ABSTRACT

ELOYAN, ANI. Semi-Parametric Models for Independent Component Analysis. (Under the direction of Dr. Sujit K. Ghosh.)

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. Some of the methods for data representation using lower dimensional structure include Principal Component Analysis, Projection Pursuit and Independent Component Analysis (ICA). In this dissertation we consider the ICA model 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 mixing matrix (the linear transform) and the densities of the independent hidden sources. However, the solution to the problem depends on the identifiability of the sources. We first discuss some characterization results of random vectors and present a set of sufficient conditions to establish the identifiability of the sources and the mixing matrix using moment restrictions (up to third order) of the hidden source variables.

Next, the problem of density estimation that conserves a given set of moments is discussed. One of the nonparametric methods for solving this problem is the kernel based method. The motivation is that any continuous density can be approximated by a mixture of densities with appropriately chosen bandwidths. In many problems we may have specific information about the moments of the density. A novel method using a mixture of known densities is proposed. A modified EM-algorithm for estimating the weights of the mixture density under the constraints is used. The proposed method also obtains an estimate of the number of components in the mixture needed for optimal approximation. The proposed method is compared with two popular density estimation methods using simulated and real datasets and it is shown that the proposed estimate outperforms the others.

Finally, based on the moment constrained mixture density estimation methodologies we obtain a semi-parametric maximum likelihood estimate of the mixing matrix using a class of finite mixture distributions. The mixing matrix and mixture weights are estimated simultaneously. We establish the consistency of our estimate under additional regularity conditions. The method is illustrated and compared with a few existing methods by simulation studies.

Future research opportunities and a possible Bayesian method for ICA are discussed. Some preliminary results are shown to compare the method with a frequentist approach using simulated datasets.

Semi-Parametric Models for Independent Component Analysis

by
Ani Eloyan

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2010

APPROVED BY:

Dr. David A. Dickey	Dr. Helen (Hao) Zhang
Dr. Lexin Li	Dr. Sujit K. Ghosh Chair of Advisory Committee

DEDICATION

To my grandfather Dr. Gevorg Tosunyan
and in memory of grandfather Zalik Eloyan

BIOGRAPHY

Ani Eloyan was born on September 6^{th}, 1984 in Yerevan, Armenia to Nelsik and Manik Eloyan and is the first of three children. She grew up in the beautiful city Yerevan. She started taking piano lessons when she was still in elementary school. She did not enjoy it very much in the beginning, but her parents insisted on her continuing the lessons. She graduated from the piano school 7 years later and was very thankful to her parents for that. She also took drawing lessons with her brother and was trained in Armenian dance. She graduated from high school number 127 in May of 2000. Even though the family wanted her to go to medical school, she realized that mathematics is a better choice for her. She took the university entrance exams and was granted a fellowship to study at the Department of Applied Mathematics and Computer Science at Yerevan State University. Ani received her Bachelor's degree with honors in May 2004 and decided to continue her education in the US. She was granted an assistantship to attend graduate school at the Department of Statistics at North Carolina State University. She moved from the highlands of Armenia to the extreme humidity of Raleigh, North Carolina to pursue her PhD in 2005. While in graduate school she worked as a student tester at the JMP group at SAS institute for 2 years. She is hoping to complete the requirements for the PhD degree in August 2010 and, among other things, restart playing the piano.

ACKNOWLEDGEMENTS

First of all I want to thank my advisor Dr. Sujit K. Ghosh for his unceasing support, guidance and patience. From my first day at the department I learned a lot from him starting from the classroom of ST521 and in different aspects of graduate studies. I am very indebted to him for showing an excellent approach to work and life through his own example. I also want to thank my committee members Dr. Dickey, Dr. Zhang and Dr. Li for their interest in my research and for many constructive remarks. I am very grateful to the Department of Statistics at NCSU for giving me the opportunity to study here. I also want to thank the JMP Division at SAS Institute for supporting me as a graduate industrial trainee during the last two years of my graduate work.

I met several people throughout graduate school who helped make these years more interesting and fun. I want to thank all my classmates for the great hours we spent studying together while taking classes. Big thanks to Huiping, Kelci, Clay and Funda for listening to me in the hard times and helping out with their suggestions. I also want to thank all of my friends that I met outside the department for countless coffee breaks and interesting conversations that made these years more memorable.

I am blessed with a wonderful family that has always been by my side whether I lived there in Armenia with them or an ocean was separating us. Thanks to my wonderful sister Tatevik and brother Taron for their infinite support. I have shared every little step of my studies and life in this foreign country with them and they both have listened to me, celebrated with me and brightened up my days. Infinite thanks to my amazing parents for their love, friendship, guidance and understanding. I would have never been where I am now without the support of my parents.

Finally, a big thanks is due to my grandparents Dr. Gevorg and Mrs. Satenik Tosunyan for their encouragement and support. My grandfather showed me the best example of an academic. I have always admired his attitude towards research and I hoped to be like him one day. I always felt his presence next to me during hard times in graduate school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 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 sources. To begin with, 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). The main difference of ICA compared to 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). A general formulation of the problem can be presented as follows.

The research problem: Given a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, where the random vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})^T$ are independent and identically distributed (iid), can we find a unique transformation $g : \mathbb{R}^n \to \mathbb{R}^m$ for some $m \leq n$ and densities $f_1(\cdot), \ldots, f_m(\cdot)$ such that

$$\boldsymbol{x}_i \stackrel{d}{=} g(\boldsymbol{s}),$$

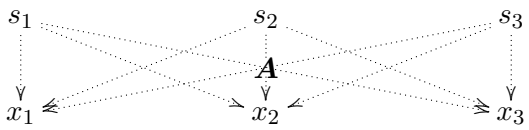
where $\boldsymbol{s} = (s_1, \ldots, s_m)^T$ is the vector of independent sources, i.e., $s_j \overset{indep}{\sim} f_j(\cdot)$ for $j = 1, \ldots, m$, $i = 1, \ldots, T$ and "$\stackrel{d}{=}$" denotes that the random quantities on either side of this equality have identical distribution. A special case emerges when the relationship is assumed to be linear, i.e., when $g(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{s}$ and the problem reduces to the estimation of the mixing matrix $\boldsymbol{A}$ and the probability densities $f_1(\cdot), \ldots, f_m(\cdot)$. In this dissertation, we derive the conditions for the uniqueness of the above linear representation and then develop estimation methodologies for the mixing matrix $\boldsymbol{A}$ and the source densities $f_1, \ldots, f_m$.

A simple example to illustrate the ICA arises from the auditory perception area and is called *the cocktail party problem* (Haykin and Chen, 2005). It refers to the ability of humans to recognize speech, which is extremely hard to accomplish by any electronic device. If there are m people in a room speaking at the same time and n recorders are placed in different parts of the room the signals recorded by each of these devices will be mixtures of the original source signals emitted by the speakers. If we define by $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ the observed mixture signals and by $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)^T$ the original source signals then the problem is to find a linear transformation of $\boldsymbol{x}$ such that

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x}, \tag{1.1}$$

where the matrix $\boldsymbol{W}$ is known as the unmixing matrix and has to be estimated based on a random sample of $\boldsymbol{x}$. The following diagram presents the cocktail party problem graphically when $m = n = 3$.

Three speakers



Three recording devices

In signal processing literature the problem is referred to as the blind separation of sources (BSS). The research in BSS showed that the problem can be reduced to finding a linear representation with independent components (Jutten and Herault, 1991).

1.1 A Review of Existing Methods

Some of the most popular classical methods for data representation and source separation such as principal component analysis (PCA) and factor analysis (FA) are based on the second-order moments of the observed variables. For instance, the PCA and FA are based on decomposing the covariance matrix via lower dimensional matrices. The main reason for the wide applicability of these methods is probably due to the simplicity of algorithms based on orthogonal decomposition. On the other hand, the justification of using only up to second order moments within these methods is the implicit assumption of gaussianity of the sources. If the data are assumed to be jointly normally distributed then its distribution is completely characterized by the moments up to the second order, hence there is no need to consider higher order moments. Throughout this chapter we are going to assume that $E(\boldsymbol{x}) = 0$, otherwise we can center the

data by its sample mean using the following location transformation.

$$\boldsymbol{x} \rightarrow \boldsymbol{x} - \widehat{E}(\boldsymbol{x}), \tag{1.2}$$

where $\widehat{E}(\cdot)$ denotes the empirical expectation of $\boldsymbol{x}$.

1.1.1 Principal Component Analysis

In principal component analysis (Jolliffe, 2002) the main goal is to find a transformation of the data to reduce the number of observed variables not losing important information contained in the data. The method is also called Hotelling transform, or Karhunen-Loève transform in the literature. We define the observed mixture variables as $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$ and the covariance matrix of $\boldsymbol{x}$ is given by

$$\boldsymbol{C}_x = E(\boldsymbol{x}\boldsymbol{x}^T), \tag{1.3}$$

where E is the expectation with respect to the density of $\boldsymbol{x}$.

The PCA procedure finds a linear transformation of $\boldsymbol{x}$ to a possibly lower dimensional vector $\boldsymbol{s} = (s_1, s_2, ..., s_m)^T$, such that the elements of $\boldsymbol{s}$ are uncorrelated and the dimension of $\boldsymbol{s}$ is $m \ll n$. The main goal for PCA is the dimension reduction. The correlation of the variables is used as a measure of redundancy in PCA. Notice, that if the components x_i of $\boldsymbol{x}$ are independent, then there is no redundancy in the data, since independence implies that the elements are uncorrelated. Hence, PCA is relevant only when the random variables x_i from the observed vector $\boldsymbol{x}$ are correlated or more generally dependent.

By spectral decomposition, the matrix $\boldsymbol{C}_x$ can be written as

$$\boldsymbol{C}_x = \sum_{i=1}^{n} \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i^T, \tag{1.4}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are the eigenvalues of $\boldsymbol{C}_x$ and $\boldsymbol{b}_i$ are the corresponding unit length eigenvectors of $\boldsymbol{C}_x$. Hence, for a small $\epsilon > 0$ if there exists m such that $\sum_{i=m+1}^{n} \lambda_i^2 < \epsilon$ then $\boldsymbol{C}_x$ can be approximated by $\sum_{j=1}^{m} \lambda_j \boldsymbol{b}_j \boldsymbol{b}_j^T$. In addition, $\boldsymbol{W} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_m)^T$ is the so-called unmixing matrix and $s_j = \boldsymbol{b}_j^T \boldsymbol{x}, j = 1, ..., m$ are the principal components. Notice, that $corr(s_i, s_j) = 0$ if $i \neq j$ but if s_i and s_j are non-gaussian they might still be dependent.

Hence, in PCA an $n \times 1$ vector of mixtures $\boldsymbol{x}$ is transformed linearly and orthogonally into an $m \times 1$ vector of original sources $\boldsymbol{s}$ such that the elements of $\boldsymbol{s}$ are uncorrelated. Notice that if we retain only $m (< n)$ sources $s_j = \boldsymbol{b}_j^T \boldsymbol{x}$ for $j = 1, ..., m$, then $\|\boldsymbol{C}_x - \sum_{j=1}^{m} \lambda_j \boldsymbol{b}_j \boldsymbol{b}_j^T\|^2 \leq \sum_{j=m+1}^{n} \lambda_j^2 < \epsilon$ by the choice of m.

As mentioned above, the main goal of PCA is the construction of the vector of principal components or sources $\boldsymbol{s} = (s_1, s_2, ..., s_m)^T$, so that s_i are uncorrelated with $m \ll n$ and $\boldsymbol{s}$

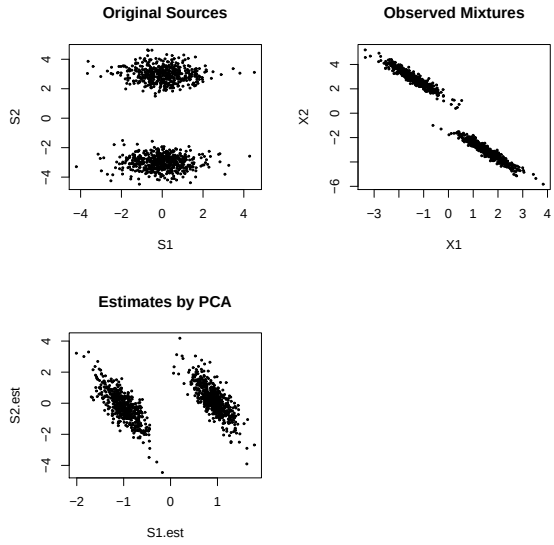


Figure 1.1: An example of the performance of PCA when the original sources are nongaussian for $T = 1000$ observations, the source s_1 is generated using a t-distribution with 10 degrees of freedom and s_2 is generated using a mixture of two normal densities.

accounts for as much of the variability in $\boldsymbol{x}$ as possible. Other common approaches for solving this problem such as minimizing the mean-squared error are discussed by Jolliffe (2002).

Example 1: To observe the performance of PCA when the sources have nongaussian densities we generate data from two such densities and apply PCA using R software. Suppose $s_1 \sim t_{10}\sqrt{10/8}$, where t_{10} is the t-distribution with 10 degrees of freedom and $s_2 \sim 0.5N(-3, 0.5) + 0.5N(3, 0.5)$, where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ. The sample size used for each component is $T = 1000$. Notice, that $E(s_1) = E(s_2) = 0$ and $\text{var}(s_1) = 1$ and $\text{var}(s_2) = 9.25$. The mixing matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} -0.5 & 0.5 \\ 0.5 & -1 \end{pmatrix}.$$

Figure 1.1 shows that if the densities of underlying sources are nongaussian then PCA does not perform well in recovering the original source densities and the mixing matrix. It is expected since by assumption only up to second order moments may be nonzero by applying PCA which may not be true if the underlying densities are nongaussian.

1.1.2 Factor Analysis

Another generative latent variable method for finding a parsimonious linear transformation of observed data is the factor analysis (Lawley and Maxwell, 1971). It was developed along with principal component analysis. The FA method may seem very similar in approach to PCA at a first glance, but they are addressing different issues.

In factor analysis the aim is to reduce the dimension of the data vector $\boldsymbol{x}$ in such a manner that the reduced vector $\boldsymbol{s}$ retains the covariances of the observed variables.

The general model for factor analysis is given by

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e}, \tag{1.5}$$

where $\boldsymbol{A}$ is the $n \times m$ mixture matrix, $\boldsymbol{s}$ is the $m \times 1$ vector of unobserved latent variables and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals with mean zero. The mixtures are usually assumed to be centered, i.e. $E(\boldsymbol{x}) = 0$. The components of $\boldsymbol{e}$ are assumed to be uncorrelated gaussian random variables and also uncorrelated with the elements in $\boldsymbol{s}$, i.e. $E(\boldsymbol{s}\boldsymbol{e}^T) = \boldsymbol{0}$. Without any loss of generality we can assume that the factors in $\boldsymbol{s}$ are uncorrelated and have unit variances $E(\boldsymbol{s}\boldsymbol{s}^T) = \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. Secondly, the covariance matrix of $\boldsymbol{e}$ is defined as $E(\boldsymbol{e}\boldsymbol{e}^T) = \boldsymbol{\Sigma}$ and the covariance matrix of the mixtures is defined as $E(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{C}_x$. Hence, the covariance matrix of the mixtures can be obtained as

$$\boldsymbol{C}_x = \boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{\Sigma}. \tag{1.6}$$

The goal in factor analysis is to estimate the parameters $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ by using the observed mixtures $\boldsymbol{x}$. However, before going into estimation issues, the first question of interest is whether the choice of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ is unique in representing $\boldsymbol{C}_x$. In other words, if we have the observed mixtures and the matrix $\boldsymbol{C}_x$, can we use $\boldsymbol{C}_x$ to find a unique positive diagonal matrix $\boldsymbol{\Sigma}$ and an $n \times m$ matrix $\boldsymbol{A}$ which will satisfy (1.6)?

Each diagonal element of $\boldsymbol{C}_x - \boldsymbol{\Sigma}$ is the part of the variance of the corresponding variate which is due to the m common factors and is called 'communality' of the variate. Since $m < n$, if there exists a unique matrix $\boldsymbol{\Sigma}$ the rank of the matrix $\boldsymbol{C}_x - \boldsymbol{\Sigma}$ should be equal to m by equation (1.6). First assume that there is a unique $\boldsymbol{\Sigma}$. If $m = 1$ then $\boldsymbol{A}$ is a column vector of size n. In that case $\boldsymbol{A}$ can be found uniquely up to the sign of all its elements. This is usually ignored in this context. Now if $m > 1$ then $\boldsymbol{A}$ is not uniquely identified. Suppose $\boldsymbol{A}$ is a solution for (1.6), then for any orthogonal matrix $\boldsymbol{F}$, $\boldsymbol{A}\boldsymbol{F}$ is also a solution. In factor analysis this is referred to as a rotation of the factors.

In practice, this issue is resolved by imposing special conditions posed in the problem at hand. For instance, in some cases we might know *a priori* that some of the elements in matrix $\boldsymbol{A}$

are zero. The rotations might change the patterns of zeros which are imposed by the problem. So we might be able to find a unique solution by considering these restrictions.

In case if there are no specific conditions stated, we will need to impose some arbitrary assumptions in order to obtain a unique solution. Suppose we rescale the covariance matrix of the mixtures as follows

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{C}_x\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-\frac{1}{2}} + \boldsymbol{I}_n,$$

and define $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{C}_x\boldsymbol{\Sigma}^{-\frac{1}{2}} = \widetilde{\boldsymbol{C}}_x$. Obviously, the matrix $\widetilde{\boldsymbol{C}}_x - \boldsymbol{I}_n$ is symmetric and, as mentioned above, it has rank m. Hence, by using the singular value decomposition of $\widetilde{\boldsymbol{C}}_x - \boldsymbol{I}_n$ we obtain

$$\widetilde{\boldsymbol{C}}_x - \boldsymbol{I}_n = \boldsymbol{Q}\Lambda\boldsymbol{Q}^T,$$

where $\boldsymbol{Q}$ is an $n \times m$ orthogonal matrix $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_m$ the columns of which are the eigenvectors of $\widetilde{\boldsymbol{C}}_x - \boldsymbol{I}_n$ and Λ is an $m \times m$ nonsingular diagonal matrix the diagonal elements of which are the eigenvalues of matrix $\widetilde{\boldsymbol{C}}_x - \boldsymbol{I}_n$. If we assume that the eigenvalues in Λ are positive and distinct and they are ordered in decreasing order of magnitude, then the matrix $\boldsymbol{Q}$ is unique and subsequently we can define $\boldsymbol{A}$ uniquely as follows

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}\Lambda^{\frac{1}{2}}. \tag{1.7}$$

In this case we obtain

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}\Lambda\boldsymbol{Q}^T\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{C}_x - \boldsymbol{\Sigma}.$$

Hence, the matrix $\boldsymbol{A}$ is defined uniquely and satisfies (1.6).

In summary, if there exists a unique positive diagonal matrix $\boldsymbol{\Sigma}$ such that the m largest eigenvalues of the matrix $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{C}_x\boldsymbol{\Sigma}^{-\frac{1}{2}} - \boldsymbol{I}_n$ are positive and distinct and the rest of them are zero then $\boldsymbol{A}$ can be uniquely defined. This implies that the factors are identified if the above conditions hold.

The next question is when such a $\boldsymbol{\Sigma}$ can be defined. By definition of $\boldsymbol{A}$ in (1.7) we obtain $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A} = \boldsymbol{Q}\Lambda^{\frac{1}{2}}$ which then implies

$$\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} = \Lambda.$$

Hence, the matrix $\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$ should be diagonal which will impose $\frac{1}{2}m(m-1)$ constraints on the parameters. Therefore, the number of parameters which do not have any constraints is equal to $n + nm - \frac{1}{2}m(m-1)$. The number of equations in (1.6) is $\frac{1}{2}n(n+1)$. If we subtract the number of free parameters from this number we obtain

$$k = \frac{1}{2}n(n+1) - n - nm + \frac{1}{2}m(m+1).$$

If $k = 0$ then we have as many equations as free parameters and we should expect to be able to solve the system and obtain a unique $\boldsymbol{\Sigma}$. If $k < 0$ then we have fewer equations than free parameters, which implies that we should expect to have infinitely many solutions for $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$. Finally if $k > 0$ then we have more equations then free parameters, so in general this might not be solvable, unless we impose some constraints on elements of $\boldsymbol{C}_x$.

1.1.3 Projection Pursuit

Projection pursuit is a statistical method of finding projections for multivariate data which are 'interesting.' It is mostly used for visualization of the clustering in data. One of the implications of the projection pursuit methods, that can be of interest here is the dimension reduction. As argued by Friedman and Tukey (1974) if the data consist of several clusters, the projections onto the principal axes may result in loosing the information about the clusters of the data. Hence, they proposed a linear mapping algorithm, which finds the optimum projections. Huber (1985) and Diaconis and Freedman (1984) gave some heuristic arguments to show that the least 'interesting' projections are the ones which are most gaussian. One of the main reasons for this is that a multivariate density is gaussian if and only if all of its one-dimensional projections are gaussian. This implies that if the 'least normal' direction is gaussian then all others are gaussian as well. Moreover, the multivariate normal density is completely specified by its location and covariance parameters, is elliptically symmetric and contains the least information for a fixed variance in the sense of maximizing the negentropy (see Lemma 1.3.1).

The problem of finding these 'interesting' projections, reduces to finding some index which will measure the nongaussianity. One of the choices presented by Jones and Sibson (1987) is the entropy of the function as a measure of nongaussianity. If the random variable $\boldsymbol{x}$ has pdf $f_x(\cdot)$ then the differential entropy of $\boldsymbol{x}$ is defined as follows

$$H(f_x) = -\int f_x(t) \log f_x(t) dt. \tag{1.8}$$

It was shown that a method of finding projection pursuit directions involves the minimization of the entropy. There are some methods using cumulants to approximate the entropy which will be discussed in Section 1.3.1.

1.2 The Independent Component Analysis Model

To give a mathematical definition of basic linear ICA, we define the observed random variables as $x_1, x_2, \ldots, x_n$. These random variables are linear combinations of some latent variables

$s_1, s_2, \ldots, s_m$ which cannot be observed. Hence, we obtain

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{im}s_m,$$

for $i = 1, \ldots, n$. Here $a_{ij}, i = 1, \ldots, n, j = 1, \ldots, m$ are real valued unknown coefficients of the linear combination. The basic assumption of ICA model is that the components s_j are statistically mutually independent. In other words, if $f(s_1, s_2, \ldots, s_m)$ is the joint probability density function of $\boldsymbol{s}$ and $f_j(s_j)$ is the marginal density function of s_j, then

$$f(s_1, s_2, \ldots, s_m) = f_1(s_1)f_2(s_2) \cdot \ldots \cdot f_m(s_m).$$

Henceforth, for simplicity we will make a conventional assumption that the number of observed random variables and the latent variables are the same (i.e., $m = n$). In case if the number of mixtures is smaller than the number of ICs the problem is called ICA with overcomplete bases, which is discussed by Hyvarinen et al. (2001).

The basic ICA model was first introduced by Jutten and Herault (1991), who coined the name ICA and gave an explicit formulation of the problem.

The ICA model is often written in matrix notation. Suppose $\boldsymbol{A}$ is the $n \times m$ matrix of coefficients a_{ij}, $\boldsymbol{x}$ is the column vector of observed random variables, and $\boldsymbol{s}$ the vector of latent variables, then the mixing model can be written as follows

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}.$$

Without any loss of generality we will also assume that the random variables x_i have zero mean $E(\boldsymbol{x}) = 0$. Otherwise, we can subtract the empirical mean value and centralize the random variables before starting the analysis.

One important assumption made in ICA is that the matrix of parameters $\boldsymbol{A}$ must be of full column rank. In the special case when $m = n$ this implies that the matrix $\boldsymbol{A}$ is invertible.

The main goal of independent component analysis is the estimation of the matrix of unknown parameters $\boldsymbol{A}$ and also the distributions of latent variables s_j having observed the variables $\boldsymbol{x}$. We discuss the identifiability of the model and the issues that are present in the currently used formulation of the ICA problem in Chapter 2.

The model described above is the *noise free* ICA model, since it does not account for any noise variables that can be present in the model.

