Model Based Clustering with Generally Dependent Data: An Exponential Power Mixture Approach

Jian Zhang and Faming Liang *

June 24, 2007

Abstract

Model based clustering methods are based on the assumption that the data under study are independent. This assumption turns out to be untrue in many recent applications, where the form of dependence is generally unknown. For example, the long range correlation structure among genes is believed to persist in microarray gene expression data even after normalisation. The potential impact of such dependence on model based clustering methods have yet to be explored. Here we consider a family of generalised Gaussian mixture models, also known as the exponential power mixture models, for clustering. The new proposal takes advantage of both leptokurtic and platykurtic deviations from the Gaussian mixture models caused by the dependence effect. The exponential power mixtures are shown to be identifiable. An expectation-conditional-maximisation algorithm is developed to calculate the maximum likelihood estimators of the unknown parameters in these mixtures. The Bayesian Information Criterion (BIC) is then employed to select the numbers of components in these mixture models. The maximum likelihood estimators are shown to be consistent under sparse dependence. Numerical performance of the proposed clustering procedure is assessed using both simulated and real data. The proposed procedure over-performs Mclust, a popular Gaussian mixture model-based clustering tool when there are strong correlations or non-Gaussian components in the data.

Key words: Exponential power mixtures, model based clustering, general dependence, sparse correlations, and expectation-conditional-maximisation algorithm.


Short title: Exponential power mixtures

*Address for correspondence: Jian Zhang, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, UK. Tel: +44- 1227- 823661; Fax: +44- 1227- 827932; E-mail: j.zhang@kent.ac.uk. Faming Liang is Associate Professor, Department of Statistics, Texas A& M University, College Station, TX 77843-3143, USA. Tel: +1- 979 8458885; Fax: +1- 979- 8453144; E-mail: fliang@stat.tamu.edu.
1 Introduction

Finite mixture distributions have been used in a variety of disciplines for the purpose of modelling data (Everitt and Hand, 1981; Titterington, Smith and Markov, 1985; McLachlan and Peel, 2000). One of recent applications has arisen from clustering gene expression data (Fraley and Raftery, 2002). In this setting, the data are represented in the form of a matrix of measurements, with the rows for the genes and the columns for the experiments or samples. Due to the amount and complexity of the data, clustering has become one of the most important statistical tools to extract useful information from gene expression data. A wide range of heuristic clustering methods have already been applied to such data, such as hierarchical clustering (Eisen et al. 1998), self-organising maps (Golub et al., 1999), CAST (Ben-Dor et al. 1999), and to mention a few (Fraley and Raftery, 2002). In particular, Yeung et al. (2001) applied the Gaussian-mixture (GM) method, a model based approach to both real and simulated gene expression datasets for which the clone groupings were known in advance. They showed that the GM method produced clusters of quality comparable to leading heuristic methods, but with key advantage of suggesting the number of clusters and an appropriate model for each cluster. Despite of this success, the application of the GM method is still hindered by its potentially invalid model assumptions:

1. \( p \)-dimensional observations \( \{x_i\} \) (i.e., the rows in the data matrix) are independently and identically distributed (iid) with the mixture density \( f(x) = \sum_{k=1}^{m} \pi_k \phi_k(x) \).

2. The \( k \)-th component density \( \phi_k(x) \) is Gaussian.

The assumption (1) is rarely true in gene expression data, where gene expression levels are usually correlated in a fashion that is difficult to estimate. In practice, taking such a dependence structure as a nuisance parameter, people often adopt a marginal likelihood model where the effect of this nuisance structure can be simply ignored. Of course, this may yield a biased result if the adopted marginal likelihood model is not appropriately selected. To demonstrate this phenomenon, we first simulated 100 datasets, each has 1000 sparsely correlated data points, say \( \{x_i, 1 \leq i \leq 1000\} \). These \( x_i \)'s are divided into six independent groups: \( (x_1, ..., x_{150}) \sim N(0, \Sigma_1), (x_{151}, ..., x_{200}) \sim N(0, \Sigma_2), (x_{201}, ..., x_{300}) \sim N(0, \Sigma_3), (x_{301}, ..., x_{400}) \sim N(0, \Sigma_4), (x_{401}, ..., x_{500}) \sim N(0, \Sigma_5), \) and \( (x_{501}, ..., x_{1000}) \sim N(0, \Sigma_6) \), where \( \Sigma \) stands for a Gaussian distribution and \( \Sigma_k \) is determined by \( \text{var}(x_i) = 1, \text{cov}(x_i, x_j) = |i-j|^{-\alpha_k}/(2 - 2^{-\alpha_k}), i < j, a_k \sim U(0, 1/64) \) (a uniform distribution over \( (0, 1/64) \)). See Gneiting (2000). The clumpy structure of these simulated data serves as a simplistic model of gene interactions within distinct pathways (Qiu, Brooks et al.,2005). Each dataset forms a single cluster with the same marginal distribution for all its members. For the purpose of comparison, we also simulated 100 datasets, each has 1000 independent data points sampled from \( N(0, 1) \). For each of these datasets, we first calculate the sample mean \( \bar{x} \), the sample standard deviation \( \hat{\sigma}_n \), and the sample shape index \( \hat{\sigma}_n^2/(\sum_{i=1}^{1000} |x_i - \bar{x}|)^2 \). Then we employ the GM
method to predict the number of clusters in each dataset. These values are compared via the boxplots in Figure 1. As we expected, for the independent datasets, these values are all sufficiently close to the true values. However for the datasets with the clumpy dependence, these values are lack of consistency. In particular, the clumpy dependence exerts a serious effect on the GM based cluster analysis by giving about four artificial clusters for each dataset on average.

This effect is partially due to the misspecification of the shape of each cluster, since applying the GM method implies that we fix the shape index of each cluster $x$ to $\frac{\text{var}(x)}{(E[x] - Ex)^2} = \pi/2 = 1.5708$. This makes the GM method vulnerable to any leptokurtic or platykurtic deviations from the Gaussian distributions because the shape index is decreasing or increasing when the distribution of $x$ is leptokurtically or platykurtically deviated from the Gaussian distributions. Consequently, although the Gaussian distribution family can handle the correlation effects on the estimation of the centre and dispersion of a cluster, it is unable to deal with the correlation effects on the distribution shape. An important theoretical question arising from this simulation is that if the independence assumption does not hold, to what extent is the clustering result wrong and how can it be corrected? This motivate us to propose the use of what are called exponential power distributions (EPDs) for clustering, a family of generalised Gaussian distributions which account for the variability of the distribution shape.

The second motivation of using the EPD mixtures comes from the fact that the normality assumption in the GM method might also be violated in some applications. For example, in unsupervised pattern recognition, where edges may be represented by points clustered uniformly, rather than normally, along a line (Banfield and Raftery, 1992). Furthermore the data in image and economics are often heavy tailed. Yeung et al. (2001) tested the Gaussian mixture assumption in the GM method and found this assumption is seldom satisfied by gene expression data. To address this issue, Chang and Walther (2006) proposed a flexible nonparametric clustering method with mixtures of log-concave distributions. However this flexibility comes at a price in terms of the identifiability problem and computational difficulties of the proposed model. In contrast, these issues can be solved by using EPD mixtures.

There are two goals of the present paper. The first is to develop a maximum likelihood estimation of the EPD mixtures and to assess their performance in cluster analysis. To prove the consistency of the model estimation, we establish the identifiability of these mixtures. The EPD based clustering method is then compared with the GM model based clustering by simulations and real data analysis. The new proposal performs favourably in comparison to the GM method. In particular, the impact of dependence on the cluster analysis is markedly reduced by using the EPD mixtures instead of the Gaussian mixtures. The large sample behaviour of the proposed estimators is investigated under general dependence. We showed that the proposed ML estimators are consistent provided the average pairwise correlation coefficient between data points converges to zero as
the sample size tends to infinity.

The second goal of the present paper is to give some feasible algorithms for calculating the maximum likelihood estimators of the unknown parameters in the models. The development of such algorithms appears to pose several challenges. The EM algorithm is currently a popular tool for calculating the maximum likelihood estimators in Gaussian mixture models. The algorithm iteratively repeats what are called E-step and M-step. In the E-step, we directly calculate the conditional expectation of the so-called complete likelihood and update the membership probabilities, given the sample and the current values of the parameters. In the M-step, we maximise the above conditional expectation of the complete likelihood with respect to the parameters of each component. The computation is very simple because we can derive the updating formulae explicitly in both steps (Titterington et al., 1985). However, in the setting of the EPD mixtures, it is impossible to derive similar closed formulae for all parameters in the M-step. Although the Newton-Raphson iterative approximation might be useful for solving the problem, it does not guarantee the convergence of the algorithm. Furthermore, the Newton-Raphson approximation might not exist because the likelihood may have no partial derivatives with respect to the shape parameters in some cases where the values of the shape parameters are less than 1. To circumvent these difficulties, we proposed an expectation-conditional-maximisation (ECM) algorithm by replacing the M step by a few of conditional maximisation steps. In some of these sub-steps, a local minorisation procedure is employed to handle the non-existence of derivatives. The convergence behaviour of our new proposal can be analysed under some regularity conditions.

Like the EM, the ECM is a local optimisation algorithm. Therefore a good initialisation is of great importance for the ECM algorithm as it can heavily influence the speed of convergence and its ability to locate the global maximum. In literature there are a numerous ways of initialisations (Karlis and Xekalaki, 2003). For example, to begin with estimates obtained by Gaussian mixture models, by partitioning around medoids (PAM) (i.e., $k$-medoids clustering, Kaufman and Rousseauw, 1990), by hierarchical clustering or by multiple random starting points. In this paper, we proposed a new method called bottom-up initialisation. Our computational experience demonstrates that it could be better than the other strategies.

The remainder of this paper is organised as follows. In Section 2, some basic notations and definitions of EPD mixtures are reviewed. The constrained maximum likelihood estimators are introduced in Section 3. The issue of model selection in clustering is also addressed in Section 3. The ECM algorithm is introduced to calculate the maximum likelihood estimator (MLE) in Section 4. The initialisation strategies and convergence of this algorithm are also studied. Section 5 discusses the application of the EPD mixtures to cluster analysis. The theoretical properties of the models such as identifiability and consistency of MLE are investigated in Section 6. The numerical studies are carried out in Section 7. The conclusions are made in Section 8. The proofs of the main theorems can be found in the Appendix I. Some extensions of the ECM algorithm to
the other models are presented in the Appendix II.

2 Mixtures of exponential power distributions

We begin with the definition of the exponential power distribution (EPD) family, \( \mathcal{G} = \{ g(\cdot | \theta) : \theta \in \Theta \} \), with the density being defined as

\[
g(x | \theta) = \frac{\beta \Gamma(p/2)}{2^{p/2} \pi^{p/2} \sigma^{p/2} \Gamma(p/\beta)} \exp \left\{ -((x - \mu)\Sigma^{-1}(x - \mu))^{\beta/2} \right\},
\]

where \( \theta = (\mu, \Sigma, \beta) \), \( \Theta = R^p \times S \times R^+ \) with \( S \) the set of all \( p \times p \) positive definite matrices, \( \mu \) and \( \Sigma \) are the location and scatter parameters of the distribution respectively, and \( \beta \) is the shape parameter, which allows flexibility in fitting the tails of \( g(\cdot | \theta) \) (Box and Tiao, 1973). Let \( \text{epd}_p(\mu, \Sigma, \beta) \) denote a \( p \)-dimensional EPD with the parameters \( \mu, \Sigma, \) and \( \beta \).

The EPD family is a sub-family of symmetric Kotz type distributions of a single mode (Fang, Kotz & Ng, 1990). This distribution family covers a wide range of distributions: the normal distribution when \( \beta = 2 \), the Laplace distribution when \( \beta = 1 \), and the uniform distribution when \( \beta \to \infty \). For \( 0 < \beta < 2 \), the distribution is heavy tailed and log-convex on \( (0, \infty) \), while for \( \beta \geq 2 \), the distribution is light tailed and log-concave on \( (0, \infty) \). In particular, the mean of this distribution is \( \mu \) and \( \text{cov}(X) = A(\beta) \Sigma \), where \( A(\beta) = \Gamma(2 + p)/\beta / (p \Gamma(p/\beta)) \) is increasing in \( \beta \), \( \lim_{\beta \to 0} A(\beta) = \infty \) and \( \lim_{\beta \to \infty} A(\beta) = 1/(2 + p) \). Note that the \( t \) distribution family does not cover distributions with the tails lighter than a normal distribution. So compared to the commonly used multivariate normal or \( t \) distributions, the exponential power distribution family is even more flexible in describing a population. Although it is not universally true, most people tend to accept that a unimodal distribution corresponds to a homogeneous, unclustered population, and, in contrast, that the existence of several distinct modes indicates a heterogeneous, clustered population, with each mode corresponding to a cluster of observations (Everitt, Landau and Leese, 2001). EPD is unimodal and thus is qualified for describing a clustering unit.

