1-2015

Semiparametric Analysis in Conditionally Independent Multivariate Mixture Models

T. Wrobel

Denis H. Y. LEUNG
Singapore Management University, denisleung@smu.edu.sg

J. Qin

T. Hettmansperger

Chapter 21
Semiparametric Analysis in Conditionally Independent Multivariate Mixture Models

Tracey W. Hammel, Thomas P. Hettmansperger, Denis H.Y. Leung, and Jing Qin

Abstract The conditional independence assumption is commonly used in multivariate mixture models in behavioral research. We propose an exponential tilt model to analyze data from a multivariate mixture distribution with conditionally independent components. In this model, the log ratio of the density functions of the components is modeled as a quadratic function in the observations. There are a number of advantages in this approach. First, except for the exponential tilt assumption, the marginal distributions of the observations can be completely arbitrary. Second, unlike some previous methods, which require the multivariate data to be discrete, modeling can be performed based on the original data.

Keywords Empirical likelihood • Exponential tilting • Repeated measures • Mixture distribution • Multivariate

21.1 Introduction

There are many applications where the interest is to classify n observations into m groups based on k measures on each observation. For example, Hettmansperger and Thomas (2000) and Cruz-Medina et al. (2004) described an experiment in developmental psychology where repeated measurements are made on children's responses to a cognitive task and the interest is to classify children into different groups based on the repeated measurements. The repeated measures data can

be considered to come from a mixture of multivariate distributions, with the components corresponding to the response distributions in the different groups of observations, the number of components corresponding to the number of groups, and the mixing proportions corresponding to the proportions in the population that belong to the different groups. Two problems are of interest. First, to determine the number of groups. Second, to estimate the underlying component distributions and the mixing proportions.

Analysis of multivariate mixture distributions is a difficult problem (see, e.g., Titterington et al. 1985; Lindsay 1995; McLachlan and Peel 2000). Computation is commonly carried out using the EM algorithm (Dempster et al. 1977), which typically requires parametric distributional assumptions. However, a number of works (Thomas and Lohaus 1993; Hettmansperger and Thomas 2000; Hall and Zhou 2003; Cruz-Medina et al. 2004; Leung and Qin 2006; Chang and Walther 2007; Benaglia et al. 2009) showed that a semiparametric or nonparametric approach might be a flexible and robust alternative to a parametric approach.

In the situation described in the first paragraph, each child who participated in the study was given a total of six tasks, each randomly selected from a large pool of similar tasks. As a result, it is unlikely for a child to predict the next task and hence the responses to different tasks can be considered independent of each other. This observation led us to make the assumption of conditional independence, which means that conditional on component membership, the multivariate component distribution is the product of its marginals; see also Sect. 21.7. Under the conditional independence assumption, the m component mixture has probability density function (PDF) or probability mass function (PMF)

$$h(x_1, \ldots, x_k) = \sum_{l=1}^{m} \lambda_l \prod_{j=1}^{k} f_{lj}(x_j), \tag{21.1}$$

where λ_l is the mixing proportion for the lth component and f_{lj} is the PDF (or PMF) for the lth component of the jth repeated measure. Later, we impose further structural assumptions. Unlike previous works (Hettmansperger and Thomas 2000; Cruz-Medina et al. 2004; Leung and Qin 2006; Chang and Walther 2007), (21.1) does not require identical marginal distributions. Conditional independence of multivariate data can also be seen as a special case of the popular random effects model with clustered data (Liu and Pierce 1994; Qu and Hadgu 1998).

A histogram of the data in the study (Cruz-Medina et al. 2004, Fig. 1) shows that the data distribution is unremarkable; there is no immediate resemblance to any well-known distribution. This observation motivates a semiparametric approach to analyzing the data. We assume the component densities are related by an exponential tilt (density-ratio) model (Anderson 1979). For a two-component mixture with PDFs f and g, our exponential tilt model assumes f and g are related by $\log\left(g(x)/f(x)\right) = \alpha + \beta x + \gamma x^2$. As a parallel to the Cox proportional hazards model and the Lehmann alternative model, the exponential tilt model is very versatile, due to its natural connection to the logistic model. Kay and Little (1986) discuss various versions

of the exponential tilt model for some common distributions. Because the normal PDF has a quadratic exponent, any two normal PDFs satisfy the condition for the exponential tilt model described above. In many situations where common parametric distributions do not fit the observed data well, the exponential tilt model still can provide excellent fits (Qin and Zhang 1997; Nagelkerke et al. 2001; Zhang 2001; Qin et al. 2002; White and Thompson 2003). Efron and Tibshirani (1996) argue that the exponential tilt is a favorable compromise between parametric and nonparametric density estimation.

The rest of this paper is organized as follows. Details of the method are described in Sect. 21.2. The exponential tilt model is formulated using an empirical likelihood (Owen 1988). Under mild conditions, the model is uniquely identifiable up to label switching, which is important for estimating the underlying mixture structure. In Sect. 21.3, we present an EM algorithm. Estimation of features of the component distributions is discussed in Sect. 21.4. In Sect. 21.5, we evaluate the method using simulations. In Sect. 21.6, we propose a model selection criterion based on the BIC (Bayesian Information Criterion; Schwarz 1978) to estimate the number of components in the mixture. In Sect. 21.7, the method is applied to the data of Cruz-Medina et al. (2004). Section 21.8 concludes with a discussion of possible future work.

21.2 Exponential Tilt Model

We consider n multivariate vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from an m component, k dimensional multivariate mixture distribution, where $\mathbf{X}_i^\top = (x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$.

Let $(x_1, \ldots, x_k)$ be a generic observation, then its joint PDF can be written as

$$h(x_1, \ldots, x_k) = \lambda_1 \prod_{j=1}^{k} f_j(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} g_{lj}(x_j), \tag{21.2}$$

where f_j and g_{lj} represent univariate PDFs, λ_1 is the mixing proportion of component one (the baseline distribution), $0 < \lambda_l < 1$ is the mixing proportion of component l and $\sum_{l=1}^{m} \lambda_l = 1$. Let H, F_j, and G_{lj} denote the CDFs corresponding to h, f_j, and g_{lj}, respectively.

Let f_j and g_{lj} be related by a quadratic exponential tilt model

$$\log\left(g_{lj}(x_j)/f_j(x_j)\right) = \alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2, \tag{21.3}$$

where α_{lj}, β_{lj}, and γ_{lj} are unknown parameters. The PDF (21.2) can be re-written as

$$h(x_1, \ldots, x_k) = \left[\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\sum_{j=1}^{k} \alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\}\right] \prod_{j=1}^{k} f_j(x_j). \tag{21.4}$$

Theorem 8 of Allman et al. (2009) states that a mixture of the form (1) is uniquely identifiable up to label switching provided that $k \geq 3$ and, for each $j = 1, \ldots, k$, the m distributions are linearly independent. This result makes sense since linear independence precludes expressing any one of the coordinate distributions as a linear combination of the other $m-1$ distributions. Since, in our case, $\sum_{l=1}^{m} \lambda_l = 1$ and $0 < \lambda_l < 1$, and for each $j = 1, \ldots, k$

$$\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\} \neq 0 \quad \text{for} \quad -\infty < x_j < \infty,$$

identifiability follows for model (21.4). For earlier results on identifiability in nonparametric mixtures, see Hall and Zhou (2003), Hall et al. (2005), and Elmore et al. (2005).

Let $\boldsymbol{\theta}_{lj}^{\top} = (\alpha_{lj}, \beta_{lj}, \gamma_{lj})$, $\tilde{\mathbf{x}}_{ij}^{\top} = (1, x_{ij}, x_{ij}^2)$, $\tilde{\mathbf{x}}_j$ the counterpart of $\tilde{\mathbf{x}}_{ij}$ for a generic observation, $\boldsymbol{\lambda}^{\top} = (\lambda_1, \ldots, \lambda_m)$, $\boldsymbol{\theta}^{\top} = (\boldsymbol{\theta}_{21}, \ldots, \boldsymbol{\theta}_{mk})$ and $\boldsymbol{\delta}^{\top} = (\boldsymbol{\lambda}^{\top}, \boldsymbol{\theta}^{\top})$, then the likelihood based on the observed data is

$$L(\boldsymbol{\delta}, F_1, \ldots, F_k) = \prod_{i=1}^{n} \left[\left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj} \right) \right\} \prod_{j=1}^{k} dF_j(x_{ij}) \right].$$

The maximizing F_j only jumps at each observed x_{ij} (Owen 1988). Let the jump sizes be p_{ij}, then the log-likelihood is

$$\ell(\boldsymbol{\delta}, p_{11}, \ldots, p_{nk}) = \sum_{i=1}^{n} \left[\log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj} \right) \right\} + \sum_{j=1}^{k} \log p_{ij} \right]. \tag{21.5}$$

For fixed $\boldsymbol{\delta}$, ℓ can be maximized with respect to the p_{ij}s subject to the constraints

$$\sum_{i=1}^{n} p_{ij} = 1, \;\; p_{ij} \geq 0, \;\; \sum_{i=1}^{n} p_{ij} \exp\left(\tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}\right) = 1, \;\; j = 1, \ldots, k, l = 2, \ldots, m. \tag{21.6}$$