Example 1 **(Continued)**: Going back to Example 1 we apply ICA to the same data and compare the results of ICA and PCA. The package `fastICA` in R software was used for computation. Figure 1.2 presents the performance of PCA and ICA. It can be easily observed that the ICA model provides a good estimate for the unmixing matrix and hence the original sources are

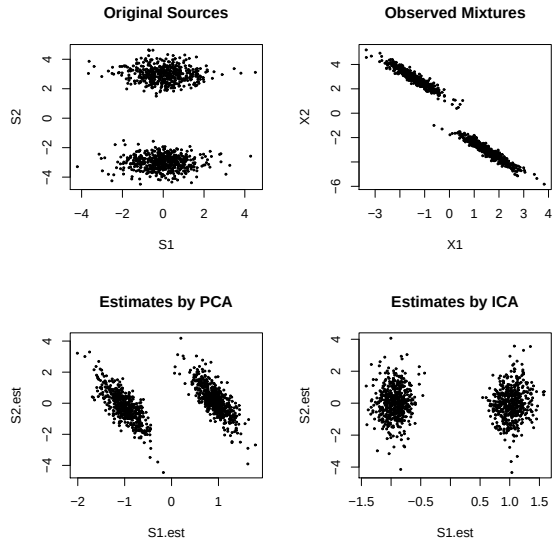


Figure 1.2: Comparison of ICA and PCA for a simulated dataset. The upper two figures show the scatterplots of the generated independent sources and the constructed mixtures. The lower two figures show the estimated sources found by PCA (on the left) and ICA (on the right) models.

recovered well as opposed to PCA. We also computed the mutual information values for the two estimation methods as defined in (1.25) by using the R package `np` for estimating the joint and marginal densities of the two estimated sources nonparametrically using kernel-based methods. The resulting mutual information values are $I_{true}(s_1, s_2) = 1.61$, $I_{obs}(x_1, x_2) = 219.52$, $I_{ICA}(\widehat{s}_1, \widehat{s}_2) = 2.99$ and $I_{PCA}(\widehat{s}_1, \widehat{s}_2) = 56$. Since the mutual information is minimized when the random variables are independent, ICA performs better then PCA by having a significantly lower mutual information value.

The performance of the method is also 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 $\boldsymbol{A}$ and estimated unmixing matrix $\widehat{\boldsymbol{W}}$ the Amari error is defined as

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

where $\boldsymbol{P} = \boldsymbol{A}\widehat{\boldsymbol{W}}$. Notice that for any two matrices $\boldsymbol{A}$ and $\widehat{\boldsymbol{W}}$, 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 $\boldsymbol{A}$ and the estimated mixing matrix $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{W}}^{-1}$ are equivalent in the sense as defined in Section 2.1.

However, the Amari error is not invariant to a constant multiplier, in other words $AE(\boldsymbol{A}, \widehat{\boldsymbol{W}}) \neq AE(\boldsymbol{A}, \widehat{\boldsymbol{W}}\boldsymbol{\Lambda})$, where Λ is a diagonal matrix with positive elements on the diagonal. Hence before computing this error we rescale the columns of the matrices $\boldsymbol{A}$ and $\widehat{\boldsymbol{W}}$ to have Euclidean norm unity.

In this example, the Amari error of the estimated mixing matrix computed by the PCA method is $AE(\boldsymbol{A}, \widehat{\boldsymbol{W}}_{PCA}) = 0.73$ and for the ICA method it is $AE(\boldsymbol{A}, \widehat{\boldsymbol{W}}_{ICA}) = 0.021$. Again this indicates the superior performance of the ICA method over the PCA method when the underlying sources are nongaussian.

1.3 Methods for finding Independent Components

There are several approaches for estimation of the mixing matrix $\boldsymbol{A}$ in the ICA model. Most of these methods are based on maximizing contrast functions defined by Comon (1994) as follows.

Definition A mapping $\phi(\cdot)$ from the set of densities $\{f_s, \boldsymbol{s} \in \mathbb{R}^n\}$ to $\mathbb{R}$ is called a contrast function if it satisfies the following requirements

- $\phi(f_s) = \phi(f_{Ps})$ where $\boldsymbol{P}$ is a permutation matrix.
- $\phi(f_s) = \phi(f_{\Lambda s})$ where Λ is a diagonal invertible matrix.
- $\phi(f_{As}) \leq \phi(f_s)$ if all elements of $\boldsymbol{s}$ are independent and the matrix $\boldsymbol{A}$ is invertible.

In this section, some commonly used contrast functions are described. The ICA models are built by maximizing the contrast functions to estimate the mixing matrix. One of the common approaches to find the contrast functions is based on the research done for projection pursuit problems. As mentioned in section 1.1.3 the 'interesting' projections are the ones which are nongaussian. This yields a simple principle of maximization of nongaussianity that can be applied for estimating the independent components in the ICA model.

1.3.1 Maximization of Nongaussianity

In basic ICA we are trying to estimate the independent components $s_j, j = 1, \ldots, m$, from the model $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$. One of the approaches to solving this problem is by maximizing some numerical measure of nongaussianity. Chapter 2 provides a formal justification for the need of nongaussianity of the sources and relates it to the identifiability of the ICA problem.

Forth-order Moments as a Measure of Nongaussianity

One of the measures of nongaussianity used in ICA estimation is the forth-order cumulant. The cumulant κ_n of a random variable s is defined as the n^{th} order derivative of the cumulant

generating function of s given by $cum_s(t) = \log[E(e^{ts})]$. If s is a zero mean random variable with unit variance, the forth-order cumulant of s is given by

$$\kappa_4(s) = E(s^4) - 3[E(s^2)]^2 = E(s^4) - 3. \tag{1.9}$$

Notice that for a normally distributed random variable z with zero mean and unit variance $\kappa_4(z) = 0$, since $E(z^4) = 3[E(z^2)]^2$.

Delfosse and Loubaton (1995) showed that the following function is a contrast function

$$\phi(f_x) = \sum_{k=1}^{m} [\kappa_4(\overline{x}_k)]^2, \tag{1.10}$$

where $\overline{\boldsymbol{x}} = E(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{x}$ and $\boldsymbol{x}$ is such that $\kappa_4(\overline{x}_k) \neq 0$ for each $k = 1, ..., m$.

To solve the ICA problem the algorithm proceeds by prewhitening the data $\boldsymbol{x}$. This implies that we can assume $E(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{I}$. Furthermore, let the function $\psi(\cdot)$ be defined on the unit sphere as

$$\psi(\boldsymbol{q}) = \frac{1}{4}\kappa_4(\boldsymbol{q}^T\boldsymbol{x})^2. \tag{1.11}$$

where $\|\boldsymbol{q}\| = 1$. By proposition 2.1 in Delfosse and Loubaton (1995), the function $\psi(\cdot)$ given in (1.11) is maximized at the vectors $\pm\boldsymbol{w}_k$, for $k = 1, ..., m$ which are the rows of the matrix $\boldsymbol{W}$.

The algorithms using kurtosis are mostly based on gradient methods as discussed by Hyvarinen et al. (2001). For instance, a gradient algorithm is shown if the absolute value of the forth order cumulant is used as a contrast function instead of the square. The gradient algorithm has two steps

$$\begin{aligned} \boldsymbol{w} &\leftarrow \boldsymbol{w} + sign(\kappa_4(\boldsymbol{w}^T\boldsymbol{x}))\hat{E}(\boldsymbol{x}(\boldsymbol{w}^T\boldsymbol{x})^3) \\ \boldsymbol{w} &\leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|. \end{aligned} \tag{1.12}$$

Relatively new approaches are described by Zhang (2005) and Taoufiki et al. (2007). The use of cumulants in ICA and related fields is most common because of the computational simplicity of the algorithms. In addition, the cumulants can be used to get global convergence results, as stated in Delfosse and Loubaton (1995). However, the estimates obtained by using the cumulants can be rather poor taking into consideration the statistical properties of the estimated components. Also as argued by Huber (1985) the cumulants are very sensitive to outliers.

Negentropy as a Measure of Nongaussianity

Another commonly used measure of nongaussianity is the negentropy of the distribution. As mentioned in Section 1.1.3 the differential entropy of a random variable $\boldsymbol{x}$ which has a pdf f_x is defined as $H(f_x) = -\int f_x(t)\log f_x(t)dt$ assuming that the integral exists.

Lemma 1.3.1 *Suppose* $\boldsymbol{z}$ *is a normally distributed random vector with* $\int \boldsymbol{z} f_z(\boldsymbol{z}) d\boldsymbol{z} = 0$ *and* $\int \boldsymbol{z}\boldsymbol{z}^T f_z(\boldsymbol{z}) d\boldsymbol{z} = \boldsymbol{\Sigma}$. *Then for any random variable* $\boldsymbol{x}$ *such that* $\int \boldsymbol{x} f_x(\boldsymbol{x}) d\boldsymbol{x} = 0$ *and* $\int \boldsymbol{x}\boldsymbol{x}^T f_x(\boldsymbol{x}) d\boldsymbol{x} = \boldsymbol{\Sigma}$ *the following holds*

$$H(f_z) \geq H(f_x) \tag{1.13}$$

where the equality holds if and only if $\boldsymbol{x} \stackrel{d}{=} \boldsymbol{z}$.

This implies that the negentropy can be used to find a measure of nongaussianity. The negentropy of $\boldsymbol{x}$ is given by

$$J(f_x) = H(f_z) - H(f_x), \tag{1.14}$$

where $\boldsymbol{z}$ is a gaussian random variable such that $E(\boldsymbol{z}) = E(\boldsymbol{x})$ and $\text{cov}(\boldsymbol{z}) = \text{cov}(\boldsymbol{x})$.

By the above lemma the negentropy of any random variable is nonnegative and it is zero if and only if the density of $\boldsymbol{x}$ is gaussian. Hence, the maximization of nongaussianity is equivalent to maximizing the negentropy.

In contrast with the higher order cumulant methods discussed in previous section, this method may result in estimates with better statistical properties. The major problem encountered here is that the calculation of the negentropy involves the unknown density of $\boldsymbol{s}$. Hence, we need to estimate the density by some nonparametric methods which can be mathematically and computationally complicated. Some methods using kernel density estimation for negentropy are discussed briefly in Section 1.3.2.

Jones and Sibson (1987) have shown that the negentropy of a real-valued random variable x with unit variance can be approximated by its higher-order cumulants as follows

$$J(f_x) \approx \frac{1}{12} E(x^3)^2 + \frac{1}{48} \kappa_4(x)^2. \tag{1.15}$$

However, similar problems as in last subsection arise when using this approximation, because of the lack of robustness in higher-order cumulants.

One of the more sophisticated approximations of the negentropy is given by nonpolynomial functions. By Hyvarinen (1997) if $G(\cdot)$ is some nonquadratic function, then the negentropy of a random variable $\boldsymbol{s}$ can be approximated by

$$J_G(f_s) = [E_s(G(s)) - E_z(G(z))]^2, \tag{1.16}$$

where z is a gaussian random variable with mean zero and variance one. Empirical studies suggest that the following choices of $G(\cdot)$ may provide good approximations

$$G_1(u) = \log \ \cosh \ au, \qquad G_2(u) = \exp(-\frac{u^2}{2}), \tag{1.17}$$

where a is a constant such that $1 \leq a \leq 2$.

Suppose $\boldsymbol{w}$ is one of the rows of the matrix $\boldsymbol{W}$. To maximize the negentropy (1.16) of $\boldsymbol{w}^T\boldsymbol{x}$ the optimum of $E(G(\boldsymbol{w}^T\boldsymbol{x}))$ can be used under the constraint $\|\boldsymbol{w}\|^2 = 1$. Hence, the following steps of a Newton-Raphson method are implemented

$$\begin{aligned} \boldsymbol{w} &\leftarrow \hat{E}(\boldsymbol{x}g(\boldsymbol{w}^T\boldsymbol{x})) - \hat{E}(g'(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{w} \\ \boldsymbol{w} &\leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|, \end{aligned} \tag{1.18}$$

where $g(\cdot)$ is the derivative function of $G(\cdot)$ and $g'(\cdot)$ is the second derivative of $G(\cdot)$. The `fastICA` package in R uses $G_1(\cdot)$ or $G_2(\cdot)$ to approximate negentropy as chosen by the user.

The algorithm described above can be used for the case when the original sources are generated using a super-gaussian (positive kurtosis) or sub-gaussian (negative kurtosis) densities. However, Gretton (2006) mentions that one should be careful about using this in practice. Indeed, as shown by example 2 in some cases the contrast function used may even be minimized (as opposed to being maximized) at the true value of the parameter.

Example 2: It was shown in Figure 1.2 that PCA does not recover the original source signals well. In fact it can be seen that the original sources are recovered up to a rotation. In many of the aforementioned ICA algorithms PCA is used as a preprocessing step for whitening the data and subsequently the algorithm is applied to recover the final rotation matrix. The resulting estimate of the unmixing matrix is found to be better than the estimate found by PCA. Now suppose we use a rotation matrix as a true unmixing matrix given by

$$\boldsymbol{W} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

where the true angle of the rotation is given by $\theta = \pi/4$. We generate two vectors of original sources by using Uniform(0,1) density (negative kurtosis) and compute the contrast function in (1.16) using the function $G_1(\cdot)$ in (1.17). Subsequently, we generate two vectors of original sources using a Laplace(0,1.9) density (positive kurtosis) and compute the contrast function. The plots are shown in Figure 1.3. The contrast function is maximized at the true value of the angle θ for the first case. However, for the second case when the kurtosis of the data is positive the contrast function is minimized at the value $\theta = \pi/4$. Similar behavior can be observed for the JADE algorithm described in Section 1.3.4.

1.3.2 Maximum likelihood approach

Another approach for finding the independent components is related to the maximum likelihood (ML) estimation. The main idea of ML estimation is that given a family of densities $\mathcal{F}_\theta = \{f_\theta :$

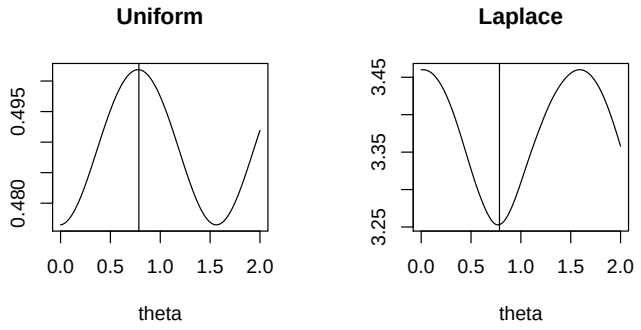


Figure 1.3: The contrast function based on the approximation of negentropy (fastICA algorithm) for data generated by using subgaussian (left) or supergaussian (right) densities.

$\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ we obtain an estimate $\widehat{\boldsymbol{\theta}}$ that minimizes the Kullback-Leibler divergence between the true density $f(\cdot)$ and a member from the family of densities $\mathcal{F}_\theta$. In other words we find

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \int f(x) \log \tfrac{f(x)}{f_\theta(x)} dx \approx \arg\min_{\boldsymbol{\theta}} \textstyle\sum_{i=1}^n \log \tfrac{f(x_i)}{f_\theta(x_i)} \\ &= \arg\max_{\boldsymbol{\theta}} \textstyle\sum_{i=1}^n \log(f_\theta(x_i)) \end{aligned} \tag{1.19}$$

Given the true densities of the hidden sources $\boldsymbol{f} = (f_1, \ldots, f_m)$, the log-likelihood of the unmixing matrix $\boldsymbol{W}$ is given by

$$L(\boldsymbol{W}, \boldsymbol{f}) = \sum_{i=1}^{T} \sum_{j=1}^{m} \log f_j(\boldsymbol{w}_j^T \boldsymbol{x}_i) + T \log |\det \boldsymbol{W}|, \tag{1.20}$$

where $\boldsymbol{x}_i, i = 1, ..., T$ are realizations of $\boldsymbol{x}$ and the matrix $\boldsymbol{W} = (\boldsymbol{w}_1^T, ..., \boldsymbol{w}_m^T)^T$ is the inverse of matrix $\boldsymbol{A}$.

The well-known asymptotic efficiency of ML approach makes it a default choice among many statistical estimation methods. Hyvarinen et al. (2001) proved that under some regularity conditions on the distribution of independent components the ML estimators are locally consistent, in other words the estimators converge in probability to the true parameter as the number of observations $T \to \infty$. However, there are some limitations of the ML approach when using the method in practice. One problem with ML estimators is that they are not robust against the model misspecification. In practice, we may not know the probability distributions of the components and hence the above mentioned consistency and asymptotic optimality of the ML approach may no longer be valid. In other words, when the densities of the ICs are not exactly

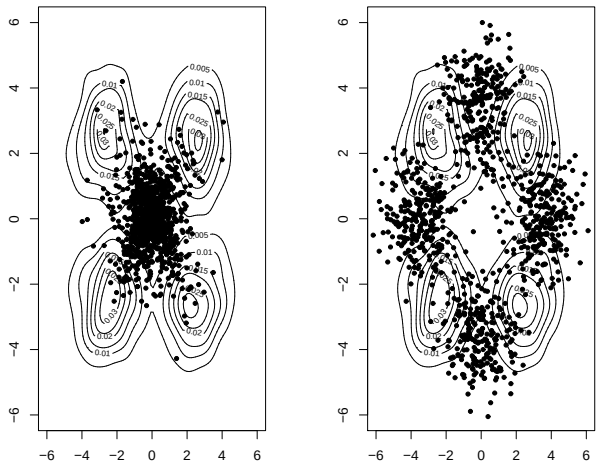


Figure 1.4: The contour plot of the joint density of the original sources given by a mixture of two normal densities overlaid by a scatterplot of the data constructed using two Laplace densities (left) and two mixture densities (right).

known to be $f_j(\cdot)$ then the MLE of $\boldsymbol{W}$ (and hence that of $\boldsymbol{A}$) will not be consistent. Without *a priori* information about the densities, we usually need to use nonparametric methods for density estimation, which require heavy computations. Gretton (2006) discussed the issues that can arise if a wrong model is used in ML estimation of ICA as illustrated in the following example.

Example 3: Suppose the original sources $\boldsymbol{S}_m$ are generated using a mixture of two normal densities $0.5N(-2.5, 1) + 0.5N(2.5, 1)$. If the correct model is used to find the ML estimate of the mixing matrix $\boldsymbol{A}$ then the ML method results in the best estimate. However, suppose the original sources $\boldsymbol{S}_l$ are generated independently using a Laplace(0,1.9) density. Further, assume that the above mixture of two normal densities is incorrectly used in the ML method to estimate the mixing matrix. Figure 1.4 shows the contour plot of the estimated joint density of original sources $\boldsymbol{S}_l$ using a mixture of gaussian densities. A rotation matrix with $\theta = \pi/4$ is used to construct the observed mixtures. Further, the scatterplots of the constructed mixtures using the $\boldsymbol{S}_l$ on the left and $\boldsymbol{S}_m$ on the right are overlaid on the contour plot. By looking at Figure 1.4 one can observe that the mixture of only two densities is a poor model if the original independent hidden sources are generated from a Laplace density. Hence, we may not obtain a good estimate of the mixing matrix using the ML method in this case. However, in this example if we would have estimated the parameters of the mixture of normals (along with the

mixing matrix), the results could be improved.

In practice the parametric models used for ML estimation are quite similar to the algorithms discussed in previous sections. The simple algorithms are obtained by gradient methods. By taking the derivative of the log-likelihood in (1.20) we obtain the score function as follows

$$\frac{\partial \log L(\boldsymbol{W}, \boldsymbol{f})}{\partial \boldsymbol{W}} = T[\boldsymbol{W}^T]^{-1} + T\sum_{i=1}^{n}[\boldsymbol{g}(\boldsymbol{W}\boldsymbol{x}_i)\boldsymbol{x}_i^T] = 0, \tag{1.21}$$

where $\boldsymbol{g}(\cdot)$ is the negative score function of densities of s_j, the components of which are calculated by $g_j = \frac{f_j'}{f_j}, j = 1, \ldots, m$. The roots of the above function can be found by the Newton-Raphson method yielding estimates for the mixing matrix $\boldsymbol{W}$. Other algorithms that also use gradients, such as natural gradient algorithm are discussed in Amari (1998).

A relatively new nonparametric technique for ML estimation of ICA is based on using the kernel density estimate for the distribution of ICs. If the random variables $s_1, s_2, ..., s_T$ have distribution function $f(\cdot)$ then by Parzen (1962) the kernel density estimator of $f(\cdot)$ is given as follows

$$\hat{f}(s) = \frac{1}{Th}\sum_{i=1}^{T} K\left(\frac{s_i - s}{h}\right), \tag{1.22}$$

where $K(\cdot)$ is the kernel density, the bandwidth $h \equiv h(T)$ is chosen so that it satisfies $h \to 0$ and $Th \to \infty$ as $T \to \infty$. Bach and Jordan (2002) give elaborate theoretical foundation for using kernel density estimation in ICA. Vlassis and Motomura (2001) discuss the use of gaussian kernel and show that in the vicinity of global minimum of the contrast function their method gives better results than the methods using approximate models of the density. Since we do not have a sample directly from the densities of the independent components we cannot directly use $s_1, ..., s_T$ in this model. Boscolo et al. (2004) proposed a pseudo-likelihood based method for ICA model (2.1) where the densities of the sources are estimated nonparametrically by using the kernel density estimate. 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 $10\log_{10}\sum_{i=1}^{T} s_{ij}^2/(\widehat{s}_{ij} - s_{ij})^2$, where s_{ij} is the original source value and $\widehat{s}_{ij}$ is the reconstructed value. Boscolo et al. (2004) show the use of KDE to construct the loglikelihood function for $\boldsymbol{W}$ as follows

$$L(\boldsymbol{W}, \widehat{\boldsymbol{f}}_n) = \sum_{i=1}^{T}\sum_{j=1}^{m} \log\left\{\frac{1}{Th}\sum_{k=1}^{T} K\left[\frac{\boldsymbol{w}_j^T(\boldsymbol{x}_i - \boldsymbol{x}_k)}{h}\right]\right\} + T\log|\det(\boldsymbol{W})| \tag{1.23}$$

where $\|\boldsymbol{w}_j\| = 1$, for $j = 1, ..., m$ and $\boldsymbol{x}_i$ is the i^{th} realization of $\boldsymbol{x}$. Hence, the function in (1.23)

can be maximized to find the MLE for $\boldsymbol{W}$. However the use of kernel methods may not provide the most efficient estimators of the mixing matrix or the source densities (e.g., see Geman and Hwang (1982)).

One common feature of most of the methods discussed above is that they are one-unit estimation methods, in other words first we are trying to estimate one linear combination of the observed variables which is equal to one of the independent components. Afterwards, we do this procedure a number of times, to get several independent components. This approach is useful mainly because we do not need to have prior knowledge about the number of independent components. Secondly, this approach connects the problem of ICA with projection pursuit. Indeed, all the methods above are dealing with maximization of nongaussianity, which is the main goal in projection pursuit problems as well.

1.3.3 Mutual Information approach

The mutual information of a random vector $\boldsymbol{s}$ is defined as the Kullback-Leibler divergence between the joint density of $\boldsymbol{s}$ and the product of the marginal densities of its components $s_j, j = 1, ..., m$ given by

$$\boldsymbol{I}(\boldsymbol{s}) = \int f_s(\boldsymbol{u}) \log \frac{f_s(\boldsymbol{u})}{\prod_{j=1}^{m} f_{s_j}(u_j)} d\boldsymbol{u} \tag{1.24}$$

Assuming $f_{s_j}(u_j) = \int f_s(\boldsymbol{u}) du_1 ... du_{j-1} du_{j+1} ... du_m$ for all $j = 1, ..., m$ the mutual information can be written as

$$\boldsymbol{I}(\boldsymbol{s}) = \sum_{j=1}^{m} H(f_{s_j}) - H(f_{\boldsymbol{s}}) \tag{1.25}$$

where H is the entropy function defined in (1.8). Comon (1994) showed that the negative mutual information is a contrast function as defined in 1.3 which can be used to find plausible estimates for ICA model.