An EPD random variable, \( x \sim \text{g}(\cdot | \theta) \) can be expressed as \( \mu + AY \) where \( \Sigma = AA' \) and \( Y \) is another EPD random variable which has a unit scatter matrix. Moreover, \( Y \) can be characterized by the product of two independent random variables, \( U \) and \( R \), where \( U \) follows a uniform distribution on \( p- \) sphere and \( R \) is a power of a Gamma random variable, say \( W^{1/\beta} \). Here \( W \) has the Gamma distribution with location parameter \( p/\beta \) and scale parameter \( 1 \). If \( \Sigma = \lambda^2 I_p \) with \( I_p \) being \( p \times p \) identity matrix, then \( x \) is spherically distributed and

\[
\frac{E||x - \mu||^2}{(E||x - \mu||)^2} = B_p(\beta)
\]
with $B_p(\beta) = \frac{\Gamma(p+2)/\beta \Gamma(p/\beta)}{((p+1)/\beta)^2}$. Here $B_p(\beta)$ is strictly decreasing in $\beta \in (0, \infty)$. In particular,

$$\lim_{\beta \to 0^+} B_p(\beta) = \infty, \quad \lim_{\beta \to \infty} B_p(\beta) = \frac{(p+1)^2}{p(p+2)}.$$ 

For a general $\Sigma$, it follows from the result of Fang, Kotz and Ng (1990, p.43) that the $j$th coordinate $x_j$ of $x$ has a univariate exponential power distribution. Therefore,

$$\frac{E(x_j - E x_j)^2}{(E| x_j - E x_j|^2)} = B_1(\beta).$$

The equations (2.1) and (2.2) indicate that the ratio between the standard deviation and the first absolute moment of $x$ determines the parameter $\beta$ uniquely. This gives rise to a natural $\Sigma$-free moment estimator for $\beta$. We call $B_p(\beta)$ a shape index.

Now an EPD mixture can be described as follows. Let the $p$-dimensional vector $x = (x_1, ..., x_p)$ contain the values of $p$ variables measured on each of $n$ entities to be clustered, and let $x_i$ denote the value of $x$ corresponding to the $i$-th entity. The data $X = (x_1, ..., x_n)$ are assumed to be iid with mixture of a finite number, say $m$, of components:

$$f(x|\psi) = \sum_{k=1}^{m} \pi_k g(x|\theta_k), x \in \mathbb{R}^p,$$

where $\psi = (\theta_1, \pi_1, ..., \theta_m, \pi_m)$ with $\theta_k = (\mu_k, \Sigma_k, \beta_k)$, $0 < \pi_k \leq 1$, $1 \leq k \leq m$, and $\pi_m = 1 - \sum_{k=1}^{m-1} \pi_k$, contains all parameters of the model; $\pi_k$ is the unknown mixing proportion of the $k$th component, namely $g(\cdot|\theta_k) \in \mathcal{G}$ with $\theta_k = (\mu_k, \Sigma_k, \beta_k)$. The unknown parameters $\mu_k, \Sigma_k$, and $\beta_k$ stand for the centre, scatter and shape of the $k$th component respectively.

### 3 Estimation

For the sample $X$ from $f(\cdot|\psi)$ defined in (2.3), the marginal likelihood can be written down as

$$L(\psi|X) = \prod_{i=1}^{n} f(x_i|\psi).$$

Note that $L \to \infty$ when $\mu_k = x_i$ and $|\Sigma_k| \to 0$ for some $k$ and $i$. Thus, the MLE of $\psi$ does not exist. To avoid such a singularity, in our implementation we replace $(x_i - \mu_k)^T \Sigma_k (x_i - \mu_k)$ by $\max\{(x_i - \mu_k)^T \Sigma_k (x_i - \mu_k), \varepsilon^2\}$ with $\varepsilon$ being a sufficiently small constant, and we impose the constraint that $\min_{k \geq 1} \sigma_k \geq \delta > 0$ for some positive $\delta$ when we solve the above optimisation problem. In what follows we will consider only the constrained MLE, say $\hat{\psi}$.

Following Banfield and Raftery (1993), Celeux and Govaert (1995), and McLachlan and Peel (2000), we view the problem of clustering model selection (parameterisation and number of clusters) as one of choosing between competing models for the same data. We use the Bayesian Information
Criterion (BIC) (Schwarz, 1978), which adds a penalty to the log-likelihood based on the number of independent parameters. For a mixture model \( \mathcal{M} \), the BIC has the form

\[
\text{BIC}(\mathcal{M}) = -\frac{2}{n} \log(L(\hat{\psi}|X)) + \text{dim}(\mathcal{M}) \frac{\log(n)}{n},
\]

where \( \log(L(\hat{\psi})) \) is the maximised log-likelihood for the model \( \mathcal{M} \), \( \text{dim}(\mathcal{M}) \) is the number of independent parameters in the model, and \( n \) is the size of the sample \( X \). The smaller BIC the better the model is. Therefore, we choose \( m \) at which BIC attains the minimum.

When all \( \beta_k = 2 \), the EPD mixtures are reduced to Gaussian mixtures. Let \( \beta_k \) be varying, we have the following non-normal mixture models according to four different parameterisations of the scatter matrices of the components:

1. PSE: EPD components with \( \Sigma_k = \lambda^2 I_p \), \( k = 1, \ldots, m \). Here \( \text{dim}(\text{PSE}) = (p + 2)m \).
2. PSV: EPD components with \( \Sigma_k = \lambda_k^2 I_p \), \( k = 1, \ldots, m \). Here \( \text{dim}(\text{PSV}) = (p + 3)m - 1 \).
3. PDE: EPD components with equal diagonal component covariance. Here \( \text{dim}(\text{PDE}) = (p + 2)m + p - 1 \).
4. PDV: EPD components with unconstrained diagonal component covariances. Here \( \text{dim}(\text{PDV}) = 2(p + 1)m - 1 \).

In practice, for a high dimensional dataset but with a low sample size, the PSV model is often more appropriate than the other models in terms of computational feasibility.

4 Algorithms

4.1 ECM

Given the sample \( X \), the observed-data likelihood can be expressed as

\[
L(\psi) = \prod_{i=1}^{n} \sum_{k=1}^{m} \pi_k g(x_i|\theta_k).
\]

The constrained maximum likelihood (ML) estimator \( \hat{\psi} \) of \( \psi \) can be calculated by the expectation-maximisation (EM) algorithm. In the EM, the “complete data” \( y_i = (x_i, z_i) \) are considered, where \( z_i = (z_{i1}, \ldots, z_{im}) \) is the unobserved group indicator of \( x_i \), with \( z_{ik} = 1 \) if \( x_i \) belongs to group \( k \); \( z_{ik} = 0 \) otherwise. The observed-data likelihood \( L(\psi) \) can be obtained by integrating \( z \) out of the complete-data likelihood

\[
L_C(\psi) = \prod_{i=1}^{n} \prod_{k=1}^{m} g(x_i|\theta_k)^{z_{ik}} p(z_i),
\]

where \( p(z) \) is the multinomial probability function

\[
p(z) \propto \prod_{i=1}^{n} \prod_{k=1}^{m} \pi_k^{z_{ik}}.
\]
The resulting complete-data log-likelihood is

\[ l(\psi, z_1, \ldots, z_n) = \sum_{i=1}^{n} \sum_{k=1}^{m} z_{ik} \log(\pi_k g(x_i | \theta_k)). \]

The \((v+1)\)th iteration in the EM algorithm consists of the following two steps, namely E and M steps. The E step is given by

\[ \hat{z}_{ik} \leftarrow w_{ik} = E[z_{ik} | \psi^{(v)}, x_1, \ldots, x_n] = \frac{\pi_k^{(v)} g(x_i | \theta_k^{(v)})}{\sum_{j=1}^{m} \pi_j^{(v)} g(x_i | \theta_j^{(v)})}, \]

given the estimate \(\psi^{(v)} = (\theta_1^{(v)}, \pi_1^{(v)}; \ldots; \theta_m^{(v)}, \pi_m^{(v)})\) in the previous step. Given \(\psi^{(v)}, x_1, \ldots, x_n, w_{ik}, 1 \leq i \leq n, 1 \leq k \leq m\), the \((v+1)\)th M step involves updating the estimates of \(\psi\) via maximizing the conditional expectation of the complete-data log-likelihood,

\[ Q(\theta) = E[l((\theta, z_1, \ldots, z_n)) | \theta^{(v)}, x_1, \ldots, x_n] \]

\[ = \sum_{i=1}^{n} w_{ik} \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{m} w_{ik} \log(g_k(x_i | \theta_k)) \]

with respect to \(\psi\), under the constraints \(0 \leq \pi_k \leq 1, 1 \leq k \leq m\), and \(\sum_{k=1}^{m} \pi_k = 1\). This only leads to simple closed-forms

\[ \pi_k^{(v+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}, 1 \leq k \leq n, \]

for the mixing proportions. There are no such kinds of forms for updating \(\theta_k\), \(1 \leq k \leq m\), i.e., for maximising

\[ Q_2(\theta) = Q_2(\theta, \theta^{(v)}) = \sum_{i=1}^{n} \sum_{k=1}^{m} w_{ik} \log(g_k(x_i | \theta_k)) \]

with respect to \(\theta = (\theta_1, \ldots, \theta_m)\). Note that \(Q_2\) has a global maximum because it is continuous. Furthermore, if \(Q_2\) has the derivatives at this maximum, then the maximum satisfies the simultaneous equations

\[ \frac{\partial Q_2}{\partial \theta_k} = 0. \] (4.1)

The updating scheme for \(\theta_k\)’s can be derived from these equations.

For the demonstration purpose, we focus on the PSV model in what follows. The extension of our method to the other parameterisations is straightforward and is summarised in the Appendix II. Under the PSV model, each cluster is spherically shaped and the equations in (4.1) become

\[ \frac{\partial Q_2}{\partial \mu_k} = 0, \quad \frac{\partial Q_2}{\partial \lambda_k} = 0, \quad \frac{\partial Q_2}{\partial \beta_k} = 0, \]

where \(\lambda_k\) and \(\beta_k\) are rescaled by \(\lambda_k^* = \log(\lambda_k)\) and \(\beta_k^* = \log(\beta_k)\), and

\[ \frac{\partial Q_2}{\partial \mu_k} = \sum_{i=1}^{n} w_{ik} \frac{\beta_k}{\lambda_k^*} \| x_i - \mu_k \|^2 \beta_k^*(x_i - \mu_k), \]
We choose 10 replace the M-step by a few steps of conditional maximisations (CM): For 1
Although the Newton-Raphson method can be used to solve these equations, here we adopt a simple

\[
\frac{\partial Q_2}{\partial \lambda_k^*} = \sum_{i=1}^{n} w_{ik} \left( -p + \frac{\beta_k}{\lambda_k^*} \|x_i - \mu_k^*\|^2 \right),
\]

\[
\frac{\partial Q_2}{\partial \beta_k^*} = \sum_{i=1}^{m} w_{ik} \left( 1 + \frac{p \Gamma(p/\beta_k)}{\beta_k \Gamma(p/\beta_k)} - \beta_k \left( \frac{||x_i - \mu_k^*||}{\lambda_k^*} \right)^{\beta_k} \log \left( \frac{||x_i - \mu_k^*||}{\lambda_k^*} \right) \right).
\]

Although the Newton-Raphson method can be used to solve these equations, here we adopt a simple

We choose 10 replace the M-step by a few steps of conditional maximisations (CM): For 1
Although the Newton-Raphson method can be used to solve these equations, here we adopt a simple

\[
\frac{\partial Q_2}{\partial \lambda_k^*} = \sum_{i=1}^{n} w_{ik} \left( -p + \frac{\beta_k}{\lambda_k^*} \|x_i - \mu_k^*\|^2 \right),
\]

\[
\frac{\partial Q_2}{\partial \beta_k^*} = \sum_{i=1}^{m} w_{ik} \left( 1 + \frac{p \Gamma(p/\beta_k)}{\beta_k \Gamma(p/\beta_k)} - \beta_k \left( \frac{||x_i - \mu_k^*||}{\lambda_k^*} \right)^{\beta_k} \log \left( \frac{||x_i - \mu_k^*||}{\lambda_k^*} \right) \right).
\]

1. Set \( \lambda_k = \lambda_k^{(v)} \) and \( \beta_k = \beta_k^{(v)} \) in \( Q_2 \), and update the estimate of \( \mu_k \) by

\[
\mu_k^{(v+1)} = \frac{\sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v)}||^{\beta_k^{(v)} - 2} x_i}{\sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v)}||^{\beta_k^{(v)} - 2}}.
\]  

(4.2)

2. Set \( \mu_k = \mu_k^{(v+1)} \) and \( \beta_k = \beta_k^{(v)} \) in \( Q_2 \) and update the estimate of \( \lambda_k \) by

\[
\lambda_k^{(v+1)} = \left( \frac{\beta_k^{(v)} \sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v+1)}||^{\beta_k^{(v)}}}{p \sum_{i=1}^{n} w_{ik}} \right)^{1/\beta_k^{(v)}}.
\]  

(4.3)

3. Set \( \mu_k = \mu^{(v+1)} \) and \( \lambda_k = \lambda_k^{(v)} \) in \( Q_2 \) and update the estimate of \( \beta_k \) in two steps: Initialize \( \beta_k \) by the solution to the equation

\[
\frac{\Gamma((p + 2)/\beta_k) \Gamma(p/\beta_k)}{\Gamma((p + 1)/\beta_k)^2} = \frac{\sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v+1)}||^2 / \sum_{i=1}^{n} w_{ik}}{\left( \sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v+1)}|| / \sum_{i=1}^{n} w_{ik} \right)^2}.
\]  

(4.4)

followed by the one-step Newton-Raphson approximation

\[
\beta_k^{(v+1)} = \beta_k \exp \left\{ - \left( \frac{\partial^2 Q_2}{\partial \beta_k^*} \right)^{-1} \frac{\partial Q_2}{\partial \beta_k^*} \right\}.
\]  

(4.5)

Here (4.4) is the weighted moment estimate based on the formula (2.1), and

\[
\frac{\partial Q_2}{\partial \beta_k^*} = \sum_{i=1}^{m} w_{ik} \left[ 1 + \frac{p}{\beta_k} \frac{\partial \log(p/\beta_k)}{\partial \beta_k} + \beta_k \left( \frac{||x_i - \mu_k^{(v+1)}||}{\lambda_k^{(v+1)}} \right)^{\beta_k} \log \left( \frac{||x_i - \mu_k^{(v+1)}||}{\lambda_k^{(v+1)}} \right) \right],
\]

\[
\frac{\partial^2 Q_2}{\partial \beta_k^*} = -\sum_{i=1}^{m} w_{ik} \left[ \frac{p}{\beta_k} \frac{\partial \log(p/\beta_k)}{\partial \beta_k} + \left( \frac{p}{\beta_k} \right)^2 \frac{\partial \log(p/\beta_k)}{\partial \beta_k} + \beta_k \left( \frac{||x_i - \mu_k^{(v+1)}||}{\lambda_k^{(v+1)}} \right)^{\beta_k} \log \left( \frac{||x_i - \mu_k^{(v+1)}||}{\lambda_k^{(v+1)}} \right) \right] + (\beta_k)^2 \left( \frac{||x_i - \mu_k^{(v+1)}||}{\lambda_k^{(v+1)}} \right)^{\beta_k} \log \left( \frac{||x_i - \mu_k^{(v+1)}||}{\lambda_k^{(v+1)}} \right)^2,
\]

where the \( \log(p/\beta_k) \) and \( \log(p/\beta_k) \) are the first and second derivatives of the log-Gamma function respectively.

