Estimation for Conditional Independence Multivariate Finite Mixture Models

Didier Chauveau, David R. Hunter, Michael Levine

Estimation for Conditional Independence Multivariate Finite Mixture Models

Didier Chauveau* David R. Hunter† Michael Levine‡

September 9, 2010

Abstract

The conditional independence assumption for nonparametric multivariate finite mixture models may be considered to be a weaker form of the well-known conditional independence assumption for random effects models for longitudinal data. After summarizing important recent identifiability results, this article describes and extends an algorithm for estimation of the parameters in these models. The algorithm works for any number of components and any dimensionality of at least three, and it possesses a descent property and can be easily adapted to situations where the data is grouped in blocks of conditionally independent variables. We discuss how to adapt this algorithm to various location-scale models that link component densities, and we even adapt it to a particular class of univariate mixture problems in which the components are assumed symmetric. We also give an example of possible bandwidth selection procedure for our algorithm. The effectiveness of the new algorithm is demonstrated in a simulation study and two psychometric datasets.

1 Introduction

The analysis of longitudinal data generally involves multivariate observations for each subject in which the correlation among observations for a

*Laboratoire MAPMO, Université d'Orléans & CNRS UMR 6628, France, didier.chauveau@univ-orleans.fr

†Department of Statistics, Pennsylvania State University, University Park PA 16801, USA, dhunter@stat.psu.edu

‡Department of Statistics, Purdue University, West Lafayette, IN 47907, USA, mlevins@purdue.edu

given subject must be taken into account. A common method for modeling this situation is the so-called "conditional-independence model" (Laird and Ware, 1982), in which each multivariate observation, say $\mathbf{X}_i$ for $1 \leq i \leq n$, consists of a subject-specific effect plus random noise. The hallmark of the conditional independence model is that the noise is independent; i.e., conditional on the subject-specific effect, the multivariate vector consists of independent observations. Furthermore, each subject-specific effect may depend on certain covariates that are observed, but it also depends on an unobserved, or latent, feature of the individual. Importantly, all aspects of the traditional random-effects model for longitudinal data—in particular, the subject-specific effects and the independent random noise—are considered to be realizations from some parametric model that is specified *a priori*, and the parameters are the objects of statistical estimation.

Here, we relax the traditional parametric assumption of the conditional-independence random effects model. The model we use retains the characteristic conditional independence assumption, but instead of subject-specific effects, we posit that the population is divided into m distinct components, each subject belonging to one of those components, and that each multivariate observation has independent measurements conditional on the *component* from which the individual comes. Trading the usual specific-subject effect for the less-specific component effect leads to a finite mixture model, and as we shall see below, it allows us to do away with the parametric assumption altogether. We are therefore led to consider nonparametric finite mixture models under an assumption of conditional independence.

Specifically, suppose the r-dimensional vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are a simple

random sample from a finite mixture density of m components $f_1, \ldots, f_m$, with $m > 1$ and known in advance. It is assumed throughout this manuscript that each one of these densities f_j is equal with probability 1 to the product of its marginal densities:

$$f_j(\mathbf{x}) = \prod_{k=1}^{r} f_{jk}(x_k). \tag{1}$$

Taking a fully nonparametric approach with regard to estimating the f_{jk}, we may therefore express the finite mixture density as

$$\mathbf{X}_i \ \sim \ g_{\boldsymbol{\theta}}(\mathbf{x}_i) \ = \ \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}), \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ must satisfy

$$\sum_{j=1}^{m} \lambda_j = 1 \quad \text{and each } \lambda_j \geq 0. \tag{3}$$

Here, we assume $\mathbf{X}_i = (X_{i1}, \ldots, X_{ir})^\top$ and we let $\boldsymbol{\theta}$ denote the vector of parameters to be estimated, including the mixing proportions $\lambda_1, \ldots, \lambda_m$ and the univariate densities f_{jk}. Furthermore, throughout this article, j and k always denote the component and coordinate indices, respectively; thus, $1 \leq j \leq m$ and $1 \leq k \leq r$.