If the s_j are uncorrelated and $E(\boldsymbol{s}\boldsymbol{s}^T) = \boldsymbol{C}_s$, then it can be shown (Comon, 1994) that the mutual information is related to the negentropy as follows

$$\boldsymbol{I}(\boldsymbol{s}) = J(f_{\boldsymbol{s}}) - \sum_{j=1}^{m} J(f_{s_j}) + \frac{1}{2} \log \frac{\prod \boldsymbol{C}_{s_{jj}}}{\det(\boldsymbol{C}_s)}, \tag{1.26}$$

where $J(\cdot)$ is the negentropy defined in (1.14).

In order to use this method in practice, we need to find an approximation for mutual information. One approach is to estimate the entropy with methods discussed in Section 1.3.1. On the other hand, nonparametric methods from Section 1.3.2 can also be applied to estimate the densities of the ICs and calculate the mutual information by using these estimates. Moreover, it may be computationally cumbersome to estimate the joint density nonparametrically.

1.3.4 Tensor-based Approach

Cumulant tensors are also used for estimating the ICA model. By Cardoso and Comon (1996) the cumulant tensor Q_x of a random variable $\boldsymbol{x}$ is a linear map $\boldsymbol{M} \rightarrow Q_x(\boldsymbol{M})$ which returns a matrix $Q_x(\boldsymbol{M})$ for any matrix $\boldsymbol{M}$. The ij^{th} element of the matrix $Q_x(\boldsymbol{M})$ is defined as

$$Q_x(\boldsymbol{M})_{ij} = \sum_{kl} cum(x_i, x_j, x_k, x_l)\boldsymbol{M}_{lk} \tag{1.27}$$

where $cum(x_i, x_j, x_k, x_l)$ is the forth-order cumulant, which can be calculated as

$$cum(x_i, x_j, x_k, x_l) = E(x_i x_j x_k x_l) - E(x_i x_j)E(x_k x_l) - E(x_i x_k)E(x_j x_l) - E(x_i x_l)E(x_k x_j). \tag{1.28}$$

The cumulant tensor is a symmetric linear operator and hence it has an eigenvalue decomposition. A matrix $\boldsymbol{M}$ is the eigenmatrix of a tensor if

$$Q_x(\boldsymbol{M}) = \lambda M$$

where λ is a scalar, which is the eigenvalue of the tensor. It has m^2 eigenvalues for the eigenmatrices.

Cardoso (1990) argued that the ICA model can be estimated by solving for the eigenvectors of these matrices. The drawback of this method is that we are once again using the cumulants for estimation. As mentioned in Section 1.3.1, there can be problems concerning the statistical properties of these estimates.

The common algorithms using cumulant tensors for ICA estimation are forth-order blind identification (FOBI) discussed in Cardoso (1990) and joint approximate diagonalization of eigenmatrices (JADE) which is introduced in Cardoso and Souloumiac (1993).

1.3.5 Bayesian Approach

In many practical problems the data contain prior information about the form of the mixing matrix. For example, this can be observed in literature on application of ICA in brain imaging, especially dealing with magnetoencophalography (MEG) data, where the classic dipole model is used to find the ICs. In order to use this *a priori* information Bayesian estimation methods are used for finding the ICs. Ali (2000) gave a unifying approach for using Bayesian estimation methods for ICA. He argued that most of the conventional methods can be shown to be special cases of *a posteriori* estimation. Here we assume that the unmixing matrix $\boldsymbol{W}$ is a random variable with a known distribution $f(\boldsymbol{W})$. Then it can be shown that because of the

independence of the components and the matrix $\boldsymbol{W}$ the posterior density is given as

$$\log p(\boldsymbol{W}, \boldsymbol{y}|\boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{W}, \boldsymbol{y}) + \log p(\boldsymbol{W}) + \log p(\boldsymbol{y}) + c, \tag{1.29}$$

where $c = -\log p(\boldsymbol{x})$ is a constant. Using this posterior density we can make inference about the parameters in matrix $\boldsymbol{W}$ and ICs. For instance, if $\boldsymbol{W}$ is integrated out from this density, then the ICs can be estimated using the resulting marginal density. Similar approach can be used to estimate $\boldsymbol{W}$.

One crucial issue that arises while using this approach is the choice of the prior density. It was shown by Hyvarinen et al. (2001) that classical noninformative priors such as Jeffrey's prior do not give good estimation results in general. In many applications a good choice of priors can be the so called sparse priors. This assumes that each row of the matrix $\boldsymbol{W}$ is populated primarily with zeros, which implies that they have supergaussian or sparse density. The conjugate sparse prior can be defined as

$$\log p(\boldsymbol{W}) = \sum_{i=1}^{T} \sum_{j=1}^{T} G(\boldsymbol{w}_i^T \boldsymbol{e}_j) + c \tag{1.30}$$

where c is a constant, $G(\cdot)$ is some nonquadratic function, which is used to measure the sparsity, $\boldsymbol{e}_j$ are normalized basis vectors, such that all the elements in $\boldsymbol{e}_j$ are zero except the j^{th} which is 1. One possible choice for $G(\cdot)$ is

$$G(s) = -|s|. \tag{1.31}$$

Thus the posterior will have the form

$$\log p(\boldsymbol{W}|\boldsymbol{x}) = \sum_{i=1}^{T} [G(\boldsymbol{w}_i^T \boldsymbol{x}) + \sum_{j=1}^{T} G(\boldsymbol{w}_i^T \boldsymbol{e}_j)] + c \tag{1.32}$$

Many nongaussian parametric densities have been used as prior densities for Bayesian estimation in ICA. A method using mixtures of gaussian densities as prior densities is discussed in Chapter 5.

1.4 Recent Developments of ICA

Most of the recent work done towards the development of ICA estimation methods is concentrated on finding more efficient estimators with better properties. As discussed above, in many efficient algorithms we need to use the probability density of the observations x_i. Because of the heavy computations needed for estimating densities, these algorithms were not practical enough for many applications, even though theoretically they gave better results. With the

development of technology and computational resources, it is becoming possible to implement these algorithms and find more trustable estimates for ICs.

For this matter, in ICA estimation problems a lot of research is being done in using nonparametric or semiparametric methods for estimating the density of the ICs which is then used in ML algorithm or mutual information algorithm. A more recent nonparametric approach to the linear model (2.1) given by Chen and Bickel (2006) is based on efficient score functions. The score functions of the sources are estimated by using B-spline approximations and the estimate of the unmixing matrix $\boldsymbol{W} = \boldsymbol{A}^{-1}$ is computed by a Newton-Raphson type optimization method. It is shown that the estimate is asymptotically efficient and performs better than some parametric methods in simulation studies in terms of minimizing the Amari error.

By the development of better algorithms for ICA estimation, the field of applications of these methods is expanding considerably. This trend is twofold; because of the issues encountered in some applied problems researchers often need to adjust the algorithms for specific cases which makes it necessary to incorporate new approaches for ICA estimation. This can be observed in such areas as brain imaging, feature extraction, source separation, etc.

1.5 Dissertation Structure

In this dissertation we review the issues concerning the identifiability of the Independent Component Analysis (ICA) model and propose a new semi-parametric method for ICA.

In Chapter 2 we discuss the identifiability of ICA. In Section 2.2 we present some characterization results from the book by Kagan et al. (1973) that lead to a set of conditions for partial identifiability of the model. Further, in Section 2.3 we discuss a set of additional constraints on the densities of the hidden sources that lead to full identifiability of ICA model.

Chapter 3 contains the research article "Smooth Density Estimation with Moment Constraints Using Mixture Distributions" that was submitted to the Journal of Nonparametric Statistics and is tentatively accepted for publication. In this Chapter we propose a novel method for density estimation using finite mixture densities. A modified EM algorithm for estimation of the mixture weights under constraints on the moments of the density is described. In addition, we propose a simple method based on a hypothesis test to obtain the number of components in the mixture density. The performance of the method is shown by using data generated from unimodal, skewed, multimodal densities and real datasets.

Based on the constraints needed for the identifiability of ICA given in Chapter 2 and the density estimation method developed in Chapter 3 we present a novel iterative semi-parametric model for ICA in Sections 4.1 and 4.2 in Chapter 4. The algorithm simultaneously estimates the densities of the hidden independent components and the mixture matrix. The method is compared with other commonly used ICA estimation algorithms by using simulated and real

datasets in Sections 4.3 and 4.4.

To conclude the dissertation we present some future research ideas in Chapter 5. A Bayesian model for ICA based on using finite gaussian mixture densities as the prior densities for the hidden sources is proposed in Section 5.1. The prior densities and consequently the posteriors are given in Sections 5.1.1 and 5.1.2. The model is implemented using the WinBUGS software and the performance is shown by using some simulated datasets in Section 5.1.3. Finally, Section 5.2 presents some ideas on applying the method to real datasets in brain imaging.

Chapter 2

Parameter Identifiability in ICA

2.1 Introduction

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. As discussed in Section 1.1.1, 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 a measure of dependence. Since in practice the data collected (e.g., in signal processing) are often nongaussian, the decorrelation approach does not usually result in adequate separation of the sources.

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

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{s} = (s_1, \ldots, s_m)^T$, $\boldsymbol{A} = [a_{ij}]_{n \times m}$ and $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ is an $n \times 1$ vector of independent gaussian noise variables each with mean 0. Writing $\boldsymbol{B} = [\boldsymbol{A}\ \boldsymbol{I}]$ and $\boldsymbol{y} = [\boldsymbol{s}^T\ \ \boldsymbol{e}^T]^T$ we can equivalently express (2.1) as

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y} \tag{2.2}$$

One of the issues that is partly unresolved in the literature on ICA is the identifiability of the model given in (2.1). 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 $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ of the same dimension are called equivalent if each column of $\boldsymbol{A}_1$ is proportional to one of the columns of $\boldsymbol{A}_2$ and vice versa. In other words, there exist an $m \times m$

permutation matrix $\boldsymbol{P}$, a diagonal matrix $\boldsymbol{\Lambda}$ with positive entries on its diagonal and a diagonal matrix $\boldsymbol{D}$ with diagonal entries equal to ± 1 such that

$$\boldsymbol{A}_2 = \boldsymbol{A}_1 \boldsymbol{P} \boldsymbol{D} \boldsymbol{\Lambda}.$$

Notice that $\boldsymbol{A}_2 \boldsymbol{s}_2 = \boldsymbol{A}_1 \boldsymbol{s}_1$ if we choose $\boldsymbol{s}_1 = \boldsymbol{P}\boldsymbol{D}\boldsymbol{\Lambda}\boldsymbol{s}_2$ for any two equivalent matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ which makes the representation (2.1) not identifiable if the goal is to estimate the matrix $\boldsymbol{A}$ and the densities of the independent components $s_1, \ldots, s_m$.

In most of the commonly used algorithms for ICA the fact that the model for ICA is not fully identifiable is often completely ignored (e.g., as in 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 $\boldsymbol{s}$ 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 $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ if the matrix $\boldsymbol{A}$ is $m \times m$ and of full column rank and the mutual information is used for estimation of the unmixing matrix $\boldsymbol{W}$.

2.2 Characterization Results

Kagan et al. (1973, p. 306) present several characterization results for random vectors having linear structure. These results are highly relevant to the ICA model and imply conditions for the existence of a solution to the model in (2.1) and its uniqueness up to matrix equivalence. To begin with, we quote some of the characterization results due to Kagan et al. (1973) in this Section. Suppose the n-dimensional random vector $\boldsymbol{x}$ has two representations given by $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ and $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y}$, where $\boldsymbol{s}$ is an $m \times 1$ and $\boldsymbol{y}$ is a $k \times 1$ random vector. We are interested in finding conditions on the matrix $\boldsymbol{A}$ and random vector $\boldsymbol{s}$ which would imply that the representation (2.1) is unique in the sense that the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are equivalent and the densities of the random vectors $\boldsymbol{y}$ and $\boldsymbol{s}$ belong to the same location-scale family. Notice, that for some cases the sizes of the vectors $\boldsymbol{s}$ and $\boldsymbol{y}$ may not be equal as shown in example 3.

First, it is shown that the random vector $\boldsymbol{x}$ is an $n \times 1$ normal vector if each column of $\boldsymbol{A}$ is not proportional to any column of $\boldsymbol{B}$ for any values of k and m. The following example shows this for a simple case.

Example 3: Suppose $n = 2$ and $\boldsymbol{x} = (x_1, x_2)^T$ has a bivariate normal distribution with

$E(\boldsymbol{x}) = \mathbf{0}$ and

$$\text{var}(\boldsymbol{x}) = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}.$$

Suppose the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are given by

$$\boldsymbol{A} = \begin{pmatrix} 2 & 0 \\ 5 & \sqrt{5} \end{pmatrix} \text{ and } \boldsymbol{B} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Then $\boldsymbol{x}$ has at least two different linear representations $\boldsymbol{x} = \boldsymbol{As}$ and $\boldsymbol{x} = \boldsymbol{By}$, where s_1, s_2 and $y_1, \ldots, y_4$ are independent standard normal random variables with mean zero and variance unity.

This in return provides a simple illustration of why the estimated sources should be non-gaussian. In the ICA model (2.1) if all of the sources s_j are gaussian then the model is not identifiable.

On the other hand, it can be shown that if all of the components of $\boldsymbol{s}$ in representation $\boldsymbol{x} = \boldsymbol{As}$ are nongaussian then there is no representation with fewer number of components, i.e. for any $\boldsymbol{y}$ such that $\boldsymbol{x} = \boldsymbol{By}$ and $k \leq m$.

By the theorem of uniqueness of decomposition we obtain that for any random vector $\boldsymbol{x} = \boldsymbol{Ay}$ if the columns of $\boldsymbol{A}$ are linearly independent then $\boldsymbol{x}$ can be written as a sum of two random vectors as follows

$$\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2,$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent vectors, $\boldsymbol{x}_1$ has unique representation and $\boldsymbol{x}_2$ is a multivariate normal vector with maximal variance. In other words, if we can find another such representation $\boldsymbol{x} = \boldsymbol{y}_1 + \boldsymbol{y}_2$ where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are independent, then $\boldsymbol{y}_2$ has a multivariate normal density and $\text{var}(\boldsymbol{x}_2) - \text{var}(\boldsymbol{y}_2)$ is a nonegative definite matrix. The following theorem follows from this result and provides a set of sufficient conditions for partial identifiability of the ICA model.

Theorem 2.2.1 *(Kagan et al. (1973)). Suppose $\boldsymbol{x}$ can be expressed as in (2.2) where the matrix $\boldsymbol{B}$ is such that the columns corresponding to the nongaussian components of $\boldsymbol{y}$ are linearly independent. Then $\boldsymbol{x}$ can be expressed as in (2.1)*

$$\boldsymbol{x} = \boldsymbol{As} + \boldsymbol{e},$$

where $\boldsymbol{e}$ has a multivariate normal distribution, $\boldsymbol{s}$ is a vector of nongaussian independent components and it is independent of $\boldsymbol{e}$ and $\boldsymbol{A}$ is of full column rank. The decomposition is unique in that if $\boldsymbol{x}$ has another representation given by $\boldsymbol{x} = \boldsymbol{A}_1\boldsymbol{s}_1 + \boldsymbol{e}_1$ then $\boldsymbol{e}$ and $\boldsymbol{e}_1$ have the same multivariate normal distribution, $\boldsymbol{s}$ and $\boldsymbol{s}_1$ have the same distribution except for change of scale

and location. The matrices $\boldsymbol{A}$ *and* $\boldsymbol{A}_1$ *are equivalent.*

2.3 Parameter Identifiability in ICA

Suppose that a vector of observed values $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ is known to be a mixture of some underlying independent sources $\boldsymbol{s} = (s_1, \ldots, s_m)^T$ as given in (2.1). The problem is the estimation of the matrix $\boldsymbol{A}$ and the densities of the underlying sources $s_1, \ldots, s_m$. The statistical estimation of the mixing matrix $\boldsymbol{A}$ (or its inverse) 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 (2.1). In this section we discuss the conditions under which the ICA model has a solution and it is unique.

By Theorem 2.2.1 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 $\boldsymbol{A}$ is of full column rank. In other words if $\boldsymbol{x}$ has two representations given by $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e}$ and $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{g} + \boldsymbol{e}_1$ then the vectors $\boldsymbol{e}$ and $\boldsymbol{e}_1$ have the same multivariate normal density, $\boldsymbol{A}$ and $\boldsymbol{B}$ are equivalent and $\boldsymbol{s}$ and $\boldsymbol{g}$ belong to the same location and scale family, i.e., if the joint density of the source vector $\boldsymbol{s}$ is given by $f_s(\cdot)$ then there exists a vector $\boldsymbol{a} \in \mathbb{R}^m$ and a scalar $b > 0$ such that the joint density of $\boldsymbol{g}$ can be expressed as $f_g(\boldsymbol{y}) = f_s[(\boldsymbol{y} - \boldsymbol{a})/b]$. Hence, there exist matrices $\boldsymbol{P}$, $\boldsymbol{D}$ and Λ, such that $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{D}\Lambda$ and $\boldsymbol{g} = \Lambda^{-1}\boldsymbol{D}^{-1}\boldsymbol{P}^T\boldsymbol{s}$, where $\boldsymbol{P}$ is a permutation matrix, Λ is a diagonal matrix with positive diagonal elements and $\boldsymbol{D}$ 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 2.3.1 *Suppose the mixing matrix* $\boldsymbol{A}$ *in noisy ICA model (2.1) is of full column rank and the independent sources are all nongaussian. Further, suppose all of the third moments of the sources exist and they satisfy the following conditions, for* $j = 1, \ldots, m$

$$E(s_j) = 0, \;\; E(s_j^2) = v_j, \;\; \textit{and } E(s_j^3) > 0, \tag{2.3}$$

where $0 < v_1 < \ldots < v_m$ *are fixed known quantities, then the ICA model is fully identifiable, in the sense if* $\boldsymbol{x}$ *has two representations given by* $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e} = \boldsymbol{B}\boldsymbol{g} + \boldsymbol{e}_1$*, where* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are each of full column rank and both* $\boldsymbol{s}$ *and* $\boldsymbol{g}$ *satisfy the conditions in (2.3), then* $\boldsymbol{A} = \boldsymbol{B}$ *and* $\boldsymbol{s} \stackrel{d}{=} \boldsymbol{g}$*.*

The details of the proof are given in Appendix A. Notice that the condition $E(s_j^3) > 0, j = 1, \ldots, m$ can be replaced by $E(s_j^3) < 0, j = 1, \ldots, m$.

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 2.3.1.

Corollary 2.3.2 *Suppose $s_1, \ldots, s_m$ in the ICA model (2.1) are positive valued random variables with $E(s_j) = 1$, for $j = 1, \ldots, m$. Then the model is fully identifiable if* $\mathrm{var}(s_1) < \ldots < \mathrm{var}(s_m)$ *are known and ordered.*

By similar arguments as in Theorem 2.3.1 another set of conditions for the identifiability of ICA model can be obtained.

Theorem 2.3.3 *Suppose the mixing matrix $\boldsymbol{A}$ in noisy ICA model (2.1) is of full column rank and the independent sources are all nongaussian each having mean zero and variance unity. Further, suppose all of the third moments of the sources exist and are ordered as follows*

$$E(s_1^3) > E(s_2^3) > \ldots > E(s_m^3).$$

Then the noisy ICA model (2.1) is fully identifiable.

Next we show that the sufficient conditions stated in Theorem 2.3.1 are minimal if we assume the existence of the third order moments for the source variables.

Theorem 2.3.4 *The conditions (2.3) given in Theorem 2.3.1 are minimal if the sources $s_1, ..., s_m$ in the ICA model (2.1) are assumed to have third order moments in the sense that any other set of source variables (with $E|s_j^3| < \infty$, for $j = 1, ..., m$) can be transformed to satisfy the sufficient conditions of Theorem 2.3.1.*

The proof of the theorem is given in Appendix B. The above result also facilitates the implementation of an algorithm where independent sources with mean zero can be transformed to satisfy the conditions (2.3).

Chapter 3

Smooth Density Estimation with Moment Constraints Using Mixture Distributions

3.1 Introduction

With the increasing amount of data being collected in many different fields of research the estimation of the true underlying density function of the data has become an important issue in statistical research. Scientists from diverse backgrounds have been trying to find good solutions for this problem, which lead to a number of different approaches discussed in the literature. One of the earliest nonparametric methods for density estimation was presented by Rosenblatt (1956). He suggested to use kernel based estimates for the unknown density. His approach was extended by Parzen (1962) who coined the term Kernel Density Estimator (KDE). If $X_1, ..., X_n$ are independent identically distributed random variables obtained from a continuous bounded density $f(\cdot)$ then the KDE is constructed as

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $K : \mathbb{R} \to [0, \infty)$ is a kernel function that integrates to one and is often assumed to be symmetric about zero. The bandwidth $h \equiv h(n)$ is the smoothing parameter which can be chosen to satisfy $h \to 0$ and $nh \to \infty$ as $n \to \infty$ to obtain a consistent estimator of $f(\cdot)$. A comprehensive overview of the properties of KDE can be found in Silverman (1986) and Devroye and Györfi (1985). In practice, the performance of the estimate highly depends on the selection of the bandwidth parameter h. Some of the commonly used methods are Silverman's rule

of thumb, least squares cross validation (Park and Marron, 1990) and the plug-in h selector proposed by Sheater and Jones (1991). One of the disadvantages of using these methods is that they are based on minimizing the mean integrated squared error (MISE), $E[\int(f-\widehat{f})^2]$. Kooperberg and Stone (1991) argue that L_2 distance may not be a suitable measure of distance between two density functions, for instance in detecting the modes of the density.

A common metric to measure the discrepancy between two densities is the Kullback-Leibler divergence (KLD) proposed by Kullback and Leibler (1951), which is defined as

$$K(f,g) = \int f(x) \log \frac{f(x)}{g(x)} dx, \tag{3.1}$$

where $f(\cdot)$ and $g(\cdot)$ are density functions. In KDE literature this metric has been used for bandwidth selection by cross-validation. However, Hall (1987) shows that under some conditions on the true density and the kernel function, the estimate of the bandwidth h does not have the correct first-order asymptotic properties.

Among the vast literature on density estimation there are methods that are based on the L_2 approximation of density using a system of orthonormal basis functions (e.g., wavelets). A major shortcoming of these methods is that often these estimates may not be proper density functions. The density function estimated by a series of orthonormal functions may not even be non-negative valued and if truncated may not integrate to unity making it unsuitable for many practical applications (e.g., random effects within linear mixed models). Moreover, the estimation methodology is based on L_2 approximation and our goal in this Chapter is to use KLD as the metric of approximation.

Another approach for solving the density estimation problem introduced by de Boore (1978) and Dierckx (1993) is based on using B-splines. Eilers and Marx (1996) extended this method to P-splines by introducing a penalty term on the likelihood function which needs to be maximized within a B-spline approach. Komárek et al. (2005) argue that instead of using a mixture of B-spline functions, a mixture of Gaussian densities can be considered for estimation. They also include a penalty term while maximizing the resulting likelihood. Eilers and Marx (1996) show that if the support of the estimated density is correctly chosen the proposed P-spline approach is not influenced by boundary effects as is for instance the KDE with compactly supported kernels and appropriately chosen bandwidths. Our method not only preserves the attractive behavior on the boundaries but also it may have infinite support as opposed to the B-spline based estimates. In addition, different densities can be used as components of the mixture density to obtain desired boundary conditions. For instance, the Epanechnikov kernel can be used to estimate a density with bounded support, or a mixture of gamma densities depending on the information available on the support of the true density (as long as conditions (A1) and (A2) of Theorem 3.2.1, in Section 3.2, are satisfied).

In this Chapter a method of estimation of the true density by using a mixture of known densities is proposed. We use the EM-algorithm to estimate the weights in the mixture density by minimizing the KLD. Dempster et al. (1977) proposed the Expectation-Maximization (EM) algorithm and showed that the algorithm converges to the maximum likelihood estimate under some regularity conditions. As shown in Section 3.3 the EM-algorithm simplifies the estimation of the weights in the mixture. One key difference of our model to other mixture models is that we use a suitable sequence of completely known densities with only the weights and the number of components being estimated from the data, under a set of moment constraints. In many real data examples different amounts of smoothness are required for different parts of the support of the original density. Since a given set of moments of the mixture densities is assumed fixed in our method, this choice can be *tailor-made* to account for different amounts of smoothing for specific intervals of the density to improve the estimation.