For each updating, we calculate the resulting value of the observed likelihood. The algorithm can be stopped when the difference between the values at two successive steps is less than a threshold. We choose \( 10^{-4} \) as a default in this paper.
4.2 Algorithm analysis

It follows from the EM theory (for example, McLachlan & Krishnan, 1997) that the above ECM algorithm will converge to a local maximum in general if the objective function is non-decreasing in the CM steps. We give the following analysis of the CM steps.

First, $Q_2$ is non-decreasing in $\mu_k$ in the CM step 1. Consider the second derivative matrix of $Q_2$ with respect to $\mu_k$,

$$
\frac{\partial^2 Q_2}{\partial \mu_k \partial \mu_k} = -\sum_{i=1}^{n} w_{ik} \frac{\beta_k}{\lambda_k^2} |x_i - \mu_k|^2 I_p \left\{ (\beta_k - 2) \frac{\langle x_i - \mu_k, (x_i - \mu_k) \rangle}{||x_i - \mu_k||^2} + I_p \right\}.
$$

Obviously when the current value of $\beta_k \geq 1$, we have $a^\top \frac{\partial^2 Q_2}{\partial \mu_k \partial \mu_k} a \leq 0$, for any $a \in \mathbb{R}^p$, $||a|| = 1$. This implies that the value of $Q_2$ is non-decreasing when the estimate of $\beta_k$ is updated by $\beta_k^{(v+1)}$. For the case that the current estimate of $\beta_k$ is in $(0,1)$, we consider the concave function (with respect to $\mu_k$)

$$
\tilde{Q}_2(\mu_k) = \sum_{i=1}^{n} \sum_{k=1}^{m} w_{ik} \left\{ \log \left( \frac{\beta_k \Gamma(p/2)}{2(\pi \lambda_k)^{p/2} \Gamma(p/\beta_k)} \right) - ||x_i - \mu_k^{(v)}||^{\beta_k-2} ||x_i - \mu_k||^2 \lambda_k^{-\beta_k} \right\}
$$

with

$$
\frac{\partial^2 \tilde{Q}_2}{\partial \mu_k \partial \mu_k} = -2 \sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v)}||^{\beta_k-2} \lambda_k^{-\beta_k} I_p < 0.
$$

Observe that when $0 < \beta_k < 1$,

$$
\frac{\partial Q_2(\mu_k^{(v)})}{\partial \mu_k} = \frac{\partial Q_2(\mu_k^{(v)})}{\partial \mu_k},
$$

$$
\frac{\partial^2 (\tilde{Q}_2 - Q_2)}{\partial \mu_k \partial \mu_k} = \sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v)}||^{\beta_k-2} \left\{ -2 ||x_i - \mu_k^{(v)}||^{\beta_k-2} \frac{\langle x_i - \mu_k, (x_i - \mu_k) \rangle}{||x_i - \mu_k||^{\beta_k-2} I_p}
+ \beta_k (\beta_k - 2) \frac{\langle x_i - \mu_k, (x_i - \mu_k) \rangle}{||x_i - \mu_k||^2} + \beta_k I_p \right\}.
$$

The last equation yields that

$$
\frac{\partial^2 (\tilde{Q}_2 - Q_2)}{\partial \mu_k \partial \mu_k} \leq 0
$$

if $p = 1$. For $p \geq 2$,

$$
\frac{\partial^2 (\tilde{Q}_2 - Q_2)}{\partial \mu_k \partial \mu_k} \bigg|_{\mu_k = \mu_k^{(v)}} = \sum_{i=1}^{n} w_{ik} ||x_i - \mu_k^{(v)}||^{\beta_k-4} (\beta_k - 2)
$$

$$
\times \left\{ ||x_i - \mu_k^{(v)}||^2 I_p + \beta_k \langle x_i - \mu_k^{(v)}, (x_i - \mu_k^{(v)}) \rangle \right\} < 0.
$$

More precisely, we have

$$
\frac{\partial^2 (\tilde{Q}_2(\mu_k) - Q_2(\mu_k))}{\partial \mu_k \partial \mu_k} \leq 0.
$$
provided that \( \| \mu_k - \mu_k^{(v)} \| \leq (1 - (\beta_k/2)^{1/(2-\beta_k)}) \min_i \| x_i - \mu_k^{(v)} \| \). This is because under this condition,
\[
-2 \frac{\| x_i - \mu_k^{(v)} \|^{\beta_k-2}}{\| x_i - \mu_k \|^{\beta_k-2}} + \beta_k \leq 0.
\]
These facts demonstrate that for \( p = 1 \), \( \tilde{Q}_2 \) minorises \( Q_2 \); and for \( p \geq 2 \), \( \tilde{Q}_2 \) at least locally minorises \( Q_2 \) in the neighbour of \( \mu_k^{(v)} \). Therefore, for \( p = 1 \),
\[
Q_2(\mu_k^{(v+1)}) \geq \tilde{Q}_2(\mu_k^{(v+1)}) \geq \tilde{Q}_2(\mu_k^{(v)}) = Q_2(\mu_k^{(v)}).
\]
For \( p \geq 2 \), these inequalities hold if \( \mu_k^{(v+1)} \) is close to \( \mu_k^{(v)} \).

Secondly, \( Q_2 \) is non-decreasing in \( \sigma_k \) in the CM step 2. This can be seen from the fact that
\[
\frac{\partial^2 Q_2}{\partial \sigma_k^2} = -\sum_{i=1}^n w_{ik} \frac{\beta_k}{\lambda_k} \frac{\beta_k}{\lambda_k} \| x_i - \mu_k \|^{\beta_k} \leq 0.
\]
Finally, \( Q_2 \) is also non-decreasing in \( \beta_k \) in the CM step 3, provided that a good starting point for \( \beta_k \) is selected (MaLachlan and Krishnan, 1997, p.155). Our practical experiences indicate that the moment estimate defined in (4.4) did work very well as a starting point.

### 4.3 Initialisation of ECM

There are numerous initialisation strategies for the ECM algorithm, such as using multiple random initial values, using the clustering results of partitioning around medoids (PAM), and using the clustering result of Mclust (Fraley and Raftery, 2006). In general PAM is more robust to outliers than Mclust while the multiple random initialisation is much time-demanding. Mclust performs relatively better compared to the other existing methods in general. When there are some heavily tailed sub-populations in data, the likelihood surface could have many local minimums and thus the maximum likelihood estimation in Mclust becomes difficult. This is because some outliers may form a what is called mini-component with a few data points. As a result, the algorithm may not produce a result at all for some of the models. This problem becomes even worse when we fit an EPD mixture to the data. To attenuate the difficulty, we are going to propose the following alternative initialisation strategy, called the bottom-up initialisation.

The strategy exploits advantages of model-based aggregating method while taking into account the computational feasibility. Set \( K = m_{\text{max}} \) with the default \( K = 12 \). Let the ECM algorithm begin with the clustering result of PAM with \( K \)-medoids. At the level of 0 (i.e. put these \( K \) clusters at the bottom), let \( m_0 = K \), for \( 1 \leq k \leq K \), set \( \mu_k^{(0)} \) equal to the \( k \)-th medoid center; \( \beta_k^{(0)} \) equal to the solution to the equation (4.4) with \( v = 0 \); estimate \( \sigma_k^{(0)} \) by setting \( A_p(\beta_k^{(0)})^2 \sigma_k^{(0)} \) equal to the median of the absolute values with respect to the medoid centre in the \( k \)-th group. Then the initial values of the up levels are set via merging the clusters in the previous step. More specifically, given
Given the minimum BIC model \( m_{v+1} = (K - v - 1) \)-component EPD mixture are formed as follows. First, for any \( k_1 \) and \( k_2 \) components, a combined component can be formed by maximising the log-likelihood

\[
\sum_{i=1}^{n} \log \left( \sum_{k \neq k_1, k_2} \pi_k^{(v)} g(x_i | \theta_k^{(v)}) + \pi_{k_1}^{(v)} g(x_i | \theta_{k_1}) + \pi_{k_2}^{(v)} g(x_i | \theta_{k_2}) \right)
\]

with respect to \( \pi_{k_1} \) and \( \theta_{k_1} \), subject to \( \sum_{k \neq k_1, k_2} \pi_k^{(v)} = 1 - \pi_{k_1} \). Let \( \pi^{(v+1)} \) and \( \theta^{(k+1)} \) denote the solution of this optimisation problem. A simple approximation can be made by setting \( \pi_{k_1}^{(v+1)} = \pi_{k_1}^{(v)} + \pi_{k_2}^{(v)} \), \( (w_{ik_1}^{(v+1)} = w_{ik_1}^{(v)} + w_{ik_2}^{(v)}) \) and

\[
\begin{align*}
\mu_{k_1}^{(v+1)} &= \frac{1}{\pi_{k_1}^{(v+1)}} \left( \frac{\pi_{k_1}^{(v)} \mu_{k_1}^{(v)}}{\pi_{k_1}^{(v)} + \pi_{k_2}^{(v)} \mu_{k_2}^{(v)}} \right), \\
\lambda_{k_1}^{(v+1)} &= \frac{1}{\pi_{k_1}^{(v+1)}} \left( \frac{\pi_{k_1}^{(v)} \lambda_{k_1}^{(v)}}{\pi_{k_1}^{(v)} + \pi_{k_2}^{(v)} \lambda_{k_2}^{(v)}} \right), \\
\beta_{k_1}^{(v+1)} &= \frac{1}{\pi_{k_1}^{(v+1)}} \left( \frac{\pi_{k_1}^{(v)} \beta_{k_1}^{(v)}}{\pi_{k_1}^{(v)} + \pi_{k_2}^{(v)} \beta_{k_2}^{(v)}} \right).
\end{align*}
\]

This, combined with the remaining \( (m_v - 2) \) components in the previous step, forms a \( m_{v+1} \)-component EPD mixture. The observed log-likelihood value \( l_{k_1,k_2} \) based on this model can be easily calculated by using the recursive formula

\[
f_{(v+1)}(x_i | \psi^{(v+1)}) = f^{(v)}(x_i | \psi^{(v)}) - \pi_{k_1}^{(v)} g(x_i | \theta_{k_1}^{(v)}) - \pi_{k_2}^{(v)} g(x_i | \theta_{k_2}^{(v)}) + \pi_{k_1}^{(v+1)} g(x_i | \theta_{k_1}^{(v+1)}) + \pi_{k_2}^{(v+1)} g(x_i | \theta_{k_2}^{(v+1)}), \quad i = 1, 2, \ldots, n.
\]

We choose \((k_1, k_2) = \arg\max l_{k_1,k_2}\). The refined \( l_{k_1,k_2} \) can be obtained by updating (4.6) via the formulae (4.2), (4.3), and (4.5).

Note that in practice, we run with both the bottom-up initialisation and the Mclust initialisation and choose the one with the smaller BIC value.