The last k constraints in (21.6) come from model (21.3) and are responsible for ensuring that the resulting g_{lj} are proper PDFs. The constrained maximization can be accomplished using a Lagrange multiplier argument, which leads to

$$p_{ij} = \frac{1}{n} \left[\frac{1}{1 + \frac{1}{n} \sum_{l=2}^{m} \eta_{lj} \{\exp(\tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}) - 1\}} \right], \;\; i = 1, \ldots, n, j = 1, \ldots, k, \tag{21.7}$$

where $\boldsymbol{\eta}^\top \equiv (\eta_{21}, \ldots, \eta_{mk})$ are Lagrange multipliers determined by

$$\sum_{i=1}^{n} \frac{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1}{1 + \frac{1}{n}\sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}} = 0, \quad j = 1, \ldots, k, l = 2, \ldots, m. \tag{21.8}$$

Note that if the exponential tilt parameters $\boldsymbol{\theta}_{lj}^\top = \mathbf{0}$, then (21.7) would simply be the weights found for the empirical distribution, namely $1/n$. Substituting the p_{ij}s back into (21.5) gives a log-profile likelihood

$$\begin{aligned} \ell_p(\boldsymbol{\delta}) = &\sum_{i=1}^{n} \log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} \right) \right\} \\ &- \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left[n + \sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\} \right]. \end{aligned} \tag{21.9}$$

Denote the maximum semiparametric likelihood estimate obtained from maximizing $\ell_p(\boldsymbol{\delta})$ by $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\eta}}$ the corresponding value of the Lagrange multipliers at the maximum likelihood. The following theorem describes the large sample behavior of the maximum semiparametric likelihood estimate.

Theorem 21.1 *Let* $\mathbf{U}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = (u_1, u_2, u_3)$, *where* $u_1(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \partial \ell_p / \partial \boldsymbol{\theta}_{lj}$, $u_2(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \partial \ell_p / \partial \eta_{lj}$, $u_3(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = \partial \ell_p / \partial \lambda_l$. *Let* $\boldsymbol{\delta}^0 \equiv (\boldsymbol{\lambda}^0, \boldsymbol{\theta}^0, \boldsymbol{\eta}^0)$ *be the true values of* $\boldsymbol{\delta} \equiv (\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\eta})$ *and let the superscript "0" denote a quantity evaluated at* $\boldsymbol{\delta}^0$. *Assume the conditions hold:*

[C1] $E\{\mathbf{U}^0(\mathbf{U}^0)^\top\}$ *is positive definite; and the rank of* $E(\partial \mathbf{U}^0 / \partial \boldsymbol{\delta})$ *is* $2(m-1)k + (m-1)$, *which is also the dimension of* $\boldsymbol{\delta}$.

[C2] $\partial^2 \mathbf{U}(\boldsymbol{\delta}) / (\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top)$ *is continuous in a neighborhood of* $\boldsymbol{\delta}^0$ *where* $\|\partial \mathbf{U}(\boldsymbol{\delta}) / \partial \boldsymbol{\delta}\|$ *is bounded,* $E(||\mathbf{U}(\boldsymbol{\delta})||)^2 < \infty$.

[C3] Functions are sufficiently smooth to allow differentiation under the integral. and $0 < \lambda_1, \ldots, \lambda_m < 1$, *then for any sufficiently smooth function g,*

$$\sqrt{n} g(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0, \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0, \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_g).$$

Furthermore, asymptotically, the estimates achieve semiparametric efficiency.

Proof *For a matrix* $\mathbf{a}$, *denote its* i, jth *element by* $a_{i,j}$ *and let* $\mathbf{A} = E(\mathbf{a})$ *where the expectation is taken under* $\boldsymbol{\delta}^0$. *Write* $w_{il}^- = \lambda_l / \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})\}$, $v_{ilj}^- = \eta_{lj} / [1 + 1/n \sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}]$. *Let* $\dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^0 = w_{il}^{-0} \partial / \partial \boldsymbol{\theta}_{l'j'} \{\lambda_l^0$

$\exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}^{0})\}$, $\dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} = v_{lj}^{-0} \partial / \partial \boldsymbol{\theta}_{l'j'} [\boldsymbol{\eta}_{lj}^{0} \{\exp(\tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}^{0}) - 1\}]$ *and similarly for* $\dot{\mathbf{w}}_{il,\eta_{l'j'}}^{0}$, $\dot{\mathbf{w}}_{il,\lambda_{l'}}^{0}$, $\dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0}$, $\dot{\mathbf{v}}_{ilj,\lambda_{l'}}^{0}$, $\dot{\mathbf{v}}_{l,\boldsymbol{\theta}_{l'j'}}^{-0}$, *and* $\dot{\mathbf{v}}_{l,\eta_{l'j'}}^{-0}$.

$$\frac{\partial \mathbf{U}(\boldsymbol{\theta}^{0}, \boldsymbol{\lambda}^{0}, \boldsymbol{\eta}^{0})}{\partial (\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\eta})^{\top}} = \mathbf{a} = \begin{pmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \mathbf{a}_{13} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{0} \\ \mathbf{a}_{31} & \mathbf{0} & \mathbf{a}_{33} \end{pmatrix} \to \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} \\ \mathbf{A}_{31} & \mathbf{0} & \mathbf{A}_{33} \end{pmatrix} = \mathbf{A}, \tag{21.10}$$

where

$$\begin{aligned}
\mathbf{a}_{11}(lj, l'j') &= \frac{\partial u_1^0}{\partial \boldsymbol{\theta}_{l'j'}} = \sum_{i=1}^{n} \Big[\dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^{0} - w_{il}^{0} \dot{\mathbf{w}}_{il',\boldsymbol{\theta}_{l'j'}}^{0} \\
&\qquad - \Big\{ \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - v_{ilj}^{0} (\dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{-0}) \Big\} \Big] \tilde{\mathbf{x}}_{ij}^{\top}, \\
\mathbf{a}_{12}(lj, l'j') &= \frac{\partial u_1^0}{\partial \eta_{l'j'}} = \sum_{i=1}^{n} \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - v_{ilj}^{0} (\dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{-0}), \\
\mathbf{a}_{13}(l, l') &= \frac{\partial u_1^0}{\partial \lambda_{l'}} = \sum_{i=1}^{n} \dot{\mathbf{w}}_{il,\lambda_{l'}}^{0} - w_{il}^{0} \dot{\mathbf{w}}_{il',\lambda_{l'}}^{0}, \\
\mathbf{a}_{21}(lj, l'j') &= \frac{\partial u_2^0}{\partial \boldsymbol{\theta}_{l'j'}} = \sum_{i=1}^{n} \frac{1}{\eta_{lj}} \Big\{ \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - (v_{ilj}^{0} - v_{ilj}^{-0}) (\dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\boldsymbol{\theta}_{l'j'}}^{-0}) \Big\}, \\
\mathbf{a}_{22}(lj, l'j') &= \frac{\partial u_2^0}{\partial \eta_{l'j'}} = \sum_{i=1}^{n} \frac{1}{\eta_{lj}} \Big\{ \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - (v_{ilj}^{0} - v_{ilj}^{-0}) (\dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{0} - \dot{\mathbf{v}}_{ilj,\eta_{l'j'}}^{-0}) \Big\}, \\
\mathbf{a}_{31}(lj, l'j') &= \frac{\partial u_3^0}{\partial \boldsymbol{\theta}_{l'j'}} = \sum_{i=1}^{n} \frac{1}{\lambda_{l}} \Big\{ \dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^{0} - (w_{il}^{0} - w_{il}^{-0}) \dot{\mathbf{w}}_{il',\boldsymbol{\theta}_{l'j'}}^{0} \Big\}, \\
\mathbf{a}_{33}(l, l') &= \frac{\partial u_3^0}{\partial \lambda_{l'}} = \sum_{i=1}^{n} \frac{1}{\lambda_{l}} \Big\{ \dot{\mathbf{w}}_{il,\boldsymbol{\theta}_{l'j'}}^{0} - (w_{il}^{0} - w_{il}^{-0}) \dot{\mathbf{w}}_{il',\boldsymbol{\theta}_{l'j'}}^{0} \Big\}.
\end{aligned}$$

Define row vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ *such that* $\mathbf{b}_1(lj) = \sum_{i=1}^{n} (w_{il}^{0} - v_{ilj}^{0}) \tilde{\mathbf{x}}_{ij}^{\top}$, $\mathbf{b}_2(lj) = \sum_{i=1}^{n} \frac{1}{\eta_{lj}} (v_{ilj}^{0} - v_{ilj}^{-0})$, $\mathbf{b}_3(l) = \sum_{i=1}^{n} \frac{1}{\lambda_{l}} (w_{il}^{0} - w_{il}^{-0})$. *Then* $\sqrt{n} \mathbf{U}(\boldsymbol{\theta}^{0}, \boldsymbol{\lambda}^{0}, \boldsymbol{\eta}^{0}) = n^{-1} (\mathbf{b}_1^{\top}, \mathbf{b}_2^{\top}, \mathbf{b}_3^{\top})^{\top} \xrightarrow{d} N(\mathbf{0}, \mathbf{W})$, *where*

$$\mathbf{W} = E \begin{pmatrix} \mathbf{b}_1^{\top} \mathbf{b}_1 & \mathbf{b}_1^{\top} \mathbf{b}_2 & \mathbf{b}_1^{\top} \mathbf{b}_3 \\ \mathbf{b}_2^{\top} \mathbf{b}_1 & \mathbf{b}_2^{\top} \mathbf{b}_2 & \mathbf{b}_2^{\top} \mathbf{b}_3 \\ \mathbf{b}_3^{\top} \mathbf{b}_1 & \mathbf{b}_3^{\top} \mathbf{b}_2 & \mathbf{b}_3^{\top} \mathbf{b}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_2^{\top} & \mathbf{W}_3 \end{pmatrix}.$$