In density estimation using finite mixtures, an important issue is the choice of the number of components, N, in the mixture. Furman and Lindsay (1994) proposed a method for choosing the number of components based on moment estimators. However, they also stated some drawbacks of the method, for instance the asymptotic distribution of the test statistic for both the likelihood based and moment estimator based methods is unknown. In addition, there are some difficulties in computing the maximum likelihood estimates. Garel (2007) presents an overview of the current asymptotic results for testing normality against a two-component normal mixture. Since the mixture densities do not satisfy regularity conditions needed for general asymptotic theory, this critical issue is still not completely resolved especially when N is large and is allowed to depend on sample size. Sisson (2005) provides an excellent overview of Bayesian approaches to the problem by using transdimensional Markov Chain methods (Green, 1995). Despite the existence of these methods, in practice the number of components is usually chosen in *ad hoc* manner because of the computational burden. We present a formal statistical test for choosing the number of components in the mixture model needed for optimal approximation in Section 3.4. In Sections 3.5 and 3.6 we illustrate finite sample performance of our proposed density estimator using simulated data and our estimator is shown to outperform two popular density estimation methods not only in terms of minimizing KLD but also in terms of minimizing MISE. In addition, our method is computationally much more stable and efficient than the penalized likelihood method proposed by Komárek et al. (2005).

3.2 Mixture Densities with Moment Restrictions

In many practical applications often we have specific knowledge about the population moments. For instance, in linear mixed effects models the random effects are assumed to have mean zero and we may want to nonparametrically estimate the densities of the random effects that conserve

the zero mean (Ghidey et al., 2004). Also in the vast and popular literature on independent component analysis (ICA) often moment constraints are required for identifiability (Comon, 1994). In such cases, suppose we observe independent identically distributed (iid) random variables $X_1, ..., X_n$ obtained from a probability density $f(\cdot)$ and it is known that $E_f(X_1) = m_0$ and $\text{var}_f(X_1) = s_0^2$, where $E_f(\cdot)$ and $\text{var}_f(\cdot)$ denote the expectation and variance respectively, with respect to the unknown density $f(\cdot)$. Given any known values of m_0 and s_0^2 we can consider the transformed variables $X_j^* = (X_j - m_0)/s_0$. Then $X_1^*, ..., X_n^*$ are also i.i.d. and satisfy the conditions $E(X_1^*) = 0$ and $\text{var}(X_1^*) = 1$. Henceforth, without any loss of generality we assume that $m_0 = 0$ and $s_0^2 = 1$.

However, in some cases we may not have specific information about the true mean and variance of the underlying density. We can replace m_0 and s_0^2 by their consistent estimators. For instance, by consistency of the sample mean ($\overline{X}$) and sample variance (s^2) our proposed method can be used on location-scale transformed data $X_j^* = (X_j - \overline{X})/s$ (see Section 3.6 for several such applications). Other moment restrictions (including no restrictions) can be considered similarly, but for simplicity we illustrate our method for the first two moments of $f(\cdot)$ only (see Section 3.7 for a generalization).

Hall and Presnell (1999) propose a method for estimating a density under constraints by using a weighted bootstrap algorithm. The constraints include restrictions on the moments, quantiles or the entropy of the density function. A weighted version of the kernel density estimate is computed by using a biased-bootstrap method. The sampling weights are obtained subject to the constraints to minimize the power-divergence between the empirical distribution of the data and the estimated density. The authors note however, that using some types of distance measures may result in negative weights of the estimated density. In addition, the use of KDE requires finding an appropriate bandwidth selection method. Our proposed method does not lead to such issues and is more easily implemented using the constrained EM-algorithm.

For any integer $N \geq 1$, consider a sequence of known densities $f_{1,N}(x), ..., f_{N,N}(x)$ and define the mixture density as

$$f_N(x|\boldsymbol{\theta}) = \sum_{j=1}^{N} \theta_{j,N} f_{j,N}(x), \tag{3.2}$$

where $\theta_{j,N} \geq 0$ for $j = 1, ..., N$ and $\sum_{j=1}^{N} \theta_{j,N} = 1$. The goal is to find the vector of weights $\boldsymbol{\theta}_N = (\theta_{1,N}, ..., \theta_{N,N})^T$ such that the mixture in (4.4) can approximate $f(\cdot)$ as $N \to \infty$ using a suitable metric. In order to avoid the well known identifiability issue with mixture density, we assume that the sequence of means of the mixture components, $\mu_{j,N} = \int x f_{j,N}(x) dx$, are known and ordered, in other words, we assume that $\mu_{1,N} < ... < \mu_{N,N}$. We further assume that the mixture components have the common known variance $\sigma_N^2 = \int (x - \mu_{j,N})^2 f_{j,N}(x) dx$ which might depend on N. Finally, the mixture densities are assumed to have a common kernel

Table 3.1: *Construction of the mean sequence. For a given sample first we let* $\mu_{1,2} = x_{(1)}$, $\mu_{2,2} = x_{(n)}$. *For each iteration the median of the former two values is added as a new mean as shown by the arrow.*

	$x_{(1)}$								$x_{(n)}$
$\mathcal{M}_2$	$\mu_{1,2}$				$\downarrow$				$\mu_{2,2}$
$\mathcal{M}_3$	$\mu_{1,3}$		$\downarrow$		$\mu_{2,3}$		$\downarrow$		$\mu_{3,3}$
$\mathcal{M}_5$	$\mu_{1,5}$	$\downarrow$	$\mu_{2,5}$		$\mu_{3,5}$		$\mu_{4,5}$	$\downarrow$	$\mu_{5,5}$
$\mathcal{M}_7$	$\mu_{1,7}$	$\mu_{2,7}$	$\mu_{3,7}$	$\downarrow$	$\mu_{4,7}$	$\downarrow$	$\mu_{5,7}$	$\mu_{6,7}$	$\mu_{7,7}$
$\mathcal{M}_9$	$\mu_{1,9}$	$\mu_{2,9}$	$\mu_{3,9}$	$\mu_{4,9}$	$\mu_{5,9}$	$\mu_{6,9}$	$\mu_{7,9}$	$\mu_{8,9}$	$\mu_{9,9}$

function symmetric around zero. In other words, there exists a density function $\psi(\cdot)$ such that

$$f_{j,N}(x) = \frac{1}{\sigma_N}\psi\left(\frac{x-\mu_{j,N}}{\sigma_N}\right), \quad j = 1, ..., N.$$

For example, we can consider a mixture of Gaussian densities: $f_{j,N}(x)$ is then the density of a Gaussian random variable with mean $\mu_{j,N}$ and variance σ_N^2 for $j = 1, ..., N$. The moments of the Gaussian mixtures can be chosen appropriately to approximate an arbitrary continuous density $f(\cdot)$ satisfying some regularity conditions given in Theorem 3.2.1. Further, if the true density is known to be compactly supported we can choose $f_{j,N}(\cdot)$ to also be compactly supported.

We choose the sequence of means by maintaining a nested structure as N grows. The idea of basing the approximation of a density function on carrying the procedure over nested subsets of a countable set of functions was suggested by Grenander (1981). This method is often called the "method of sieves." Let $\mathcal{M}_N$ be a nested family of mixture densities with N mixture components where the mixture means are chosen as shown in Table 3.1. Then the sequence of nested families $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq ... \subseteq \mathcal{M}_N \subseteq ...$ forms the "sieve" in the case where the means are required to satisfy some regularity conditions (see Theorem 3.2.1 for more details).

A more challenging issue here is the choice of the number of components in the mixture. If the number of components, N, is too small then the mixture will obviously not be a good approximation of the density and will result in oversmoothing. On the contrary, having too many mixture components will result in having to estimate more parameters than necessary, which will result in an overparameterized model and hence undersmoothing. We provide a formal statistical test to resolve this issue in Section 3.4. Komárek et al. (2005) argue that a good choice for the standard deviation can be $\sigma_N = 2(\mu_{N,N} - \mu_{1,N})/[3(N-1)]$, which acts as a bandwidth parameter in the context of KDE.

In order to conserve the first two moments of the true density, we impose the same restrictions on the moments of the approximating mixture density. In general suppose X is a random

variable having density (4.4). For a fixed N, we consider another representation of X as follows.

$$\begin{aligned} X|Z=j &\sim f_{j,N}(\cdot) \\ P(Z=j) &= \theta_{j,N},\ j=1,...,N. \end{aligned} \tag{3.3}$$

Here the random variable Z acts as an index that identifies the components of the mixture. By using this representation we can derive the moments of the variable X as follows:

$$E(X) = E[E(X|Z)] = \sum_{j=1}^{N} \theta_{j,N}\mu_{j,N}$$

and the variance can be obtained as

$$\text{var}(X) = E[\text{var}(X|Z)] + \text{var}[E(X|Z)] = \sigma_N^2 + \sum_{j=1}^{N} \theta_{j,N}\mu_{j,N}^2 - \left(\sum_{j=1}^{N} \theta_{j,N}\mu_{j,N}\right)^2.$$

Later in Section 3.3, we use the representation (3.3) to develop EM-algorithm to estimate $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_N$. Hence, we can consider the following constraints on the weights of the mixture to conserve the moment restrictions.

$$\begin{aligned} &\textstyle\sum_{j=1}^{N} \theta_{j,N} = 1 \ \text{ and } \ \theta_{j,N} \geq 0, j=1,...,N, \\ &\textstyle\sum_{j=1}^{N} \theta_{j,N}\mu_{j,N} = 0, \ \text{and} \\ &\textstyle\sum_{j=1}^{N} \theta_{j,N}\mu_{j,N}^2 = 1-\sigma_N^2. \end{aligned} \tag{3.4}$$

In order to estimate the density $f(\cdot)$ by a mixture $f_N(\cdot|\boldsymbol{\theta})$ one approach could be to minimize the discrepancy between these two densities. The KLD of the true density and the estimated density can be written as $K(f, f_N) = \int f(x)\log[f(x)/f_N(x|\boldsymbol{\theta})]dx$. Notice that, for a fixed $N \geq 1$, $K(f, f_N)$ is a function of the mixing weights $\boldsymbol{\theta}$ which we require to satisfy the constraints (3.4). However, in practice this function cannot be minimized directly, because we do not know the true density function $f(\cdot)$. Instead, we generally have a random sample of observations $\boldsymbol{X} = (X_1, ..., X_n)^T$ obtained from the true density. Then we can consider an empirical estimate of $K(f, f_N)$,

$$D_n(N, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i)}{f_N(X_i|\boldsymbol{\theta})} \tag{3.5}$$

Since $D_n(N, \boldsymbol{\theta})$ is a strongly consistent estimator of $K(f, f_N)$ for any $\boldsymbol{\theta}$, then in practice, for a given N, we can minimize $D_n(N, \boldsymbol{\theta})$ to estimate the weights, $\boldsymbol{\theta}$, which is equivalent to

maximizing the log-likelihood function,

$$L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_N(x_i|\boldsymbol{\theta}). \tag{3.6}$$

It is well known that many other metrics measuring discrepancy between two densities are bounded above by $K(f, f_N)$. For example, consider the Hellinger distance between any two density functions f and g which is defined as

$$H(f,g) = \left[\int (\sqrt{f} - \sqrt{g})^2\right]^{1/2}.$$

The Hellinger distance and the KLD are well known to satisfy the inequality

$$H^2(f,g) \leq K(f,g).$$

Another commonly used metric is the L_p distance defined as $L_p(f,g) = \left(\int |f(x) - g(x)|^p dx\right)^{1/p}$, $p \geq 1$. Devroye and Györfi (1985) showed the equivalence of L_1 distance and Hellinger distance by using the following inequality

$$\left[\frac{1}{2} L_1(f,g)\right]^2 \leq H^2(f,g) \leq L_1(f,g).$$

Thus, the results obtained in terms of KLD can possibly be used to extend the method to other metrics.

The next question of interest is whether a mixture density of the form (4.4) can be used to approximate any continuous density function using the KLD. This is shown by the following result.

Theorem 3.2.1 *Let $f(\cdot)$ and $\psi(\cdot)$ be any continuous densities with respect to Lebesgue measure defined on $\mathbb{R}$ satisfying the following conditions:*

- $(A1)$ $0 \leq f(x) \leq M$ *for some* $M > 0$ *and* $x \in S = \{x \in \mathbb{R} : f(x) > 0\}$.
- $(A2)$ $|\int_S f(x) \log f(x) dx| < \infty$ *and* $|\int_S f(x)\{-\log \psi[(x-\mu)/\sigma]\}dx| < \infty$ *for any* $\mu \in S$ *and* $\sigma > 0$.
- $(A3)$ *There exists a sequence of known quantities* $\{\mu_{j,N} \in S,\ j = 1, ..., N\}$ *such that:*
  - (i) $\{\mu_{1,N} < ... < \mu_{N,N}\} \subseteq \{\mu_{1,N+1} < ... < \mu_{N+1,N+1}\},\ N = 2, 3, ...$
  - (ii) $\max\limits_{1 \leq j < N} (\mu_{j+1,N} - \mu_{j,N}) = o(1)$ *as* $N \to \infty$

Let $\boldsymbol{\Theta}_N = \{\boldsymbol{\theta} \in [0,1]^N : \sum_{j=1}^N \theta_j = 1\}$, *define* $f_N(x|\boldsymbol{\theta}_N) = \sum_{j=1}^N \theta_{j,N} \psi[(x - \mu_{j,N})/\sigma_N]/\sigma_N$,

where $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N$. *Further assume that* $\sigma_N > 0$ *satisfies* $\sigma_N = o(1)$ *as* $N \to \infty$. *Then*

$$K[f, f_N(\cdot|\boldsymbol{\theta}_N)] < \infty \text{ for } N = 2, 3, ... \text{ and } \min_{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N} K[f, f_N(\cdot|\boldsymbol{\theta}_N)] \downarrow 0 \text{ as } N \to \infty, \tag{3.7}$$

where $K[f, f_N(\cdot|\boldsymbol{\theta}_N)]$ *is the Kullback-Leibler divergence as defined in (3.1).*

The proof of the theorem is presented in the Appendix C. Thus any continuous density $f(\cdot)$ satisfying (A1) and (A2) can be approximated in the sense of (3.7) by using any sequence of $\mu_{j,N}, j = 1, ..., N$ satisfying (A3). Notice that $\sigma_N = c(\mu_{N,N} - \mu_{1,N})/(N-1)$ satisfies $\sigma_N = o(1)$ for any $c > 0$ such that $\sigma_N < 1$.

3.3 Mixture Weight Estimation Using the EM-algorithm

In this section, we develop an EM-algorithm to maximize the log-likelihood function given in (3.6) satisfying the constraints (3.4). Throughout this section we assume that $N \geq 1$ is an arbitrary but fixed integer and write $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_N$ to simplify notations. Suppose that $Z_1, ..., Z_n$ are unobserved random variables indicating the index of the component in the mixture as defined in (3.3), where $P(Z_i = j) = \theta_j$ for $j = 1, ..., N$ and $i = 1, ..., n$. Then letting the vector of observed values be $\boldsymbol{x} = (x_1, ..., x_n)^T$ and the vector of index variables be $\boldsymbol{z} = (z_1, ..., z_n)^T$ we can construct the likelihood function for the complete data $(\boldsymbol{x}^T, \boldsymbol{z}^T)^T$ as,

$$L_C(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \theta_{z_i} f_{z_i}(x_i).$$

Now we can implement the two steps of the EM-algorithm with a given starting value denoted by $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_N^{(0)})^T$. For each $k = 0, 1, ...$ we assume that we have the value of $\boldsymbol{\theta}^{(k)}$ and we calculate $\boldsymbol{\theta}^{(k+1)}$ by performing the E and M steps of the EM-algorithm.

E-step: Find the expectation of the log-likelihood given the observed values and the values of the coefficients from the previous iteration. By Bayes rule,

$$P(Z_i = j|\boldsymbol{x}, \boldsymbol{\theta}^{(k)}) = \frac{\theta_j^{(k)} f_j(x_i)}{\sum_{j=1}^{N} \theta_j^{(k)} f_j(x_i)} = w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x}).$$

Since $\sum_{j=1}^{N} P(Z_i = j|\boldsymbol{x}, \boldsymbol{\theta}^{(k)}) = 1$, it follows that for any $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{x}$, $\sum_{j=1}^{N} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x}) = 1$, for $i = 1, ..., n$. Hence, we can find the expectation of the log-likelihood of the complete data

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E_z\{\log[L_C(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z})]|\boldsymbol{x}, \boldsymbol{\theta}^{(k)}\} = E_z\left\{\sum_{i=1}^{n}[\log[f_{z_i}(x_i)] + \log(\theta_{z_i})]|\boldsymbol{x}, \boldsymbol{\theta}^{(k)}\right\}.$$

In the above, since the first term is constant with respect to $\boldsymbol{\theta}$ and depends only on the value calculated from the last iteration we will denote its expectation by

$$C(\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{N} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x}) \log[f_j(x_i)].$$

Then the Q function becomes

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = C(\boldsymbol{\theta}^{(k)}) + \sum_{j=1}^{N}\sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x}) \log(\theta_j). \tag{3.8}$$

M-step: Next we need to maximize the function Q in (3.8) under the constraints (3.4). We can use the Lagrange Multiplier method to achieve this. Incorporating constants λ_1, λ_2 and λ_3 we can build the Lagrangian function to be maximized as follows

$$Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + \lambda_1 \left(1 - \sum_{j=1}^{N} \theta_j\right) - \lambda_2 \left(\sum_{j=1}^{N} \theta_j \mu_j\right) + \lambda_3 \left(1 - \sigma_N^2 - \sum_{j=1}^{N} \theta_j \mu_j^2\right).$$

By taking the derivatives with respect to θ_j for $j = 1, ..., N$, λ_1, λ_2 and λ_3 we obtain

$$\frac{\partial Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})}{\partial \theta_j} = \frac{1}{\theta_j}\sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x}) - \lambda_1 - \lambda_2 \mu_j - \lambda_3 \mu_j^2 = 0 \tag{3.9}$$

$$\frac{\partial Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})}{\partial \lambda_1} = 1 - \sum_{j=1}^{N} \theta_j = 0, \quad \frac{\partial Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})}{\partial \lambda_2} = -\sum_{j=1}^{N} \theta_j \mu_j = 0$$

$$\frac{\partial Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})}{\partial \lambda_3} = 1 - \sigma_N^2 - \sum_{j=1}^{N} \theta_j \mu_j^2 = 0.$$

By finding θ_j from equation (3.9) and substituting the expression in the next three equations we obtain the following nonlinear system of equations for finding the values of λ_1, λ_2 and λ_3.

$$\begin{aligned}
&\textstyle\sum_{j=1}^{N} \frac{\sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x})}{\lambda_1 + \lambda_2 \mu_j + \lambda_3 \mu_j^2} = 1 \\
&\textstyle\sum_{j=1}^{N} \frac{\sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x})\mu_j}{\lambda_1 + \lambda_2 \mu_j + \lambda_3 \mu_j^2} = 0 \\
&\textstyle\sum_{j=1}^{N} \frac{\sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x})\mu_j^2}{\lambda_1 + \lambda_2 \mu_j + \lambda_3 \mu_j^2} = 1 - \sigma_N^2.
\end{aligned} \tag{3.10}$$

We can also note that by adding (3.9) for all j and considering the next three equations we obtain $\widehat{\lambda}_3 = (n - \widehat{\lambda}_1)/(1 - \sigma_N^2)$ and thus we need only to solve for λ_1 and λ_2 in (3.10). Since

there is no analytical solution for this system of equations, we will need to use some numerical method for finding the values $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ and $\widehat{\lambda}_3$ which satisfy (3.10) subject to the constraint $\widehat{\lambda}_1 + \widehat{\lambda}_2\mu_j + \widehat{\lambda}_3\mu_j^2 \geq 0$ for $j = 1, ..., N$ which are satisfied for any sequence of $\{\mu_1, ..., \mu_N\}$ if $\lambda_2^2(1-\sigma_N^2) \leq 4(n-\lambda_1)$. Then we substitute these values in the expression for $\boldsymbol{\theta}^{(k+1)}$ and obtain

$$\widehat{\theta}_j^{(k+1)} = \frac{\sum_{i=1}^{n} w_{ij}(\boldsymbol{\theta}^{(k)}, \boldsymbol{x})}{\widehat{\lambda}_1 + \widehat{\lambda}_2\mu_j + \widehat{\lambda}_3\mu_j^2}. \tag{3.11}$$

Notice that $\widehat{\theta}_j^{(k+1)} \geq 0$ for $j = 1, ..., N$ and satisfy the set of constraints (3.4). The iterations (3.11) are repeated until the algorithm converges, in other words, until the values of $\boldsymbol{\theta}^{(k)}$ in two consecutive iterations are very close, for instance, if $\max_{1\leq j\leq N} |\widehat{\theta}_j^{(k+1)} - \widehat{\theta}_j^{(k)}| \leq \epsilon$ for some small ϵ, say, for $\epsilon = 10^{-4}$. Once we find the converged value of $\widehat{\boldsymbol{\theta}}$ from the EM-algorithm we can construct the estimated density by using these coefficients. This density is represented as $f_N(x|\widehat{\boldsymbol{\theta}}) = \sum_{j=1}^{N} \widehat{\theta}_j f_j(x)$. Notice that $\widehat{\boldsymbol{\theta}}$ depends on the choice of N and hence we denote it by $\widehat{\boldsymbol{\theta}}_N$. Next we discuss the estimation of N by using a formal statistical test procedure.

3.4 A Test for the Number of Mixture Components

One of the harder issues to address here is the determination of the number of components in the mixture density which is denoted by N. Garel (2007) shows that if testing homogeneity against a two-component mixture there are well developed methods available, though these methods are computationally intensive and complicated. If the complexity of the mixture is larger the theory behind the existing testing methods is not well developed.

Heuristically, since our goal is to minimize $D_n(N)$ in (3.5) to find the coefficients, we can consider the minimum values of this function for different values of N and choose N which gives the minimum $D_n(N)$. In practice the function $\widehat{D}_n(N) = 1/n \sum_{i=1}^{n} \log[f(x_i)/f_N(x_i|\widehat{\boldsymbol{\theta}}_N)]$ cannot be directly minimized over N, since it depends on the true density $f(\cdot)$. But we can consider the successive differences,

$$\nabla\widehat{D}_n(N) = \widehat{D}_n(N) - \widehat{D}_n(N+1) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{N+1}(x_i|\widehat{\boldsymbol{\theta}}_{N+1})}{f_N(x_i|\widehat{\boldsymbol{\theta}}_N)}. \tag{3.12}$$

In general, it is expected that $\widehat{D}_n(N)$ is a decreasing function of N, since as the number of components increases the approximation could only improve (see Theorem 3.2.1) and hence the expected value of $\widehat{D}_n(N)$ would approach a limit, for example zero. Moreover, a proper choice of the sequence of known means of mixture components can give precise results for the behavior of this random variable. As mentioned in Section 3.2 we are choosing the mean sequences by maintaining the nested structure as N increases. This implies that the sequences of the means

of the mixture densities satisfy (A3).

Then $\nabla D_n(N)$ defined in (3.12) is a non-negative function of N which approaches zero as N increases. Hence, if the value of N is sufficiently large then $\nabla D_n(N) \approx 0$ and such a random value of N would be an optimal value of the number of mixture components. It is worth to mention that since the sequence of means is chosen instead of being estimated it is simple to achieve the nested structure of the means as described in Section 3.2. Starting from the first set in the sieve we find the best approximating density function for the observed data within the functions in the set and continue the process for each subsequent set until we find the density estimate that is close enough to the true density in terms of KLD which is then equivalent to $E[\nabla D_n(N)] \approx 0$.

The problem is to find the minimum N for which $E[\nabla D_n(N)]$ is not significantly different from zero. In other words we find, $\widehat{N} = \inf\{N \geq N_0 : E[\nabla D_n(N)]$ is not significantly different from $0\}$, where N_0 is an initial estimate of N such that $\sigma_N \in (0, 1)$ so that the third constraint in (3.4) is satisfied. For instance, if we choose $\sigma_N^2 = 2(\mu_{N,N} - \mu_{1,N})/[3(N-1)]$, we need $N_0 \geq 1 + 2(\mu_{N,N} - \mu_{1,N})/3$.