5 Cluster analysis

5.1 Cluster assignments

Given the minimum BIC model

\[
\sum_{k=1}^{\hat{m}} \hat{\pi}_k g(x_i | \hat{\theta}_k),
\]

\( w_{ik} \) can be estimated by

\[
w_{ik} = \frac{\hat{\pi}_k g(x_i | \hat{\theta}_k)}{\sum_{j=1}^{\hat{m}} \hat{\pi}_j g(x_i | \hat{\theta}_j)}.
\]

Assign the \( i \)-th observation to the \( k \)-th cluster \( C_k \) if

\[
w_{ik}^* = \max_{1 \leq l \leq \hat{m}} w_{il}.
\]

The uncertainty of this assignment is defined as \( 1 - w_{ik}^* \). The expected misclassification rate of the \( k \)-th cluster is defined \( 1 - \sum_{i \in C_k} w_{ik}^*/|C_k| \), where \(|C_k|\) is the number of elements in \( C_k \).
5.2 Super-clusters

It is often the case that one is interested in not only the clusters but also the relationship among the clusters. For example, in multiple testing problems, one is more interested in grouping the \( z_i \)'s into two super-clusters. Each of them is made of several clusters. The null and non-null components are then described by these super-clusters respectively. Here we present a bottom-up hierarchical description of the \( \hat{m} \) components (clusters) defined by (5.1), i.e., a series of partitions of these components: the bottom consists of \( \hat{m} \) single-member groups of components; the top consists of a single group containing all \( \hat{m} \) components. Each partition is achieved by selecting and merging two super-clusters from the previous partition in order to maximise the merged likelihood. The details are as follows. To begin with, let \( s^{(1)}_1 = \{1\}, ..., s^{(1)}_{\hat{m}} = \{\hat{m}\} \) denote \( \hat{m} \) single component groups at the level 1. They have the EPD parameters \((\pi^{(1)}_1, \theta^{(1)}_1), ..., (\pi^{(1)}_{\hat{m}}, \theta^{(1)}_{\hat{m}})\). For any \( j < k \), the proximity of \( s^{(1)}_j \) and \( s^{(1)}_k \) is measured by the likelihood \( L_{j,k} \) for the sample \( X \) based on a new \((\hat{m} - 1)\) component EPD mixture model. This new mixture model is produced by inheriting \((m - 2)\) components from the previous step, i.e., for \( 1 \leq t \leq j - 1 \), set \((\pi^{(2)}_t, \theta^{(2)}_t) = (\pi^{(1)}_t, \theta^{(1)}_t)\), for \( j + 1 \leq t \leq \hat{m} \), set \((\pi^{(2)}_t, \theta^{(2)}_t) = (\pi^{(1)}_{t-1}, \theta^{(1)}_{t-1})\), and by forming a new component with

\[
\pi^{(2)}_j = \pi^{(1)}_j + \pi^{(1)}_k, \quad \theta^{(2)}_j = \frac{\pi^{(1)}_j \theta^{(1)}_j + \pi^{(1)}_k \theta^{(1)}_k}{\pi^{(2)}_j}, \quad \beta^{(2)}_j = \frac{\pi^{(1)}_j \beta^{(1)}_j + \pi^{(1)}_k \beta^{(1)}_k}{\pi^{(2)}_j}.
\]

The closer the cluster \( s^{(1)}_j \) to \( s^{(1)}_k \), the higher the resulting value of likelihood \( L_{j,k} \). We select \( j_1 < k_1 \) to maximise \( L_{j,k} \). The partition at level 2 is produced by merging the \( j_1 \) and \( k_1 \) components. We repeat this procedure to the partition at each level until the level \( \hat{m} \) is reached. In particular, at the level \( \hat{m} - 1 \) we produced a two-group partition of the \( \hat{m} \)-components, say

\[
\{j_1, j_2, ..., j_{m_0}\} \cup \{k_1, k_2, ..., k_{m_1}\} = \{1, 2, ..., \hat{m}\}.
\]

This yields estimators

\[
\hat{\pi}_0 = \sum_{t=1}^{m_0} \hat{\pi}_j, \quad \hat{f}_0 = \sum_{t=1}^{m_0} \frac{\hat{\pi}_j}{\hat{\pi}_0} g(\cdot|\hat{\theta}_j),
\]

\[
\hat{\pi}_1 = 1 - \hat{\pi}_0, \quad \hat{f}_1 = \sum_{t=1}^{m_1} \frac{\hat{\pi}_k}{\hat{\pi}_1} g(\cdot|\hat{\theta}_k),
\]

for \( \pi_0, f_0, \) and \( f_1 \) in the model

\[
f(\cdot) = \pi_0 f_0(\cdot) + (1 - \pi_0) f_1(\cdot), \quad (5.2)
\]

5.3 Measure of clustering quality

Let \( X = (x_1, ..., x_n) \) denote a set of \( n \) observations with the underlying (non-overlapping) grouping \( C = \{c_1, ..., c_l\} \). Let \( \hat{C} = \{\hat{c}_1, ..., \hat{c}_s\} \) denote a clustering result. Viewing \( C \) and \( \hat{C} \) as two partitions
for \( \mathbf{X} \), we use the adjusted Rand index of Hubert and Arabie (1985) to assess the degree of agreement between \( \hat{C} \) and \( C \). Let \( n_{ij} \) be the number of observations that are in both group \( c_i \) and cluster \( \hat{c}_j \), \( n_i \) be the number of observations in group \( c_i \), and \( n_j \) be the number of observations in clustering \( \hat{c}_j \). The adjusted Rand index is defined as

\[
\rho = \frac{\sum_{ij} (\binom{n_{ij}}{2}) - |\sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})|/\binom{n}{2}}{(|\sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})|/\binom{n}{2} - |\sum_i (\binom{n_i}{2}) \sum_j (\binom{n_j}{2})|)/\binom{n}{2}}.
\]

The larger value of \( \rho \) the higher degree of agreement between two partitions. \( \rho \) takes the maximum value of 1 when two partitions are identical.

In a simulation study, the adjusted Rand index can be directly used to assess the quality of a clustering result, because the true grouping is known. However, this is untrue in a real data analysis, where the following resampling based adjusted Rand index could be useful: Given the grouping \( \hat{C} \) of \( \mathbf{X} \), we draw \( v_0 \) random sub-samples from \( \mathbf{X} \), each of size \( n_1 < n \). We set \( n_1 \) to be around 70\% of \( n \). Since the estimated groupings of these samples are known, we can apply the clustering method to each sub-sample, resulting in a grouping for it. Then for each sub-sample, calculate the adjusted Rand index \( \rho \) between the new grouping and the grouping inherited from \( \hat{C} \). The average of these \( \rho \)'s is called a resampling based estimate of the unknown adjusted Rand index.

6 Theoretical properties

6.1 Identifiability

The identifiability, the existence of a unique characterization for any one of the class of models being considered, lies at the heart of most statistical theory and practice. Without it, the above model is not to be well defined. Finite mixtures from the location-scatter family \( \{g(\cdot|\mu, \Sigma, \beta) : \mu \in \mathbb{R}^p, \Sigma \in D_p, \beta \in (0, \infty)\} \) are called identifiable if a relation of the form

\[
\sum_{k=1}^m \pi_k g(x|\mu_k, \Sigma_k, \beta_k) = \sum_{k=1}^{m^*} \pi_k^* g(x|\mu_k^*, \Sigma_k^*, \beta_k^*), \quad x \in \mathbb{R}^p,
\]

where \( m \) and \( m^* \) are positive integers, \( \sum_{k=1}^m \pi_k = \sum_{k=1}^{m^*} \pi_k^* = 1 \) and \( \pi_k > 0, 1 \leq k \leq m, \pi_k^* > 0, 1 \leq k \leq m^* \), implies that \( m = m^* \) and that there exists a permutation \( \nu \) on \( 1, 2, \ldots, m \) such that \( (\pi_k^*, \mu_k^*, \Sigma_k^*, \beta_k^*) = (\pi_{\nu(k)}^*, \mu_{\nu(k)}, \Sigma_{\nu(k)}, \beta_{\nu(k)}) \). We have

**Theorem 1** The above multivariate exponential power mixtures are identifiable for \( 1 \leq p < \infty \).

We now investigate the identifiability of the super-cluster model (5.2). To begin with, let \( \mu_g \) denote the mean of the density function \( g \). Consider the following set of EPD mixture pairs,

\[
\mathcal{F}_2 = \{(h_0, h_1) : \quad h_0 = \sum_{k=1}^{m_0} \pi_{0k} g_{0k}, h_1 = \sum_{j=1}^{m_1} \pi_{1j} g_{1j}, \quad \pi_{0k} > 0, 1 \leq k \leq m_0, \sum_{k=1}^{m_0} \pi_{0k} = 1; \quad \pi_{1j} > 0, 1 \leq j \leq m_1, \sum_{j=1}^{m_1} \pi_{1j} = 1; \quad \{g_{0k}\} \text{ and } \{g_{1j}\} \text{ are EPDs and } \max_k \mu_{g0k} < \min_j \mu_{g1j}\}
\]
Proposition 1 Suppose that the true pair of null and non-null densities in (5.2) belongs to $\mathcal{F}_2$. Then the model (5.2) is identifiable subject to $(f_0, f_1) \in \mathcal{F}_2$, i.e., if $\pi_0 f_0 + (1 - \pi_0) f_1 = \pi_0^* f_0^* + (1 - \pi_0^*) f_1^*$, and $(f_0, f_1), (f_0^*, f_1^*) \in \mathcal{F}_2$, then $\pi_0 = \pi_0^*$ and $f_j = f_j^*$, $j = 0, 1$.

Note that the mixture model in (5.2) is not identifiable if $f_1$ is left unspecified, even $f_0$ is assumed to have normality. See Bordes, Delmas and Vandekerkhove (2006).

6.2 Asymptotics

To ease of presentation, we rewrite the finite mixture distribution $\sum_{k=1}^m \pi_k g(x|\theta_k)$ as

$$p_\psi(x) = \int g(x|\theta) dF(\theta),$$

where $F(\theta) = \sum_{k=1}^m I(\theta = \theta_k)$ and $I(\theta = \theta_k)$ is an indicator function that it is equal to 1 if $\theta = \theta_k$ and equal to 0 otherwise. For $\psi \in \Psi_m$ and $\phi \in \Psi_m$, the Kullback-Leibler divergence between two mixtures $p_\psi$ and $p_\phi$ is defined by

$$K(p_\psi, p_\phi) = \int p_\psi \log \left( \frac{p_\psi}{p_\phi} \right).$$

Note that

$$-K(p_\psi, p_\phi) \leq \int p_\psi (p_\phi/p_\psi - 1) = 0.$$

And $K(p_\psi, p_\phi) = 0$ if and only if $\int p_\psi p_\phi(y) = p_\psi(y) p_\phi = 0$. Let $\mathcal{F}_m = \{ \sum_{k=1}^m \pi_k g(x|\theta_k) : \pi_k > 0, 1 \leq k \leq m, \sum_{k=1}^m \pi_k = 1, \theta \in \Theta \}$.

Since the EPD family is identifiable, for any two finite mixture distributions $p_\psi = \int g(x|\theta) dF_\psi(\theta) \in \mathcal{F}_{m_1}$ and $p_\phi = \int g(x|\theta) dF_\phi(\theta) \in \mathcal{F}_{m_2}$ with $m_1 \leq m_2$, the equality $K(p_\psi, p_\phi) = 0$ holds if and only if $F_\phi$ and $F_\psi$ have the same supporting points of non-zero masses up to a permutation of these points. In particular, $K(p_\phi, p_\psi) < 0$ when $F_\psi$ and $F_\phi$ has the different number of supporting points of non-zero masses.

For simplicity, we consider only the constrained estimator $(\hat{\psi}_m, \hat{m})$ of $(\psi_m, m)$ over $\psi_m \in \Psi_m$, $1 \leq m \leq m_{\text{max}}$ based on the BIC. Here

$$\Psi_m = \{ \psi_m : \pi_k \geq 0, \sum_{k=1}^m \pi_k = 1, \| \mu_k \| \leq b_1, \lambda_{\min}(\Sigma_k) > \delta_1, \delta_2 \leq \beta_k \leq b_2 \},$$

for some sufficiently large constants $b_1$, $b_2$, and $m_{\text{max}}$ and some sufficiently small constants $\delta_1$, $\delta_2$. To show the consistency of the constrained MLE, we introduce the following conditions:

(C1) For $\psi_m \in \Psi_m$,

$$C(\psi_m) = \frac{2}{n(n-1)} \sum_{i<j} \text{corr}(\log(p_{\psi_m}(x_i)), \log(p_{\psi_m}(x_j))) = o(1),$$

where $\text{corr}(\log(p_{\psi_m}(x_i)), \log(p_{\psi_m}(x_j)))$ is the correlation coefficient between $\log(p_{\psi_m}(x_i))$ and $\log(p_{\psi_m}(x_j))$. 

15
(C2) \( E \sup_{\psi \in \Psi_m} |\log(p_{\psi})| < \infty. \)

(C3) The underlying model is a \( m_0 \)-component EPD mixture with \( 1 \leq m_0 \leq m_{\max} \) and the true value \( \psi_{\theta m_0} \).

Note that the condition (C1) means that the correlations are sparse. It covers many interesting cases of dependence, such as the data with clumpy correlations, and the data with long memory. We refer to a dataset as correlated in clumps if the data points can be divided into groups \( S_1, \ldots, S_{n_1} \) with the maximum size \( n_2 = \max_j |S_j| \) so that the data points are only correlated within each group (Lumley and Mayer-Hamblett, 2003). In this case, we have

\[
|C(\psi_m)| = \left| \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \in S_i} \text{corr}(\log(p_{\psi_m}(x_i)), \log(p_{\psi_m}(x_j))) \right| 
\leq O\left( \frac{2}{n(n-1)} \right) \sum_{i=1}^{n} n_2 = O\left( \frac{2n_2}{n-1} \right) = O\left( \frac{1}{n_1} \right) = o(1)
\]

provided the dependency is sparse in the sense that \( n_1 \to \infty \) and \( n_1 n_2 = O(n) \). This yields that the average correlation coefficient over all pairs of genes is quite low even if the genes involved in the same group are heavily correlated.