It then follows that

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\\ \hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^0\\ \hat{\boldsymbol{\eta}}-\boldsymbol{\eta}^0\end{pmatrix}=n^{-1}(\mathbf{a}^\top)^{-1}(\mathbf{b}_1^\top,\mathbf{b}_2^\top,\mathbf{b}_3^\top)^\top+o_p(\sqrt{n})\xrightarrow{d}N\left\{\mathbf{0},(\mathbf{A}^T)^{-1}\mathbf{W}\mathbf{A}^{-1}\right\}.$$

We can also study the behavior of particular parameters of interest. In particular,

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^0)\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Sigma}\equiv\mathbf{C}^\top\mathbf{W}_{11}\mathbf{C}+\mathbf{D}\mathbf{W}_{12}^\top\mathbf{C}+\mathbf{C}^\top\mathbf{W}_{12}\mathbf{C}+\mathbf{D}\mathbf{W}_{22}\mathbf{D}),$$

where

$$\begin{pmatrix}\mathbf{V}_{11}&\mathbf{V}_{12}&\mathbf{V}_{13}\\ \mathbf{V}_{21}&\mathbf{V}_{22}&\mathbf{V}_{23}\\ \mathbf{V}_{31}&\mathbf{V}_{23}&\mathbf{V}_{33}\end{pmatrix}=\mathbf{A}^{-1},\ \ \mathbf{D}=\left(\mathbf{A}_{33}-\mathbf{A}_{13}^\top\mathbf{V}_{11}\mathbf{A}_{13}\right),\ \ \mathbf{C}=\begin{pmatrix}\mathbf{V}_{11}\mathbf{A}_{13}\\ \mathbf{V}_{21}\mathbf{A}_{13}\end{pmatrix}\mathbf{D}.$$

We now prove the semiparametric efficiency of the proposed method. Write $\psi(x_1,\ldots,x_k,\boldsymbol{\delta})=\lambda_1+\sum_{l=2}^m\lambda_l\exp(\sum_{j=1}^k\tilde{\mathbf{x}}_j^\top\boldsymbol{\theta}_{lj})$. *For a finite dimensional parameter* $\boldsymbol{\phi}$, *consider a parametric submodel of* $h(x_1,\ldots,x_k)$

$$h(x_1,\ldots,x_k,\boldsymbol{\delta},\boldsymbol{\phi})=\psi(x_1,\ldots,x_k,\boldsymbol{\delta})\prod_{j=1}^k f_j(x_j,\boldsymbol{\phi}).\tag{21.11}$$

The profile likelihood $L_p(\boldsymbol{\delta})$ *is proportional to*

$$\frac{h(x_1,\ldots,x_k,\boldsymbol{\delta},\boldsymbol{\phi})}{h_1(x_1,\boldsymbol{\delta},\boldsymbol{\phi})\cdots h_k(x_k,\boldsymbol{\delta},\boldsymbol{\phi})},\tag{21.12}$$

where $h_j(x_j,\boldsymbol{\delta},\boldsymbol{\phi})=\{\lambda_1+\sum_{l=2}^m\lambda_l\exp(\tilde{\mathbf{x}}_j^\top\boldsymbol{\theta}_{lj})\}f_j(x_j,\boldsymbol{\phi})$. *Let* $\mathbf{S}_{\boldsymbol{\delta}},\mathbf{S}_{\boldsymbol{\phi}}$ *be the score functions of* $\boldsymbol{\delta}$ *and* $\boldsymbol{\phi}$ *based on (21.11) and (21.12). Write* $\psi(\boldsymbol{\delta})=\psi(x_1,\ldots,x_k,\boldsymbol{\delta})$, $h(\boldsymbol{\delta},\boldsymbol{\phi})=h(x_1,\ldots,x_k,\boldsymbol{\delta},\boldsymbol{\phi})$ $h_j(\boldsymbol{\delta},\boldsymbol{\phi})=h_j(x_j,\boldsymbol{\delta},\boldsymbol{\phi})$, $f_j(\boldsymbol{\phi})=f_j(x_j,\boldsymbol{\phi})$, $\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})=\partial h/\partial\boldsymbol{\delta}$, $\dot{\_}_{\boldsymbol{\delta}}(\boldsymbol{\delta})=\partial\psi/\partial\boldsymbol{\delta}$ *and* $\dot{\mathbf{f}}_{\boldsymbol{\phi}}(\boldsymbol{\phi})=\partial f/\partial\boldsymbol{\phi}$. *Then*

$$\mathbf{S}_{\boldsymbol{\delta}}=\frac{\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h(\boldsymbol{\delta},\boldsymbol{\phi})}-\sum_{j=1}^k\frac{\dot{\mathbf{h}}_{j,\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h_j(\boldsymbol{\delta},\boldsymbol{\phi})}=\mathbf{S}_{\boldsymbol{\delta}}^A+\sum_{j=1}^k\mathbf{S}_{j,\boldsymbol{\delta}}^B,$$

$$\mathbf{S}_{\boldsymbol{\phi}}=\frac{\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h(\boldsymbol{\delta},\boldsymbol{\phi})}=-\sum_{j=1}^k\frac{\dot{\mathbf{f}}_{j,\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_j(\boldsymbol{\phi})}=\sum_{j=1}^k\mathbf{S}_{j,\boldsymbol{\phi}}.$$

We will show that $\mathbf{S}_{\boldsymbol{\delta}}$ *and* $\mathbf{S}_{\boldsymbol{\phi}}$ *are orthogonal by showing* $E(\mathbf{S}^A_{\boldsymbol{\delta}}\mathbf{S}_{j,\boldsymbol{\phi}}) = \mathbf{0}$ *and* $E(\mathbf{S}^B_{j,\boldsymbol{\delta}}\mathbf{S}_{j',\boldsymbol{\phi}}) = \mathbf{0}, j, j' = 1, \ldots, k$ *and hence, the estimator is efficient. Denote* $\int \cdot d\mathbf{x} \equiv \int \cdots \int \cdot dx_1 \cdots dx_k$ *and* $\int \cdot d\mathbf{x}_{-1} \equiv \int \cdots \int \cdot dx_2 \cdots dx_k$.

$$
\begin{aligned}
E(\mathbf{S}^A_{\boldsymbol{\delta}}\mathbf{S}_{j,\boldsymbol{\phi}}) &= \int \frac{\dot{\mathbf{h}}_{\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h(\boldsymbol{\delta},\boldsymbol{\phi})} \frac{\dot{\mathbf{f}}_{j,\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_j(\boldsymbol{\phi})} h(\boldsymbol{\delta},\boldsymbol{\phi}) d\mathbf{x} \\
&= \int \dot{\llcorner}_{\boldsymbol{\delta}}(\boldsymbol{\delta}) \prod_{j=1}^{k} f_j(\boldsymbol{\phi}) \frac{\dot{\mathbf{f}}_{j,\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_j(\boldsymbol{\phi})} d\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\delta}} \int \frac{\partial}{\partial \boldsymbol{\phi}} \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=2}^{k} \tilde{\mathbf{x}}_j^\top \boldsymbol{\theta}_{lj})\} \prod_{j=2}^{k} f_j(\boldsymbol{\phi}) d\mathbf{x}_{-1} = \mathbf{0}. \\
E(\mathbf{S}^B_{j,\boldsymbol{\delta}}\mathbf{S}_{j',\boldsymbol{\phi}}) &= \int \frac{\dot{\mathbf{h}}_{j,\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h_j(\boldsymbol{\delta},\boldsymbol{\phi})} \frac{\dot{\mathbf{f}}_{j',\boldsymbol{\phi}}(\boldsymbol{\phi})}{f_{j'}(\boldsymbol{\phi})} \psi(\boldsymbol{\delta}) \prod_{j=1}^{k} f_j(\boldsymbol{\phi}) d\mathbf{x} \\
&= \int \frac{\dot{\mathbf{h}}_{j,\boldsymbol{\delta}}(\boldsymbol{\delta},\boldsymbol{\phi})}{h_j(\boldsymbol{\delta},\boldsymbol{\phi})} \frac{\partial}{\partial \boldsymbol{\phi}} \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j \neq j'} \tilde{\mathbf{x}}_j^\top \boldsymbol{\theta}_{lj})\} \prod_{j \neq j'}^{k} f_j(\boldsymbol{\phi}) d\mathbf{x}_{-1} = \mathbf{0}.
\end{aligned}
$$

□

The theorem allows us to draw inference about the mixture parameter $\boldsymbol{\lambda}$, as well as other quantities, such as component moments, that are smooth functions of the distribution parameters.