Thus the problem reduces to the testing of the following hypothesis

$$H_0 : E[\nabla D_n(N)] = 0, \tag{3.13}$$

for each $N \geq N_0$.

In order to derive a test statistic for (3.13), we can consider the random variables

$$\widehat{\Delta}_i = \log \frac{f_{N+1}(X_i|\widehat{\boldsymbol{\theta}}_{N+1})}{f_N(X_i|\widehat{\boldsymbol{\theta}}_N)}, \tag{3.14}$$

for $i = 1, ..., n$ and then we can calculate the sample mean $\overline{\Delta}_n = \sum_{i=1}^n \widehat{\Delta}_i/n$ and the sample variance $S^2_{\Delta_n} = \sum_{i=1}^n (\widehat{\Delta}_i - \overline{\Delta}_n)^2/(n-1)$. If the random variables $\widehat{\Delta}_1, ..., \widehat{\Delta}_n$ were i.i.d. then to test the hypothesis in (3.13) we could use the test statistic

$$z_n \equiv z_n(N) = \sqrt{n}\frac{\overline{\Delta}_n}{S_{\Delta n}} \tag{3.15}$$

and reject the null hypothesis in (3.13) if $z_n > z_\alpha$, where z_α denotes the $\alpha\%$ upper percentile of the standard normal density, where for example, we can take $\alpha = 0.10$.

However, each $\widehat{\Delta}_i$ depends on $\widehat{\boldsymbol{\theta}}_N$ which is a function of all the observations $X_1, ..., X_n$. So these random variables cannot be assumed independent. Theorem 3.4.1 shows how to overcome this issue.

Theorem 3.4.1 *If $\widehat{\Delta}_i$ are defined as in (3.14) then under the assumed regularity conditions (A1)-(A2) on $f(\cdot)$ and $\psi(\cdot)$ as in Theorem* 3.2.1*, for each $N \geq 1$, there exists a vector $\boldsymbol{\theta}_{0,N}$*

such that

$$\widehat{\Delta}_i = \log \frac{f_{N+1}(X_i|\boldsymbol{\theta}_{0,N+1})}{f_N(X_i|\boldsymbol{\theta}_{0,N})} + o_p(1) \ \textit{as} \ n \to \infty,$$

where

$$\boldsymbol{\theta}_{0,N} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_N}{\arg\max} \int f(x) \log f_N(x|\boldsymbol{\theta}) dx.$$

The proof is presented in Appendix D.

Next we define

$$\Delta_i^0 = \log \frac{f_{N+1}(X_i|\boldsymbol{\theta}_{0,N+1})}{f_N(X_i|\boldsymbol{\theta}_{0,N})}.$$

It follows that the random variables $\Delta_1^0, ..., \Delta_n^0$ are i.i.d. So by the Central Limit Theorem and by Theorem 3.4.1 we can use the test statistic given in (3.15) for testing the hypothesis of interest. Thus, our proposed estimate of N is given by

$$\widehat{N} \equiv \widehat{N}(\alpha) = \inf\{N \geq N_0 : z_n(N) \leq z_\alpha\}, \tag{3.16}$$

for some $\alpha \in (0, 1)$. Next we present finite sample performance of our estimator $\widehat{f}_N(x|\widehat{\boldsymbol{\theta}}_{\widehat{N}}) = \sum_{j=1}^{\widehat{N}} \widehat{\boldsymbol{\theta}}_{j,\widehat{N}} \psi[(x - \mu_{j,\widehat{N}})/\sigma_{\widehat{N}}]/\sigma_{\widehat{N}}$ where $\widehat{N}$ is obtained by (3.16) and $\widehat{\boldsymbol{\theta}}_{\widehat{N}}$ is obtained by (3.11).

3.5 Simulation Studies

In order to evaluate the performance of the algorithm described in the last section we use simulated datasets to explore how well the estimated mixture density approximates the underlying true density. First suppose we have n observations from the known density $f(\cdot)$ with mean 0 and variance 1. Then we can estimate the coefficients of the mixture density by the EM-algorithm described in Sections 3.3 and 3.4 and construct the estimated density. As discussed in Section 3.2 the KLD can be used as a metric of discrepancy between two densities. So here we calculate the KLD between the true density and the estimated density to measure the performance of the proposed method. Simultaneously, we compute the commonly used KDE as defined in Section 3.1 by using the option "bw.SJ" in the R software to find the bandwidth parameter using the algorithm in Sheater and Jones (1991). The Gaussian-spline estimate (GSE) using the method described in Komárek et al. (2005) is also computed by the R package `smoothSurv`. The comparison of the three methods shows that the proposed mixture density estimate (MDE) performs much better than the KDE or GSE not only in terms of minimizing KLD but also in terms of minimizing the L_2 distance between the true density and the estimated density.

For illustrating this we first generate 250 different samples each of size $n = 150$ from the true density $f(\cdot)$ and compute $f_{\widehat{N}}(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_{\widehat{N}})$ as the MDE for each sample using $\alpha = 0.05$ and then compute the KLD and MISE between $f_{\widehat{N}}(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}_{\widehat{N}})$ and $f(\cdot)$ using the "integrate" function in R

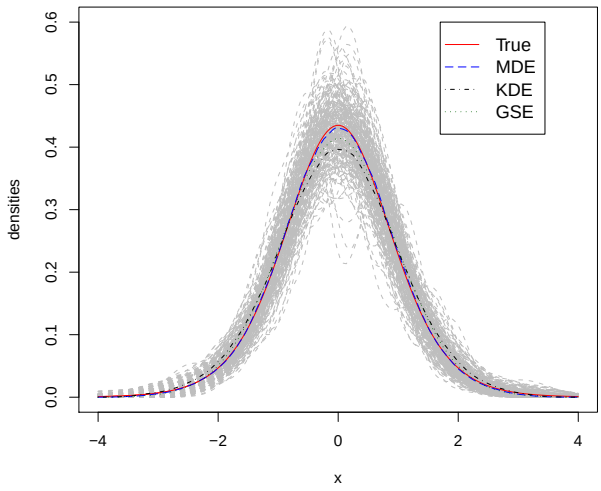


Figure 3.1: True and estimated densities for 250 samples of size 150 from scaled t density. The true density in solid (red) is overlaid with the median of the proposed density estimator MDE in dashed (blue), the median of KDE in dot-dashed (black) and the median of GSE in dotted (green). All the short-dashed (gray) lines are the estimated MDEs for 250 datasets.

software

$$KLD(f, \widehat{f}) = \int f(x) \log \frac{f(x)}{\widehat{f}(x)} dx \quad \text{and} \quad L_2(f, \widehat{f}) = \int [f(x) - \widehat{f}(x)]^2 dx$$

where $\widehat{f}$ denotes one of the three density estimates (MDE, KDE or GSE). To compute the KDE we used Gaussian kernel and "bw.SJ" method for bandwidth selection and the GSE is computed using the package `smoothsurv` in R software. The results are summarized in Figure 3.5 by presenting the boxplots of the two discrepancy measures over 250 Monte Carlo runs. In Figures 3.1-3.4 we present the median values of the estimated densities along with the true density and the cloud of all 250 MDEs. In almost all cases the EM-algorithm converged within 25 iterations on average.

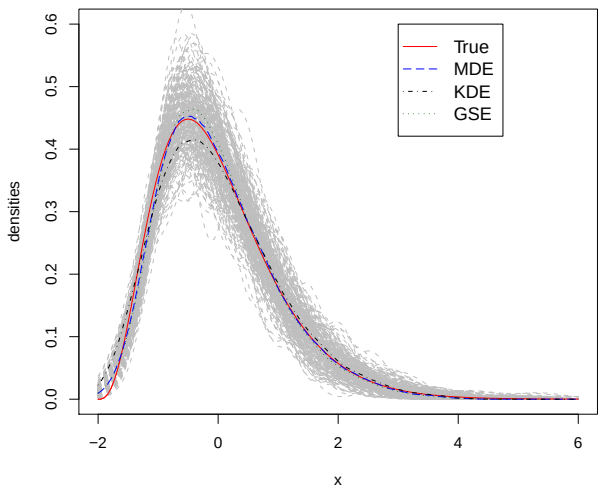


Figure 3.2: True and estimated densities for 250 samples of size 150 from shifted and scaled gamma density. The true density in solid (red) is overlaid with the median of the proposed density estimator MDE in dashed (blue), the median of KDE in dot-dashed (black) and the median of GSE in dotted (green). All the short-dashed (gray) lines are the estimated MDEs for 250 datasets.

We have used the following choices for the true density:

(a) Scaled t-distribution: $f(x) = \frac{\Gamma(5.5)}{(10\pi)^{1/2}\Gamma(5)}\left(1+\frac{x^2}{8}\right)^{-5.5}\left(\frac{10}{8}\right)^{1/2}$, $x \in \mathbb{R}$

(b) Shifted and Scaled Gamma distribution: $f(x) = \frac{4^4}{2\Gamma(4)}\left(\frac{1}{2}x+1\right)^3 e^{-4(x/2+1)}$

(c) Mixture of normal densities:
$f(x) = 0.4\frac{1}{(0.02\pi)^{1/2}}e^{-\frac{(x+1)^2}{0.02}} + 0.55\frac{1}{(0.5\pi)^{1/2}}e^{-\frac{(x-0.5)^2}{0.5}} + 0.05\frac{1}{(0.16\pi)^{1/2}}e^{-\frac{(x-2.5)^2}{0.16}}$, $x \in \mathbb{R}$

(d) Mixture of shifted and scaled Gamma densities:
$f(x) = 0.4\frac{5.4^4}{\Gamma(4)}(5x+6)^3e^{-4(5x+6)} + 0.55\frac{11.3}{\Gamma(2)}(0.71x+0.15)e^{-4(0.71x+0.15)}$
$+0.05\frac{20.2}{\Gamma(2)}(1.26x-2.65)e^{-4(1.26x-2.65)}$, $x \in (0,\infty)$

Notice that $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ for any $x > 0$ is the well known Gamma function and each of the above four densities satisfies the mean(=0) and variance(=1) constraints. The above choices in (c) and (d) are motivated by estimated densities obtained from real data that we present in Section 3.6.

It can be noted that when the true density is unimodal and symmetric as in (a) our method performs nominally better than the KDE both in terms of KLD and MISE and the GSE is

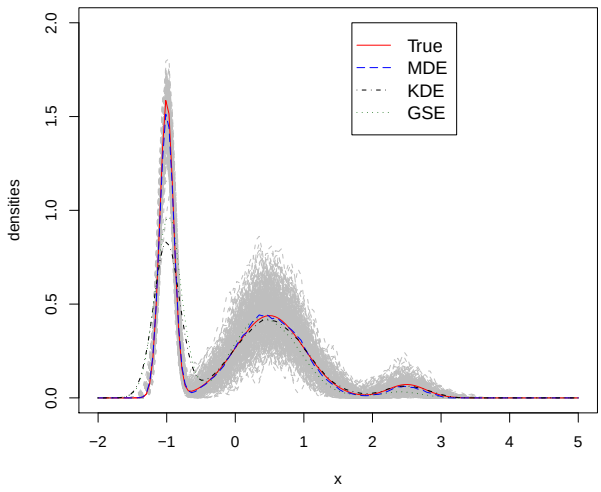


Figure 3.3: True and estimated densities for 250 samples of size 150 from a mixture of Normal densities. The true density in solid (red) is overlaid with the median of the proposed density estimator MDE in dashed (blue), the median of KDE in dot-dashed (black) and the median of GSE in dotted (green). All the short-dashed (gray) lines are the estimated MDEs for 250 datasets.

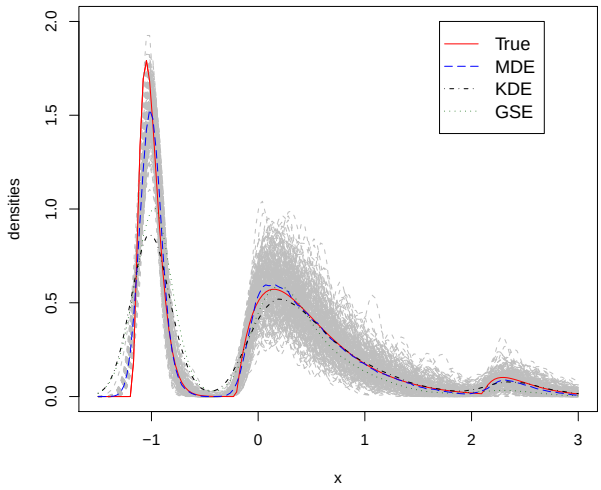


Figure 3.4: True and estimated densities for 250 samples of size 150 from a mixture of shifted and scaled gamma densities. The true density in solid (red) is overlaid with the median of the proposed density estimator MDE in dashed (blue), the median of KDE in dot-dashed (black) and the median of GSE in dotted (green). All the short-dashed (gray) lines are the estimated MDEs for 250 datasets.

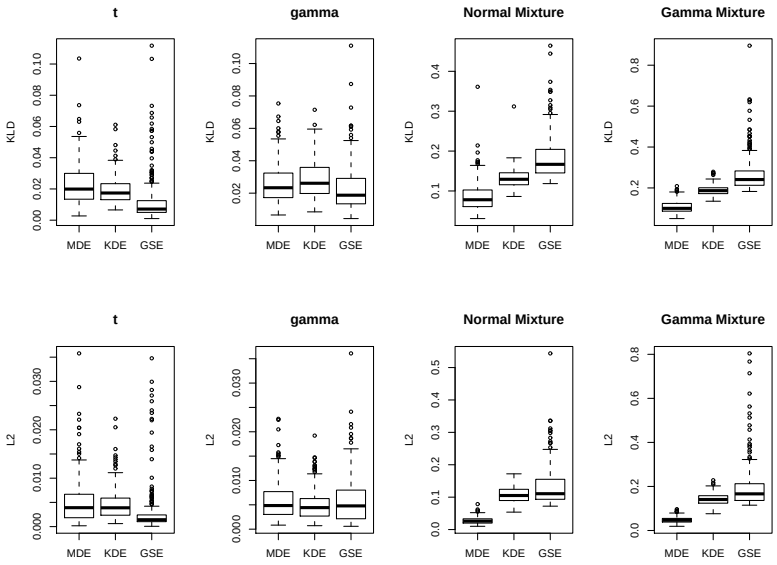


Figure 3.5: Boxplots of the KLD and MISE for the density estimation methods: the proposed MDE, the KDE and the GSE.

performing better than the other two methods. When the true density is unimodal and skewed as in case (b) the three methods seem to perform similarly in terms minimizing both KLD and MISE. However, if the true density has several modes as is the case with many real applications we observe that our proposed estimate of the density performs almost uniformly better than the KDE and the GSE in terms of both KLD and MISE. For instance, when the true density is as in (c), then median KLD between the true and our proposed density estimates is about 0.08 which lies below the 2.5 percentile (0.09) of the KLD between the true density and the KDE and the 2.5 percentile (0.12) of KLD between the true density and the GSE. Further in terms of MISE the median distance between true density and MDE is only 0.03 which is again lower than the 2.5 percentile of the MISE between true density and KDE (0.07) and the MISE between the true density and GSE (0.08). The same is true in case (d) where the proposed MDE outperforms the KDE and the GSE in terms of minimizing both KLD and MISE by large magnitudes (see Figure 3.5). The GSE seems to oversmooth the density in cases (c) and (d).

3.6 Real Data Application

Here we consider two different datasets from practice and use our method to estimate the underlying density for each case. We compute the approximate KLD and MISE as numerical measures and also compare the results visually by overlaying the three estimated densities on a histogram.

First dataset that we consider here is related to the enzymatic activity in the blood. This dataset contains $n = 245$ observations from 245 unrelated people. The enzymatic activity is measured for an enzyme which has a part in metabolism of carcinogenic substances. This dataset has been analyzed by Bechtel et al. (1993), where they suggested that the underlying density might have a bimodal shape. It was subsequently analyzed by Richardson and Green (1997) using Bayesian density estimation approach. We obtained the dataset from the following website:

`http://www.stats.bris.ac.uk/~peter/mixdata`

We first calculate the mean and the standard deviation of the data and rescale the data to have an empirical mean of zero and variance one. By looking at the histogram of the data in Figure 3.6, we observe that the underlying density might be asymmetrical having one global mode at a negative value and a few local modes at positive values. Next we estimate the density using our method MDE, the KDE and the GSE using the rescaled data. For this dataset our estimated number of components is $\widehat{N} = 38$ using $\alpha = 0.15$. It is clear that our proposed MDE

is capturing the peaks of the histogram better than that of the KDE (with bandwidth selected by "bw.SJ" method) and GSE.

Next we analyze one of the commonly used datasets in density estimation problems. It contains the velocities of galaxies diverging from Milky Way with $n = 82$ observations. The histogram of these data in Figure 3.7, looks approximately symmetric with two global modes and possibly two local modes at the two tails. In Figure 3.7, we plot the three density estimates overlaid on the histogram. It can be observed that again our proposed mixture density estimate is capturing the peaks of the histogram very well. In this case the estimated number of components is $\widehat{N} = 30$ using $\alpha = 0.15$, starting with initial $N_0 = 20$.

In general, since we do not know the underlying true density of these data, we cannot calculate the KLD of the true density and the estimated density to evaluate the fit. However, the use of KLD as a measure of discrepancy allows us to compare two different density estimates of the true density. If $\widehat{f}_n$ denotes the KDE/GSE and $f_{\widehat{N}}$ denotes our proposed MDE then it follows that

$$\begin{aligned} KLD(f, \widehat{f}_n) - KLD(f, f_{\widehat{N}}) &= \int f(x) \log \frac{f_{\widehat{N}}(x|\widehat{\boldsymbol{\theta}})}{\widehat{f}_n(x)} dx \\ &= \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_{\widehat{N}}(X_i|\widehat{\boldsymbol{\theta}})}{\widehat{f}_n(X_i)} + o_p(1), \text{ as } n \to \infty. \end{aligned} \tag{3.17}$$

Thus we may compare two density estimates $\widehat{f}_n$ and $f_{\widehat{N}}$ by calculating the first term in (3.17). For the galaxy data this quantity is 0.08 when $\widehat{f}_n$ is the KDE and 0.15 when $\widehat{f}_n$ is GSE and for enzyme data we have 0.06 against KDE and 0.11 against GSE which suggest that the proposed estimator $f_{\widehat{N}}$ provides a better fit as compared to both KDE and GSE.

In addition, we can find an approximation for the difference of MISE of the two density estimates from the true density based on the observed data. Notice that

$$L_2(f, \widehat{f}_n) - L_2(f, f_{\widehat{N}}) = \int [f(x) - f_{\widehat{N}}(x|\widehat{\boldsymbol{\theta}})]^2 dx - \int [f(x) - \widehat{f}_n(x)]^2 dx.$$

Hence by empirical approximation we obtain

$$L_2(f, \widehat{f}_n) - L_2(f, f_{\widehat{N}}) = \int [\widehat{f}_n(x) - f_{\widehat{N}}(x|\widehat{\boldsymbol{\theta}})]^2 dx - \frac{2}{n} \sum_{i=1}^{n} [\widehat{f}_n(X_i) - f_{\widehat{N}}(X_i|\widehat{\boldsymbol{\theta}})] + o_p(1).$$

For the enzyme data this quantity is equal to 0.32 when $\widehat{f}_n$ is chosen to be KDE and 0.21 when $\widehat{f}_n$ is taken to be GSE and for galaxy data this difference is 0.13 against KDE and 0.11 against GSE. Hence both KLD and MISE between the true density and KDE/GSE are expected to be larger than that of the KLD and MISE between the true density and our estimate MDE for both datasets respectively. Thus, in real applications even if we do not have exact knowledge of the

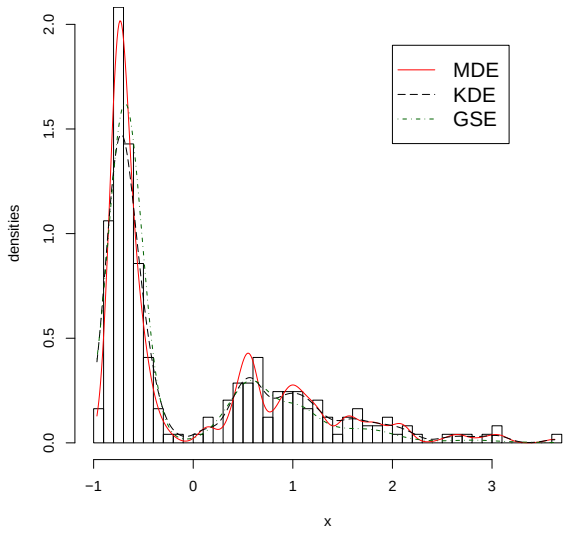


Figure 3.6: For the enzyme dataset, the solid (red) line shows the proposed MDE and the dashed (black) line shows the estimated density by using KDE, the dotted (green) shows the GSE.

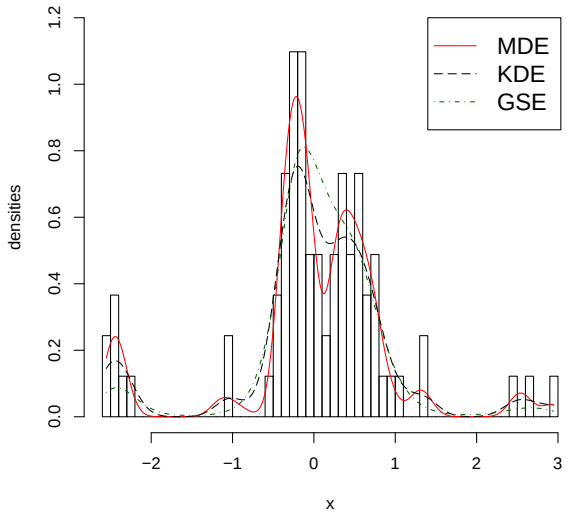


Figure 3.7: For the galaxy dataset, the solid (red) line shows the proposed MDE and the dashed (black) line shows the estimated density by using KDE, the dotted (green) shows the GSE.

true mean and variance of the population, we can replace those moments by their consistent estimates and apply our mixture method to obtain a better estimate of the density that is expected to have lower KLD and MISE.

3.7 Conclusions

In this Chapter we considered the density estimation problem in the case when we may have *a priori* information about the moments of the underlying density of the data. The idea was motivated by the conjecture that if we are using some moment information about the data in estimating the density then we can possibly obtain a better estimate for the true density. In particular, we illustrated the use of a mixture density with known component densities. The coefficients or weights of the mixture components were estimated by a constrained EM algorithm which made the estimation much faster and more efficient, for instance convergence was obtained within 25 iterations. The problem of estimating the number of components in the mixture density was also addressed. We proposed an automatic and iterative algorithm for finding the number of components of the mixture density by conducting a hypothesis test to check whether additional components were necessary or not. At the convergence we obtain an estimate for the number of components and subsequently avoid the problem of choosing a bandwidth of KDE.

The method was validated by applying it to several simulated datasets. The data were generated from various densities having symmetric or multimodal shapes. For each case the mixture density estimate was computed and compared with the KDE widely used in density estimation problems. The simulation studies indicate that our method performed almost uniformly better than the KDE method. In fact we observed that when the true density has more than one mode and when the mean and variance of the true underlying density are known our method is almost uniformly better than the KDE in terms of minimizing of KLD and the L_2 distance between the densities.

Our approach remains applicable even when true moments are not known. By applying our method to the real datasets we showed that this method can be extended to the case where we do not have exact information about the moments of the density as long as we assume that these moments exist. Instead of using the true mean and variance we may use their consistent estimators in practice. This will result in a more universally applicable solution for density estimation problem.

In the Chapter we considered the case when the first two moments of the density are known. However, we might want to estimate the density under a more general moment constraint given by $E_f\{g(X)\} = g_0$, where $g(\cdot)$ is a known function such that $E_f(|g(X)|) < \infty$ and g_0 is a given constant. Our algorithm can be easily extended to this case, with the following modification in

the system of equations (3.10). Notice that

$$E_{f_N}\{g(X)\} = E[E\{g(X)|Z\}] = \sum_{j=1}^{N} \theta_j \int g(x) f_j(x) dx = g_0. \tag{3.18}$$

Consequently, we can find the system of equations (3.10) for determining $\boldsymbol{\theta}_N$ by minimizing the loglikelihood in (3.6) subject to the constraints (3.4) and (3.18). Hence, the estimation method described in Section 3.4 can be easily extended to obtain the estimate of $\boldsymbol{\theta}_N$ subject to any type of generalized moment constraints.