We refer to a stationary dataset as having a long memory if there exists a permutation, say \( \{y_1, \ldots, y_n\} \) of \( \{x_1, \ldots, x_n\} \) such that

\[
\sup_{1 \leq i \leq n-k} \max_{f_1 \in L^2(\sigma^i_j), f_2 \in L^2(\sigma^i_{i+k})} |\text{corr}(f_1, f_2)| = O(k^{-\alpha}), 1 > \alpha > 0,
\]

where \( \sigma^i_j \) is the \( \sigma \)-field generated by \( y_1, \ldots, y_i \), \( \sigma^i_{i+k} \) is the \( \sigma \)-field generated by \( y_{i+k}, \ldots, y_n \), and \( L^2(\sigma^i_j) (L^2(\sigma^i_{i+k})) \) is all the \( \sigma^i_j (\sigma^i_{i+k}) \) measurable and square integrable functions. See Beran (1994).

In this case, we also have

\[
C(\psi_m) \leq \frac{2}{n(n-1)} \sum_{i<j} r_{j-i} = O\left( \frac{1}{n^2} \right) \sum_{i<j} (j-i)^{-\alpha} = O\left( \frac{1}{n^2} \right) \sum_{k=1}^{n-1} \frac{(1-k/n)^{k-\alpha}}{k^{\alpha}} = O(n^{-\alpha}) = o(1).
\]

A similar result holds for the data with short memory defined in Beran (1994). The following theorem shows the consistency of the proposed estimators for EPD mixture models.

**Theorem 2** Under the conditions (C1)–(C3), \( P(\liminf \hat{m} \geq m_0) = 1 \) and \( \|\hat{\psi}_m - \psi_{\theta m}\| = o_p(1), \)

where \( \psi_{\theta m} = (\psi_{\theta m_0}, 0, \ldots, 0) \) with the last \( (\hat{m} - m_0)(p(p+3)/2 + 1) \) components being 0.

**Remark 6.1** Theorem 2 implies that the estimated number of the asymptotically non-negligible components are close to the true number of the components.
Remark 6.2 A first-order Taylor series approximation to the covariance between \( \log(p_{\psi_m}(x_i)) \) and \( \log(p_{\psi_m}(x_j)) \) can be made as follows:

\[
\frac{\partial p_{\psi_m}(E[x])}{\partial(x)}(p_{\psi_m}(E[x]))^{-1}\text{cov}(x_i - E[x], x_j - E[x]) \frac{\partial p_{\psi_m}(E[x])}{\partial(x)}(p_{\psi_m}(E[x]))^{-1},
\]

where \( \frac{\partial p_{\psi_m}(E[x])}{\partial(x)} \) is the partial derivative of \( p_{\psi_m} \) at \( E[x] \). Similarly for \( 1 \leq i \leq n, \)

\[
\text{var}(\log(p_{\psi_m}(x_i))) \approx \frac{\partial p_{\psi_m}(E[x])}{\partial(x)}(p_{\psi_m}(E[x]))^{-1}\text{cov}(x_i, x_j) \frac{\partial p_{\psi_m}(E[x])}{\partial(x)}(p_{\psi_m}(E[x]))^{-1}.
\]

If \( \text{cov}(x_i, x_i), 1 \leq i \leq n, \) are unit matrices, then \( C(\psi_m) \) can be approximated by

\[
c_0(\psi_m)^\tau \frac{2}{n(n-1)} \sum_{i < j} \text{corr}(x_i, x_j) c_0(\psi_m),
\]

where \( c_0(\psi_m) \) is a constant vector. This means that \( C(\psi_m) \) is approximately proportional to \( \frac{2}{n(n-1)} \sum_{i < j} \text{corr}(x_i, x_j) \). Therefore as \( n \) tends infinity, \( C(\psi_m) = o(1) \) if \( \frac{2}{n(n-1)} \sum_{i < j} \text{corr}(x_i, x_j) = o(1) \).

Remark 6.3 Note that the slowly decaying correlations may lead to a rather slow convergence of the MLE. In particular, if \( |C(\psi_m)| > 0 \) (i.e., the average correlation doesn’t decay as \( n \to \infty \)), then MLE may not converge at all. For example, let \( \{x_i, 1 \leq i \leq n\} \) be identically distributed as \( N(0, 1) \) but strongly correlated in the sense that \( \rho(x_i, x_j) = \rho_o > 0 \) for all \( i < j \). If the ML estimates \( \hat{\mu}, \hat{\sigma} \) and \( \hat{\beta} \) are consistent, then \( \hat{\mu} \) is approximately equal to the sample mean \( \bar{x} \). However, \( n(\bar{x} - E[x])^2 \to \chi^2_1 \) weakly (Beran, 1994), suggesting the \( \bar{x} \) does not converge to any value at all. This shows that the strong correlation can potentially deteriorate the model based cluster analysis.

7 Numerical studies

7.1 Simulations

In this sub-section, we will investigate the performance of the EPD procedure under a variety of different scenes. Using simulations we attempt to answer the following questions:

1. What is the impact of general dependence on the EPD and GM (implemented as Mclust by Fraley and Raftery, 2006) procedures?

2. When will the EPD procedure overperform the GM?

Example 1: This example is modified from the dataset SIMU3 in Qiu, Brooks et al.(2005). We simulated 30 datasets, each has \( n = 1000 \) data points, say \( \{x_i, 1 \leq i \leq 1000\} \), correlated in six independent groups: \( x_{11}, \ldots, x_{150} \sim N(s \times 1_{150}, \Sigma_1), (x_{151}, \ldots, x_{200}) \sim N(0, \Sigma_2), (x_{201}, \ldots, x_{300}) \sim N(0, \Sigma_3), (x_{301}, \ldots, x_{400}) \sim N(0, \Sigma_4), (x_{401}, \ldots, x_{500}) \sim N(0, \Sigma_5), \) and \( (x_{501}, \ldots, x_{1000}) \sim N(-s \times \)
where $1_{150} = (1, ..., 1)^T$, $1_{500} = (1, ..., 1)^T$, and $\Sigma_k$ is determined by $\text{var}(x_i) = 1$, $\text{cov}(x_i, x_j) = |j - i|^{-a_k}/(2 - 2^{-a_k})$, $i < j$, $a_k \sim U(\alpha_1, \alpha_2)$ (a uniform distribution over $(\alpha_1, \alpha_2)$). We are interested in finding differentially expressed genes, while the correlation pattern between these genes is not primary interest. We consider two cases with $s = 2$ and 6 respectively. The former is a case with not well separated clusters while the latter is a case with well separated clusters. In both cases we examine the performances of the EPD and GM methods when $(\alpha_1, \alpha_2) = (1/4, 1/2), (1/16, 1/8), (1/64, 1/32)$ and $(0, 1/64)$ respectively. For the above four values of $(\alpha_1, \alpha_2)$, the average correlation coefficient $C = \frac{2}{n(n-1)} \sum_{i<j} \text{corr}(x_i, x_j)$ satisfies $0.032 \leq C \leq 0.090$, $0.163 \leq C \leq 0.221$, $0.259 \leq C \leq 0.281$, and $0.281 \leq C \leq 0.304$. They stand for four scenarios with increasing but not very high degrees of dependence. The results are reported in Figure 2 and summarised in Table 1, showing that the impact of the clumpy dependence on the model based clustering could be serious even with a moderate value of average correlation coefficient such as $C = 0.304$. The results also show that the EPD procedure performs much better than the GM under clumpy dependence.

Example 2: This example is modified from an example in Qiu,Klebanov et al.(2005). We simulate 60 datasets as follows. First, generate a $1255 \times 40$ matrix and denote this matrix by $X = (x_{ij})_{1255 \times 40}$. All the elements $x_{ij}$ of this matrix are stochastically independent; but the elements with $1 \leq i \leq 125$ and $1 \leq j \leq 21$ are drawn from $N(2, 1)$, and other elements are drawn from $N(0, 1)$. The first 125 rows model the 125 differentially expressed genes. The remaining genes belong to another class. Next, generate a 40-dimensional random vector $\mathbf{a} = (a_1, ..., a_{40})$ with $a_i$ standing for an $i$-th array specific noise and being drawn from $N(0, 1)$ independently. Define

$$y_{ij} = r^{1/2}a_j + (1 - r)^{1/2}x_{ij}, 1 \leq i \leq 1255, 1 \leq j \leq 40.$$  

Then, $y_{i,j}$ and $y_{i'j}$ have correlation coefficient $r$ for any $i_1 \neq i_2$ and $j$. Here we consider three cases with $r = 0, 0.3, 0.6$, and 0.9 respectively. The first 21 columns represent the gene expressions under a control condition, while the remaining 19 columns show the gene expressions when a treatment is introduced. For each gene, a two-sample $t$-statistic was applied to measure the differences between the log gene expression levels under two experimental conditions. The resulting P-values are transformed via the $z_i = \Phi^{-1}(1 - P_i)$. We apply both the GM and the EPD procedures to these $z_i$’s. The results, described in Table 2 and Figure 3, again indicate that the EPD procedure has a better performance than the GM method under array-specific dependence.
Example 3: We generate the following 40 datasets to assess the performance of our EPD approach when there is a uniformly distributed sub-population. Each data set has 375 observations from the uniform distribution $U(0, 4)$ and 125 realisations from the normal distribution $N(6, 1)$. We apply the EPD and GM approaches to these data sets. The results summarised in Table 3 show that the EPD approach has 2.5 times higher adjusted Rand index than the GM does.

Example 4: We test our EPD clustering in the presence of heavy tailed clusters as well as low kurtosis clusters in this example. Four clusters centred at 0, -4, 6 and 4 are generated with 2600 data points in total. Among them 1000 are sampled from $\text{epd}_1(0, 1, 1.5)$ which is slightly deviated from the standard normal, another 1000 are from $\text{epd}_1(-4, 2, 0.8)$, 400 observations are from the normal $N(6, 1)$, and 200 observations are from $\text{epd}_1(4, 1, 4)$. The distribution $\text{epd}_1(-4, 2, 0.8)$ has heavy tails and a sharp peak at the mean, while $\text{epd}_1(4, 1, 4)$ has low coefficient of kurtosis 2.18844, tending to have a flat top near the mean rather than a sharp peak. We simulate 70 times. The histogram of one of these datasets is plotted in Figure 4. The results are summarised in Table 3. Compared with the GM, the EPD method shows considerably higher accuracy in terms of both adjusted Rand index and error rate of selecting correct number of clusters.

Example 5: To examine the behaviour of our EPD procedure in multivariate settings, we simulate 59 datasets. Each dataset has 130 subjects sampled from $\text{epd}_6(\mu_1, I_6, 1.5)$, 50 subjects sampled from $\text{epd}_6(\mu_2, I_6, 4)$, 70 subjects sampled from $\text{epd}_6(\mu_3, 2^2 I_6, 0.8)$, 110 subjects sampled from $\text{epd}_6(\mu_4, I_p, 2)$, and 140 subjects sampled from $\text{epd}_6(\mu_5, I_p, 1.6)$. Here $I_6$ is a 6 by 6 unit matrix, and for $1 \leq k \leq 5$, $\mu_k$ is a normalisation of the vector $v_k$ across its components, i.e.,

$$
\mu_k = \frac{v_k - \text{mean}(v_k)}{\sqrt{\text{var}(v_k)}}, \quad \text{mean}(v_k) = \frac{1}{6} \sum_{j=1}^{6} v_{kj}, \quad \text{var}(v_k) = \frac{1}{5} \sum_{j=1}^{6} (v_{kj} - \text{mean}(v_k))^2,
$$

where

$$
\mathbf{a} = (-0.75, -0.50, 0.25, 0.65, 0.85)^T, \quad \mathbf{b} = (0.25, 0.5, 1.25, 3.50, 3.80)^T, \quad \mathbf{d} = (1, 2, 3, 4, 5, 6)^T; \quad v_{1j} = \cos(2 * a_1 * d_j * 3.1416 / 5 + b_1), \quad v_{2j} = \cos(2 * a_2 * d_j * 3.1416 / 5 + b_2),
$$

Example 3

Example 4

Example 5
\[
v_{3j} = \cos(2 \ast a_3 \ast d_j \ast 3.1416/5 + b_3), \quad v_{4j} = \cos(2 \ast a_4 \ast d_j \ast 3.1416/5 + b_4), \\
v_{5j} = \cos(2 \ast a_5 \ast d_j \ast 3.1416/5 + b_5), \quad 1 \leq j \leq 6.
\]

These datasets mimic the log-expression levels of proteins produced by 2D-gel experiments, where 500 protein spots are assumed to be observed in each cell line. We investigate the expression patterns of these proteins under 6 different treatments. The 6 coordinates of \( \mu_k \) represent the group effects of 6 different treatments on the subjects belong to the \( k \)-th group respectively. These \( \mu_k \)'s are plotted in Figure 5(a). We apply both the GM and EPD procedure to these datasets. The results displayed in Figure 5(b) show that the EPD procedure performs better than the GM.

7.2 Real data analysis

In this sub-section, we are going to test our proposed procedure on a number of real data arising from some genetic studies.