21.3 Estimation

Estimates of the parameters can be found by differentiating (21.9) with respect to each of the parameters and solving the simultaneous equations:

$$
\frac{\partial \ell_p}{\partial \boldsymbol{\theta}_{lj}} \Rightarrow \sum_{i=1}^{n} w_{il} \tilde{\mathbf{x}}_{ij}^\top - \sum_{i=1}^{n} v_{ilj} \tilde{\mathbf{x}}_{ij}^\top = 0, \tag{21.13}
$$

$$
\frac{\partial \ell_p}{\partial \eta_{lj}} \Rightarrow \sum_{i=1}^{n} \frac{\frac{1}{n}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}}{1 + \frac{1}{n}\sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}} = 0, \tag{21.14}
$$

$$
\frac{\partial \ell_p}{\partial \lambda_l} \Rightarrow \sum_{i=1}^{n} \frac{\exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1}{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})} = 0, \tag{21.15}
$$

for $l = 2, \ldots, m$ and $j = 1, \ldots, k$ and w_{il} and v_{ilj} are defined by

$$w_{il} = \frac{\lambda_l \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})}{\lambda_1 + \sum_{l=2}^m \lambda_l \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})},$$
$$v_{ilj} = \frac{\frac{1}{n}\eta_{lj} \exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})}{1 + \frac{1}{n}\sum_{l=2}^m \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1\}}. \tag{21.16}$$

Notice that in (21.14), we have used the fact that $\lambda_1 = 1 - \sum_{l=2}^m \lambda_l$. There is no explicit solution for λ_l, $l = 1, \ldots, k$. Therefore, we develop an EM type algorithm (Dempster et al. 1977). Define the latent variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ where $\mathbf{Z}_i^\top = (z_{i1}, \ldots, z_{im})$ for the component membership for the ith observation in the data set. If the ith observation belonged to the lth component, then $\mathbf{Z}_i^\top$ is a vector of $m - 1$ 0s and a single 1 in the lth position. Furthermore, $\sum_{l=1}^m z_{il} = 1$. Of course, $\mathbf{Z}_i, i = 1, \ldots, n$ are not observed. We define the "complete data" as $\{(\mathbf{X}_1, \mathbf{Z}_1), \ldots, (\mathbf{X}_n, \mathbf{Z}_n)\}$; then, a complete data semiparametric likelihood is

$$\begin{aligned} L_c(\boldsymbol{\delta}, F_1, \ldots, F_k) = \prod_{i=1}^n \Bigg[& \left\{ \lambda_1 \prod_{j=1}^k F_j(x_{ij}) \right\}^{z_{i1}} \\ & \prod_{l=2}^m \left\{ \lambda_l \exp\left(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} \right) \prod_{j=1}^k dF_j(x_{ij}) \right\}^{z_{il}} \Bigg] \\ = \prod_{i=1}^n \Bigg[& \lambda_1^{z_{i1}} \prod_{l=2}^m \lambda_l^{z_{il}} \exp\left(z_{il} \sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} \right) \prod_{j=1}^k dF_j(x_{ij})^{z_{il}} \Bigg]. \end{aligned}$$

Using p_{ij} as the jump size of F_j at x_{ij}, the complete data log-likelihood is

$$\begin{aligned} & \ell_c(\boldsymbol{\delta}, p_{ij}, i = 1, \ldots, n, j = 1, \ldots, k) \\ & = \sum_{i=1}^n \sum_{1=1}^m z_{il} \log \lambda_l + \sum_{i=1}^n \sum_{1=2}^m z_{il} \sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj} + \sum_{i=1}^n \sum_{j=1}^k \log p_{ij}, \end{aligned} \tag{21.17}$$

where (21.6) still hold and p_{ij}s can be profiled out using (21.7) and (21.8).

Let the parameter estimates at iteration t of the EM algorithm be $[\boldsymbol{\delta}^{(t)}]^\top = (\boldsymbol{\theta}_{21}^{(t)}, \ldots, \boldsymbol{\theta}_{mk}^{(t)}, \lambda_1^{(t)}, \ldots, \lambda_m^{(t)})$ and write

$$w_{il}^{(t)} = \mathrm{E}(z_{il} | \boldsymbol{\delta}^{(t)}, x_1, \ldots, x_n) = \frac{\lambda_l^{(t)} \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}^{(t)})}{\lambda_1^{(t)} + \sum_{l=2}^m \lambda_l^{(t)} \exp(\sum_{j=1}^k \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}^{(t)})}. \tag{21.18}$$

Since (21.17) is linear in z_{il}s, substituting (21.18) for z_{il}s and (21.7) for p_{ij}s in (21.17) gives the expected complete data profile log-likelihood (E-step) at iteration $t+1$,

$$Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)}) = \mathrm{E}(\ell_c|\boldsymbol{\delta}^{(t)}, x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{1=1}^{m} w_{il}^{(t)} \log \lambda_l + \sum_{i=1}^{n}\sum_{1=2}^{m} w_{il}^{(t)} \sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top} \boldsymbol{\theta}_{lj}$$
$$- \sum_{i=1}^{n}\sum_{j=1}^{k} \log[n + \sum_{l=2}^{m} \eta_{lj}\{\exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}) - 1\}].$$

The M-step maximizes $Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)})$ with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$. Since $\lambda_l, l = 1, \ldots, m$, satisfy the constraint $\sum_{l=1}^{m} \lambda_l = 1$, we immediately obtain

$$\hat{\lambda}_l^{(t+1)} = \frac{\sum_{i=1}^{n} w_{il}^{(t)}}{n}. \tag{21.19}$$

Differentiating $Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)})$ with respect to the other parameters gives exactly the same equations as (21.13) to (21.14), but with $w_{il}^{(t)}$s replacing w_{il}s. Using (21.19) and replacing w_{il}s by $w_{il}^{(t)}$s in (21.13) and (21.14) now gives

$$n\left(\frac{\eta_{lj}^{(t+1)}}{n}\right) - \sum_{i=1}^{n} w_{il}^{(t)} = 0 \Rightarrow \frac{\eta_{lj}^{(t+1)}}{n} = \frac{\sum_{i=1}^{n} w_{il}^{(t)}}{n} = \lambda_l^{(t+1)}. \tag{21.20}$$

Using (21.20) in (21.13) now gives

$$\sum_{i=1}^{n} \frac{\lambda_l^{(t)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}^{(t)})\tilde{\mathbf{x}}_{ij}^{\top}}{\lambda_1^{(t)} + \sum_{l=2}^{m} \lambda_l^{(t)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}^{(t)})} - \frac{\lambda_l^{(t+1)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj})\tilde{\mathbf{x}}_{ij}^{\top}}{\lambda_1^{(t+1)} + \sum_{l=2}^{m} \lambda_l^{(t+1)} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj})} = 0, \tag{21.21}$$

which can be used to easily solve for $\boldsymbol{\theta}_{lj}^{(t+1)}$ by a Newton-Raphson procedure.

To show that our EM algorithm increases $\ell_p(\boldsymbol{\delta})$ at every step, we note that

$$Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)}) \leq \sum_{i=1}^{n} \log \left\{ w_{i1}^{(t)}\lambda_1 + \sum_{1=2}^{m} w_{il}^{(t)}\lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}\right)\right\} + \sum_{i=1}^{n}\sum_{j=1}^{k} \log p_{ij}$$
$$< \sum_{i=1}^{n} \log \left\{ \lambda_1 + \sum_{1=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^{\top}\boldsymbol{\theta}_{lj}\right)\right\} + \sum_{i=1}^{n}\sum_{j=1}^{k} \log p_{ij} = \ell_p(\boldsymbol{\delta}),$$

Since $\boldsymbol{\delta}^{(t+1)}$ maximizes $Q(\boldsymbol{\delta}, \boldsymbol{\delta}^{(t)})$, therefore, $Q(\boldsymbol{\delta}^{(t+1)}, \boldsymbol{\delta}^{(t)}) \geq Q(\boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t)}) \Rightarrow \ell(\boldsymbol{\delta}^{(t+1)}) \geq \ell(\boldsymbol{\delta}^{(t)})$. We suggest using different starting values for the EM algorithm to check that the algorithm did not stop at a local maximum. Since the exponential

tilt parameters may be hard to interpret, it may be difficult to find initial values for them. We recommend generating an $n \times m$ matrix of initial values of the $z_{il}^{(t)}$s and starting the algorithm with the M-step.

