimates obtained by FICA, PICA and JADE are comparable to our proposed method MICA. We compute the logarithms of the efficiencies for each estimation method with respect to our proposed MICA method as

$$leff(\widehat{W}_1, \widehat{W}_2) = \log \frac{AE(A, \widehat{W}_1)}{AE(A, \widehat{W}_2)},$$

where $\widehat{W}_2$ is the estimate obtained by our proposed MICA method and $\widehat{W}_1$ is the estimated unmixing matrix using one of the three methods FICA, JADE or PICA. Clearly, if $leff(\widehat{W}_1, \widehat{W}_2) >$

0 then MICA performs better than its competitor. The larger the value of $leff(\widehat{W}_1, \widehat{W}_2)$, the more efficient MICA is compared to its competitor. For each of the simulation scenarios we compute the log-efficiency based on several simulated data sets.

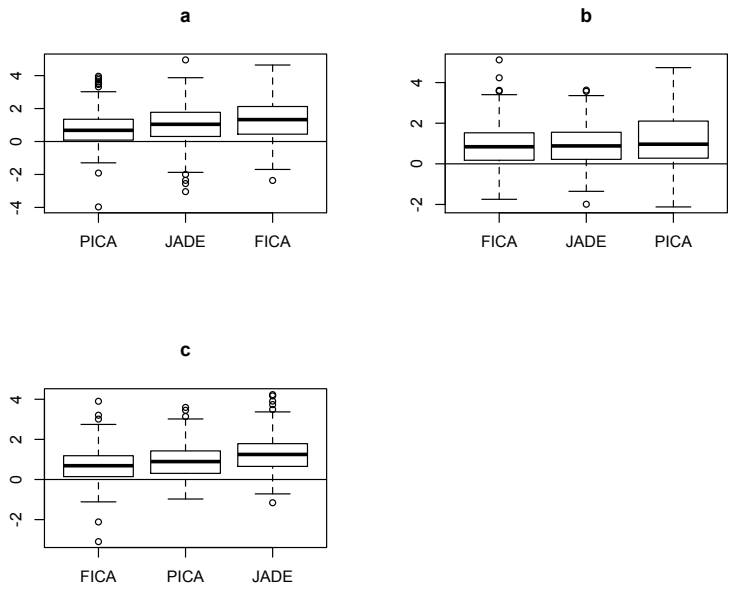

Figure 1: Boxplots of the log-efficiencies of our proposed MICA as compared to the three commonly used methods FICA, JADE and PICA for Case I: $m = 2$.

Figure 1 shows the boxplots of the log-efficiencies of Amari errors for the three different estimates compared with our proposed MICA estimate for 200 simulated datasets for each of the three scenarios (a)-(c) under Case I: $m = 2$ sources. It can be noted that our method performs significantly better than all three competing methods. In particular, the 25th percentiles of the log-efficiencies are above zero against all three methods indicating a superior performance of MICA in at least 75% of the test cases (out of 200 runs). The median Amari error value for our proposed MICA method in case (c) is $0 \cdot 009$ which is smaller than that of FICA with median $AE = 0 \cdot 021$, PICA with median $AE = 0 \cdot 027$ and JADE with median $AE = 0 \cdot 034$. Thus, it appears that our proposed method performs substantially better than all three methods when one of the sources has multimodal or skewed distribution.

Next we consider the case when there are $m = 3$ hidden sources again generated from various nongaussian distributions:

Case II: $m = 3$

(a) Shifted and scaled Gamma and Weibull densities

$$f_1(s) = \tfrac{4^4}{2\Gamma(4)} \left(\tfrac{1}{2}s + 1\right)^3 e^{-4(s/2+1)},$$

$$f_2(s) = \tfrac{1}{4} \left(\sqrt{2}s + 4\right) e^{-(\sqrt{2}s+4)/2},$$

$$f_3(s) = 3 \left\{ s\sqrt{\Gamma(5/3) - \Gamma(4/3)^2} + \Gamma(4/3) \right\}^2 e^{-\{s\sqrt{\Gamma(5/3)-\Gamma(4/3)^2}+\Gamma(4/3)\}^3}.$$