When describing the mixture density used for estimation in Section 3.2 we mentioned that any sequence of densities satisfying (A2)-(A3) can be used as the components of the mixture density if the mean and variance structure is preserved. However, in the simulation studies we only considered a sequence of Gaussian densities as components of the mixture density. While this is a reasonable choice clearly this might not be the best sequence for some cases when it is known to have finite support. Some other densities with restricted support can be used as mixture components such as the Epanechnikov kernel, gamma density, etc if we have information about the support of the true density. In addition, while proving the theoretical results we assumed that the underlying density of the data is continuous and univariate. It would be interesting to consider the extension when the underlying density is discrete or multivariate and to extend the method to such mixed multivariate density estimation.

Chapter 4

Semi-Parametric Model for Independent Component Analysis

4.1 A Semiparametric ICA Model

As defined in Chapter 2 in matrix notation, a model for ICA can be written as,

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e}, \tag{4.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{s} = (s_1, \ldots, s_m)^T$, $\boldsymbol{A} = [a_{ij}]_{n\times m}$ and $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ is an $n \times 1$ vector of independent gaussian noise variables each with mean 0.

By Theorem 2.3.1 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 $\boldsymbol{A}$ identifiable. In Chapter 3 we developed a flexible class of models based on a mixture of densities (for instance gaussian densities) 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 $\boldsymbol{s}$.

Again, without any loss of generality we assume throughout this chapter that $E(\boldsymbol{x}) = \boldsymbol{0}$ and $E(\boldsymbol{s}) = \boldsymbol{0}$. Also, following the previous works on ICA, for the rest of the chapter we assume that $n = m$ (but see our remarks in Section 4.5). In many ICA algorithms such as the FastICA described by Hyvarinen and Oja (2000) it is further assumed that $\boldsymbol{e} = 0$ and the model is called *noise free*. We will make this assumption as well.

Following the discussions in Chapter 3, we propose to estimate each of the source densities

f_j by the following mixture of densities

$$f_j(s) = \sum_{k=1}^{N_j} \theta_{jk} \phi \left(\frac{s - \mu_{jk}}{\sigma_{N_j}} \right) \frac{1}{\sigma_{N_j}}, \tag{4.2}$$

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 that satisfies a set of regularity conditions stated in Theorem 3.2.1. 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 (2.3) in Theorem 2.3.1). In particular, in order to satisfy the set of three conditions given in (2.3), we estimate the θ_{jk}'s subject to the following necessary conditions:

$$\begin{aligned} &(i) \quad \sum_{k=1}^{N_j} \theta_{jk} \mu_{jk} = 0, \ \sum_{k=1}^{N_j} \theta_{jk} \mu_{jk}^2 = v_j - \sigma_{N_j}^2 \text{ and } \sum_{k=1}^{N_j} \theta_{jk} \mu_{jk}^3 > 0, \\ &(ii) \quad \sum_{k=1}^{N_j} \theta_{jk} = 1 \text{ and } \theta_{jk} \geq 0. \end{aligned} \tag{4.3}$$

Clearly the first set of restrictions (i) correspond to the three conditions in (2.3) and the last set of conditions (ii) in (4.3) is needed to ensure that the mixture density in (4.2) is a legitimate probability density function. A constrained EM algorithm (see Section 3.3) can be used to estimate the θ_{jk}'s subject to restrictions given in (4.3). One interesting feature of this density estimation methodology is that not only the weights θ_{jk} are estimated but also the number of mixture components N_j for each of the source densities are also being estimated. This feature makes the density estimation method adaptive to various shapes of the estimated densities of underlying nongaussian sources. Next, we describe a method to simultaneously estimate the unmixing matrix $\boldsymbol{W}$ and the weights θ_{jk} for a *completely known* sequence of the knots μ_{jk} and bandwidth 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 and not requiring any tuning parameter selection.

In matrix notation the *noise free* ICA model is given by

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A},$$

where $\boldsymbol{X}$ is the $T \times m$ matrix of observed values (signals), $\boldsymbol{S}$ is the $T \times m$ matrix of underlying (hidden) sources and $\boldsymbol{A}$ is the $m \times m$ unknown mixing matrix. In addition, suppose that the mixing matrix is nonsingular and define its inverse as $\boldsymbol{W} = \boldsymbol{A}^{-1}$. We further assume that the densities of the independent components are given by $s_{ij} \sim f_j$, for each $j = 1, \ldots, m$ and hence

the column $\boldsymbol{s}_j$ is a sample of T iid variates from the density f_j.

Suppose $\widehat{\boldsymbol{W}}^{(0)}$ is an initial estimate for $\boldsymbol{W}$ found by some fast estimation method (e.g. SVD, fastICA or JADE). 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, $\{\mu_{j1}^{(0)}, \ldots, \mu_{jN_j^{(0)}}^{(0)}\}$ be a starting set of known means and $\sigma_{N_j}^{(0)}$ a known common variance for the mixture density.

Our goal is to estimate the true unknown unmixing matrix $\boldsymbol{W}_0$ and the densities of the sources f_j simultaneously using an iterative method. For the iteration step $M \in \{1, 2, \ldots\}$, we find an estimate of the matrix of independent sources as $\widehat{\boldsymbol{S}}^{(M)} = \boldsymbol{X}\widehat{\boldsymbol{W}}^{(M-1)}$. 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 $\mathcal{M}_{N_j^{(M-1)}} \subset \mathcal{M}_{N_j^{(M)}}$. Further details for constructing the means that conserve the sieve structure can be found in Table 3.1. Further, the constrained EM-algorithm as described in Section 3.3) 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

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

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

$$\widehat{f}_j^{(M)}(s) = \sum_{k=1}^{N_j^{(M)}} \widehat{\theta}_{jk}^{(M)} \phi\left(\frac{s - \mu_{jk}^{(M)}}{\sigma_{N_j}^{(M)}}\right) \frac{1}{\sigma_{N_j}^{(M)}}, \tag{4.4}$$

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 (4.4).

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

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

where $\boldsymbol{f} = (f_1, \ldots, f_m)$ is the vector of densities of hidden sources. By using the estimates given in (4.4) and writing $\widehat{\boldsymbol{f}} = (\widehat{f}_1, \ldots, \widehat{f}_m)^T$ the loglikelihood function of the unmixing matrix

is given by

$$L(\boldsymbol{W}, \widehat{\boldsymbol{f}}) = \sum_{i=1}^{T} \sum_{j=1}^{m} \log \left[\sum_{k=1}^{N_j^{(M)}} \widehat{\theta}_{jk}^{(M)} \phi \left(\frac{\sum_{l=1}^{m} x_{il} w_{lj} - \mu_{jk}^{(M)}}{\sigma_{N_j}^{(M)}} \right) \frac{1}{\sigma_{N_j}^{(M)}} \right] + T \log |\det \boldsymbol{W}|. \quad (4.5)$$

Notice that by the choice of the estimating densities of original sources the gradient vector $\nabla L(\boldsymbol{W}, \widehat{\boldsymbol{f}})$ and hessian matrix $\nabla^2 L(\boldsymbol{W}, \widehat{\boldsymbol{f}})$ of the loglikelihood above can be computed analytically and are given in Appendix F. Hence by using a hill-climbing version of the Newton-Raphson algorithm (see Section 4.2 for more computational details) an update of the unmixing matrix can be computed as follows

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

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 $x \in \text{supp}(f) = \{x \in \mathbb{R} : f(x) > 0\}$.

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

We assume that the true source densities f_j satisfy the identifiability conditions (2.3) given in Theorem 2.3.1. The following result provides a set of regularity conditions under which we establish the consistency of the estimator of $\boldsymbol{W}$ obtained by maximizing the log-likelihood function given by (4.5) and simultaneously estimating the densities $\widehat{f}_1, \ldots, \widehat{f}_m$.

Theorem 4.1.1 *Suppose in the ICA model (4.1) 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. *A sequence of known quantities σ_{N_j} satisfies $\sigma_{N_j} = o(1)$ as $N_j \to \infty$.*

4. *The estimated densities $\widehat{f}_j \in \mathcal{F}_N$ and satisfy the constraints (2.3).*

5. *The true mixing matrix $\boldsymbol{A}$ is nonsingular.*

If $KLD(f_j, \widehat{f}_j) \to 0$ *as* $N_j \to \infty$ *and* $\widehat{\boldsymbol{W}} = \arg\max L(\boldsymbol{W}, \widehat{f})$ *then*

$$\widehat{\boldsymbol{W}} \to \boldsymbol{W}_0 \text{ almost surely, as } T \to \infty.$$

where $\boldsymbol{W}_0$ *is the true value of the inverse of mixing matrix, i.e.* $\boldsymbol{W}_0 = \boldsymbol{A}_0^{-1}$.

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

4.2 An Iterative Method to Compute the MLE of $\boldsymbol{W}$

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

$$\boldsymbol{C}_x = \boldsymbol{A}\boldsymbol{C}_s\boldsymbol{A}^T,$$

where $\boldsymbol{C}_s = \text{cov}(\boldsymbol{S}) = \text{diag}(v_1, \ldots, v_m)$ given by constraints (2.3) and $v_1, \ldots, v_m$ are known variance components. Let $\boldsymbol{A}_1 = \boldsymbol{A}\boldsymbol{C}_s^{1/2}$, hence $\boldsymbol{C}_x = \boldsymbol{A}_1\boldsymbol{A}_1^T$. By spectral decomposition we obtain $\boldsymbol{C}_x = \boldsymbol{Q}\Lambda\boldsymbol{Q}$, where Λ is the diagonal matrix of eigenvalues of $\boldsymbol{C}_x$ and $\boldsymbol{Q}$ is an orthogonal matrix of corresponding eigenvectors satisfy $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. This implies that a good choice for $\boldsymbol{A}_1$ can be obtained as $\boldsymbol{A}_1 = \boldsymbol{Q}\Lambda^{1/2}\boldsymbol{Q}^T$. We choose a starting value for $\boldsymbol{W}$ as

$$\widehat{\boldsymbol{W}}^{(0)} = [\boldsymbol{A}_1\boldsymbol{C}_s^{-1/2}]^{-1}.$$

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

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

1. Let $\boldsymbol{S}^{(M)} = \boldsymbol{X}\boldsymbol{W}^{(M)}$.
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 in Chapter 3 estimates the density subject to constraints on the mean and variance of the random variable. However, we may transform the $\boldsymbol{S}^{(M)}$ to satisfy the third moment constraints by Theorem 2.3.4.

4. Compute the gradient $\nabla L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{f}})$ and hessian matrix $\nabla^2 L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{f}})$ (see Appendix F for exact analytical expressions).

5. Update the unmixing matrix by setting $\widehat{\boldsymbol{W}}^{(M+1)} = \widehat{\boldsymbol{W}}^{(M)} - \nabla^2 L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{f}})^{-1} \nabla L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{f}})$.

6. If $L(\widehat{\boldsymbol{W}}^{(M+1)}, \widehat{\boldsymbol{f}}) < L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{f}})$, set $\widehat{\boldsymbol{W}}^{(M+1)} = \widehat{\boldsymbol{W}}^{(M)}$ and repeat steps 3-6. In other words increase the number of mixture components and implement steps 3-6 until a new value of $\widehat{\boldsymbol{W}}$ is obtained. Otherwise return to step 1.

Repeat the steps 1-6 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 $\epsilon = 10^{-3}$.

4.3 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 non-gaussian 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
$f_1(s) = 3 \left(s\sqrt{\Gamma(5/3) - \Gamma(4/3)^2} + \Gamma(4/3)\right)^2 e^{-(s\sqrt{\Gamma(5/3)-\Gamma(4/3)^2}+\Gamma(4/3))^3}$,
$f_2(s) = 3 \left(\frac{s}{2}\sqrt{\Gamma(5/3) - \Gamma(4/3)^2} + \Gamma(4/3)\right)^2 e^{-(s\sqrt{\Gamma(5/3)-\Gamma(4/3)^2}/2+\Gamma(4/3))^3}$.

(c) Shifted and scaled Gamma and a mixture of normals
$f_1(s) = \frac{4^4}{2\Gamma(4)} \left(\frac{1}{2}s + 1\right)^3 e^{-4(s/2+1)}$,
$f_2(s) = 1.2\phi(2(x+2)) + 0.8\phi(2(x-3))$,

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 (2.3) required for the identifiability of ICA, in other words the means are equal to 0, the variances are increasing for each case and the third moments of all of the above densities are positive. 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 performances of each of the four methods are evaluated by a commonly used error criterion in signal processing literature called the Amari error (Amari, 1998). As discussed in Section 1.2, for a given known mixing matrix $\boldsymbol{A}$ and estimated unmixing matrix $\widehat{\boldsymbol{W}}$ the Amari error is defined as

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

where $\boldsymbol{P} = \boldsymbol{A}\widehat{\boldsymbol{W}}$. Again, since the Amari error is not invariant to a constant multiplier before computing this error we rescale the columns of the matrices $\boldsymbol{A}$ and $\widehat{\boldsymbol{W}}$ 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

$$leff(\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2) = \log\frac{AE(\boldsymbol{A}, \widehat{\boldsymbol{W}}_1)}{AE(\boldsymbol{A}, \widehat{\boldsymbol{W}}_2)}, \tag{4.6}$$

where $\widehat{\boldsymbol{W}}_2$ is the estimate obtained by our proposed MICA method and $\widehat{\boldsymbol{W}}_1$ is the estimated unmixing matrix using one of the three methods FICA, JADE or PICA. Clearly, if $leff(\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2) > 0$ then MICA performs better than its competitor. The larger the value of $leff(\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2)$, 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 4.1 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 has similar efficiency compared to PICA for the case (b) where both source vectors are generated using weibull density function and performs nominally better than FICA and JADE. However, if the sources have more skewed underlying densities (as in (a)) or a bimodal density (as in (c)) the MICA algorithm performs significantly better than all three competing methods. In particular, in cases (a) and (c) 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 0.009 which is smaller than that of FICA with median $AE = 0.021$, PICA with median $AE = 0.032$ and JADE with median $AE = 0.043$. Thus, it appears that our proposed method performs substantially better than all three methods when one of the sources has non-unimodal or highly skewed distribution.

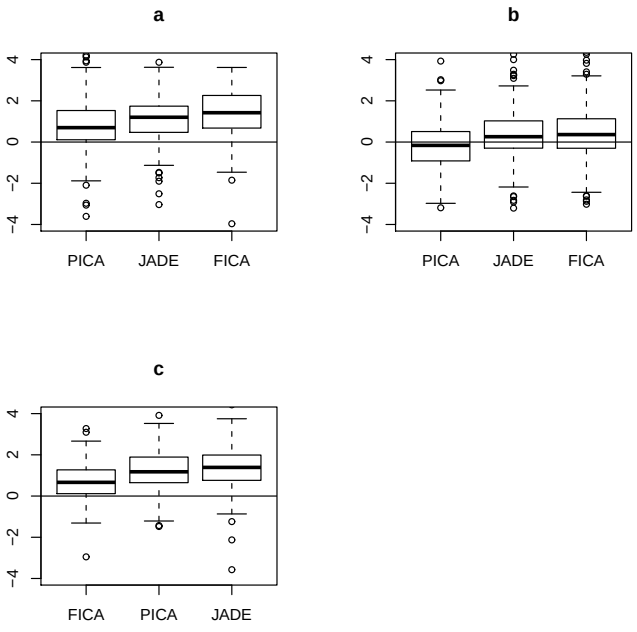


Figure 4.1: Boxplots of the log-efficiencies of our proposed MICA as compared to the three commonly used methods FICA, JADE and PICA for Case I: $m = 2$.

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

Case II: $m = 3$

(a) Shifted and scaled Gamma and Weibull densities

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

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

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

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

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

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

$f_3(s) = 1.2\phi(2(x + 2)) + 0.8\phi(2(x - 3))$.

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

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

$f_2(s) = 1.2\phi(2(x + 2)) + 0.8\phi(2(x - 3))$,

$f_3(s) = 0.8\phi(2(x + 2.5)) + 0.8\phi(2x) + 0.6\phi(2(x - 5))$.

The mixing matrix used in this case is as follows

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \sqrt{3}/2 & 0.5 \\ 0.1 & -0.5 & \sqrt{3}/2 \end{pmatrix}$$

Here again the chosen densities satisfy the identifiability conditions as stated in Theorem 2.3.1. The resulting boxplots of the log-efficiencies based on 200 simulated data sets for each of the cases (a)-(c) under case II: $m = 3$ are presented in Figure 4.2. For case (a) the efficiency of MICA is similar to that of PICA, our method is performing better then FICA and JADE. However, in cases (b) and (c) the proposed MICA method substantially outperforms the others. It can be noted that the 25th percentile for case (b) and the 10th percentile for case (c) of the log-efficiencies of the three methods compared with our proposed MICA estimate are above zero in Figure 4.2 showing that our method outperforms the others in terms of minimizing the Amari error criterion. Tables 4.1 and 4.2 present some selected summary values of the Amari errors of the estimates of $\boldsymbol{A}$ found by four different estimation methods corresponding to cases (b) and (c) with $m = 3$.

Table 4.1: *Summary of Amari errors for the four methods for Case II: $m = 3$, (b) where two sources are generated by using shifted and scaled gamma 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	0.029	0.039	0.058	0.086	0.135	1.6 (0.9)
FICA	0.049	0.078	0.103	0.139	0.193	2.3 (0.9)
JADE	0.088	0.129	0.178	0.229	0.284	3.4 (0.7)
PICA	0.057	0.085	0.118	0.173	0.239	2.6 (0.9)

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 4.1 and 4.2 respectively. Since the mean rank of the AE found by MICA is close to 1 for both cases (b) and

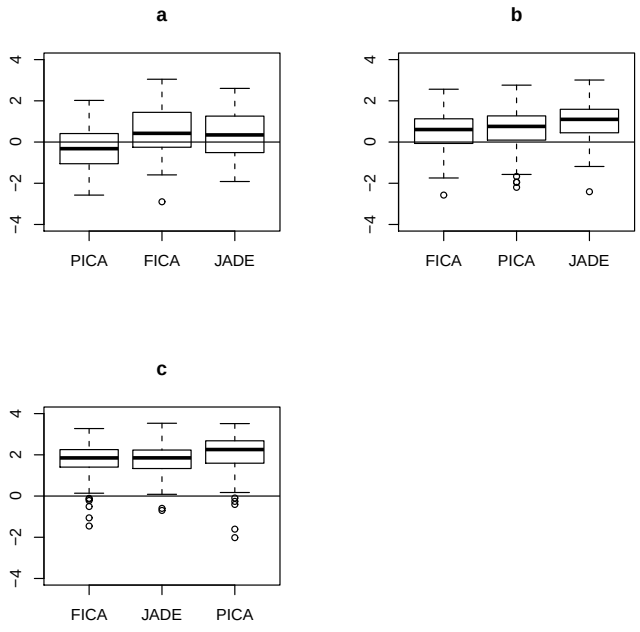


Figure 4.2: Boxplots of the log-efficiencies of our proposed MICA as compared to the three commonly used methods FICA, JADE and PICA for Case II: $m = 3$.

(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 4.1 and 4.2 are ordered by the rank of the corresponding method.

4.4 Application to Microarray Data Analysis

We evaluate the performance of our method by applying it to a microarray gene expression data. Golub et al. (1999) proposed a generic approach to the problem of cancer classification and applied the proposed method on a human acute leukemia dataset. The DNA microarray expression data analyzed by Golub et al. (1999) is available online in the R software package `golubEsets`. Two types of leukemia were considered in the experiment, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). For a total of 7129 probe sets the expression levels were measured using Affimetrix high-density oligonuleotide arrays for 72 patients. The annotation of the chip is hu6800. Filzmoser et al. (2008) proposed a cancer classification method using outlier analysis based on a robust version of the Mahalanobis distance. However, since

Table 4.2: *Summary of Amari errors for the four methods for Case II: $m = 3$, (c) where the sources are generated using a shifted and scaled gamma 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	0.018	0.025	0.033	0.045	0.054	1.1 (0.3)
FICA	0.075	0.130	0.236	0.266	0.297	2.7 (0.7)
JADE	0.092	0.141	0.226	0.297	0.331	2.9 (0.8)
PICA	0.085	0.167	0.302	0.438	0.518	3.3 (0.9)

the gene expression data are often very large, preprocessing tools need to be applied to reduce the dimension of the data. Filzmoser et al. (2008) use PCA to reduce the dimension of the data before classification. Naturally, a better preprocessing method would lead to a better classification, however in this article we also used PCA as a preliminary preprocessing tool. Suppose the 7129 genes are the observations and the 72 patients in the trial are the mixtures, i.e. $T = 7129$ and $n = 72$. We use a preliminary analysis following these steps:

1. Preprocess the data by truncating the values of the expression levels between 100 and 16,000, take a $\log_2$ transformation of the data and transform the data to have empirical mean zero and variance one across the genes.

2. Computing the pairwise correlations between the 72 subjects shows that the data are highly correlated between subjects. Apply PCA with $m = 3$ principal components (which preserve 86% of the variation) to reduce the dimension of the data to 7129×3 using the subjects as columns. Now apply ICA by using each of the estimation methods MICA, FICA, PICA and JADE to find the underlying independent sources.

3. Use the sum of squared independent scores to find a subset of outlying genes. In other words compute the statistic $I_i^2 = S_i^T S_i$ for each $i = 1, ..., T$ and choose a subset of the genes which have I^2 values above the 95th percentile of the I^2.

4. For this subset of the genes using the data for all subjects apply a hierarchical clustering algorithm to find the different clusters within the data. In this example we used the R function `hclust` with `average` method.

5. Plot the clusters using the different cancer types to visualize the performance of the method in identifying the cancer type.

When applying the different ICA methods one of the measures that can be of interest to investigate is the Kendall's τ between the pairs of estimated ICs as a nonparametric measure of dependence for the extracted components. We applied the four ICA estimation methods to identify the $(m =)3$ independent source variables. The corresponding Kendall's τ coefficients between the three extracted sources show that our method results in lower values of τ coefficients $(-0.038, -0.057, 0.001)$ compared with the other three methods $(0.027, 0.110, -0.005$ for FICA, $0.119, -0.023, -0.062$ for PICA and $0.050, 0.068, 0.020$ for JADE) which indicates that our method provides more separation of the sources compared with the three other competing methods. The estimated densities of the three underlying sources are presented in Figure 4.3. The densities are clearly non-gaussian, two of the independent sources seem to have bimodal distributions.

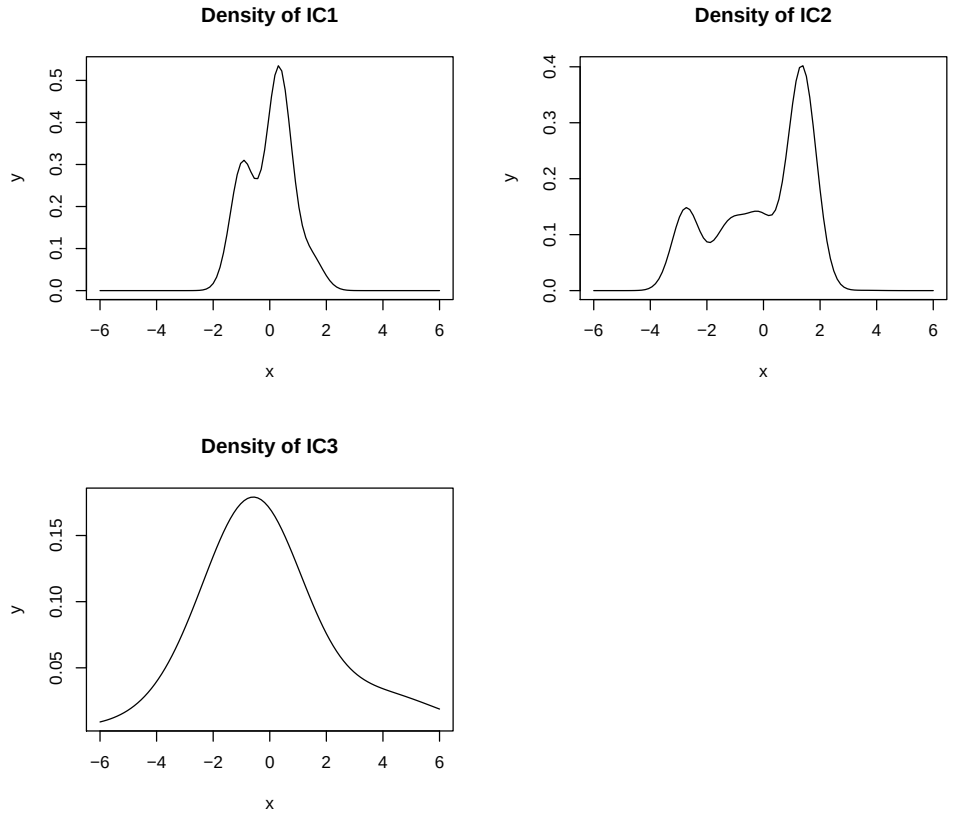


Figure 4.3: The estimated densities of the three independent components extracted by using MICA from the DNA expression data.