This finite-mixture version of the conditional independence assumption has appeared in a growing body of literature on non- and semi-parametric multivariate mixture models. Hettmansperger and Thomas (2000) introduced a more restrictive version of (2) in which the f_{jk} depended only on j. This conditionally i.i.d. (independent *and* identically distributed) finite mixture model was later examined by Elmore and Wang (2003) and Cruz-Medina and Hettmansperger (2004). Hall and Zhou (2003) considered (2)

in its full generality, establishing some rudimentary results concerning the identifiability of the parameters in this model. Other articles (Elmore et al., 2005; Hall et al., 2005) explored this identifability question further, until Allman et al. (2009) established the fundamental result that we elucidate fully in Section 2. Benaglia et al. (2009a) proposed an estimation algorithm for (2), which was later modified and put on more solid theoretical ground by Levine et al. (2010), who showed that the modified algorithm possesses a descent property, much like any EM algorithm. In Section 3 of this article, we extend the algorithm of Levine et al. (2010), and in Section 5, we summarize numerical tests of the extended algorithm.

2 Identifiability

The fundamental result concerning identifiability of finite mixtures of non-parametric measure products is due to Allman et al. (2009). It is based on a fundamental algebraic result of Kruskal (1976, 1977) that we need to present first. J. B. Kruskal studied contingency tables in the context of his interest in psychometrics. His work describes a 3-way contingency table that cross-classifies a sample of n individuals with respect to 3 polytomous variables, the kth of which has a state space $\{1, \ldots, \kappa_k\}$. This classification can also be described in terms of the latent structure model. Assume that there is a latent (unobservable) variable Z with values in $\{1, \ldots, m\}$. Let us suppose that each of the individuals is known to belong to one of m latent classes and, conditionally on knowing the exact class j, $j = 1, \ldots, m$, the 3 observed variables are mutually independent. Then latent class structure explains relationships among the categorical variables that we observe

through the contingency table.

For more detailed explanation, some algebraic notation is needed. For $k = 1, 2, 3$, let A_k be a matrix of size $m \times \kappa_k$, with $\mathbf{a}_j^k = (a_j^k(1), \ldots, a_j^k(\kappa_k))$ being the jth row of A_k Later, we will see that $a_j^k(\ell)$ is the probability that the kth variable is in the ℓth state, conditional on the observation coming from the jth mixture component. Let $A_1 \times A_2 \times A_3$ be the $\kappa_1 \times \kappa_2 \times \kappa_3$ tensor defined by

$$[A_1, A_2, A_3] = \sum_{j=1}^{m} \mathbf{a}_j^1 \bigotimes \mathbf{a}_j^2 \bigotimes \mathbf{a}_j^3. \tag{4}$$

Using simpler language, the tensor $[A_1, A_2, A_3]$ is a three-dimensional array whose element with coordinates (u, v, w) is a sum of products of elements of matrices A_k, $k = 1, 2, 3$, with column numbers u, v, and w, respectively, added up over all of the m rows:

$$[A_1, A_2, A_3]_{u,v,w} = \sum_{j=1}^{m} a_j^1(u) a_j^2(v) a_j^3(w).$$

Such a tensor describes exactly the probability distribution in a finite latent-class model with three observed variables. To see why this is the case, imagine that there is some latent variable Z that takes positive integer values from 1 to some $m > 1$ and each of the n individuals belongs to one of m latent classes. If the 3 observed variables are mutually independent when the specific latent class j, $1 \leq j \leq m$, is known, we have a mixture of m components with each component being a product of finite measures and probabilities $\lambda_j \stackrel{\text{def}}{=} P(Z = j)$, $j = 1, \ldots, m$ being the mixing probabilities. Now, let the jth row of the matrix A_k be the vector of probabilities of the kth variable conditioned on belonging to jth class $\mathbf{p}_{jk} = P(X_k = \cdot \mid Z = j)$.

Choose one of the three matrices (say, A_1) and define $\tilde{A}_1 = \text{diag}(\boldsymbol{\lambda})A_1$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ is a vector describing the distribution of the latent class variable Z. Then, the (u, v, w) element of the tensor $[\tilde{A}_1, A_2, A_3]$ is the unconditional probability $P(X_1 = u, X_2 = v, X_3 = w)$ and, therefore, the joint probability distribution in such a model is exactly described by the tensor (4).

Define the Kruskal rank of a matrix A, $rank_K A$, as the largest number I of rows such that every set of I rows of A is independent. The following result was established by Kruskal in the mid-1970s.

Theorem 1. *Let $I_k = rank_K A_k$. If*

$$I_1 + I_2 + I_3 \geq 2m + 2,$$

then $[A_1, A_2, A_3]$ uniquely determines the A_j, up to simultaneous permutation and rescaling of rows.