(b) Shifted and scaled Weibull densities and a mixture of normals

$$f_1(s) = 3 \left\{ s\sqrt{\Gamma(5/3) - \Gamma(4/3)^2} + \Gamma(4/3) \right\}^2 e^{-\{s\sqrt{\Gamma(5/3)-\Gamma(4/3)^2}+\Gamma(4/3)\}^3},$$

$$f_2(s) = 3\phi(5x + 4) + 2\phi(5x - 6) \ ,$$

$$f_3(s) = e^{-(s+1)^2}.$$

(c) Shifted and scaled Weibull and mixtures of gaussian densities

$$f_1(s) = 3 \left\{ s\sqrt{\Gamma(5/3) - \Gamma(4/3)^2} + \Gamma(4/3) \right\}^2 e^{-\{s\sqrt{\Gamma(5/3)-\Gamma(4/3)^2}+\Gamma(4/3)\}^3},$$

$$f_2(s) = 3\phi(5x + 4) + 2\phi(5x - 6) \ ,$$

$$f_3(s) = \sqrt{7 \cdot 5}(0 \cdot 8\phi\{2(\sqrt{7 \cdot 5}x + 2 \cdot 5)\} + 0 \cdot 8\phi(2\sqrt{7 \cdot 5}x) + 0 \cdot 4\phi\{2(\sqrt{7 \cdot 5}x - 5)\}).$$

The mixing matrix used in this case is as follows

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \sqrt{3}/2 & 0 \cdot 5 \\ 0 \cdot 1 & -0 \cdot 5 & \sqrt{3}/2 \end{pmatrix}$$

Here again the chosen densities satisfy the identifiability conditions as stated in Theorem 2.2. The resulting boxplots of the log-efficiencies based on 200 simulated data sets for each of the scenarios (a)-(c) under case II: $m = 3$ are presented in Figure 2. For case (a) the efficiency of MICA is similar to that of PICA, but performs significantly better than FICA and JADE. However, in cases (b) and (c) the proposed MICA method substantially outperforms the others. It can be noted that the 25th percentiles for cases (b) and (c) of the log-efficiencies of the three methods compared with our proposed MICA estimate are above zero in Figure 2 showing that our method outperforms the others in terms of minimizing the Amari error criterion. Tables 1

and 2 present some selected summary values of the Amari errors of the estimates of A found by four different estimation methods corresponding to cases (b) and (c) with $m = 3$.

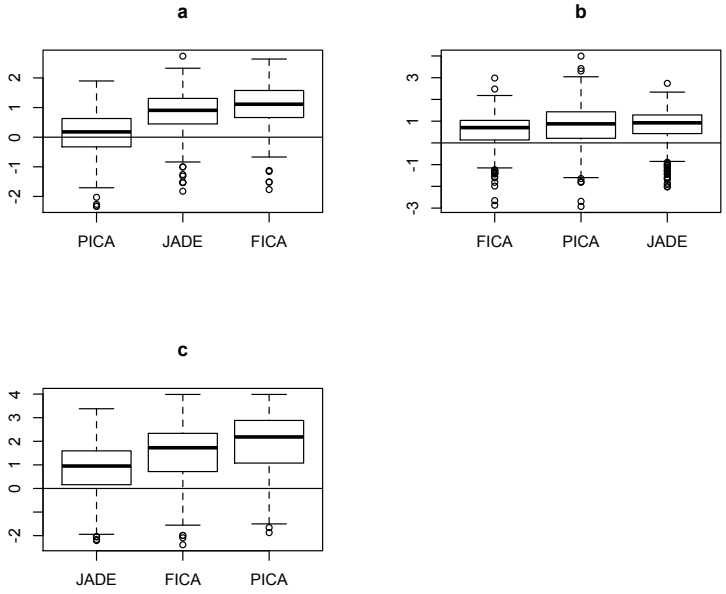

Figure 2: Boxplots of the log-efficiencies of our proposed MICA as compared to the three commonly used methods FICA, JADE and PICA for Case II: $m = 3$.