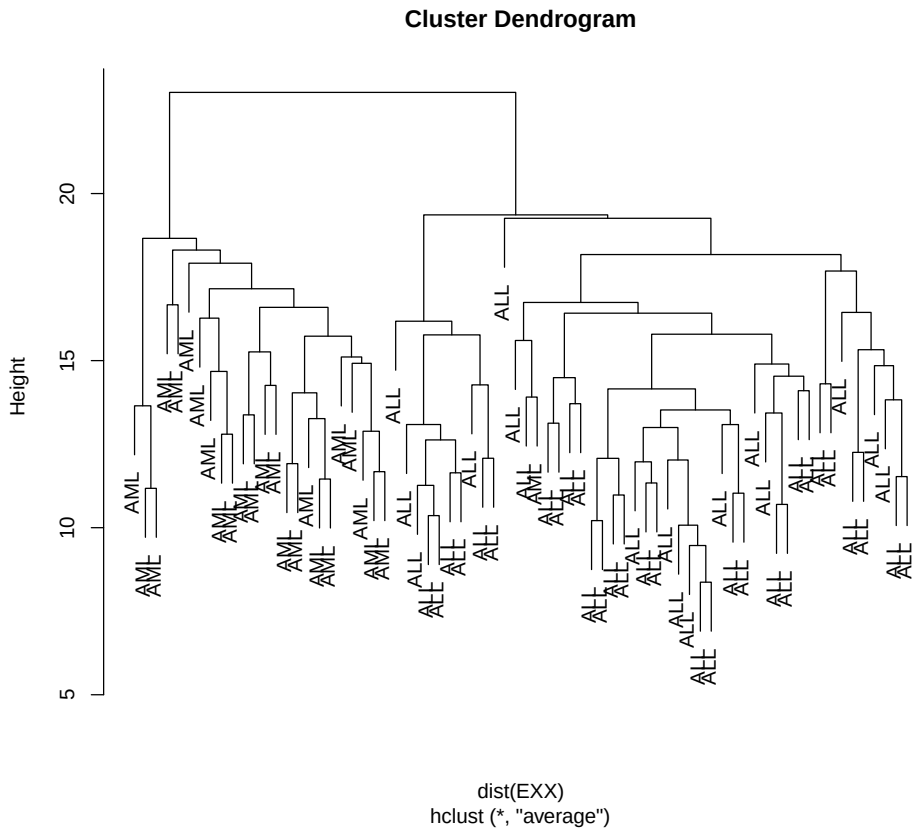


Figure 4.4: Clusters created using the MICA method for the acute leukemia dataset.

Figure 4.4 shows that the clusters obtained by using our proposed MICA are identifying the two types of cancer quite well. The two cancer types are clearly well clustered, only one subject is misclassified using no information about the cancer types. The clusters found by using FICA, PICA and JADE were similar, 3 subjects were misclassified using these methods.

4.5 Conclusion and Discussions

In this chapter, a new semiparametric approach for Independent Component Analysis (ICA) for source separation was proposed. 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 in Chapter 2. 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 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 "more symmetric". 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 this chapter 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 $\boldsymbol{A}$ 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, $\boldsymbol{A}$ and the densities of the ICs $(f_1, \ldots, f_m)$ would be of utmost interest.

Chapter 5

A Bayesian Model for Noisy ICA

5.1 Introduction

As discussed in Chapter 1 in most of the frequentist approaches to ICA it is assumed that there is no noise in the observed data. However, in most practical situations the noise is bound to be present in the model. The addition of the noise vector in the frequentist framework makes the problem more complicated. We propose the use of a Bayesian model for estimation in noisy ICA using the mixture densities described in Chapter 3 as prior densities.

Suppose $(x_{1j}, x_{2j}, ..., x_{Tj})^T$ has been observed for $j = 1, 2, ..., m$ and it is known that the observations $x_{.j}$ are mixtures of hidden independent sources $(\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_m)^T$. If we assume that the mixing is linear then the relationship can be expressed in matrix notation as

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A} + \boldsymbol{e}, \tag{5.1}$$

where $\boldsymbol{X}$ is a $T \times m$ matrix of the observed data, $\boldsymbol{S}$ is the $T \times m$ matrix of underlying hidden sources, $\boldsymbol{A}$ is the $m \times m$ mixing matrix and $\boldsymbol{e}$ is a $T \times m$ matrix corresponding to the gaussian noise in the model. Recall that by Theorem 2.2.1 if the columns of $\boldsymbol{A}$ are linearly independent and the variables $\boldsymbol{S}$ are independent and nongaussian there exists a representation of $\boldsymbol{X}$ given by (5.1). Suppose the gaussian noise variables are modeled as

$$e_{ij} \sim N(0, \sigma_e^2).$$

The problem of interest is the estimation of the mixing matrix $\boldsymbol{A}$, the densities of the hidden sources $(f_1, \ldots, f_m)$ and the variance σ_e^2 given the data $\boldsymbol{X}$. The following two subsections present the proposed prior and posterior densities. It is worth to mention that for the proposed set of priors the posterior densities can be found as commonly used known density functions.

5.1.1 Prior Distribution for $\boldsymbol{A}$

Since the noise vector is usually assumed to be normally distributed we can impose the following model for the data

$$x_{ij}|\boldsymbol{S}, \boldsymbol{A}, \sigma_e \sim N(\boldsymbol{S}_{i.}\boldsymbol{A}_{.j}, \sigma_e^2), \text{ for } i = 1, 2, \ldots, T, \; j = 1, 2, \ldots, m,$$

where $\boldsymbol{S}_{i.}$ represents the ith row of matrix $\boldsymbol{S}$ and $\boldsymbol{A}_{.j}$ is the jth column of matrix $\boldsymbol{A}$. We use a conjugate prior for the variance of the noise vector

$$\sigma_e^2|\alpha_e, \beta_e \sim InverseGamma(\alpha_e, \beta_e).$$

Further, we specify the prior densities of the parameters in the model. We use the gaussian mixture density with known means and variance as described in Chapter 3 as a prior for the hidden independent components.

$$\boldsymbol{S}_{ij}|z_{ij} = k \sim N(\mu_k, \sigma_N^2),$$

$$P[\boldsymbol{Z}_{ij} = k] = \theta_{jk}, \text{ for } k = 1, 2, \ldots, N, j = 1, \ldots, m, i = 1, \ldots, T,$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$ is a sequence of known means and σ_N is a fixed number. Subsequently a conjugate prior is used for the weights of the mixture densities for each independent hidden source respectively.

$$\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jN})|\alpha \sim Dirichlet(\alpha, \alpha, ..., \alpha), j = 1, \ldots, m.$$

We use a gaussian density as the prior for the elements of the mixing matrix $\boldsymbol{A}$ given by

$$\boldsymbol{A}_{lj}|\mu_a, \sigma_a^2 \sim N(\mu_a, \sigma_a^2), \tag{5.2}$$

with hyperparameters

$$\mu_a|\mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0^2), \tag{5.3}$$

$$\sigma_a^2|\alpha_a, \beta_a \sim InverseGamma(\alpha_a, \beta_a), \tag{5.4}$$

for $l = 1, 2, ..., m$ and $j = 1, 2, ..., m$.

Remark : Notice that when choosing the prior density for the mixing matrix $\boldsymbol{A}$ we used independent gaussian densities. However, the mixing matrix $\boldsymbol{A}$ has to be nonsingular for the model to be identifiable. It can be shown that $P[\det(A) = 0] = 0$, however in practice the sample values of the matrix generated from such a prior distribution may have a determinant value

very close to zero (numerically). We may use the LU decomposition of a matrix to generate $\boldsymbol{A}$ such that $\det(\boldsymbol{A}) \neq 0$ for all sampled values. Suppose $\boldsymbol{p} = (p_1, \ldots, p_m)$ and $\boldsymbol{g} = (g_1, \ldots, g_m)$ such that

$$p_j \sim Binomial(0.5) \text{ and } g_j \sim Gamma(1, 1),$$

and hence $g_j > 0, j = 1, \ldots, m$. Further, let $\boldsymbol{L}$ and $\boldsymbol{U}$ be two triangular matrices given by

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ l_{21} & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ l_{m1} & l_{m2} & \ldots & l_{m,m-1} & 1 \end{pmatrix} \text{ and } \boldsymbol{U} = \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1,m-1} & u_{1m} \\ 0 & u_{22} & \ldots & u_{2,m-1} & u_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & u_{mm} \end{pmatrix}$$

where

$$l_{jk} \sim N(0, 1), j = 2, \ldots, m, k = 1, \ldots, j - 1$$

$$u_{jk} \sim N(0, 1), k = 2, \ldots, m, j = 1, \ldots, k - 1$$

$$u_{jj} = g_j p_j - g_j(1 - p_j), j = 1, \ldots, m.$$

Finally, defining $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{LU} \tag{5.5}$$

we obtain a matrix $\boldsymbol{A}$ such that $\det(\boldsymbol{A}) = (-1)^d \prod_{j=1}^m g_j$, where $d = \sum_{j=1}^m (1 - p_j)$. Hence, $\det(\boldsymbol{A}) \neq 0$ for any sampled value of $\boldsymbol{A}$ generated by (5.5).

As a part of future research we would like to develop Bayesian methodologies that would enable us to obtain a better posterior estimate of the mixing $\boldsymbol{A}$ but also an estimate of the uncertainty in estimating $\boldsymbol{A}$. One major goal would be to develop a reasonable prior for the mixing matrix using the concept of Amari error which would act as a penalization if one views the problem from a frequentist perspective.

Another issue that we want to point out is that the prior densities used for the independent hidden sources do not retain the constraints on the moments of the densities of the hidden sources. By Theorem 2.3.1 moment constraints on the densities of the sources are required for the identifiability of the ICA model.

As in Chapter 4 we assume that the means $\mu_1, \mu_2, \ldots, \mu_N$ are known as well as the standard deviation σ_N. The parameters to be estimated are then the weights of the mixture density $\theta_{j1}, \theta_{j2}, \ldots, \theta_{jN}$ under the assumption that the mean of the density is 0 and the standard deviation is a fixed number $v_j, j = 1, \ldots, m$. In that case the weights should satisfy the

following constraints

$$\begin{array}{l} \sum_{k=1}^{N} \theta_{jk} = 1 \ \text{ and } \ \theta_k \geq 0, k = 1, 2, \ldots, N, \\ \sum_{k=1}^{N} \theta_{jk}\mu_{jk} = 0, \ \text{and} \\ \sum_{k=1}^{N} \theta_{jk}\mu_{jk}^2 = v_j^2 - \sigma_N^2. \end{array} \tag{5.6}$$

Another issue for future discussion is the development of these mixture priors which would maintain the moment constraints.

Also within Bayesian methodologies we can consider mixture of normals for the sources as before but now we can consider a reasonable prior for the number of mixtures (defined as N, in Chapter 3) and use reversible jump MCMC (Green, 1995) methods to obtain samples from the posterior distribution of $\boldsymbol{A}$ and the weight parameters (defined as θ's in Chapter 3).

5.1.2 Posterior Distribution of $\boldsymbol{A}$

Consider the model

$$\boldsymbol{X}_{ij}|\boldsymbol{A}, \boldsymbol{S}, \sigma_e^2 \sim N(\boldsymbol{S}_{i\cdot}\boldsymbol{A}_{\cdot j}, \sigma_e^2), \text{ for } j = 1, 2, \ldots, m, \ i = 1, \ldots, T.$$

Here we show the full conditional posterior densities can be obtained by using the priors given in Section 5.1.1 which enables us to implement a Gibbs sampling algorithm to obtain samples from the posterior distribution of $\boldsymbol{A}$ given the data $\boldsymbol{X}$.

$$\sigma_e^2|\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{S}, \alpha_e, \beta_e \sim IG\left(\alpha_e + \frac{Tm}{2}, \beta_e + \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{T}(x_{ij} - \sum_{l=1}^{m} s_{il}a_{lj})^2\right). \tag{5.7}$$

Since

$$\boldsymbol{A}_{ij}|\mu_a, \sigma_a \sim N(\mu_a, \sigma_a^2),$$

we obtain

$$\boldsymbol{A}_{\cdot j}|\boldsymbol{X}, \boldsymbol{S}, \mu_a, \sigma_a^2, \sigma^2 \sim N(\boldsymbol{\mu}_{pa}, \boldsymbol{\Sigma}_{pa}),$$

where

$$\boldsymbol{\mu}_{pa} = (\sigma_e^2\mu_a\mathbf{1}_m + \sigma_a^2\boldsymbol{S}^T\boldsymbol{X}_{\cdot j})(\sigma_e^2\boldsymbol{I}_{mm} + \sigma_a^2\boldsymbol{S}^T\boldsymbol{S})^{-1},$$

$$\boldsymbol{\Sigma}_{pa} = \sigma_a^2\sigma_e^2(\sigma_e^2\boldsymbol{I}_{mm} + \sigma_a^2\boldsymbol{S}^T\boldsymbol{S})^{-1}.$$

By (5.2) and (5.3)

$$\mu_a|\boldsymbol{A}, \mu_0, \sigma_0^2 \sim N\left(\frac{\overline{\boldsymbol{A}}\sigma_0^2 + \mu_0\sigma_a^2/m^2}{\sigma_0^2 + \sigma_a^2/m^2}, \frac{\sigma_0^2\sigma_a^2/m^2}{\sigma_0^2 + \sigma_a^2/m^2}\right).$$

From (5.2) and (5.4)

$$\sigma_a^2|\boldsymbol{A},\alpha_a,\beta_a \sim InverseGamma\left(\alpha_a+\frac{m^2}{2},\beta_a+\frac{1}{2}\sum_{l,j=1}^{m}(\boldsymbol{A}_{lj}-\mu_a)^2\right).$$

To find the posterior density of $\boldsymbol{S}$ let $\boldsymbol{\mu}_{Zi}=(\mu_{Z_{i1}},\mu_{Z_{i2}},...,\mu_{Z_{im}})$ for a given value of $\boldsymbol{Z}_{i.}, i=1,2,\ldots,T$.

$$\boldsymbol{S}_{i.}|\boldsymbol{Z}_{i.},\boldsymbol{X}_{i.},\boldsymbol{A},\sigma_e^2 \sim N\left((\sigma_e^2\boldsymbol{\mu}_{Zi}+\sigma_N^2\boldsymbol{A}^T\boldsymbol{X}_{i.})(\sigma_e^2\boldsymbol{I}_{mm}+\sigma_N^2\boldsymbol{A}\boldsymbol{A}^T)^{-1},\sigma_N^2\sigma_e^2(\sigma_e^2\boldsymbol{I}_{mm}+\sigma_N^2\boldsymbol{A}\boldsymbol{A}^T)^{-1}\right),$$

$$P(\boldsymbol{Z}_{ij}=k|\boldsymbol{S},\boldsymbol{\theta},\boldsymbol{X}) \propto \theta_{jk}e^{-\frac{(\boldsymbol{S}_{ij}-\mu_k)^2}{2\sigma_N^2}},$$

$$\boldsymbol{\theta}_j|\boldsymbol{Z},\boldsymbol{S},\boldsymbol{X} \sim Dirichlet(\alpha+n_{j1},\alpha+n_{j2},\ldots,\alpha+n_{jN}),$$

where $n_{jk}=\sharp\{i:\boldsymbol{Z}_{ij}=k\}$.

5.1.3 Preliminary Simulation Results

To illustrate the performance of the Bayesian model which we will call BICA and compare the results with the FICA method described in Chapter 4 we used a few simulated datasets. As in the examples in Chapter 4 the mixing matrix is given by

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

The number of hidden sources used is $m=2$ with sample size of $T=1000$. The hidden sources are generated using a shifted and scaled Gamma density and a mixture of two normal densities as in Case I: (c) in Section 4.3.

$$S_1 \sim 2(Gamma(4,0.25)-1) \text{ and } S_2 \sim 0.6N(-2,0.5)+0.4N(3,0.5). \tag{5.8}$$

Finally, the true value of the standard deviation of the gaussian error is chosen as $\sigma_e=0.05$. Notice, that for this choice of the densities $E(S_1)=E(S_2)=0$, $\text{var}(S_1)=1$, $\text{var}(S_2)=6.25$ and $E(S_1^3)=1$, $E(S_2^3)=6$. For this example we used a fixed number of components of the estimated mixture densities as $N=21$. We chose the fixed vectors of means for the densities of the two hidden sources by equally dividing the intervals $[-2,2]$ and $[-4,4]$ by N for each of the sources respectively, the standard deviations σ_N are chosen as $\sigma_N=2(\mu_{N,N}-\mu_{1,N})/(3(N-1))$.

First suppose that the prior density of the mixing matrix $\boldsymbol{A}$ is given as in (5.2). In terms of prior specification we use $\alpha_e=2$, $\beta_e=1$ and $\alpha_0=2$, $\beta_0=3$. We assume that $\mu_a=0$ and do not use a prior density for this as in (5.3). Further, a uniform prior for the weights $\boldsymbol{\theta}$ is

used, i.e. $\alpha = \mathbf{1}_m$, where $\mathbf{1}_m$ is an m dimensional vector of ones. The WinBUGS code is shown in Appendix G, the results are processed by using the R package `R2WinBUGS`. The posterior estimation is based on 5000 MCMC runs from single chain after 1000 burn-ins.

We then change the settings to use the prior density for $\boldsymbol{A}$ based on the LU decomposition (5.5) in Section 5.1.1. The modified WinBUGS code is presented in Appendix H.

Figure 5.1 shows the Amari errors computed using FICA and BICA for the two different prior densities of $\boldsymbol{A}$ for 10 simulation runs. Admittedly, this method does not conserve the moment constraints required for the identifiability of the model, however, the obtained results of the hidden sources can be transformed as mentioned in Theorem 2.3.4. The Amari errors computed using the BICA method are nominally better than the errors computed using FICA for these few examples. The MCSE values of the elements of the mixing matrix $\boldsymbol{A}$ are at most 0.766 and those for the elements of the weight vector are at most 0.217.

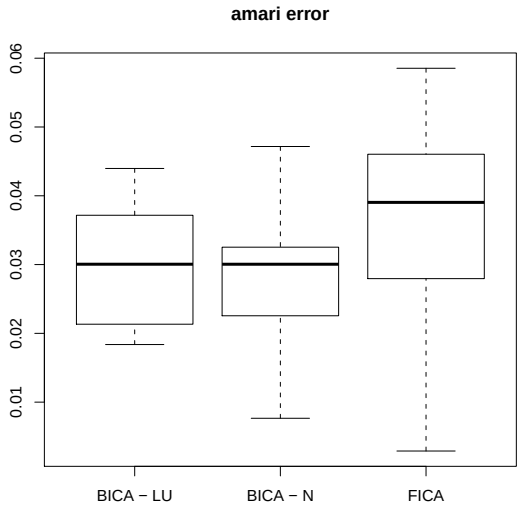


Figure 5.1: The boxplots of the Amari errors computed by using BICA and FICA. The prior density of $\boldsymbol{A}$ based on the LU decomposition is shown at the left. For the plot at the center a Normal prior is used for each element of $\boldsymbol{A}$. The plot at the right shows the amari errors computed by using FICA.

5.2 Future Research Directions

One of the future research developments is the application of the method developed in Chapter 4 to some real datasets. For instance, ICA has been widely used to analyze data from electroencephalogams (EEG) and magnetoencephalograms (MEG) in brain imaging literature (Vigario et al., 1998). These are recordings of electric or magnetic signals which result in neural currents within the brain. The data can be recorded by a magnetometer through channels while the brain of the person is triggered, for example, the person is asked to blink. The response is recorded for several minutes by milliseconds. It is assumed that the observations are independent across time. The data are analyzed to find the underlying source signals emitted by brain during different triggers. The research in this area has shown that ICA models usually perform better then PCA for these data because of the nature of underlying processes in the brain. We may apply our method for one of the brain imaging datasets and compare the results with the existing methods.

REFERENCES

Ali, M.-J. "A Bayesian approach to source separation." *Int. worskhop on Bayesian and Maximum Entropy methods (MaxEnt 1999), Boise, Idaho, USA* (2000).

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

Apostol, T. M. *Mathematical Analysis*. Addison Wesley Publishing Company, 2 edition (1986).

Bach, F. and Jordan, M. "Kernel independent component analysis." *Journal of Machine Learning Research*, 3:1–48 (2002).

Bechtel, Y. M., Bonaiti-Pellie, C., Poisson, N., Magnette, J., and Bechtel, P. R. "A population and family study of N-acetyltransferase using caffeine urinary metabolites." *Clin. Pharm. Therp.*, 54:134–141 (1993).

Boscolo, R., Pan, H., and Roychowdhury, V. "Beyond Comon's Identifiability Theorem for Independent Component Analysis." *ICANN 2002, LNCS 2415*, 1119–1124 (2002).

—. "Independent Component Analysis Based on Nonparametric Density Estimation." *IEEE Transactions on Neural Networks*, 15(1):55–65 (2004).

Cardoso, J.-F. "Eigenstructure of the fourth-order cumulant tensor with application to the blind source separation problem." *Proc. ICASSP'90, Albuquerque, NM USA*, 2655–2658 (1990).

Cardoso, J.-F. and Comon, P. "Independent component analysis, a survey of some algebraic methods." *ISCAS'96*, 2:93–96 (1996).

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

Chen, A. and Bickel, P. "Efficient independent component analysis." *The Annals of Mathematical Statistics*, 34(6):2825–2855 (2006).

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

de Boore, C. *A Practical Guide to Splines*. Springer-Verlag, New York (1978).

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

Dempster, A., Liard, N., and Rubin, D. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, 39:1–38 (1977).

Devroye, L. and Györfi, L. *Nonparametric Density Estimation The L_1 View*. John Wiley and Sons, New York (1985).

Diaconis, P. and Freedman, D. "Asymptotics of Graphical Projection Pursuit." *The Annals of Statistics*, 12(3):793–815 (1984).

Dierckx, P. *Curve and Surface Fitting with Splines*. Oxford Science Publications (1993).

Eilers, P. and Marx, B. "Flexible Smoothing with B-splines and Penalties." *Statistical Science*, 11:89–121 (1996).

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

Ferger, D. "A Continuous Mapping Theorem for the Argmax-functional in the Non-unique Case." *Statistica Neerlandica*, 58(1):83–96 (2004).

Filzmoser, P., Maronna, R., and Werner, M. "Outlier Identification in High Dimensions." *Computational Statistics & Data Analysis*, 52:1694–1711 (2008).

Friedman, J. H. and Tukey, J. W. "A projection pursuit algorithm for exploratory data analysis." *IEEE Transactions of Computers*, c-23(9):881–890 (1974).

Furman, W. and Lindsay, B. "Testing for the Number of Components in a Mixture of Normal Distributions Using Moment Estimators." *Computational Statistics & Data Analysis*, 17:473–492 (1994).

Garel, B. "Recent Asymptotic Results in Testing for Mixtures." *Computational Statistics & Data Analysis*, 51:5295–5304 (2007).

Geman, S. and Hwang, C.-R. "Nonparametric Maximum Likelihood Estimation by the Method of Sieves." *The Annals of Statistics*, 10(2):401–414 (1982).