21.4 Estimation of Features in the Component Distributions

In this section, we discuss estimation of features in the component distributions. We also identify a moment matching property similar to that found by Efron and Tibshirani (1996) in the univariate non-mixture case. For any quantity a, let $\hat{a}$ denote its estimate based on the final values of the EM algorithm at convergence. Define

$$\hat{p}_{ij} = \frac{1}{n}\left(\frac{1}{\hat{\lambda}_1 + \sum_{l=2}^{m} \hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^\top \hat{\boldsymbol{\theta}}_{lj})}\right) \qquad \text{and} \qquad \hat{q}_{lij} = \exp(\tilde{\mathbf{x}}_{ij}^\top \hat{\boldsymbol{\theta}}_{lj})\hat{p}_{ij}.$$

The CDF of the mixture distribution, H, can be estimated by

$$\hat{H}(x_1, \ldots, x_k) = \hat{\lambda}_1 \prod_{j=1}^{k} \hat{F}_j(x_j) + \sum_{l=2}^{m} \hat{\lambda}_l \prod_{j=1}^{k} \hat{G}_{lj}(x_j),$$

where the marginal CDF estimates of F_j and G_{lj} are

$$\hat{F}_j(x_j) = \sum_{i=1}^{n} I(x_{ij} \le x_j)\hat{p}_{ij}, \quad \hat{G}_{lj}(x_j) = \sum_{i=1}^{n} I(x_{ij} \le x_j) \exp(\tilde{\mathbf{x}}_{ij}^\top \hat{\boldsymbol{\theta}}_{lj})\hat{p}_{ij}. \quad (21.22)$$

The estimates resemble the empirical CDF with the weights given by the estimated jumps. In Sect. 21.5, we give examples that show how well these estimates match the true marginal CDFs. We can also find estimates of the marginal PDFs using a weighted kernel density estimate with the posterior probabilities, $\hat{w}_{il}$, as the weights. The estimated PDFs are

$$\hat{g}_{lj}(u) = \frac{1}{\kappa} \sum_{i=1}^{n} \frac{\hat{w}_{il}}{\sum_{i=1}^{n} \hat{w}_{il}} \xi\left(\frac{u - x_{ij}}{\kappa}\right), \quad l = 2, \ldots, m \quad (21.23)$$

where κ is a bandwidth, ξ is the standard normal PDF. The R package `mixtools` contains a function, `wkde`, that allows us to do this quite easily (Young et al. 2008). This function also has the ability to choose different bandwidths for the k coordinates.

Writing $\hat{q}_{1ij} = \hat{p}_{ij}$, the mean and variance of the jth measurement in the lth component distribution, m_{lj} and s_{lj}^2 can be estimated by

$$\hat{m}_{lj} = \sum_{i=1}^{n} x_{ij}\hat{q}_{lij}, \qquad \hat{s}_{lj}^2 = \sum_{i=1}^{n} x_{ij}^2\hat{q}_{lij} - \hat{m}_{lj}^2, \tag{21.24}$$

for $l = 1, \ldots, m$ and $j = 1, \ldots, k$. An interesting result from the EM algorithm is a moment matching property. For example, we can write:

$$\begin{aligned}
\sum_{i=1}^{n} x_{ij}\hat{q}_{lij} &= \sum_{i=1}^{n} x_{ij} \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj}) \left\{ \frac{1}{n} \left(\frac{1}{\hat{\lambda}_1 + \sum_{l=2}^{m} \hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj})} \right) \right\} \\
&= \frac{1}{n\hat{\lambda}_l} \sum_{i=1}^{n} x_{ij} \left\{ \frac{\hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj})}{\hat{\lambda}_1 + \sum_{l=2}^{m} \hat{\lambda}_l \exp(\tilde{\mathbf{x}}_{ij}^{\top}\hat{\boldsymbol{\theta}}_{lj})} \right\} \\
&= \frac{\sum_{i=1}^{n} \hat{w}_{il} x_{ij}}{\sum_{i=1}^{n} \hat{w}_{il}},
\end{aligned}$$

where the last quantity is due to (21.18) and (21.19) from the EM algorithm. This expression matches the weighted first moment using the posterior probabilities to the tilted component first moment; see Efron and Tibshirani (1996) for an example of the moment matching property in the univariate non-mixture exponential tilt model. They argue that moment matching reduces the bias.

It should be noted that non-identifiability due to label switching (e.g., McLachlan and Peel 2000) can affect bootstrap estimation in the exponential tilt model. Suppose observations come from the following mixture

$$H(x_1, x_2, x_3) = 0.3N(0, 1)N(0, 1)N(0, 1) + 0.7N(2, 1.5)N(2.5, 2)N(3, 1).$$

Consider the first coordinate, then one possible baseline distribution is $N(0, 1)$ and the parameters in the exponential tilt would be $\boldsymbol{\theta}_{21}^{\top} = (-1.53, 1.33, 0.16)$. Another possible baseline may be $N(2, 1.5)$ in which case the parameters would be $\boldsymbol{\theta}_{21}^{T*} = (1.53, -1.33, -0.16)$. The result is that in a bootstrap, for example, the signs of the coefficients in the quadratic exponent may change. We resolve this ambiguity by designating the component corresponding to the smallest proportion as the baseline distribution and make the adjustment after the EM algorithm converges. We then have identifiable estimates of the coefficients in the quadratic exponents. The estimates of the marginal means and standard deviations are not affected by this label switching.

21.5 Simulation Results

In this section, we give simulation results for different models. The data were generated from two component mixture distributions of the following form:

$$H(x_1, x_2, x_3) = \lambda F_1(x_1)F_2(x_2)F_3(x_3) + (1-\lambda)G_1(x_1)G_2(x_2)G_3(x_3).$$

We focus on the following parameters: λ, and m_{lj} and s_{lj}, the mean and standard deviation of the jth measurement in the lth component distribution.

21.5.1 Mixtures with Normal Component Distributions

The first model is a trivariate normal mixture model, such that F_1, F_2, F_3 are CDFs of $N(0,1)$ and G_1, G_2, G_3 are CDFs of $N(\mu, \sigma^2)$ with $(\mu, \sigma^2) = (2, 1.5), (2.5, 2), (3, 1)$, respectively. Three values of $\lambda = 0.3, 0.5, 0.8$ and two different sample sizes $n = 50$ and 500 were used. For each combination of λ and n, 500 simulations were carried out. The results using $\lambda = 0.3, 0.5, 0.8$ are similar and therefore, only those under $\lambda = 0.3$ are shown. The results are given in Table 21.1, where the parameter estimates using an exponential tilt and a conditional independence normal model are given under the columns "ET" and "Normal," respectively. The exponential tilt model performs very well, its estimates are comparable to those from the normal mixture. For small samples ($n = 50$), the standard errors for the estimates using the normal model are smaller. However,

Table 21.1 Mean (standard error) of parameter estimates based on 500 simulations from a normal mixture model

		$n = 50$		$n = 500$	
	True	ET	Normal	ET	Normal
λ	0.3	0.30 (0.09)	0.31 (0.07)	0.30 (0.02)	0.30 (0.02)
m_{11}	0	0.11 (0.56)	0.02 (0.36)	0.00 (0.08)	0.00 (0.08)
m_{12}	0	0.15 (0.69)	0.01 (0.37)	−0.01 (0.09)	−0.01 (0.09)
m_{13}	0	0.22 (0.77)	0.02 (0.45)	0.02 (0.09)	0.00 (0.09)
m_{21}	2	1.97 (0.31)	1.99 (0.27)	2.00 (0.06)	1.99 (0.06)
m_{22}	2.5	2.46 (0.39)	2.49 (0.33)	2.50 (0.08)	2.49 (0.08)
m_{23}	3	2.92 (0.39)	2.98 (0.31)	2.99 (0.05)	3.00 (0.05)
s_{11}	1	0.90 (0.29)	0.92 (0.21)	0.99 (0.06)	0.99 (0.06)
s_{12}	1	0.92 (0.28)	0.93 (0.20)	0.99 (0.06)	0.99 (0.06)
s_{13}	1	1.01 (0.37)	0.95 (0.25)	1.02 (0.09)	0.99 (0.06)
s_{21}	1.22	1.18 (0.17)	1.18 (0.15)	1.21 (0.04)	1.22 (0.04)
s_{22}	1.41	1.37 (0.20)	1.37 (0.18)	1.41 (0.05)	1.41 (0.05)
s_{23}	1	1.01 (0.21)	0.97 (0.13)	0.99 (0.04)	1.00 (0.04)

the advantage of using a normal model effectively disappears for large samples ($n = 500$).

21.5.2 Mixtures with Gamma Component Distributions

Exponential tilt modeling can be thought of as a density estimation method (Efron and Tibshirani 1996). Hence we can use exponential tilt even for data that do not satisfy the exponential tilt assumption. We illustrate using mixtures of gamma distributions with different shape parameters (for application, see Dey et al. 1995; Wiper et al. 2001). We let F_1, F_2, F_3 be CDFs from a gamma(k, ζ) distribution with $(k, \zeta) = (2, 2)$, $(\mu = 4, \sigma^2 = 8)$, and G_1, G_2, G_3 are CDFs corresponding to gamma(k, ζ) distributions with $(k, \zeta) = (5, 2), (10, 1),$ and $(10, 0.5)$, respectively (with $(\mu, \sigma^2) = (10, 20), (10, 10),$ and $(5, 2.5)$, respectively). The results are also similar for different λ values, hence, we only present $\lambda = 0.4$ here. We use 1000 simulations of sample sizes $n = 50$ and 300 were carried out. We computed the estimates of the component means and standard deviations using the conditional independence normal mixture model and the conditional independence nonparametric mixture (NP) model proposed by Benaglia et al. (2009) and Levine et al. (2011) for comparison. The estimates from all three methods are shown in Table 21.2. When the sample size is small, the performance of the exponential tilt method is similar to the normal mixture model. For larger sample size ($n = 300$), the tilted method does much better than the normal model and follows more closely to the nonparametric method.