Avian pineal gland gene expression dataset: This dataset arises from an experiment made by Dr. Cassone’s lab at Texas A &M university. The expression levels of 7400 pineal gland genes were measured under light-dark (LD) and constant darkness (DD) conditions. The birds were euthanized at Zeitgeber time (ZT) 2, 6, 10, 14, 18, 22 hour to obtain mRNA to produce adequate cDNA libraries. Four microarray chips per time point were produced, and there were two replicates for each gene in chip. Throughout the experiment, samples from LD ZT18 were used as controls. The goal is to identify genes that are differentially expressed at different time points. Some mixed effect analysis presented a sequence of P-values, \( P_i, 1 \leq i \leq 7400 \), which are transformed to z-scores \( x_i = \Phi^{-1}(1 - P_i) \) (Liang, Liu, and Wang, 2007). The histogram of these z-scores under the DD condition is plotted at the top left of Figure 6. Two peaks could be spotted, where the main one, corresponding to the null-component, won’t fit in a single normal distribution. The second peak is very small, reflecting the small number of differentially expressed genes. With the EPD approach, the best model is a two-component mixture, \( 0.8717 \text{epd}_1(0.3614, 0.8816^2, 1.3202) + 0.1283 \text{epd}_1(1.0788, 3.1646^2, 6.1992) \). It suggests two clusters for these genes, most of genes (with the proportion of 0.8717) are in the cluster of non-differentially expressed genes and a small number belongs to the cluster of differentially expressed genes. By contrast, with the GM, the best model is a three-component Gaussian mixture,

\[
0.4822N(0.1845, 0.8581) + 0.2589N(0.4080, 0.1430) + 0.2588N(0.9171, 2.0826).
\]

The three associated clusters are displayed in Figure 6, indicating at least 25.88% of genes are differentially expressed. This probably exaggerates the proportion of the differentially expressed
genes, since this proportion is believed to be less than 10% by biologists in large-scale microarray studies (Efron, 2004). This biased result may be caused by the correlation effects as demonstrated in the previous simulations.

Breast cancer gene expression dataset: Efron (2004) investigated a well-known microarray experiment concerning differences between two types of genetic mutations causing increased breast cancer risk, BRCA1 and BRCA2. The experiment made by Hedenfalk et al. (2001) included 15 microarrays for 15 breast cancer patients respectively, seven with the BRCA1 mutation and eight with BRCA2. Each microarray is reporting on the same set of 3226 genes. For each gene, a P-value was calculated using the classical t-statistics under the null assumption that there is no difference between BRCA1 and BRCA2. These P-values are then transformed to z-scores as we did in the previous example. The histogram of these scores is displayed in Figure 7.

We begin with the assumption of Gaussian mixture distributions on the data and that the component with the smallest centre is associated with non-differential expressed genes. We run the GM method on these z-scores, resulting in the two-component fit,

$$0.5055686N(-0.97408, 1.18057) + 0.49443N(0.91852, 1.18057).$$

This entails 0.49443 as the potential estimate of the proportion of differentially expressed cases. This value is far large than 10%, the empirical bound for the percentage of interesting cases assumed in a large scale microarray study (Efron, 2004). This cluster of differentially expressed cases may be generated by the correlation effects.

Unlike the GM method, the EPD method gives rise to the best fit to these z-scores, epd\(_1\)(−0.03677, 2.23129\(^2\), 2.48574), meaning that all these genes are non-differentially expressed. This is in agreement with the fact observed by Efron (2004).

2D gel proteomic dataset: Ahmad et al. (2006) investigated the proteome of murine myeloma NS0 cells stably producing a recombinant chimeric monoclonal antibody at different specific production rates. Specifically, the cell lines under investigation (transfectant blank, 4O, 4R, 2X, 2P, 2N2) produce antibody in increasing amounts with the control (blank) producing no antibody and 2N2 producing the most. Two-dimensional electrophoresis separation of NS0 cell protein extracts was performed. Protein spots were then detected using the resulting gel images. After background subtraction and normalisation, a 2817 × 6 data matrix was obtained, where
the rows are corresponding to 2817 spots and the columns to the 6-cell lines in the orders of control, 4O, 4R, 2X, 2P, and 2N2. The goal is to understand why some cell lines secrete more recombinant antibody than others. In the dataset 839 spots were completely observed across all 6-cell lines. Here we focus on these completely observed spots. Since the variability patterns of these spots have been studied in Ahmad et al. (2006), we intend to extract non-linear patterns of the spots across the 6-cell lines from this $839 \times 6$ sub-data matrix. For this purpose, the log-expression level of each spot is standardised across the cell lines to have mean 0 and variance 1. The resulting data matrix is denoted by $X = (x_{ij})$. Note that the rows are degenerate in the sense that $\sum_{j=1}^{6} x_{ij} = 0, 1 \leq i \leq 839$. So we apply both the GM and the EPD approaches to only the first 5 columns of $X$. The GM yields 5 clusters with the sizes 146, 332, 99, 98, and 164 respectively. The EPD procedure also produces 5 clusters with the sizes 51, 401, 121, 112, and 154 respectively. The normalised log-expression curves of the protein spots at the 6 cell lines are depicted for each cluster in Figure 8. These two clustering results are different as their adjusted Rand index is lower than 0.56. The EPD approach suggests a group of 51 protein spots, which first slightly up-expressed when treatments are changed from Control to 4O, then down-expressed from 4O to 4R to 2X to 2P, and finally up-expressed when the treatments are changed from 2P to 2N2. The GM also identifies a group of 141 protein spots, which showed significant down-expression and up-expression when the treatments are changed from 2X to 2P to 2N2. See Figure 8.

8 Conclusion

In this paper we have studied the impact of general dependence on the model-based methods. The study provides direct evidence that the correlation structure of microarray data cannot be ignored when designing methods for clustering genes. The misspecification of the distribution shapes has been found to be the main effect of dependence on the performance of the GM method. Based on this finding, we have proposed a family of EPD mixture models for clustering, taking advantage of the varying distribution shapes in the underlying clusters. The numerical results have indicated that compared to the GM method, the new proposal has a nice feature in the robustness to the violation of the model assumptions. The EPD mixtures have been shown to be identifiable. The large sample theory for the BIC based estimator has been established under sparse dependence. To solve the optimisation problem in calculating the maximum likelihood estimators, we have developed an expectation-conditional-maximisation algorithm with a bottom-up model-based initialisation strategy. The Bayesian Information Criterion (BIC) has been chosen to select the number of components in the models. The above algorithm has been proved to converge to a
local maximum of the observed likelihood under some conditions.

We note that Fraley and Raftery (2002) considered various parameterisations of the scatter matrices in their software Mclust. For the ease of programming, we only consider the spherical case. The extension of our procedure to other parameterisations is straightforward and thus provided in the Appendix II.

**Appendix I: Proofs of the main theorems**

To prove Theorem 1, we need the following two facts of Holzmann, Munk and Gneiting (2006):

1. the above finite mixtures is identifiable if restrict \( \beta_k, 1 \leq k \leq m \) by \( 1 < \beta_k < \infty \); (2) the identifiability of finite mixtures from the univariate exponential power family \( \{ g(.|\theta) : \theta = (\mu, \sigma, \beta) \in R^1 \times R^+ \times R^+ \} \) implies the identifiability of finite mixtures from the multivariate exponential power family. These two facts imply that to prove Theorem 1, it suffices to show that finite mixtures of the univariate exponential power distributions are identifiable. The proof of the identifiability of the univariate exponential power mixtures is based on the following two lemmas.

**Lemma 8.1** Let \( \Gamma(\cdot) \) denote the Gamma function. Then, for \( k \geq v \geq \beta_j - 1 \), we have

\[
\frac{\Gamma((v+1)/\beta_j)}{\Gamma((k+1)/\beta_j)} \leq e^{\beta_j/12} \left( \frac{e\beta_j}{k+1} \right)^{(k-v)/\beta_j}.
\]

For \( v < \beta_j - 1 < k \), we have

\[
\frac{\Gamma((v+1)/\beta_j)}{\Gamma((k+1)/\beta_j)} \leq 2^{3/2} e^{\beta_j/12} \frac{\beta_j}{k+1} \left( \frac{e\beta_j}{k+1} \right)^{(k-v)/\beta_j}.
\]

For \( k + 1 \geq \beta_j \geq \beta_m \),

\[
\frac{\Gamma((k+1)/\beta_j)}{\Gamma((k+1)/\beta_m)} \leq e^{\beta_j/12} \frac{\beta_j}{\beta_m} \left( \frac{\beta_m}{\beta_j} \right)^{1/\beta_j} \left( \frac{e\beta_m}{k+1} \right)^{1/\beta_m-1/\beta_j} \left( \frac{e\beta_j}{k+1} \right)^{k+1}.
\]

**Proof of Lemma 8.1:** It directly follows from the following inequality for the Gamma function (Devroye, 1986, p490): For \( x \geq 0 \),

\[
\left( \frac{x+1}{e} \right)^{x+1} \sqrt{\frac{2\pi}{x+1}} \leq \Gamma(x+1) \leq \left( \frac{x+1}{e} \right)^{x+1} \sqrt{\frac{2\pi}{x+1}} e^{1/(12(x+1))}.
\]

The proof is completed.

**Lemma 8.2** For \( \mu_j \neq \mu_m \), \( k \geq \beta_j - 1 \), \( \beta_j > 0 \), \( \sigma_j > 0 \), and \( \sigma_m > 0 \), we have

\[
\left( \frac{\mu_j - \mu_m}{\sigma_m} \right)^k \sum_{v=0}^k \binom{k}{v} \left( \frac{\sigma_j}{\mu_j - \mu_m} \right)^v \frac{\Gamma((v+1)/\beta_j)}{\Gamma((k+1)/\beta_j)} \leq c(\beta_j) \left( \frac{e\beta_j}{k+1} + \frac{\sigma_j}{|\mu_j - \mu_m|} \right)^k \left( \frac{\mu_j - \mu_m}{\sigma_m} \right)^{1/\beta_j-1} \left( \frac{e\beta_j}{k+1} \right)^{k+1},
\]

23
where for \( \beta_j \leq 1 \), \( c(\beta_j) = e^{\beta_j/12} \), and for \( \beta_j > 1 \), \( c(\beta_j) = e^{\beta_j/12}\left(\beta_j^{1/2}2^{3/2}\sqrt{k+1}+1\right) \).

**Proof of Lemma 8.2:** Define \( 0 \leq q \leq 1 \) by

\[
(1 - q)^{-1} = 1 + \frac{\sigma_j}{|\mu_j - \mu_m|} \frac{k + 1}{e^{\beta_j}}.
\]

It follows from Lemma 8.1 that

\[
\left(\frac{|\mu_j - \mu_m|}{\sigma_m}\right)^k \sum_{v=0}^k \binom{k}{v} \left(\frac{\sigma_j}{|\mu_j - \mu_m|}\right)^v \frac{\Gamma((v+1)/\beta_j)}{\Gamma((k+1)/\beta_j)}
\]

\[
\leq c(\beta_j) \left(\frac{|\mu_j - \mu_m|}{\sigma_m}\right)^k \left(\frac{e\beta_j}{k+1}\right)^{k/\beta_j} \sum_{v=0}^k \binom{k}{v} \left(\frac{\sigma_j(k+1)}{|\mu_j - \mu_m|e\beta_j}\right)^v
\]

\[
= c(\beta_j) \left(\frac{|\mu_j - \mu_m|}{\sigma_m}\right)^k \left(\frac{e\beta_j}{k+1}\right)^{k/\beta_j} (1 - q)^{-k}
\]

\[
= c(\beta_j) \left(\frac{|\mu_j - \mu_m|}{\sigma_m}\right)^k \left(\frac{e\beta_j}{k+1}\right)^{k/\beta_j} \left(1 + \frac{\sigma_j(k+1)}{|\mu_j - \mu_m|e\beta_j}\right)^k
\]

The proof is completed.

**Proof of Theorem 1:** Since the linear independence of a parametric distribution family is a necessary and sufficient for identifiability of mixtures from this family (Yakowitz and Spragins, 1968), we are going to show the linear independence of the univariate exponential power distribution family, namely \( \{g(\cdot|x) : \theta = (\mu, \sigma^2, \beta) \in R^1 \times R^+ \times R^+\} \).

Suppose that for \( a_1, ..., a_m \in R^1 - \{0\} \) and \( \theta_k \in R^1 \times R^+ \times R^+, \theta_k \neq \theta_j, k \neq j, \)

\[
\sum_{k=1}^m a_k g(x|\theta_k) = 0, \quad x \in R^1.
\]  \hspace{1cm} (8.1)

In the following, we try to find a contradiction. For this purpose, we define a lexicographic ordering on \( R \times R^+ \times R^+ \) as follows: \( \theta_1 < \theta_2 \) if \( \beta_1 > \beta_2 \), or if \( \beta_1 = \beta_2 \) and \( \sigma_1 < \sigma_2 \), or if \( \beta_1 = \beta_2 \) and \( \sigma_1 = \sigma_2 \) and \( \mu_1 < \mu_2 \). Without loss of generality, assume that \( \theta_1 < \theta_2 < \cdots < \theta_m \). Then \( \beta_m \leq \beta_k, 1 \leq k \leq m-1 \).