As another performance measure we also computed the mean ranks of the Amari errors for each method over 200 simulated cases. In other words, for each simulated data we compute the AE corresponding to each of the four methods and rank them as $1, 2, 3, 4$ by the increasing order of their AEs. E.g., if $AE(MICA) < AE(FICA) < AE(JADE) < AE(PICA)$ then $rank(MICA) = 1$ while $rank(PICA) = 4$, in case of a tie we use equal ranks. The mean ranks for the cases (b) and (c) with $m = 3$ are shown in column seven of Tables 1 and 2 respectively with standard deviations of the ranks in the parenthesis. Since the mean rank of the AE for MICA is close to 1 for both cases (b) and (c) we can claim that our method results in a lower value of the Amari error in majority of the cases (out of 200 MC runs) when compared with the other three methods. The boxplots of the Amari errors for each case in Figures 1 and 2 are

Table 1: *Summary of Amari errors for the four methods for Case II: $m = 3$, (b) where two sources are generated by using shifted and scaled weibull densities and the third is generated by using a mixture distribution. The last column presents the results for the mean ranks of Amari errors for 200 simulation runs.*

	10%	25 %	50 %	75 %	90 %	mean rank (SD)
MICA	0·020	0·028	0·041	0·052	0·259	1·5 (0·9)
FICA	0·037	0·058	0·080	0·106	0·137	2·52 (0·7)
JADE	0·042	0·058	0·094	0·169	0·366	3·055 (0·9)
PICA	0·051	0·069	0·104	0·134	0·183	2·86 (0·9)

ordered by the rank of the corresponding method. The mean rank of our proposed MICA was found to be close to 1 in all simulation cases above for $m = 2$ and $m = 3$.

6 Application to Iris Data

We use the historical Iris dataset to illustrate the performance of our method compared with other commonly used ICA algorithms. Three types of iris plants are studied in the experiment. The sepal length and width and petal length and width (hence $n = 4$) are measured for 50 iris plants in each plant category. A total of $T = 150$ observations are collected with no missing values. The dataset has been widely used in the literature for testing different clustering algorithms or classification methods. However, here we illustrate the use of ICA in visualizing the clusters within the data. We first apply PCA to these data to find the principal directions of variability within the data. The $99 \cdot 5\%$ of variability in the data is explained by the first three principal components. This implies that the data can be reduced to a $T \times m$ dimensional matrix X constructed using the first m principal components, where $m = 3$. We choose the values for the fixed standard deviation as $v_j = \sqrt{j}$. By applying the proposed algorithm we find

Table 2: *Summary of Amari errors for the four methods for Case II: $m = 3$, (c) where the sources are generated using a shifted and scaled weibull distribution and mixture distributions. The last column presents the results for the mean ranks of Amari errors for 200 simulation runs.*

	10 %	25 %	50 %	75 %	90 %	mean rank (SD)
MICA	0·020	0·028	0·041	0·052	0·250	1·4 (0·7)
FICA	0·037	0·058	0·080	0·106	0·137	2·9 (0·6)
JADE	0·042	0·058	0·094	0·169	0·366	2·0 (0·4)
PICA	0·051	0·069	0·204	0·134	0·183	3·6 (0·5)

an estimate of the mixing matrix given by $\widehat{W}$ and estimates of the densities of the independent hidden sources. The computations are done using the R software. By using the resulting unmixing matrix we compute $\widehat{S} = X\widehat{W}$. The scatterplot matrices of the estimated hidden sources are plotted in Figure 3. The plots also include the Kendall's τ coefficient for each pair of the sources. We observe that our method results in the smallest values of the Kendall's τ coefficients compared with the other methods, which implies that the estimates computed by our method are better separated as independent sources. The scatterplots also show that one of the clusters is clearly separated, however, the other two types do not seem to be so well separated. Next, the estimated densities of the hidden sources using MICA are plotted in Figure 4 (first row). The densities are clearly nongaussian and in fact the first estimated density has three modes. The first independent component seems to show the clustering structure within the data. By looking at the densities computed by using the estimated source vectors obtained by FICA and applying KDE in Figure 4 (second row) we observe that the clustering structure is not captured so well compared with MICA.