Table 21.2 Mean (standard error) of parameter estimates based on 1000 simulations from a gamma mixture model

		$n = 50$			$n = 300$		
	True	ET	Normal	NP	ET	Normal	NP
λ	0.4	0.39 (0.12)	0.37 (0.13)	0.38 (0.10)	0.37 (0.04)	0.32 (0.05)	0.36 (0.04)
m_{11}	4	3.86 (1.02)	3.66 (0.99)	3.86 (0.85)	3.85 (0.35)	3.41 (0.36)	3.78 (0.31)
m_{12}	4	4.78 (2.52)	4.53 (2.68)	4.47 (2.15)	3.78 (0.53)	3.32 (0.45)	3.71 (0.34)
m_{13}	4	4.03 (0.89)	4.11 (0.93)	4.02 (0.79)	3.97 (0.30)	4.06 (0.35)	3.96 (0.29)
m_{21}	10	9.99 (1.12)	10.03 (1.38)	9.93 (1.19)	9.84 (0.38)	9.56 (0.39)	9.76 (0.36)
m_{22}	10	9.04 (1.77)	9.04 (1.66)	9.27 (1.61)	9.86 (0.41)	9.59 (0.32)	9.79 (0.28)
m_{23}	5	4.84 (0.44)	4.80 (0.42)	4.86 (0.42)	4.96 (0.13)	4.84 (0.13)	4.95 (0.12)
s_{11}	2.82	2.47 (1.05)	2.11 (0.86)	2.63 (0.96)	2.65 (0.44)	1.96 (0.28)	2.62 (0.40)
s_{12}	2.82	2.47 (1.00)	2.11 (0.87)	2.61 (0.91)	2.53 (0.50)	1.90 (0.33)	2.54 (0.38)
s_{13}	2.82	2.42 (0.90)	2.40 (0.95)	2.49 (0.79)	2.84 (0.32)	2.91 (0.39)	2.80 (0.32)
s_{21}	4.47	4.23 (0.80)	4.23 (0.87)	4.31 (0.72)	4.48 (0.29)	4.58 (0.28)	4.50 (0.28)
s_{22}	3.16	3.20 (0.62)	3.27 (0.58)	3.15 (0.51)	3.21 (0.22)	3.42 (0.23)	3.27 (0.20)
s_{23}	1.58	1.69 (0.46)	1.75 (0.43)	1.70 (0.35)	1.60 (0.14)	1.71 (0.13)	1.67 (0.12)

Table 21.3 Mean (standard error) of parameter estimates based on 1000 simulations from a mixture model of normal and gamma

		$n = 50$			$n = 300$		
	True	ET	Normal	NP	ET	Normal	NP
λ	0.4	0.33(0.08)	0.34(0.11)	0.36(0.09)	0.30 (0.03)	0.35 (0.04)	0.35 (0.04)
m_{11}	0	0.91(1.61)	1.57(2.04)	1.50(1.90)	0.04 (0.16)	0.27 (0.23)	0.39 (0.22)
m_{12}	4	3.23(1.60)	2.55(1.79)	2.70(1.74)	3.94 (0.31)	3.45 (0.41)	3.64 (0.33)
m_{13}	0	−0.04(1.48)	−0.15(3.69)	−0.07(3.37)	0.00 (0.11)	−0.01 (0.15)	0.00 (0.23)
m_{21}	4	3.69(0.98)	3.49(1.23)	3.52(1.22)	4.00 (0.20)	4.13 (0.23)	4.04 (0.20)
m_{22}	0	0.23(0.86)	0.48(0.93)	0.39(1.05)	0.01 (0.10)	0.02 (0.11)	−0.07 (0.10)
m_{23}	0	−0.01(1.04)	−0.01(0.98)	−0.01(1.06)	0.00 (0.38)	0.01 (0.40)	0.01 (0.42)
s_{11}	1	1.61(1.07)	1.63(1.02)	1.97(1.00)	1.15 (0.41)	1.12 (0.18)	1.75 (0.48)
s_{12}	2.82	2.45(0.91)	2.43(1.19)	2.40(0.91)	2.78 (0.33)	3.03 (0.34)	2.82 (0.31)
s_{13}	1	1.85(1.85)	2.54(2.53)	2.88(2.12)	0.99 (0.11)	1.07 (0.56)	2.06 (0.81)
s_{21}	2.82	2.61(0.62)	2.62(0.69)	2.56(0.63)	2.79 (0.23)	2.87 (0.23)	2.75 (0.22)
s_{22}	1.41	1.46(0.59)	1.61(0.76)	1.55(0.69)	1.42 (0.15)	1.33 (0.12)	1.33 (0.13)
s_{23}	5.65	5.10(1.37)	4.69(1.63)	4.76(1.43)	5.64 (0.44)	5.79 (0.52)	5.60 (0.44)

21.5.3 Mixtures with Different Component Distributions

The third set of simulations studied the situation where the marginals are from different families of distributions (see, e.g., Khalili et al. 2007). We let F_1, F_2, F_3 be CDFs from N(0,1), gamma($k = 2, \zeta = 2$), ($\mu = 4, \sigma^2 = 8$), and N(0,1) distributions, respectively, and G_1, G_2, G_3 are CDFs corresponding to gamma($k = 2, \zeta = 2$), Laplace distributions with location and scale parameters (0,1), ($\mu = 0, \sigma^2 = 2$), and a Laplace with parameters (0,4), ($\mu = 0, \sigma^2 = 32$), respectively. The results under different values of λ are similar and hence only results for $\lambda = 0.3$ are given. One thousand simulations of sample sizes $n = 50$ and 300 were carried out. The results are given in Table 21.3.

It can be observed that the tilted method produces the best results for nearly all the parameters. We also plotted the estimated marginal CDFs and PDFs for one of the simulations in Fig. 21.1.

Again, even though the exponential tilt assumption does not hold here, the exponential tilt estimates of the component means, standard deviations, CDF, and PDF are very good.

21.6 Model Selection

In this section, we show how to estimate the number of components in the mixture. We use a modified BIC (Bayesian Information Criterion, Schwarz 1978) model selection criterion pBIC $\equiv -2 \ln L_p + s \ln(n)$, where L_p is the maximized

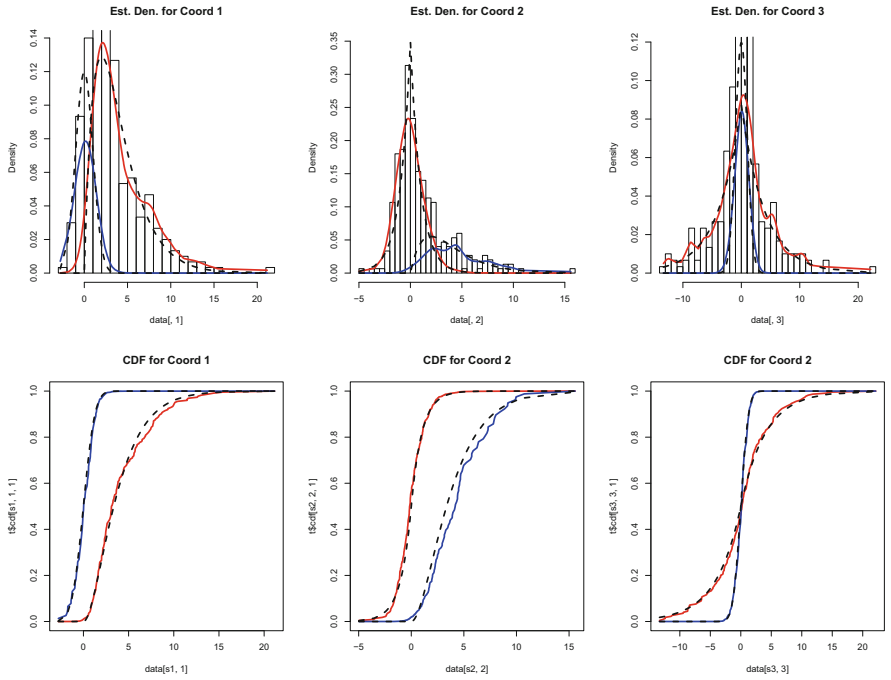

Fig. 21.1 Semiparametric estimation for a randomly selected dataset from the simulations with different distributions with $n = 300$. The *dotted line* represents the true CDFs and PDFs. In this dataset, the estimates are: $\hat{\lambda} = 0.25$, $(\hat{m}_{11}, \hat{m}_{12}, \hat{m}_{13}) = (0.10, 4.49, 0.06)$, $(\hat{m}_{21}, \hat{m}_{22}, \hat{m}_{23}) = (4.20, -0.04, 0.14)$, $(\hat{s}_{11}, \hat{s}_{12}, \hat{s}_{13}) = (1.08, 2.86, 1.02)$, $(\hat{s}_{21}, \hat{s}_{22}, \hat{s}_{23}) = (3.28, 1.37, 5.05)$

semiparametric profile likelihood and s is the number of parameters in the model. Since mixture models do not satisfy all the regularity conditions in Schwarz (1978) we turn to simulations to study the criterion. We use three models for simulations:

Model 1: Normal location mixtures with $m = 2, 3, 4$ components. There are $k = 7$ repeated measures with $(m_{1j}, m_{2j}, m_{3j}, m_{4j}) = (0, 2, 4, 6)$ and $s_{lj} = 1$ for $l = 1, \ldots, m; j = 1, \ldots, 7$.