If \( \beta_m \geq 1 \), then according to the results of Holzmann, Munk, and Gneiting (2006) and Yakowitz and Spragins (1968), (8.1) implies that \( a_k = 0, 1 \leq k \leq m \), which is in contradict with the assumption.
Therefore in the remaining part of this proof we focus on the case where $\beta_m < 1$. First, performing the Fourier transformation on both sides of the equation (8.1), we obtain

$$\sum_{j=1}^{m} e^{i\mu_j t} \phi_j(\sigma_j t) = 0, t \in R^1,$$

which yields

$$\sum_{j=1}^{m} e^{i(\mu_j - \mu_m)t} \phi(\sigma_j t) = 0, t \in R^1. \tag{8.2}$$

Here $i = \sqrt{-1}$ and the characteristic function $\phi_j(t) = \int_{-\infty}^{\infty} e^{itx} \frac{\beta_j}{2\pi(1/\beta_j)} \exp(-|x|^{\beta_j})dx$. Note that $\phi_j(t)$ has the $v$-th derivative $i^v \sigma_j^v \phi_j^{(v)}(0)$ at zero, which is equal to $i^v \sigma_j^v \Gamma((v+1)/\beta_j)/\Gamma(1/\beta_j)$ when $v$ is even, and equal to zero when $v$ is odd. Then, evaluating the $k$-th derivatives for both sides of the equation (8.2) at $t = 0$, we have

$$\sum_{j=1}^{m} a_j \left\{ \sum_{0 \leq v \leq k, v \text{ is even}} \binom{k}{v} (\mu_j - \mu_m)^{k-v} \sigma_j^v \frac{\Gamma((v+1)/\beta_j)}{\Gamma(1/\beta_j)} \right\} = 0.$$

Letting $k$ be even and dividing both sides of the above equation by $\sigma_m^k \frac{\Gamma((k+1)/\beta_m)}{\Gamma(1/\beta_m)}$, we have

$$\sum_{j=1}^{m} a_j w_j = 0, \tag{8.3}$$

where

$$w_j = \sigma_m^{-k} \frac{\Gamma((k+1)/\beta_j)\Gamma(1/\beta_m)}{\Gamma((k+1)/\beta_m)\Gamma(1/\beta_j)} \sum_{0 \leq v \leq k, v \text{ is even}} \binom{k}{v} (\mu_j - \mu_m)^{k-v} \sigma_j^v \frac{\Gamma((v+1)/\beta_j)}{\Gamma(1/\beta_j)}.$$

Note that when $\mu_j = \mu_m$,

$$w_j = \left( \frac{\sigma_j}{\sigma_m} \right)^k \frac{\Gamma((k+1)/\beta_j)\Gamma(1/\beta_m)}{\Gamma((k+1)/\beta_m)\Gamma(1/\beta_j)},$$

and when $\mu_j \neq \mu_m$,

$$w_j = \left( \frac{|\mu_j - \mu_m|}{\sigma_m} \right)^k \frac{\Gamma((k+1)/\beta_j)\Gamma(1/\beta_m)}{\Gamma((k+1)/\beta_m)\Gamma(1/\beta_j)} \sum_{0 \leq v \leq k, v \text{ is even}} \binom{k}{v} \left( \frac{\sigma_j}{\mu_j - \mu_m} \right)^v \frac{\Gamma((v+1)/\beta_j)}{\Gamma(1/\beta_j)}.$$

Moreover, it follows from Lemmas 8.1 and 8.2 that when $\mu_j = \mu_m$ and $\beta_m < \beta_j$ or when $\mu_j = \mu_m$, $\beta_m = \beta_j$ and $\sigma_j \leq \sigma_m$,

$$w_j \leq \left( \frac{\sigma_j}{\sigma_m} \right)^k \frac{\Gamma(1/\beta_m)\beta_j^{1/2}}{\Gamma(1/\beta_j)\beta_m^{1/2} e^{\beta_j/12}} \left( \frac{\beta_m}{\beta_j} \right)^{1/\beta_j} \left( \frac{e\beta_m}{k+1} \right)^{1/\beta_m-1/\beta_j} \left( \frac{\beta_j}{k+1} \right)^{k+1} \to 0, \tag{8.4}$$

25
as $k \to \infty$. When $\mu_j \neq \mu_m$ and $1 \geq \beta_m > \beta_j$, or when $\mu_j \neq \mu_m$, $\beta_m = \beta_j$ and $\beta_m < 1$,

$$w_j \leq c(\beta_j) \left( \frac{e^{\beta_j}}{k+1} + \frac{\sigma_j}{\mu_j - \mu_1} \right)^k \left( \frac{|\mu_j - \mu_m|}{\sigma_m} \left( \frac{e^{\beta_j}}{k+1} \right)^{1/\beta_j} \right)^k$$

$$\times \frac{\Gamma(1/\beta_m)\beta_j^{1/2}}{\Gamma(1/\beta_j)} e^{\beta_j/12} \left( \frac{\beta_m}{\beta_j} \right)^{1/\beta_j - 1/2} \left( \frac{e^{\beta_m}}{k+1} \right)^{1/\beta_m - 1/\beta_j}$$

$$= c(\beta_j)e^{\beta_j/12} \frac{\Gamma(1/\beta_m)}{\Gamma(1/\beta_j)} \left( \frac{\beta_m}{\beta_j} \right)^{1/\beta_j - 1/2} \left( \frac{e^{\beta_m}}{k+1} \right)^{1/\beta_m - 1/\beta_j}$$

$$\times \left( \frac{e^{\beta_j}}{k+1} + \frac{\sigma_j}{|\mu_j - \mu_m|} \right) \left( \frac{e^{\beta_j}}{k+1} \right)^{1/\beta_j - 1} \left( \frac{\beta_m}{\beta_j} \right)^{1/\beta_m}$$

$$\rightarrow 0\quad (8.5)$$

as $k \to \infty$. Note that by the definition $\theta_j < \theta_m, j < m$ implies that either $\beta_m < \beta_j$, or $\beta_m = \beta_j$ and $\sigma_j < \sigma_m$, or $\beta_m = \beta_j, \sigma_j = \sigma_m$ and $\mu_m > \mu_j$. Combining this with (8.3), (8.4) and (8.5) and letting $k \to \infty$ yields that $a_m = 0$. This is in contradict with the non-zero assumption of $a_m$. The proof is completed.

**Proof of Proposition 1:** It follows directly from Theorem 1.

**Proof of Theorem 2:** First under the condition (C1),

$$\text{var} \left( \frac{1}{n} \sum_{i=1}^{n} \log(p_{\psi_m}(x_i)) \right) = \frac{\text{var}(\log(p_{\psi_m}(x)))}{n^2} \left( n + 2n^2C(\psi_m) \right) = o_p(1),$$

which implies

$$\frac{1}{n} \sum_{i=1}^{n} \log(p_{\psi_m}(x_i)) = E \log(p_{\psi_m}(x)) + o_p(1). \quad (8.6)$$

Let $h_m(x) = \sup_{\psi \in \Psi_m} |\log(p_{\psi_m})|$. Let $P_n$ be the empirical distribution and for any function $f$ let $(P_n - P)f(x) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - Ef(x))$. Then we have

$$(P_n - P)\log(p_{\psi}(x)) = D_{1n}(\psi) + D_{2n}(\psi)$$

with

$$D_{1n}(\psi) = (P_n - P)[\log(p_{\psi_m}(x))I(||x|| \leq B)],$$

$$D_{2n}(\psi) = (P_n - P)[\log(p_{\psi_m}(x))I(||x|| > B)].$$

For any $\varepsilon > 0$, there exists a large value $B$ such that

$$\lim \max_{\psi \in \Psi_m} |D_{2n}(\psi)| \leq (P_n + P)h(x)I(||x|| > B) < \varepsilon.$$

Note that $\log(p_{\psi_m}(x))$ is uniformly continuous when $||x|| \leq B$ and $\psi \in \Psi_m$. Therefore, there is a finite number of spheres, say $N_j, j = 1, ..., J$, such that $\Psi_m \subset \bigcup_{j=1}^{J} N_j$ and for any two points $\psi, \phi$
\[ \epsilon \in N_j \sup_{||x|| \leq B} | \log(p_\psi(x))I(||x|| \leq B) - \log(p_{\bar{\psi}}(x))I(||x|| \leq B) | < \varepsilon. \] Let \( \psi^{(j)} \) be the centres of these \( N_j \). Then for each \( \psi \), there exists \( j \) such that

\[ D_{1n}(\psi) = (P_n - P)[\log(p_{\psi^{(j)}}(x))]I(||x|| \leq B)] + D_{12n}, \]

where

\[ |D_{12n}| = |(P_n - P)[\log(p_\psi(x))]I(||x|| \leq B) - \log(p_{\psi^{(j)}}(x))]I(||x|| \leq B) | < \varepsilon. \]

This yields for large \( n \),

\[ |D_{1n}(\psi) - (P_n - P)[\log(p_{\psi^{(j)}}(x))]| \leq 2\varepsilon \]

and therefore

\[ |(P_n - P)[\log(p_\psi(x))] - (P_n - P)[\log(p_{\psi^{(j)}}(x))]| \leq 3\varepsilon. \]

This together with (8.6) shows that

\[ P(\sup_{\psi \in \Psi_m} |(P_n - P)\log(p_\psi(x))| > 4\varepsilon) \]

\[ \leq P(\sum_{j=1}^{J} |(P_n - P)\log(p_{\psi^{(j)}}(x))| + 3\varepsilon > 4\varepsilon) \]

\[ \to 0 \]

as \( n \to \infty \). Consequently

\[ \max_{1 \leq m \leq m_{\max}} \sup_{\psi \in \Psi_m} P_n[\log(p_\psi(x))] - \sup_{\psi \in \Psi_m} E[\log(p_\psi(x))] = o_p(1). \quad (8.7) \]

Note that if \( p_{\psi_{m_0}} \) is the underlying mixture model, then \( E[\log(p_\psi(x))] \) can attain the maximum value \( E[\log(p_{\psi_{m_0}}(x))] \) only when \( m \geq m_0 \). Moreover, for each \( m \geq m_0 \), \( E[\log(p_{\psi_m}(x))] \) attains the maximum if and only if \( p_{\psi_m} \) and \( p_{\psi_{m_0}} \) have the same supporting points of non-zero masses, up to a permutation on the supporting points. The desired conclusion follows directly from these facts and the equation (8.7). The proof is completed.

**Appendix II: Iterative formulae for the other models**

**PSE model**

Under the PSE model,

\[ g(x|\theta_k) = \frac{\beta_k \Gamma(p/2)}{2\pi^{p/2}\lambda^p \Gamma(p/\beta_k)} \exp(-||x| - \mu_k||/\lambda)^{\beta_k}. \]

The equations (4.1) becomes

\[ 0 = \frac{\partial Q_2}{\partial \mu_k} = \sum_{i=1}^{n} w_{ik} \beta_k \lambda^{\beta_k} ||x_i - \mu_k||^{\beta_k} - 2(\mu_i - \mu_k), \]
0 = \frac{\partial Q_2}{\partial \lambda_k} = \sum_{i=1}^{n} \sum_{k=1}^{m} w_{ik} (-p + \frac{\beta_k}{\lambda_k} \|\mathbf{x}_i - \mu_k\|^{\beta_k}),

0 = \frac{\partial Q_2}{\partial \beta_k^2} = \sum_{i=1}^{n} w_{ik} \left\{ 1 + \frac{p \Gamma'(p/\beta_k)}{\beta_k \Gamma(p/\beta_k)} - \beta_k \left( \frac{\|\mathbf{x}_i - \mu_k\|}{\lambda} \right)^{\beta_k} \log \left( \frac{\|\mathbf{x}_i - \mu_k\|}{\lambda} \right) \right\},

which yields the updating formulae,

\mu_k^{(v+1)} = \frac{\sum_{i=1}^{n} w_{ik} \|\mathbf{x}_i - \mu_k^{(v)}\|^2 - 2 \sum_{i=1}^{n} w_{ik} \|\mathbf{x}_i - \mu_k^{(v)}\|^2}{\sum_{i=1}^{n} w_{ik}}

and

\beta_k^{(v+1)} = \beta_k \exp \left\{ - \left( \frac{\partial^2 Q_2}{\partial \beta_k^2} \right)^{-1} \frac{\partial Q_2}{\partial \beta_k^2} \right\},

where

\frac{\partial Q_2}{\partial \beta_k^2} = \sum_{i=1}^{m} w_{ik} \left\{ 1 + \frac{p \Gamma'(p/\beta_k)}{\beta_k \Gamma(p/\beta_k)} - \beta_k \left( \frac{\|\mathbf{x}_i - \mu_k^{(v+1)}\|}{\lambda^{(v+1)}} \right)^{\beta_k} \log \left( \frac{\|\mathbf{x}_i - \mu_k^{(v+1)}\|}{\lambda^{(v+1)}} \right) \right\},

\frac{\partial^2 Q_2}{\partial \beta_k^2} = - \sum_{i=1}^{n} w_{ik} \left\{ \frac{p}{\beta_k} \log(p/\beta_k) + (p/\beta_k)^2 \log(p/\beta_k) \right\}

+ \beta_k \left( \frac{\|\mathbf{x}_i - \mu_k^{(v+1)}\|}{\lambda^{(v+1)}} \right)^{\beta_k} \log \left( \frac{\|\mathbf{x}_i - \mu_k^{(v+1)}\|}{\lambda^{(v+1)}} \right)

+ (\beta_k)^2 \left( \log \left( \frac{\|\mathbf{x}_i - \mu_k^{(v+1)}\|}{\lambda^{(v+1)}} \right) \right)^2.