Kruskal's result is very general and is a cornerstone of several subsequent results establishing identifiability criteria for various latent structure models with multiple observed variables. The one that follows most directly is the identifiability result of finite mixtures of finite measure products. Mixtures of that type have been widely used to model data in biological taxonomy, medical diagnosis or classification of text documents (for some practical examples, see Glick, 1973; Nigam et al., 2000). It was understood long ago that finite mixtures of Bernoulli products are not identifiable in a strict sense (see Gyllenberg et al., 1994); however, these mixtures are known to be well behaved in practice with respect to statistical parameter inference (see,

for example, Carreira-Perpiñán and Renals, 2000). Allman et al. (2009) explained this seeming contradiction by providing exact sufficient conditions for *generic* identifiability of these mixtures, up to the label swapping. *Generic* identifiability here is understood to mean identifiability on the entire parameter set except a subset of Lebesgue measure zero. The subset can be precisely described using terminology from algebraic geometry. For more details, see Allman et al. (2009).

Models that can also be viewed from the same latent structure viewpoint include random graph mixture models, hidden Markov models, and finite mixtures of nonparametric measure products. An important contribution of Allman et al. (2009) is that, for the first time, all of these various latent class models have been shown to be generically identifiable and that all of these identifiability results are derived using just one fundamental result from algebraic geometry—Kruskal's theorem 1.

Let us recall that we are specifically interested in finite mixtures of nonparametric measure products. We consider a nonparametric model of finite mixtures of m probability distributions. Each distribution is specified as a measure μ_j on R^r, $1 \leq j \leq m$. Assume that the dimensionality r (the number of classification variables) is at least 3. The kth marginal of μ_j is denoted μ_j^k. As before, let Z be the variable defining the latent structure of the model with values in $\{1, \ldots, m\}$ and $P(Z = j) = \lambda_j$ for any $j = 1, \ldots, m$. Then, the mixture model becomes

$$\mathcal{P} = \sum_{j=1}^{m} \lambda_j \mu_j = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} \mu_j^k. \tag{5}$$

This model implies that the r variates are, yet again, independent condi-

tional on a latent structure. The next theorem can be proved by using cut points to discretize the continuous distribution described by the measure $\mathcal{P}$ and using Kruskal's theorem. The details can be found in Allman et al. (2009).

Theorem 2. *Let $\mathcal{P}$ be a mixture of nonparametric measure products as defined in (5) and, for every variate $k \in \{1, \ldots, r\}$, the marginal measures $\{\mu_j^k\}_{1 \le j \le m}$ are linearly independent in the sense that the corresponding (univariate) distribution functions satisfy no nontrivial linear relationship. Then, if the number of variates $r \ge 3$, the parameters $\{\lambda_j, \mu_j^k\}_{1 \le j \le m, 1 \le k \le r}$ are uniquely identifiable from $\mathcal{P}$, up to label swapping.*

3 The algorithm and its extension

3.1 Notational conventions

Let Ω be a compact subset of R^r and define the linear vector function space

$$\mathcal{F} = \{\mathbf{f} = (f_1, \ldots, f_m)^\top : 0 < f_j \in L_1(\Omega), \log f_j \in L_1(\Omega), j = 1, \ldots, m\}.$$

Take $K(\cdot)$ to be a kernel density function on the real line and, with a slight abuse of notation, define the product kernel function $K(\mathbf{u}) = \prod_{k=1}^r K(u_k)$. For a row-vector $\mathbf{h} = (h_1, \ldots, h_r)$, define the rescaled version of K by $K_{\mathbf{h}}(\mathbf{u}) = \prod_{k=1}^r h_k^{-1} K(h_k^{-1} u_k)$. For $f \in L_1(\Omega)$, the smoothing operator $\mathcal{S}_{\mathbf{h}}$ is defined by

$$\mathcal{S}_{\mathbf{h}} f(\mathbf{x}) \;=\; \int_\Omega K_{\mathbf{h}}(\mathbf{x} - \mathbf{u}) f(\mathbf{u})\, d\mathbf{u}$$

and its corresponding nonlinear operator $\mathcal{N}_{\mathbf{h}}$ by

$$\mathcal{N}_{\mathbf{h}} f(\mathbf{x}) \;=\; \exp\{(\mathcal{S}_{\mathbf{h}} \log f)(\mathbf{x})\} \;=\; \exp \int_\Omega K_{\mathbf{h}}(\mathbf{x} - \mathbf{u}) \log f(\mathbf{u})\, d\mathbf{u}.$$