Model 2: Normal location mixtures with $m = 2, 3, 4$ components. There are $k = 10$ repeated measures with $(m_{1j}, m_{2j}, m_{3j}, m_{4j}) = (0, 2, 4, 6)$ and $s_{lj} = 1$ for $l = 1, \ldots, m; j = 1, \ldots, 10$.

Model 3: Normal scale mixtures with $m = 2, 3$ components. There are $k = 5$ repeated measures with $(m_{1j}, m_{2j}, m_{3j}) = (0, 0, 0)$ and $(s_{1j}, s_{2j}, s_{3j}) = (0, 10, 50)$ for $j = 1, \ldots, 5$.

Table 21.4 gives the proportion of times pBIC selected the correct number of components. For each model considered, the mixing proportions of the components are equal, i.e., for a model with m components, $\lambda_1 = \lambda_2 = \ldots = \lambda_m$. Included in the table is the number of parameters estimated in each model, $N_p = 3k(m-1)+(m-1)$, which includes the exponential tilt parameters for each of k dimensions in the $m-1$

Table 21.4 pBIC simulations results for Models 1–3 where n is the number of observations, k is the number of repeated measures, and m is the true number of components

		$m = 2$			$m = 3$			$m = 4$		
Model	k	N_p	$n = 100$	$n = 200$	N_p	$n = 100$	$n = 200$	N_p	$n = 100$	$n = 200$
1	7	15	1.00	1.00	30	1.00	1.00	45	0.96	0.98
2	10	21	1.00	1.00	42	0.95	0.99	63	0.91	0.97
3	5	11	0.94	0.97	22	0.65	0.67	—	—	—

components and the $m - 1$ mixing proportions λ_l. For Model 3 with $m = 3$ the success rate for pBIC was roughly $2/3$. However, when the sample size increased to 500, the success rate increased to 0.90. As a check, we compared pBIC to a modified Akakie Information Criterion, which gives similar results. We conclude that pBIC is effective for estimating the number of components in the semiparametric mixture.

21.7 Example

We applied the proposed method to a real data problem. The data comes from a cognitive experiment discussed in Cruz-Medina et al. (2004) and is available at http://www.blackwellpublishing.com/rss. The experiment was used to demonstrate children fall into different groups in their approach to solve cognitive tasks. The experiment recruited normally developing 9-year-old children. Each child was given a set of different task conditions, which is a visual stimulus that involves two images on a computer monitor. The left image is the target stimulus and the right image is either identical to the target image or the mirror image of the target stimulus. The child pressed one key indicating if he/she thought the right image was identical or another key if they thought it was the mirror image. The outcome of interest is the reaction time (RT), in milliseconds, for a child to give a response to the visual stimulus. Each child was given $k = 6$ different task conditions and the RT for the child to choose the correct response on each task was recorded. We focused on the subset of $n = 197$ children who gave correct responses to all the task conditions. Since the six task conditions were embedded in a random sequence of tasks, the children could not have anticipated which task condition would appear. Therefore, given that a child was in a particular group, it would not be unreasonable to assume that their reaction times were independent and the conditional independence assumption seems valid. Longer response times may indicate reading comprehension problems. See Miller et al. (2001) for additional background on this experiment.

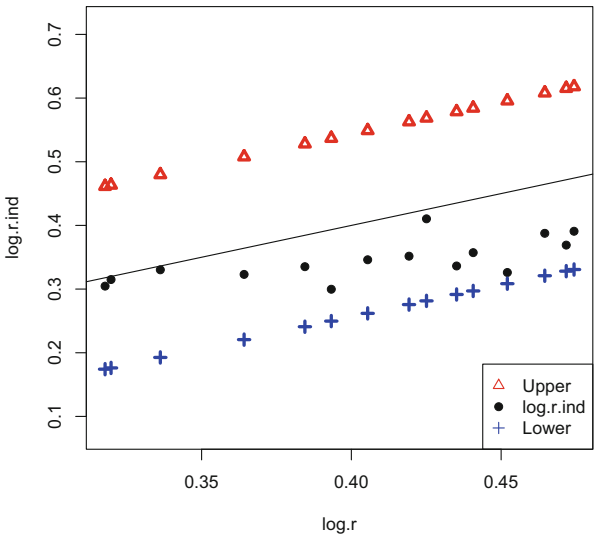

Fig. 21.2 Plot of the transformed sample correlations (log.r) against the transformed sample correlations calculated under conditional independence (log.r.ind). The *upper* and *lower* confidence bands of the transformed sample correlations are shown. The *solid line* is the 45° line

To further examine whether the conditional independence assumption is reasonable, we looked at the sample correlations between the coordinates and those under conditional independence. The correlations were calculated and Fisher's transformations were performed. Figure 21.2 plots the transformed correlations under conditional independence against the transformed correlations with no assumptions.

As a rough check, we included upper and lower points computed using $2/\sqrt{n-3}$ as an estimate of the standard errors of the transformed correlations. All estimates assuming conditional independence fall within these bounds.

We compared $m = 1, 2, 3, 4$ component models for this dataset using pBIC and selected $m = 3$ based on its lowest pBIC value ($\ell_P = -6081.6$, pBIC$= 12300.6$ and corresponding number of parameters $= 26$). The data based on $m = 3$ can be written as $x_{ij}, i = 1, 2, \ldots, 197; j = 1, \ldots, 6$ with corresponding CDF

$$H(x_1, \ldots, x_6) = \lambda_1 \prod_{j=1}^{6} F_j(x_j) + \sum_{l=2}^{3} \lambda_l \prod_{j=1}^{6} G_{lj}(x_j).$$

The estimated marginal CDFs, means and standard deviations of $F_j, G_{2j}, G_{3j}, j = 1, \ldots, 6$ using (21.22) and (21.24) are given in Table 21.5.

It appears that the distribution of RTs for the first task condition may well be different from the distributions of RTs for the other task conditions. From the results, the smallest group of children composed of about 20 % appear to have the shortest RTs and also the smallest variation. This might suggest that these children

Table 21.5 Estimated component means for the RT data with $m = 3$ components

Component					
1		2		3	
λ_1	0.49	λ_2	0.20	λ_3	0.31
m_{11}	2024	m_{21}	1577	m_{31}	3024
m_{12}	1712	m_{22}	1456	m_{32}	2776
m_{13}	1864	m_{23}	1265	m_{33}	2761
m_{14}	1799	m_{24}	1312	m_{34}	2771
m_{15}	1870	m_{25}	1171	m_{35}	2729
m_{16}	1957	m_{26}	1216	m_{36}	2661
s_{11}	691	s_{21}	420	s_{31}	1074
s_{12}	469	s_{22}	337	s_{32}	907
s_{13}	609	s_{23}	200	s_{33}	1101
s_{14}	516	s_{24}	332	s_{34}	1097
s_{15}	777	s_{25}	402	s_{35}	1162
s_{16}	636	s_{26}	261	s_{36}	1180

understand the concept and are quick to choose correctly. The next group composed of about 30 % of the children have the longest RTs as well as the largest variation. For the children in this group, a possible explanation is that they look longer to react to certain tasks and quicker for other tasks. It would be interesting to break up the trials based on which was the correct answer, the identical image or the mirror image. The last and largest group, about 50 %, are the children in the middle.

Figure 21.3 shows the semiparametric estimates of the component CDFs. Similar analyses were carried out using the log transformed data with similar results. Note the variation in the coordinate means and standard deviations again suggests that the component marginal distributions differ. The data were originally analyzed by Cruz-Medina et al. (2004) by discretizing the data and assuming that the repeated measures were identically distributed. The estimated proportions were 0.55, 0.16, and 0.29 in the order given in Table 7. The common coordinate medians were 1689, 1273, and 2523 for the three components and are a bit lower than the reported sets of five means for each component.

21.8 Discussion and Modifications

Walther (2002) introduces a univariate mixture of log-concave densities. He gives a representation theorem, and based on this theorem, develops a test for the presence of a mixture model. Chang and Walther (2007) extend this model to the multivariate case in (21.1). However, lack of identifiability is a difficulty for their model (Walther 2002, pp. 509); a mixture of log-concave densities may itself be log-concave and identifiability fails.