Ghidey, W., Lesaffre, E., and Eilers, P. "Smooth Random Effects Distribution in a Linear Mixed Model." *Biometrics*, 60:945–953 (2004).

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*, 286 (1999).

Green, P. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, 82(4):711–732 (1995).

Grenander, U. *Abstract Inference*. John Wiley and Sons. (1981).

Gretton, A. "ICA and Kernel Distribution Testing." *Machine Learning Summer School, Canberra* (2006).

Hall, P. "On Kullback-Leibler Loss and Density Estimation." *The Annals of Statistics*, 15(4):1491–1519 (1987).

Hall, P. and Presnell, B. "Density Estimation Under Constraint." *Journal of Computational and Graphical Statistics*, 8:259–277 (1999).

Haykin, S. and Chen, Z. "The cocktail party problem." *Neural Computation*, 17:1875–1902 (2005).

Huber, P. *Robust Statistics*. John Wiley and Sons. (1981).

—. "Projection Pursuit." *The Annals of Statistics*, 13(2):435–475 (1985).

Hyvarinen, A. "One-unit contrast functions for independent component analysis: a statistical analysis." *Neural Networks for Signal Processing VII (Proc. IEEE Workshop on Neural Networks for Signal Processing), Amelia Island, Florida*, 388–397 (1997).

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

Hyvarinen, A. and Oja, E. "Independent Component Analysis: Algorithms and Applications." *Neural Neworks*, 13 (2000).

Jolliffe, I. *Principal Component Analysis, 2nd ed.*. Springer NY (2002).

Jones, M. and Sibson, R. "What is projection pursuit?" *Journal of the Royal Statistical Sociaty, Series A*, 150(1):1–37 (1987).

Jutten, C. and Herault, J. "Blind separation of sources, Part I: An adaptive algorithm based on neuromimethic arcitecture." *Signal Processing*, 24:1–10 (1991).

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

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

Komárek, A., Lesaffre, E., and Hilton, J. F. "Accelerated Failure Time Model for Arbitrarily Censored Data with Smoothed Error Distribution." *Journal of Computational and Graphical Statistics*, 14:726–745 (2005).

Kooperberg, C. and Stone, C. J. "A Study of Logspline Density Estimation." *Computational Statistics & Data Analysis*, 12:327–347 (1991).

Kullback, S. and Leibler, R. A. "On Information and Sufficiency." *The Annals of Mathematical Statistics*, 22(1):79–86 (1951).

Lawley, D. and Maxwell, A. *Factor Analysis as a Statistical Method*. Second edition. Butterworths (1971).

Park, B. U. and Marron, J. S. "Comparison of Data-Driven Bandwidth Selectors." *JASA*, 85(409):66–72 (1990).

Parzen, E. "n Estimation of a Probability Density Function and Mode." *The Annals of Mathematical Statistics*, 33(3):1065–1076 (1962).

Richardson, S. and Green, P. J. "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Sociaty* (1997).

Rosenblatt, M. "Remarks on Some Nonparametric Estimates of a Density Function." *Annals of Mathematical Statistics*, 27:832–837 (1956).

Sheater, S. and Jones, M. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation." *Journal of the Royal Statistical Sociaty* (1991).

Silverman, B. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall (1986).

Sisson, S. "Transdimensional Markov Chains: A Decade of Progress and Future Perspectives." *Journal of the American Statistical Association*, 100:1077–1089 (2005).

Taoufiki, M., Adib, A., and Aboutajdine, D. "Blind separation of any source distributions via high-order statistics." *Signal Processing*, 87:1882–1889 (2007).

Vigario, R., Jousmaki, V., Hamalainen, M., and Oja, E. "Independent component analysis for identification of artifacts in magnetoencephalographic recordings." *In Advances in Neural Information Processing Systems, MIT Press*, 10:229–235 (1998).

Vlassis, N. and Motomura, Y. "Efficient source adaptivity in independent component analysis." *IEEE Trans. Neural Networks*, 12(3):559–565 (2001).

Zhang, F. "A high order cumulants based multivariate nonlinear blind source separation method." *Machine Learning*, 61:105–127 (2005).

APPENDICES

APPENDIX A. Proof of Theorem 2.3.1 in Section 2.3

Proof Suppose $\boldsymbol{x}$ has two representations $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e}$ and $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{g} + \boldsymbol{e}_1$ where $\boldsymbol{s}$ and $\boldsymbol{g}$ satisfy (2.3). Then it follows from Theorem 2.2.1 that $\boldsymbol{g} = \Lambda \boldsymbol{D}\boldsymbol{P}^T\boldsymbol{s}$ with

$$\text{cov}(\boldsymbol{s}) = \boldsymbol{C}_s = \text{diag}(v_1, v_2, \ldots, v_m). \tag{9}$$

where Λ, $\boldsymbol{D}$ and $\boldsymbol{P}$ are as defined in Section 2.3. On the other hand

$$\text{cov}(\boldsymbol{g}) = \boldsymbol{C}_g = \text{cov}(\Lambda \boldsymbol{D}\boldsymbol{P}^T\boldsymbol{s}) = \Lambda \boldsymbol{D}\boldsymbol{P}^T\boldsymbol{C}_s\boldsymbol{P}\boldsymbol{D}\Lambda.$$

Suppose the columns of the permutation matrix $\boldsymbol{P}$ are defined as follows

$$\boldsymbol{P} = (\epsilon_1 \ \epsilon_2 \ \ldots \ \epsilon_m).$$

Since $\boldsymbol{P}$ is a permutation matrix, the columns of $\boldsymbol{P}$ should be orthonormal and hence $\epsilon_i^T\epsilon_j = 0$ if $i \neq j$ and $\epsilon_j^T\epsilon_j = 1$. Thus it follows that,

$$\boldsymbol{P}^T\boldsymbol{C}_s\boldsymbol{P} = \sum_{j=1}^{m} v_j\epsilon_j\epsilon_j^T = \text{diag}(v_1, v_2, \ldots, v_m) = \boldsymbol{C}_s.$$

By its definition we can write $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_m)$, where $d_j^2 = 1$, for $j = 1, 2, \ldots, m$ and so it follows that

$$\boldsymbol{D}\boldsymbol{C}_s\boldsymbol{D} = \text{diag}(v_1d_1^2, v_2d_2^2, \ldots, v_md_m^2) = \boldsymbol{C}_s.$$

Hence, the covariance of $\boldsymbol{g}$ can be obtained as

$$\boldsymbol{C}_g = \Lambda \boldsymbol{D}\boldsymbol{P}^T\boldsymbol{C}_s\boldsymbol{P}\boldsymbol{D}\Lambda = \Lambda \boldsymbol{C}_s\Lambda.$$

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

$$\boldsymbol{C}_g = \Lambda \boldsymbol{C}_s\Lambda = \text{diag}(v_1\lambda_1^2, v_2\lambda_2^2, \ldots, v_m\lambda_m^2). \tag{10}$$

Since $\boldsymbol{s}$ and $\boldsymbol{g}$ both satisfy (2.3) it follows from (9) and (10) that $v_j\lambda_j^2 = v_j$ for $j = 1, 2, \ldots, m$, and hence $\Lambda = \boldsymbol{I}$ as $\lambda_j > 0$ for $j = 1, \ldots, m$.

Now suppose the columns of the identity matrix are defined as $I_j, j = 1, 2, \ldots, m$ which contains 1 at the jth entry and 0 elsewhere. Any permutation matrix can be obtained by permuting columns or rows of the identity matrix. Suppose $\boldsymbol{P}$ is a permutation matrix different

from $\boldsymbol{I}$ and is constructed by permuting two columns i and j of $\boldsymbol{I}$ as follows

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

Since $\boldsymbol{g} = \boldsymbol{D}\boldsymbol{P}^T\boldsymbol{s}$

$$= \begin{pmatrix} d_1 s_1 & \dots & d_{i-1}s_{i-1} & d_i s_j & d_{i+1}s_{i+1} & \dots & d_{j-1}s_{j-1} & d_j s_i & d_{j+1}s_{j+1} & \dots & d_m s_m \end{pmatrix}^T,$$

for the jth element of $\boldsymbol{g}$ the variance is given by

$$v_j = \text{var}(g_j) = \text{var}(d_j s_i) = \text{var}(s_i) = v_i$$

which contradicts the assumption that the variances of the sources are distinct and hence $\boldsymbol{P} = \boldsymbol{I}$.

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

$$E(g_j^3) = E(-s_j^3) = -E(s_j^3) < 0 \tag{11}$$

which contradicts the assumption that the third moments are positive (see (2.3)). Thus, it follows that $\boldsymbol{D} = \boldsymbol{I}$.

As a result if $\boldsymbol{x}$ has two representations given by $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{e}$ and $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{g} + \boldsymbol{e}_1$, then $\boldsymbol{s} \stackrel{d}{=} \boldsymbol{g}$, $\boldsymbol{e}_1$ and $\boldsymbol{e}$ have the same gaussian densities which in tern implies $\boldsymbol{A} = \boldsymbol{B}$ and therefore the model is fully identifiable.

APPENDIX B. Proof of Theorem 2.3.4 in Section 2.3

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

$$\boldsymbol{s} = \left(\sqrt{v_1} \frac{sgn[E(z_1^3)]z_1}{\sqrt{\text{var}(z_1)}}, \sqrt{v_2} \frac{sgn[E(z_2^3)]z_2}{\sqrt{\text{var}(z_2)}}, \ldots, \sqrt{v_m} \frac{sgn[E(z_m^3)]z_m}{\sqrt{\text{var}(z_m)}} \right)^T,$$

then we obtain $\text{var}(s_j) = v_j$ and $E(s_j^3) > 0$.

APPENDIX C. Proof of Theorem 3.2.1 in Section 3.2

Proof First notice that by Jensen's inequality we have

$$-\int f(x)\log f_N(x|\boldsymbol{\theta}_N)dx \leq \sum_{j=1}^{N}\theta_{j,N}\int f(x)\left\{-\log\left[\psi\left(\frac{x-\mu_{j,N}}{\sigma_N}\right)\frac{1}{\sigma_N}\right]\right\}dx. \tag{12}$$

Hence, (A2) and (12) imply that $K[f, f_N(\cdot|\boldsymbol{\theta}_N)] < \infty$ for $N = 2, 3, ...$

Let $a = \inf_{x\in S} x$ and $b = \sup_{x\in S} x$. Notice that if the support of $f(\cdot)$ is an unbounded subset of $\mathbb{R}$ it is possible to have $a = -\infty$ or $b = \infty$. Let $a = x_{0,N} < x_{1,N} < ... < x_{N,N} = b$ be a partition of S such that $\mu_{j,N} \in [x_{j-1,N}, x_{j,N}]$, for $j = 1, ..., N$. Notice that $\max_{1\leq j\leq N-1}(x_{j+1,N} - x_{j,N}) \leq 2\max_{1\leq j\leq N-1}(\mu_{j+1,N} - \mu_{j,N})$. Hence, using (A3)(ii) we have

$$\max_{1\leq j\leq N-1}(x_{j+1,N} - x_{j,N}) = o(1) \text{ as } N \to \infty.$$

Since $K[f, f_N(\cdot|\boldsymbol{\theta}_N)] < \infty$ we can choose $x_{1,N} = a + o(1)$ and $x_{N-1,N} = b + o(1)$ as $N \to \infty$ to satisfy

$$K[f, f_N(\cdot|\widehat{\boldsymbol{\theta}}_N)] = \int_{x_{1,N}}^{x_{N-1,N}} f(x)\log\frac{f(x)}{f_N(x|\widehat{\boldsymbol{\theta}}_N)}dx + o(1), \tag{13}$$

where $\widehat{\theta}_{j,N} = \int_{x_{j-1,N}}^{x_{j,N}} f(\mu)d\mu$ for $j = 1, ..., N$. Clearly $\widehat{\boldsymbol{\theta}}_N \in \boldsymbol{\Theta}_N$. Next, by the first mean-value theorem for Riemann-Stieltjes integrals (Apostol, 1986, p. 160) there exists $\widetilde{x} \in [x_{1,N}, x_{N-1,N}]$ such that

$$\begin{aligned}\int_{x_{1,N}}^{x_{N-1,N}} f(x)\log\frac{f(x)}{f_N(x|\widehat{\boldsymbol{\theta}}_N)}dx &= \log\frac{f(\widetilde{x})}{f_N(\widetilde{x}|\widehat{\boldsymbol{\theta}}_N)}\int_{x_{1,N}}^{x_{N-1,N}} f(x)dx \\ &= \log\frac{f(\widetilde{x})}{f_N(\widetilde{x}|\widehat{\boldsymbol{\theta}}_N)}[1 + o(1)] \text{ as } N \to \infty.\end{aligned} \tag{14}$$

Finally, to complete the proof we show that

$$\frac{f_N(\widetilde{x}|\widehat{\boldsymbol{\theta}}_N)}{f(\widetilde{x})} = 1 + o(1) \text{ as } N \to \infty.$$

Notice that

$$f_N(\widetilde{x}|\widehat{\boldsymbol{\theta}}_N) = \sum_{j=2}^{N-1}\widehat{\theta}_{j,N}\psi\left(\frac{\widetilde{x}-\mu_{j,N}}{\sigma_N}\right)\frac{1}{\sigma_N} + o(1).$$

By using (A3)(ii) and the definition of Riemann-Stieltjes integral we have

$$\sum_{j=2}^{N-1}\widehat{\theta}_{j,N}\psi\left(\frac{\widetilde{x}-\mu_{j,N}}{\sigma_N}\right)\frac{1}{\sigma_N} = \int_{x_{1,N}}^{x_{N-1,N}} f(\mu)\psi\left(\frac{\widetilde{x}-\mu}{\sigma_N}\right)\frac{1}{\sigma_N}d\mu + o(1).$$

Hence

$$f_N(\widetilde{x}|\widehat{\boldsymbol{\theta}}_N) = \int_S f(\mu)\psi\left(\frac{\widetilde{x}-\mu}{\sigma_N}\right)\frac{1}{\sigma_N}d\mu + o(1) = f(\widetilde{x})\int_S \frac{f(\widetilde{x}-\sigma_N z)}{f(\widetilde{x})}\psi(z)dz + o(1).$$

By using (A1) and the fact $\sigma_N = o(1)$ it follows by an application of the dominated convergence theorem that

$$\log\frac{f(\widetilde{x})}{f_N(\widetilde{x}|\widehat{\boldsymbol{\theta}}_N)} = o(1) \text{ as } N \to \infty. \tag{15}$$

Combining (13), (14) and (15) we obtain

$$\min_{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N} K[f, f_N(\cdot|\boldsymbol{\theta}_N)] \leq K[f, f_N(\cdot|\widehat{\boldsymbol{\theta}}_N)] = o(1) \text{ as } N \to \infty. \tag{16}$$

Finally, by using (A3)(i) it follows that $K_N = \min_{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N} K[f, f_N(\cdot|\boldsymbol{\theta}_N)]$ is a decreasing function of N and hence by (16) given any $\epsilon > 0$ there exists an $N_0 \geq 2$ such that $K_N \leq K_{N_0} < \epsilon$ for all $N \geq N_0$. Hence we can find a known sequence of $\mu_{j,N}$'s satisfying (A3) and a σ_{N_0} such that $K[f, \widetilde{f}_N(\cdot|\widehat{\boldsymbol{\theta}}_N)] < \epsilon,$ for all $N \geq N_0$, where $\widetilde{f}_N(x|\widehat{\boldsymbol{\theta}}_N) = \sum_{j=1}^N \widehat{\theta}_{j,N}\psi[(\widetilde{x}-\mu_{j,N})/\sigma_{N_0}]/\sigma_{N_0}$. This completes the proof.

Remark The assumption (A2) provides just a sufficient condition for KLD to be finite and can be relaxed by weaker conditions without affecting the rest of the results. In fact, when $\psi(\cdot)$ is chosen to be the density of a standard normal distribution, the condition $|\int_S f(x)\{-\log\psi[(x-\mu)/\sigma]\}dx| < \infty,$ for all $\mu \in S$ and $\sigma > 0$ reduces to the existence of the second moment of $f(\cdot)$ which is assumed to exist when we use the variance constraint.

APPENDIX D. Proof of Theorem 3.4.1 in Section 3.4

Proof By adding and subtracting terms, we can write $\widehat{\Delta}_i$ for each $i = 1, ..., n$ as

$$\widehat{\Delta}_i = \log \frac{f_{N+1}(X_i|\boldsymbol{\theta}_{0,N+1})}{f_N(X_i|\boldsymbol{\theta}_{0,N})} + \log \frac{f_{N+1}(X_i|\widehat{\boldsymbol{\theta}}_{N+1})}{f_{N+1}(X_i|\boldsymbol{\theta}_{0,N+1})} + \log \frac{f_N(X_i|\boldsymbol{\theta}_{0,N})}{f_N(X_i|\widehat{\boldsymbol{\theta}}_{0,N})}.$$

In order to prove that the second and third additives in the above equation are of order $o_p(1)$ we need to find a sequence of vectors $\widehat{\boldsymbol{\theta}}_N$ which are consistent for $\boldsymbol{\theta}_{0,N}$. Since $\widehat{\boldsymbol{\theta}}_N$ maximizes the loglikelihood function we have

$$\widehat{\boldsymbol{\theta}}_N = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_N}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \log f_N(X_i|\boldsymbol{\theta})$$

which implies that $\widehat{\boldsymbol{\theta}}_N$ is an M-estimate as defined in (Huber, 1981, p. 43) and it follows that

$$\widehat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}_{0,N} \text{ almost surely as } n \to \infty, \tag{17}$$

where $\boldsymbol{\theta}_{0,N}$ is as defined in the statement of the theorem. Hence, by continuity theorem and by (17) we conclude

$$\log \frac{f_N(X_i|\widehat{\boldsymbol{\theta}}_N)}{f_N(X_i|\boldsymbol{\theta}_{0,N})} \to 0 \text{ almost surely as } n \to \infty$$

for each N, which proves the theorem.

APPENDIX E. Proof of Theorem 4.1.1 in Section 4.1

Proof Suppose the sequence of the true densities of the hidden sources is defined as $\boldsymbol{f}_0 = (f_1, \ldots, f_m)$. Since any continuous and bounded density function can be approximated by an infinite mixture of gaussian densities (see Chapter 3 for other regularity conditions and metric of convergence) then there exists a sequence of weights $\boldsymbol{\theta}_\infty = (\boldsymbol{\theta}_{1\infty}, \ldots, \boldsymbol{\theta}_{m\infty})$ such that for any $\epsilon > 0$

$$|L(\boldsymbol{W}_0, \boldsymbol{\theta}_\infty) - L(\boldsymbol{W}_0, \boldsymbol{f}_0)| < \frac{\epsilon}{2}, \tag{18}$$

which follows from the fact that $L(\boldsymbol{W}_0, \boldsymbol{f})$ is a continuous functional of $\boldsymbol{f}$. Note that we use the notation $L(\boldsymbol{W}, \boldsymbol{\theta}) = L(\boldsymbol{W}, \widehat{f}_{\boldsymbol{\theta}})$ as defined in (4.5). By the nested structure of the sets of means of estimated densities and by construction of $\widehat{\boldsymbol{W}}^{(M)}$ we obtain

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

Hence the monotone sequence $L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{\theta}}^{(M)}_{N^{(M)}})$ has a limit as $M, N^{(M)} \to \infty$. Notice that any estimated weight vector at any stage of iteration (e.g. $\widehat{\boldsymbol{\theta}}^{(M)}_{N^{(M)}}$) belongs to $\boldsymbol{\Theta}_N = \{\boldsymbol{\theta}_N \in [0,1]^N : \sum_{j=1}^N \theta_{jN} = 1\}$, where $\boldsymbol{\Theta}_N$ is a compact set. Hence, for a compact set $\boldsymbol{\Omega} \subset \mathbb{R}^{m \times m}$ by continuity of the function $L(\cdot)$ there exist $\boldsymbol{W}_0 \in \boldsymbol{\Omega}$ and $\boldsymbol{\theta}_\infty = (\boldsymbol{\theta}_{1\infty}, \ldots, \boldsymbol{\theta}_{m\infty})$ such that for any $\delta \in (0,1)$

$$|L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{\theta}}^{(M)}_{N^{(M)}}) - L(\boldsymbol{W}_0, \boldsymbol{\theta}_\infty)| < \frac{\epsilon}{2} \text{ with probability } \geq 1 - \delta. \tag{19}$$

for sufficiently large T and M, where

$$L(\boldsymbol{W}_0, \boldsymbol{\theta}_\infty) = \sum_{i=1}^{T} \sum_{k=1}^{m} \log \left[\sum_{j=1}^{\infty} \theta_{jk} \phi \left(\frac{\sum_{l=1}^m x_{il} w_{lj} - \mu_{jk}}{\sigma} \right) \frac{1}{\sigma} \right] + T \log |det \boldsymbol{W}|.$$

Hence, by (18) and (19) we obtain for any $\delta \in (0,1)$,

$$|L(\widehat{\boldsymbol{W}}^{(M)}, \widehat{\boldsymbol{\theta}}^{(M)}_{jN_j^{(M)}}) - L(\boldsymbol{W}_0, \boldsymbol{f}_0)| < \epsilon \text{ with probability } \geq 1 - \delta. \tag{20}$$

This implies that

$$\widehat{\boldsymbol{W}} = \underset{W \in \mathbb{R}^{m \times m}}{\operatorname{argmax}} L(\boldsymbol{W}, \boldsymbol{f}_0) + o_p(1)$$

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

$$\widehat{\boldsymbol{W}} \to \boldsymbol{W}_0 \text{ a.s. as } M, T \to \infty.$$

which completes the proof of the theorem.

APPENDIX F. The gradient vector and the Hessian matrix of the loglikelihood function of $\boldsymbol{W}$.

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

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

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

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

If $\delta \neq \beta$

$$\frac{\partial^2 L(\boldsymbol{w}, \widehat{\boldsymbol{f}})}{\partial w_{\gamma\beta} \partial w_{\alpha\delta}} = -T[\boldsymbol{W}^{-1}]_{\beta\gamma} [\boldsymbol{W}^{-1}]_{\delta\alpha}.$$

APPENDIX G. WinBUGS Code for BICA Without the Moment Constraints.

```
model{
for(i in 1:T){
for(j in 1:m){
muX[i,j] <- inprod(S[i,],A[,j])
X[i,j] ~ dnorm(muX[i,j],taue)
}}

taue ~ dgamma(alphae,betae)

for(j in 1:m){
theta[j,1:N]~ddirch(alpha[])}

for(i in 1:T){
for(j in 1:m){
S[i,j] ~ dnorm(mu[j,Z[i,j]],tauN[j])
Z[i,j]~dcat(theta[j,])}}

for(k in 1:m){
for(j in 1:m){
A[k,j]~dnorm(mua,taua)}}

taua ~ dgamma(alpha0,beta0)

sigmaa <- 1/taua
sigmae <- 1/taue
}
```

APPENDIX H. WinBUGS Code for BICA Without the Moment Constraints. The Prior Density of A is Constructed Based on the LU Decomposition.

```
model{

for(i in 1:T){
for(j in 1:m){
muX[i,j] <- inprod(S[i,],A[,j])
X[i,j] ~ dnorm(muX[i,j],taue)
}}

taue ~ dgamma(alphae,betae)

for(j in 1:m){
theta[j,1:N]~ddirch(alpha[])}

for(i in 1:T){
for(j in 1:m){
S[i,j] ~ dnorm(mu[j,Z[i,j]],tauN[j])
Z[i,j]~dcat(theta[j,])}}

for(j in 2:m){
for(k in 1:(j-1)){
L[k,j]<-0
U[j,k]<-0}}
for(j in 1:m){
p[j]~dbern(0.5)
g[j]~dgamma(3,0.5)
U[j,j]<-g[j]*p[j]-g[j]*(1-p[j])
L[j,j]<-1}
for(j in 2:m){
for(k in 1:(j-1)){
L[j,k] ~ dnorm(0,0.5)
U[k,j] ~ dnorm(0,0.5)
}}
```

```
for(k in 1:m){
for(j in 1:m){
A[k,j]<-inprod(L[k,1:m],U[1:m,j])}}

sigmae <- 1/taue
}
```