This $\mathcal{N}_{\mathbf{h}}$ operator is strictly concave (Eggermont, 1999, Lemma 3.1) and also multiplicative in the sense that $\mathcal{N}_{\mathbf{h}} f_j = \prod_k \mathcal{N}_{h_k} f_{jk}$ for f_j defined as in Equation (1). Letting H denote the $m \times r$ bandwidth matrix $(\mathbf{h}_1^\top, \ldots, \mathbf{h}_m^\top)^\top$, we may extend $\mathcal{S}$ (or $\mathcal{N}$) to $\mathcal{F}$ by defining $\mathcal{S}_H \mathbf{f} = (\mathcal{S}_{\mathbf{h}_1} f_1, \ldots, \mathcal{S}_{\mathbf{h}_m} f_m)^\top$.

Define the finite mixture operator

$$\mathcal{M}_{\boldsymbol{\lambda}} \mathbf{f}(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}),$$

whence we also obtain $\mathcal{M}_{\boldsymbol{\lambda}} \mathbf{f}(\mathbf{x}) = g_{\boldsymbol{\theta}}(\mathbf{x})$ as in Equation (2), and

$$\mathcal{M}_{\boldsymbol{\lambda}} \mathcal{N}_H \mathbf{f}(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{j=1}^{m} \lambda_j \mathcal{N}_{\mathbf{h}_j} f_j(\mathbf{x}).$$

3.2 The Descent Property

Let $g(\mathbf{x})$ now represent a known target density function. Following Levine et al. (2010), we define the functional

$$\ell_H(\boldsymbol{\theta}, g) = \int_\Omega g(\mathbf{x}) \log \frac{g(\mathbf{x})}{[\mathcal{M}_{\boldsymbol{\lambda}} \mathcal{N}_H \mathbf{f}](\mathbf{x})} \, d\mathbf{x}, \tag{6}$$

which can be viewed as a penalized Kullback-Leibler distance between $g(\mathbf{x})$ and $(\mathcal{M}_{\boldsymbol{\lambda}} N_H \mathbf{f})(\mathbf{x})$. Letting $\boldsymbol{\theta}^0 = (\boldsymbol{\lambda}^0, \mathbf{f}^0)$, define

$$\hat{f}_{jk}(u) \;=\; \alpha_{jk} \int K_{h_{jk}}(x_k - u) g(\mathbf{x}) w_j^0(\mathbf{x}) \, d\mathbf{x}, \tag{7}$$

where

$$w_j^0(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\lambda_j^0 \mathcal{N}_{\mathbf{h}_j} f_j^0(\mathbf{x})}{\mathcal{M}_{\boldsymbol{\lambda}^0} \mathcal{N}_H \mathbf{f}^0(\mathbf{x})}, \tag{8}$$

which implies $\sum_{j=1}^m w_j^0(\mathbf{x}) = 1$, and α_{jk} is a constant chosen so that $\int \hat{f}_{jk}(u) \, du = 1$. Furthermore, let

$$\hat{\lambda}_j = \frac{\int g(\mathbf{x}) w_j^0(\mathbf{x}) \, d\mathbf{x}}{\sum_{a=1}^m \int g(\mathbf{x}) w_a^0(\mathbf{x}) \, d\mathbf{x}} = \int g(\mathbf{x}) w_j^0(\mathbf{x}) \, d\mathbf{x}. \tag{9}$$

The newly updated $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\lambda}}, \hat{\mathbf{f}})$ then satisfies the following "descent property":

$$\ell_H(\hat{\boldsymbol{\theta}}, g) \;\leq\; \ell_H(\boldsymbol{\theta}^0, g). \tag{10}$$

This fact relies on a so-called MM algorithm, which stands for majorization-minimization algorithm, and its proof follows the proof of an analogous result in Levine et al. (2010) almost exactly except for the presence of the different bandwidth values H. For a general introduction to MM algorithms, which generalize the well-known class of iterative maximum likelihood algorithms known as EM algorithms, see Hunter and Lange (2004).