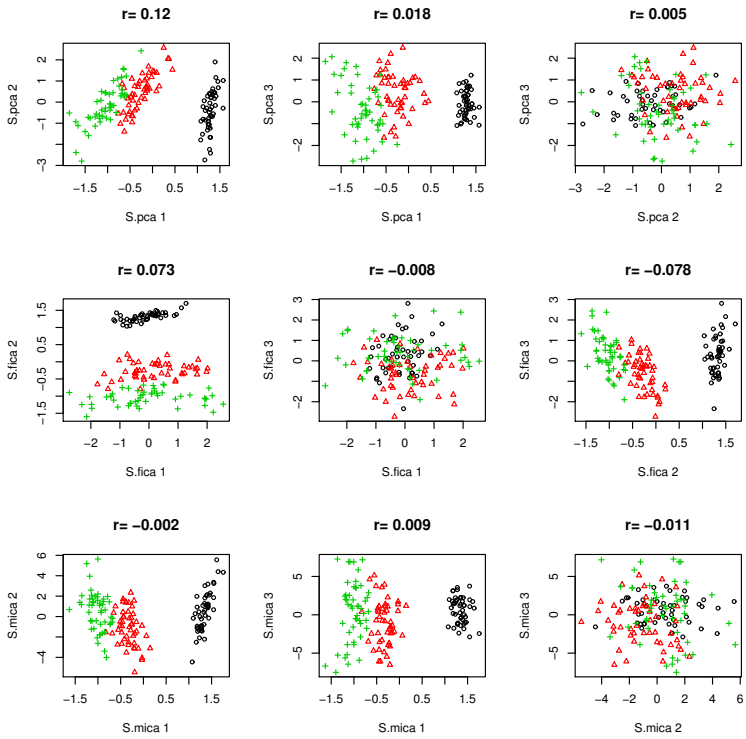

Figure 3: Scatterplots of the estimated hidden sources using PCA (first row), FICA (second row) and MICA (third row). The scatterplots are colored according to the iris type. The titles of the matrices show the Kendall's τ coefficients for the corresponding two source vectors. The coloring and point types in the scatterplots show the three different iris plants.

7 Conclusion and Discussions

In this paper, a new semiparametric approach for Independent Component Analysis (ICA) for source separation is proposed. We first discussed the identifiability issues related to the ICA and introduced a new semiparametric method to estimate the so-called mixing matrix within ICA. Even though ICA is gaining more popularity in different fields of statistical research there is still some ambiguity in the identifiability of the model used. We derived some sufficient conditions for the densities of the hidden sources which guarantee that the ICA model is fully identifiable. Based on these sufficient conditions we proposed a semi-parametric likelihood based method for the estimation of the unmixing matrix while making no parametric assumptions about the

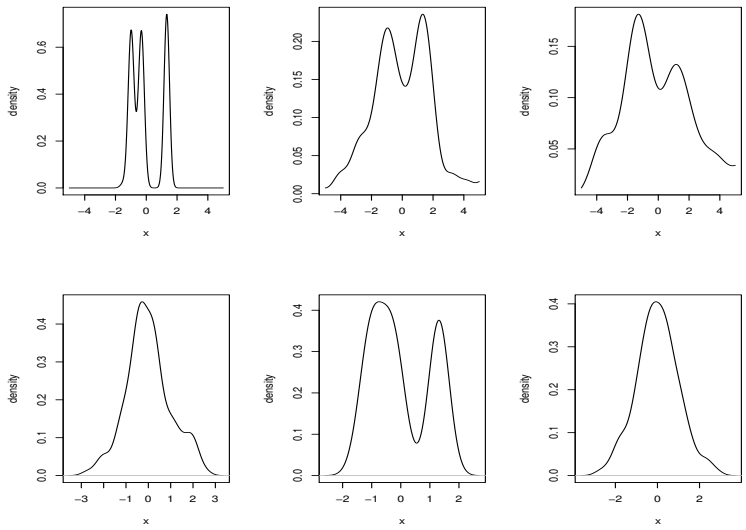

Figure 4: Estimated densities of the hidden sources using MICA (top row) and FICA (bottom row).

independent hidden source distributions.