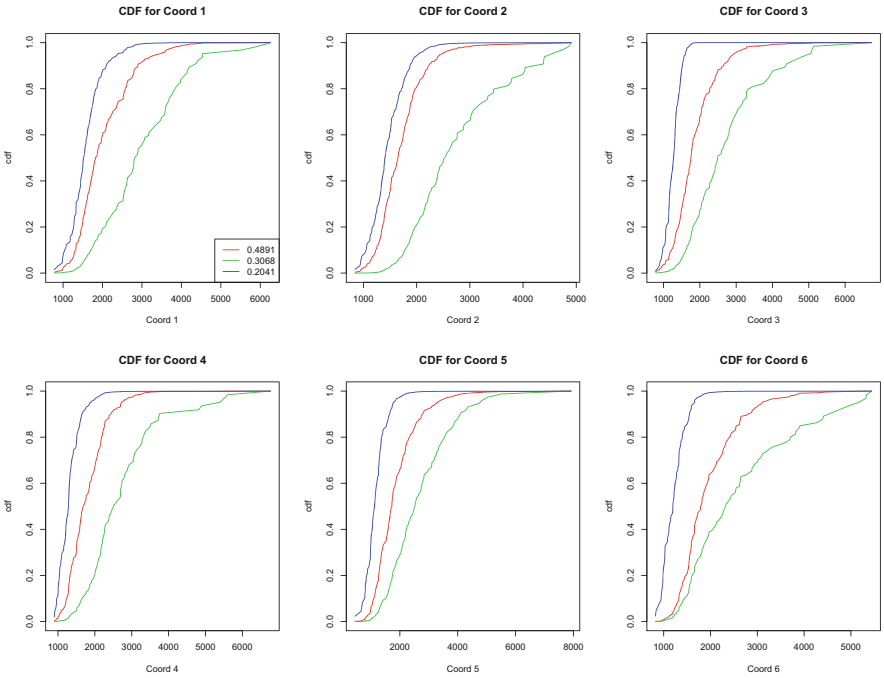

Fig. 21.3 Semiparametric estimates of $F_j, G_{lj},\ j = 1, \ldots, 6,\ l = 2, 3$ under the exponential tilt model for the RT data

The method we proposed assumes all repeated measures are related by (21.3). Our model could be modified to handle situations where some of the dimensions are modeled by (21.3) while the others are modeled parametrically. For example, let the first j_1 dimensions be modeled by (21.3), then

$$
\begin{aligned}
&h(x_1, \ldots, x_k, \boldsymbol{\delta}, \boldsymbol{\Omega}) \\
&= \left\{ \lambda_1 \prod_{j=j_1+1}^{k} f_j(x_j, \omega_j^f) + \sum_{l=2}^{m} \lambda_l \exp\Big(\sum_{j=1}^{j_1} \tilde{\mathbf{x}}_j^\top \boldsymbol{\theta}_{lj}\Big) \prod_{j=j_1+1}^{k} g_j(x_j, \omega_j^g) \right\} \prod_{j=1}^{j_1} f_j(x_j) \\
&= \psi(x_1, \ldots, x_k, \boldsymbol{\delta}, \boldsymbol{\Omega}) \prod_{j=1}^{j_1} f_j(x_j),
\end{aligned}
$$

where $f_j, g_j, j = j_1 + 1, \ldots, k$ are parametrized by $\boldsymbol{\Omega} = (\omega_j^f, \omega_j^g)$, which leads to

$$
\ell_p(\boldsymbol{\delta}, \boldsymbol{\Omega}) = \log \psi(x_1, \ldots, x_k, \boldsymbol{\delta}, \boldsymbol{\Omega}) - \sum_{i=1}^{n} \sum_{j=1}^{j_1} \log \left[n + \sum_{l=2}^{m} \eta_{lj} \{ \exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj}) - 1 \} \right].
$$

Our formulation of a multivariate mixture can also be interpreted as a copula. Replacing η_{lj}/n by λ_l, a simple rearrangement yields the profile likelihood as

$$n^{-kn} \prod_{i=1}^{n} \left[\frac{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\sum_{j=1}^{k} \tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})}{\prod_{j=1}^{k} \{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp(\tilde{\mathbf{x}}_{ij}^\top \boldsymbol{\theta}_{lj})\}} \right].$$

If we multiply the numerator and denominator by $\prod_{j=1}^{k} f_j(x_j)$, then the profile likelihood is proportional to $c\,(H_1(x_1), \ldots, H_k(x_k)) \equiv h(x_1, \ldots, x_k) / \prod_{j=1}^{k} h_j(x_j)$, the joint mixture density divided by a product of the marginal densities. This can be viewed as a semiparametric copula density evaluated at the marginal CDFs. We can also interpret our exponential tilt mixture as $h(x_1, \ldots, x_k) = c\,(H_1(x_1), \ldots, H_k(x_k)) \prod_{j=1}^{k} h_j(x_j)$. Hence we begin with a product of (independent) marginals and model the correlation and mixture structure via the copula based on the mixture of exponential tilts. Further motivation can be found in Chen et al. (2006). This approach also avoids the curse of dimensionality problem associated with estimation in high dimensional distributions.

Acknowledgements Tracey Wrobel Hammel and Thomas Hettmansperger were partially supported by NSF Grant SES-0518772. Denis Leung was supported by SMU Research Center.

References

Allman, E.S., Matias, C., Rhodes, J.A.: Identifiability of parameters in latent class models with many observed variables. Ann. Stat. **37**, 3099–3132 (2009)

Anderson, J.A.: Multivariate logistic compounds. Biometrika **66**, 17–26 (1979)

Benaglia, T., Chauveau, D., Hunter, D.R.: An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. J. Comput. Graph. Stat. **18**, 505–526 (2009)

Chang, G.T., Walther, G.: Clustering with mixtures of log-concave distributions. Comput. Stat. Data Anal. **51**, 6242–6251 (2007)

Chen, X., Fan, Y., Tsyrennikov, V.: Efficient estimation of semiparametric multivariate copula models. J. Am. Stat. Assoc. **101**, 1228–1240 (2006)

Cruz-Medina, I.R., Hettmansperger, T.P., Thomas, H.: Semiparametric mixture models and repeated measures: the multinomial cut point model. Appl. Stat. **53**, 463–474 (2004)

Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)

Dey, D., Kuo, L., Sahu, S.: A bayesian predictive approach to determining the number of components in a mixture distribution. Stat. Comput. **5**, 297–305 (1995)

Efron, B., Tibshirani, R.: Using specially designed exponential families for density estimation. Ann. Stat. **24**, 2431–2461 (1996)

Elmore, R.T., Hall, P., Neeman, A.: An application of classical invariant theory to identifiability in nonparametric mixtures. Ann. I. Fourier **55**, 1–28 (2005)

Hall, P., Neeman, A., Pakyari, R., Elmore, R.T.: Nonparametric inference in multivariate mixtures. Biometrika **92**, 667–678 (2005)

Hall, P., Zhou, X.H.: Nonparametric estimation of component distributions in a multivariate mixture. Ann. Stat. **31**, 201–224 (2003)

Hettmansperger, T.P., Thomas, H.: Almost nonparametric inference for repeated measure in mixture models. J. R. Stat. Soc. Ser. B **62**, 811–825 (2000)
Kay, R., Little, S.: Assessing the fit of the logistic model: A case study of children with the haemolytic uraemic syndrome. Appl. Stat. **35**, 16–30 (1986)
Khalili, A., Potter, D., Yan, P., Li, L., Gray, J., Huang, T., Lin, S.: Gamma-normal-gamma mixture model for detecting differentially methylated loci in three breast cancer cell lines. Cancer Informat. **3**, 43–54 (2007)
Leung, D., Qin, J.: Semi-parametric inference in a bivariate (multivariate) mixture model. Stat. Sin. **16**, 153–163 (2006)
Levine, M., Hunter, D.R., Chauveau, D.: Maximum smoothed likelihood for multivariate mixtures. Biometrika **98** 403–416 (2011)
Lindsay, B.G.: Mixture Models: Theory, Geometry, and Applications. Institute of Mathematical Statistics, Hayward, CA (1995)
Liu, Q., Pierce, D.A.: A note on gauss-hermite quadrature. Biometrika **81**, 624–629 (1994)
McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
Miller, C.A., Kail, R., Leonard, L.B., Tomblin, J.B.: Speed of processing in children with specific language impairment. J. Speech Lang. Hear. Res. **44**, 416–433 (2001)
Nagelkerke, N.J.D., Borgdorff, M.W., Kim, S.J.: Logistic discrimination of mixtures of m. tuberculosis and non-specific tuberculin reactions. Stat. Med. **20**, 1113–1124 (2001)
Owen, A.: Empirical likelihood ratio confidence intervals for a single functional. Biometrika **75**, 237–49 (1988)
Qin, J., Berwick, M., Ashbolt, R., Dwyer, T.: Quantifying the change of melanoma incidence by breslow thickness. Biometrics **58**, 665–670 (2002)
Qin, J., Zhang, B.: A goodness-of-fit test for logistic regression models based on case-control data. Biometrika **84**, 609–618 (1997)
Qu, Y.S., Hadgu, A.: A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. J. Am. Stat. Assoc. **93**, 920–928 (1998)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **5**, 461–464 (1978)
Thomas, H., Lohaus, A.: Modeling growth and individual differences in spatial tasks. Monogr. Soc. Res. Child Dev. **58**, 1–191 (1993)
Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester (1985)
Walther, G.: Detecting the presence of mixing with multiscale maximum likelihood. J. Am. Stat. Assoc. **97**, 508–513 (2002)
White, I.R., Thompson, S.G.: Choice of test for comparing two groups, with particular application to skewed outcomes. Stat. Med.**21**, 1205–1215 (2003)
Wiper, M., Rios Insua, D., Ruggeri, F.: Mixtures of gamma distributions with applications. J. Comput. Graph. Stat. **10**, 440–454 (2001)
Young, D.S., Benaglia, T., Chauveau, D., Elmore, R.T., Hettmansperger, T.P., Hunter, D.R., Thomas, H., Xuan, F.: mixtools: Tools for mixture models. R package version 0.3.2 (2008)
Zhang, B.: An information matrix test for logistic regression models based on case-control data. Biometrika **88**, 921–932 (2001)