3.3 Estimation of Parameters

We now assume that we observe a simple random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ distributed according to some r-dimensional density $g(\mathbf{x})$. One may posit that $g \equiv g_{\boldsymbol{\vartheta}}$, where $\boldsymbol{\vartheta}$ represents the "true" parameter values and $g_{\boldsymbol{\vartheta}}$ is defined as in Equation (2), or one may instead take the view that the truth is not contained in our model class and that the goal of estimation is merely to minimize the criterion function $\ell_H(\boldsymbol{\theta})$, thereby finding in some sense a "best" vector $\boldsymbol{\theta}$ to approximate the truth by a density of the form (2). Since we do not discuss any notion of consistency in the current article, either point of view will work here.

Letting $\tilde{G}_n(\cdot)$ denote the empirical distribution function of the sample and ignoring the term $\int g(\mathbf{x}) \log g(\mathbf{x})\, d\mathbf{x}$ that does not involve any parame-

ters, a discrete version of (6) is

$$\begin{aligned} \ell_H(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \int \log \frac{1}{[\mathcal{M}_{\boldsymbol{\lambda}}\mathcal{N}_H\mathbf{f}](\mathbf{x})}\, d\tilde{G}_n(\mathbf{x}) \\ &= -\sum_{i=1}^{n} \log\left\{[\mathcal{M}_{\boldsymbol{\lambda}}\mathcal{N}_H\mathbf{f}](\mathbf{x}_i)\right\}. \end{aligned} \tag{11}$$

For the sake of notational simplicity, we drop the explicit dependence of ℓ_H on $\tilde{G}_n(\cdot)$ here; we trust that this re-definition of ℓ_H will not cause confusion, as it is essentially the same function as in Equation (6). In its new form, Equation (11), it resembles a loglikelihood function except for the presence of the nonlinear smoothing operator $\mathcal{N}_H$ and the fact that with the negative sign preceding the sum, our goal is minimization rather than maximization.

Here, we recall the maximum smoothed likelihood (MSL) algorithm from Levine et al. (2010): In that algorithm, it is possible to fix some of the coordinates in the $\mathbf{x}$ vectors to be identically distributed, in additional to being conditionally independent. We say that groups of conditionally independent and identically distributed coordinates belong to the same "block". Let b_k denote the block index of the kth coordinate, where $1 \leq b_k \leq B$ and B is the total number of such blocks, so that the model is

$$g_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_{ik}). \tag{12}$$

A simplification is possible when $b_k = k$ for all k, whereby (12) becomes (2). Assuming model (12) and letting $h_{j\ell}$ be the bandwidth used in the jth component and the ℓth block, the objective function of Equation (11) may be written

$$\ell_H(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j \exp\left\{\sum_{k=1}^{r} \int K_{h_{jb_k}}(x_{ik} - u) \log f_{jb_k}(u)\, du\right\}. \tag{13}$$

With initial parameter values $\boldsymbol{\theta}^0 = (\mathbf{f}^0, \boldsymbol{\lambda}^0)$, our modified MSL algorithm iterates the following steps for $t = 0, 1, \ldots$:

- **Majorization step:** Define, for each i and j,

$$w_{ij}^t = \frac{\lambda_j^t \mathcal{N}_{\mathbf{h}_j} f_j^t(\mathbf{x}_i)}{\sum_{a=1}^m \lambda_a^t \mathcal{N}_{\mathbf{h}_a} f_a^t(\mathbf{x}_i)} = \frac{\lambda_j^t \prod_{k=1}^r \mathcal{N}_{h_{jb_k}} f_{jb_k}^t(x_{ik})}{\sum_{a=1}^m \lambda_a^t \prod_{k=1}^r \mathcal{N}_{h_{ab_k}} f_{ab_k}^t(x_{ik})}. \tag{14}$$