PDE model
Under the PDE model,

g(\mathbf{x}; \beta_k) = \frac{\beta_k \Gamma(p/2)}{2^{p/2} \Gamma(p/\beta_k) \prod_{i=1}^{p} \lambda_t} \exp \left\{ - \left( \frac{\sum_{i=1}^{p} (x_{it} - \mu_k)^2}{\lambda_t^2} \right)^{\beta_k/2} \right\},

0 = \frac{\partial Q_2}{\partial \mu_k} = \sum_{i=1}^{n} w_{ik} \beta_k \left( \mathbf{x}_i - \mu_k \right) \Sigma^{-1} \left( \mathbf{x}_i - \mu_k \right) \left( \mathbf{x}_i - \mu_k \right)^T \Sigma^{-1} \left( \mathbf{x}_i - \mu_k \right) \beta_k^{2-1} \Sigma^{-1} \left( \mathbf{x}_i - \mu_k \right).
which yields a updating formula for $\mu_k$. Let $\lambda^* = (\lambda^*_1, ..., \lambda^*_p)^\tau$ with

$$
\lambda^*_t = \log(\lambda_t), 1 \leq t \leq p.
$$

Then

$$
0 = \frac{\partial Q_2}{\partial \lambda^*_t} = \sum_{i=1}^n \sum_{k=1}^m w_{ik}(-1_p + \beta_j((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k))^{\beta_k/2-1}\Sigma^{-1}(x_i - \mu_k)^{\odot 2}),
$$

where

$$(x_i - \mu_j)^{\odot 2} = ((x_{i1} - \mu_{j1})^2, ..., (x_{ip} - \mu_{jp})^2)^\tau, \quad 1_p = (1,...,1)^\tau.$$ 

This will entail a updating formula for $\lambda_t$’s. Likewise, let $\beta^*_k = \log(\beta_k), k = 1, ..., m$. Then

$$
\frac{\partial Q_2}{\partial \beta^*_k} = \sum_{i=1}^n w_{ik}(1 + \frac{p}{\beta_k} \text{dg}(p/\beta_k)((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k))^{\beta_k/2} \log((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k)).
$$

$$
\frac{\partial^2 Q_2}{\partial \beta^*_k^2} = -\sum_{i=1}^n w_{ik} \left\{ \frac{p}{\beta_k} \text{dg}(p/\beta_k) + \frac{p^2}{\beta_k^2} \text{tg}(p/\beta_k)
+ \frac{\beta_k}{2}((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k))^{\beta_k/2} \log((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k))
+ \frac{\beta_k^2}{4}((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k))^{\beta_k/2}(\log((x_i - \mu_k)^\tau \Sigma^{-1}(x_i - \mu_k)))^2 \right\}.
$$

A Newton-Rapson updating formula can be made.

**PDV model**

Under the PDV model, Then $|\Sigma_k| = \prod_{t=1}^p \lambda^2_{kt}$ and

$$
g(x_i|\mu_k, \Sigma_k, \beta_k) = \frac{\beta_k \Gamma(p/2)}{2\pi^{p/2} \prod_{t=1}^p \lambda_{kt} \Gamma(p/\beta_k)} \exp \left\{ -\frac{p}{\beta_k} (x_i - \mu_k)^2 \lambda_{kt}^{-2} \beta_k^{\beta_k/2} \right\}.
$$

In the CM-steps: Note that

$$
\frac{\partial Q_2}{\partial \mu_k} = \sum_{i=1}^n w_{ik}(\psi^{(u)}) \beta_j((x_i - \mu_k)^\tau \Sigma^{-1}_j(x_i - \mu_k))^{\beta_k/2-1}\Sigma_k^{-1}(x_i - \mu_k).
$$

Let $\lambda^*_k = (\lambda^*_{k1}, ..., \lambda^*_{kp})^\tau$ with

$$
\lambda^*_kt = \log(\lambda_{kt}), 1 \leq t \leq p.
$$

Then

$$
\frac{\partial Q_2}{\partial \lambda^*_k} = \sum_{i=1}^n w_{ik}(-1_p + \beta_j((x_i - \mu_k)^\tau \Sigma_k^{-1}(x_i - \mu_k))^{\beta_k/2-1}\Sigma_k^{-1}(x_i - \mu_k)^{\odot 2}),
$$

where

$$(x_i - \mu_k)^{\odot 2} = ((x_{i1} - \mu_{k1})^2, ..., (x_{ip} - \mu_{kp})^2)^\tau, \quad 1_p = (1,...,1)^\tau.$$
Likewise, let $\beta_k^* = \log(\beta_k)$, $k = 1, \ldots, m$. Then

$$\frac{\partial Q_2}{\partial \beta_k^*} = \sum_{i=1}^n w_{ik}(1 + \frac{pP_p(p/\beta_k)}{\beta_k \Gamma(p/\beta_k)}) \left( (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right)^{\beta_k/2} \log((x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k)).$$

$$\frac{\partial^2 Q_2}{\partial \beta_k^*} = -\sum_{i=1}^n w_{ik} \left\{ \frac{p}{\beta_k^*} \text{dg}(p/\beta_k) + \frac{p^2}{\beta_k^*} \text{tg}(p/\beta_k) \\
+ \frac{\beta_k}{2} ((x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k))^{\beta_k/2} \log((x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k)) \\
+ \frac{\beta_k^2}{4} ((x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k))^{\beta_k/2} \log((x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k))^2 \right\}.$$

Setting $\frac{\partial Q_2}{\partial \mu_k} = 0$, we have the updated estimate of $\mu_k$,

$$\mu_k = \frac{\sum_{i=1}^n w_{ik}(\psi^{(v)}(x_i - \mu_k^{(v)})^\top (\Sigma_k^{(v)})^{-1} (x_i - \mu_k^{(v)}) \beta_k^{(v)/2} - 1)}{\sum_{i=1}^n w_{ik}(\psi^{(v)}(x_i - \mu_k^{(v)})^\top (\Sigma_k^{(v)})^{-1} (x_i - \mu_k^{(v)}) \beta_k^{(v)/2} - 1)}.$$

Setting $\frac{\partial Q_2}{\partial \lambda_k^{\circ2}} = 0$, we have the updated estimate of $\lambda_k^{\circ2}$,

$$\lambda_k^{\circ2} = \frac{\sum_{i=1}^n w_{ik}(\psi^{(v)})(x_i - \mu_k^{(v)})^\top (\Sigma_k^{(v)})^{-1} (x_i - \mu_k^{(v)}) \beta_k^{(v)/2} - 1)}{\sum_{i=1}^n w_{ik}(\psi^{(v)})}.$$

Finally, we initialize $\beta_k$ by the solution to the equation

$$\frac{\Gamma(3/\beta_k) \Gamma(1/\beta_k)}{\Gamma(2/\beta_k)^2} = \frac{1}{p \sum_{l=1}^p (\sum_{i=1}^n w_{ik}(x_{it} - \mu_{kl^{(v)+1})}^2 / \sum_{i=1}^n w_{ik})^2},$$

followed by the one-step Newton-Raphson approximation

$$\beta_j = \beta_j^{(v)} \exp \left\{ -\left( \frac{\partial^2 Q_2}{\partial \beta_j^{\circ2}} \right)^{-1} \frac{\partial Q_2}{\partial \beta_j^{\circ2}} \right\}.$$

References


Figure 1: Comparison between 100 simulated datasets with clumpy correlations and 100 simulated iid datasets: Panels (a), (b), (c), and (d) are, respectively, the boxplots for the sample means, the sample standard deviations, the sample shape indices, and the numbers of clusters predicted by the GM method. In each panel, the boxplot on the left is for the datasets with clumpy dependence while the boxplot on the right is for the iid datasets.

Table 1: Comparison of the EPD and GM methods via ERRm and ave(ρ) under clumpy dependence. The error rate in estimating the number of clusters over the $K$ datasets is defined as $ERRm = \frac{1}{K} \sum_{i=1}^{K} |m_i - m_t|$, where $m_i$ is the estimated number of clusters for the $i$-th dataset and $m_t$ is the true number of clusters in these datasets. The average adjusted Rand index $ave(\rho)$ is defined over $K = 30$ datasets. In the table the number in the bracket is the standard error.

<table>
<thead>
<tr>
<th></th>
<th>(α₁, α₂) = (1/4, 1/2)</th>
<th>(α₁, α₂) = (1/16, 1/8)</th>
<th>(α₁, α₂) = (1/64, 1/32)</th>
<th>(α₁, α₂) = (0, 1/64)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ERRm</td>
<td>ave(ρ)</td>
<td>ERRm</td>
<td>ave(ρ)</td>
</tr>
<tr>
<td>s = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPD</td>
<td>0.77</td>
<td>0.42 (0.028)</td>
<td>0.4</td>
<td>0.68 (0.034)</td>
</tr>
<tr>
<td>GM</td>
<td>0.73</td>
<td>0.42 (0.030)</td>
<td>0.4</td>
<td>0.66 (0.035)</td>
</tr>
<tr>
<td>s = 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPD</td>
<td>0</td>
<td>1 (0.001)</td>
<td>0.3</td>
<td>0.97 (0.010)</td>
</tr>
<tr>
<td>GM</td>
<td>0.1</td>
<td>0.97 (0.013)</td>
<td>1.47</td>
<td>0.83 (0.032)</td>
</tr>
</tbody>
</table>
Figure 2: Comparison between the EPD and GM methods under clumpy dependence using simulated datasets: Boxplots of the differences $\rho_e - \rho_g$, where $\rho_e$ and $\rho_g$ are the adjusted Rand indices of the EPD and GM methods. In each panel, from left to right, these plots are corresponding to $(\alpha_1, \alpha_2) = (1/4, 1/2), (1/16, 1/8), (1/64, 1/32), (0, 1/64)$. The left panel is for the case $s = 2$ with not well separated clusters while the right panel is for the case $s = 6$ with well separated clusters.

Figure 3: Comparison of the EPD and GM methods via ERRm and $\text{ave}(\rho)$. Here we adopt the notations used in Figure 1. The left panel is the boxplots (from the left to the right) for Example 2 with $r = 0, 0.3, 0.6, 0.9$ respectively. The panel on the right is the boxplot for Example 3.
Figure 4: Comparison of the EPD and GM methods in presence of heavy tailed clusters. Panel (a): Histogram of one dataset. Panel (b): The boxplot of the difference $\rho_e - \rho_g$, where $\rho_e$ and $\rho_g$ stand for the adjusted Rand indices of the clustering results derived from the EPD and GM approaches respectively. This plot suggests that the EPD performs favourably in comparison with the GM.

Figure 5: Comparison of the EPD and GM methods. This comparison is based on 60 datasets of size 500 each, generated from $\frac{13}{50}\text{epd}_6(\mu_1, 1, 1.5) + \frac{17}{50}\text{epd}_6(\mu_2, 1, 4) + \frac{7}{50}\text{epd}_6(\mu_3, 2^2, 0.8) + \frac{11}{50}\text{epd}_6(\mu_4, 1, 2) + \frac{14}{50}\text{epd}_6(\mu_5, 1, 1.6)$. Here we adopt the notations used in Figure 4. Panel (a): The depiction of the means of the five components. Panel (b): Boxplot of $\rho_e - \rho_g$ over the above 60 datasets.
Figure 6: z-scores of avian pineal gland gene expression data. Top row: left panel, histogram and kernel density estimate; right panel, the normal Q-Q plot for the cluster 1. Bottom row: left panel, the normal Q-Q plot for the cluster 2; right panel, the normal Q-Q plot for the cluster 3. These three clusters are derived from the GM method (Fraley and Raftery, 2006). This plot indicates that the main peak in the histogram does not fit to a normal density. The Gaussian mixture may not be the best model for these scores as the corresponding clusters are not normally distributed.

Table 2: Comparison of the EPD and GM methods via ERRm and ave(\(\rho\)) over \(K = 60\) datasets under array specific dependences. The notations are the same as in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>(r = 0)</th>
<th>(r = 0.3)</th>
<th>(r = 0.6)</th>
<th>(r = 0.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{ERRm} \quad \text{ave}(\rho))</td>
<td>(\text{ERRm} \quad \text{ave}(\rho))</td>
<td>(\text{ERRm} \quad \text{ave}(\rho))</td>
<td>(\text{ERRm} \quad \text{ave}(\rho))</td>
<td></td>
</tr>
<tr>
<td>EPD</td>
<td>0.02 (0.001)</td>
<td>0.08 (0.017)</td>
<td>0.20 (0.041)</td>
<td>0.45 (0.046)</td>
</tr>
<tr>
<td>GM</td>
<td>0 (0.001)</td>
<td>0.12 (0.023)</td>
<td>0.51 (0.042)</td>
<td>0.75 (0.040)</td>
</tr>
</tbody>
</table>

37
Figure 7: z-scores of breast cancer gene expression data. Histogram and density curves of a two-component Gaussian mixture (dashed) and a single EPD fit (solid).

Table 3: Comparison between the EPD and GM methods. Adopt the notations in Table 1

<table>
<thead>
<tr>
<th>Example 3</th>
<th>Example 4</th>
<th>Example 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERRm</td>
<td>ave((\rho))</td>
<td>ERRm</td>
</tr>
<tr>
<td>EPD</td>
<td>0.075</td>
<td>0.39</td>
</tr>
<tr>
<td>GM</td>
<td>2</td>
<td>1.84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3</th>
<th>Example 4</th>
<th>Example 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERRm</td>
<td>ave((\rho))</td>
<td>ERRm</td>
</tr>
<tr>
<td>EPD</td>
<td>0.93</td>
<td>0.69</td>
</tr>
<tr>
<td>GM</td>
<td>0.37</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Figure 8: The clustering results of the EPD and GM methods on the 2D gel proteomic dataset with 839 protein spots being observed at the six cell lines. These cell lines, Control, 4O,4R,2X,2P, 2N2, coded by 1,2,3,4, 5, and 6 respectively and under different treatments, produce antibody in increasing amounts with Control producing no antibody and 2N2 producing the most. (a)-(e): The 5 clusters derived from the EPD. (f)-(j): The 5 clusters derived from the GM. Two methods identify two slightly different patterns that might be associated with the antibody producing rates.