Mixtures of gaussian densities were used for modeling the underlying densities of the hidden sources. The simulation studies showed that our method performs similar to the existing methods for some cases when the underlying densities of the hidden sources are unimodal. Our method outperforms some of the competitors when the underlying densities are possibly multimodal and/or skewed. Different kernel densities can be used for the mixture densities to obtain a better estimate of the densities of underlying sources and such possibilities will be explored in the future. Finally, the problem of estimating the minimum number of independent sources remains unresolved. Throughout our paper we have assumed $n = m$ for simplicity, however in practice m could be significantly smaller than n. The estimation of m appears to be a non-trivial problem as in that case A is no longer a square matrix and definition of Amari error and unmixing matrix may need to be modified suitably possibly using some version of g-inverse. In practice, often PCA or other dimension reduction methods are first used to "estimate" m and then ICA is used on the extracted PCs. Admittedly, such a two-step approach is sub-optimal

and hence simultaneous estimation of m, A and the densities of the ICs $(f_1, \ldots, f_m)$ would be of utmost interest.

References

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

R. Boscolo, H. Pan, and V.P. Roychowdhury. Beyond comon's identifiability theorem for independent component analysis. *ICANN 2002, LNCS 2415*, pages 1119–1124, 2002.

R. Boscolo, H. Pan, and V.P. Roychowdhury. Independent component analysis based on nonparametric density estimation. *IEEE Transactions on Neural Networks*, 15(1):55–65, 2004.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEE-Proceedings-F*, 140(6):362–370, 1993.

A. Chen and P.J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, pages 2825–2855, 2006.

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

P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121, 1996.

A. Eloyan and S.K. Ghosh. Smooth density estimation with moment constraints using mixture densities. *Journal of Nonparametric Statistics*, page (in press), 2011.

P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependency in risk management: properties and pitfalls. *In Risk Management: Value at Risk and Beyond, Eds. M.-Dempster and H.-Moffatt*, pages 176–223, 2001.

D. Ferger. A continuous mapping theorem for the argmax-functional in the non-unique case. *Statistica Neerlandica*, 58(1):83–96, 2004.

A. Hyvarinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Neworks*, 15, 2000.

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

C. Jutten and J. Herault. Blind seperation of sources, part i: An adaptive algorithm based on neuromimethic arcitecture. *Signal Processing*, 24:1–10, 1991.

A.M. Kagan, Y.V. Linnik, and C.R. Rao. *Characterisation Problems in Mathematical Statistics*. New York: Wiley, 1973.

J. Karvanen and V. Koivunen. Blind separation methods based on pearson system and its extensions. *Signal Processing*, 82(4):663–673, 2002.

A. Komarek, E. Lessafre, and J. F. Hilton. Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14:726–745, 2005.

W.B. Silverman. *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall, 1985.

Appendix A: Proof of Theorem 2.2 in Section 2

Proof. Suppose X has two representations $X = AS+E$ and $X = BG+\widetilde{E}$ where S and G satisfy (3). Then it follows from Theorem 2.1 that $G = \Lambda D P^T S$ with

$$\text{cov}(S) = C_s = \text{diag}(1, \ldots, 1). \tag{8}$$

where Λ, D and P are as defined in Section 2. On the other hand

$$\text{cov}(G) = C_g = \text{cov}(\Lambda D P^T S) = \Lambda D P^T C_s P D \Lambda.$$

Suppose the columns of the permutation matrix P are defined as follows $P = (p_1 \ \ldots \ p_m)$. Since P is a permutation matrix, its columns should be orthonormal and hence $p_i^T p_j = 0$ if $i \neq j$ and $p_j^T p_j = 1$. Thus it follows that,

$$P^T C_s P = \sum_{j=1}^{m} p_j p_j^T = \text{diag}(1, \ldots, 1) = C_s.$$

By its definition we can write $D = \text{diag}(d_1, \ldots, d_m)$, where $d_j^2 = 1$, for $j = 1, \ldots, m$ and so it follows that $D C_s D = \text{diag}(d_1^2, \ldots, d_m^2) = C_s$. Hence, the covariance of G can be obtained as

$$C_g = \Lambda D P^T C_s P D \Lambda = \Lambda C_s \Lambda.$$

Finally, by its definition, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$ with $\lambda_j > 0$ for $j = 1, \ldots, m$ and so

$$C_g = \Lambda C_s \Lambda = \text{diag}(\lambda_1^2, \ldots, \lambda_m^2). \tag{9}$$

Since S and G both satisfy (3) it follows from (8) and (9) that $\lambda_j^2 = 1$, for $j = 1, \ldots, m$, and hence $\Lambda = I$.

Consider again the matrix D with diagonal elements satisfying $d_j^2 = 1$ for $j = 1, \ldots, m$. Now suppose $d_j = -1$ for some $j \in 1, \ldots, m$, this implies $g_j = -s_j$ and hence it follows that

$$E(g_j^3) = E(-s_j^3) = -E(s_j^3) < 0$$

which contradicts the assumption that the third moments are positive (see (3)). Thus, it follows that $D = I$.

Now suppose the columns of the identity matrix are defined as $I_j, j = 1, \ldots, m$ with 1 at the jth entry and 0 elsewhere. Any permutation matrix can be obtained by permuting columns or rows of the identity matrix. Suppose P is a permutation matrix different from I and is constructed by permuting two columns i and j of I as follows

$$P = \begin{pmatrix} I_1 & \ldots & I_{i-1} & I_j & I_{i+1} & \ldots & I_{j-1} & I_i & I_{j+1} & \ldots & I_m \end{pmatrix}.$$

Since $G = P^T S = \begin{pmatrix} s_1 & \ldots & s_{i-1} & s_j & s_{i+1} & \ldots & s_{j-1} & s_i & s_{j+1} & \ldots & s_m \end{pmatrix}^T$, the third moment of the ith element of vector G satisfies

$$E(g_i^3) = E(s_j^3) > E(s_i^3) = E(g_j^3), \text{ where } i < j.$$

which contradicts the assumption that the third moments of the sources are increasing and hence $P = I$.

As a result if X has two representations given by $X = AS + E$ and $X = BG + \widetilde{E}$, then $S \stackrel{d}{=} G$, $\widetilde{E}$ and E have the same gaussian densities which in tern implies $A = B$ and therefore the model is fully identifiable. □

Appendix B: Proof of Theorem 2.4 in Section 2

Proof. Suppose $Z = (z_1, \ldots, z_m)^T$ is an $m \times 1$ vector, where $E(|z_j|^3) < \infty$ and $E(z_j) = 0$, for $j = 1, \ldots, m$. Now if we consider

$$\widetilde{S} = \left[\frac{sgn\{E(z_1^3)\} z_1}{\sqrt{\text{var}(z_1)}}, \ldots, \frac{sgn\{E(z_m^3)\} z_m}{\sqrt{\text{var}(z_m)}} \right]^T,$$

then we obtain $\text{var}(s_j) = 1$ and $E(s_j^3) > 0$. Define S as the vector constructed by reordering the $\widetilde{S}$ according to the values of the third moments of its elements. The vector S will satisfy the conditions in Theorem 2.2. □

Appendix C: Proof of Theorem 3.1 in Section 3

Proof. Suppose the sequence of the true densities of the hidden sources is defined as $F_0 = (f_1, \ldots, f_m)$. Since any continuous and bounded density function can be approximated by an infinite mixture of gaussian densities (see Eloyan and Ghosh (2011) for other regularity conditions and metric of convergence) then there exists a sequence of weights $\Theta_\infty = (\Theta_{1\infty}, \ldots, \Theta_{m\infty})$ such that for any $\epsilon > 0$

$$|L(W_0, \Theta_\infty) - L(W_0, F_0)| < \frac{\epsilon}{2}, \tag{10}$$

which follows from the fact that $L(W_0, F)$ is a continuous functional of F. Note that we use the notation $L(W, \Theta) = L(W, \widehat{f}_\Theta)$ as defined in (7). By the nested structure of the sets of means of estimated densities and by construction of $\widehat{W}^{(M)}$ we obtain

$$L(\widehat{W}^{(M)}, \widehat{\Theta}^{(M)}_{N^{(M)}}) \leq L(\widehat{W}^{(M)}, \widehat{\Theta}^{(M+1)}_{N^{(M+1)}}) \leq L(\widehat{W}^{(M+1)}, \widehat{\Theta}^{(M+1)}_{N^{(M+1)}}).$$

Hence the monotone sequence $L(\widehat{W}^{(M)}, \widehat{\Theta}^{(M)}_{N^{(M)}})$ has a limit as $M, N^{(M)} \to \infty$. Notice that any estimated weight vector at any stage of iteration (e.g. $\widehat{\Theta}^{(M)}_{N^{(M)}}$) belongs to $\Delta_N = \{\Theta_N \in [0,1]^N : \sum_{j=1}^N \theta_{jN} = 1\}$, where Δ_N is a compact set. Hence, for a compact set $\Omega \subset \mathbb{R}^{m \times m}$ by continuity of the function $L(\cdot)$ there exist $W_0 \in \Omega$ and $\Theta_\infty = (\Theta_{1\infty}, \ldots, \Theta_{m\infty})$ such that for any $\delta \in (0,1)$

$$|L(\widehat{W}^{(M)}, \widehat{\Theta}^{(M)}_{N^{(M)}}) - L(W_0, \Theta_\infty)| < \frac{\epsilon}{2} \text{ with probability } \geq 1 - \delta. \tag{11}$$

for sufficiently large T and M, where

$$L(W_0, \Theta_\infty) = \sum_{i=1}^{T} \sum_{k=1}^{m} \log \left\{ \sum_{j=1}^{\infty} \theta_{jk} \phi \left(\frac{\sum_{l=1}^m x_{il} w_{lj} - \mu_{jk}}{\sigma} \right) \frac{1}{\sigma} \right\} + T \log |\det W|.$$

Hence, by (10) and (11) we obtain for any $\delta \in (0,1)$,

$$|L(\widehat{W}^{(M)}, \widehat{\Theta}^{(M)}_{jN_j^{(M)}}) - L(W_0, F_0)| < \epsilon \text{ with probability } \geq 1 - \delta.$$

This implies that

$$\widehat{W} = \underset{W \in \mathbb{R}^{m \times m}}{\operatorname{argmax}} L(W, F_0) + o_p(1).$$

Finally, by using the argmax theorem stated in Ferger (2004) we obtain

$$\widehat{W} \to W_0 \text{ a.s. as } M, T \to \infty.$$

which completes the proof of the theorem. □

Appendix D: The gradient vector and Hessian matrix of loglikelihood function.

For $\alpha, \beta = 1, \ldots, m$ the first derivative of $L(W, \widehat{F})$ can be found as

$$\frac{\nabla L(W, \widehat{F})}{\nabla W_{\alpha\beta}} = \sum_{i=1}^{T} \frac{f'_\alpha(\sum_{j=1}^m x_{ij} w_{j\beta}) x_{i\alpha}}{f_\alpha(\sum_{j=1}^m x_{ij} w_{j\beta})} + T[W^{-1}]_{\beta\alpha}.$$

Now suppose $\alpha, \beta, \delta, \gamma = 1, \ldots, m$ and $\delta = \beta$ then

$$\begin{aligned} \frac{\nabla^2 L(W,\widehat{F})}{\nabla w_{\gamma\beta} \nabla w_{\alpha\beta}} = & \sum_{i=1}^{T} \frac{\left[f''_\alpha(\sum_{j=1}^m x_{ij} w_{j\beta}) f_\alpha(\sum_{j=1}^m x_{ij} w_{j\beta}) - \{f'_\alpha(\sum_{j=1}^m x_{ij} w_{j\beta})\}^2\right] x_{i\alpha} x_{i\gamma}}{f^2_\alpha(\sum_{j=1}^m x_{ij} w_{j\beta})} \\ & + T(-[W^{-1}]_{\beta\gamma}[W^{-1}]_{\beta\alpha}). \end{aligned}$$

$$\frac{\nabla^2 L(W, \widehat{F})}{\nabla w_{\gamma\beta} \nabla w_{\alpha\delta}} = -T[W^{-1}]_{\beta\gamma}[W^{-1}]_{\delta\alpha}, \text{ if } \delta \neq \beta$$