- **Minimization step, part 1:** Set, for $j = 1, \ldots, m$,

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n w_{ij}^t \tag{15}$$

- **Minimization step, part 2:** For each component j and block $\ell \in \{1, \ldots, B\}$, let

$$f_{j\ell}^{t+1}(u) = \frac{1}{n h_{j\ell} \lambda_j^{t+1} C_\ell} \sum_{k=1}^r \sum_{i=1}^n w_{ij}^t I_{\{b_k = \ell\}} K\left(\frac{u - x_{ik}}{h_{j\ell}}\right), \tag{16}$$

where $C_\ell = \sum_{k=1}^r I_{\{b_k=\ell\}}$ is the number of coordinates in the ℓth block, and $h_{j\ell}$ is the bandwidth for the kernel density estimate corresponding to the ℓth block in the jth component. It appears at first glance that the bandwidths $h_{j\ell}$ in the second M-step (16) need not be the same as those in the E-step (14). However, in order to prove that our new algorithm retains the desirable descent property, we require an analogue of Equation (7), which means that these bandwidths must indeed match. We demonstrate in the Appendix how to adapt a method of proof given by Levine et al. (2010) to show that $\ell_H(\boldsymbol{\theta}^t)$ is nonincreasing in t using the algorithm in this section. In other words, equations (14) through (16) ensure that

$$\ell_H(\boldsymbol{\theta}^{t+1}) \leq \ell_H(\boldsymbol{\theta}^t). \tag{17}$$

3.4 Bandwidth Selection

As discussed in Benaglia et al. (2009a) in the case of the similar npEM algorithm, the selection of a bandwidth in a mixture setting like (12) can be an intricate problem, and there are several reasons for which using a single, fixed bandwidth as in (16) is not always appropriate. An iterative bandwidth scheme adapting the well-known rule of Silverman (Silverman, 1986, p. 46) has been proposed in Benaglia et al. (2010) for the npEM algorithm. Briefly, it amounts to replacing, in Silverman's rule

$$h = 0.9 \min \left\{ \mathrm{SD}, \frac{\mathrm{IQR}}{1.34} \right\} n^{-1/5} \tag{18}$$

for a simple random sample, the sample size (n), interquartile range (IQR) and standard deviation (SD) by corresponding block- and component-wise versions. These estimates are to be iteratively defined using the posterior probabilities. This scheme can be applied straightforwardly in the npMSL algorithm and gives estimated bandwiths at $(t+1)$th iteration,

$$h_{j\ell}^{t+1} = 0.9 \min \left\{ \sigma_{j\ell}^{t+1}, \frac{IQR_{j\ell}^{t+1}}{1.34} \right\} (nC_\ell \lambda_j^{t+1})^{-1/5}, \tag{19}$$

where $nC_\ell \lambda_j^{t+1}$ estimates the sample size for the ℓth block of coordinates in the jth component, and $\sigma_{j\ell}^{t+1}$ and $IQR_{j\ell}^{t+1}$ are the weighted standard deviation and empirical interquartile range for the jth component and ℓth block, as introduced in Benaglia et al. (2010), but using here the w_{ij}^t to weight the data.

However, the major difference between the npEM algorithm and our MSL algorithm is that the latter satisfies a descent property when the bandwidths $h_{j\ell}$ are fixed throughout. It remains an open question whether there

is any sort of descent property that is satisfied by a modified MSL in which the bandwidths are iteratively updated. A deeper question is whether there is some sense in which the iteratively updated $h_{j\ell}$ converge in some sense to, say, the "oracle" bandwidths that would result if somehow the true parameter vector $\boldsymbol{\vartheta}$ were known and the modified Silverman rule (19) were applied using the true parameter values. We do not tackle these difficult questions in the current article.

Nonetheless, it is possible in theory to implement a two-stage algorithm in which the bandwidths are allowed to change for several iterations (until a reasonable estimate of the mixture structure and thus the set of bandwidths is achieved), then the bandwidths are fixed and the algorithm allowed to converge. Such a scheme allows for both a reasonable set of bandwidth estimates and the guaranteed descent property beginning from the point at which the bandwidths are fixed. In practice, however, note that this is no different from a scheme in which the first stage is allowed to run until some convergence criterion is satisfied, since fixing the bandwidths at that stage and continuing to run the algorithm does not result in any further changes because the algorithm has already achieved convergence according to the original criterion. The downside to such a scheme is that the inability to verify any descent property removes one possible method to check that the algorithm is coded correctly. In our numerical examples in Section 5, we find that the scheme appears to work well.

4 Extensions of the estimation algorithm

Here, we discuss two extensions of the basic idea of the algorithm of Section 3.3 to situations distinct from, but related to, model (2). The first is a univariate case in which a more stringent assumption is required for identifiability. The second is multivariate but with an assumption that the components and/or the coordinates differ only by a location or a scale parameter. Proofs of the descent properties of the algorithms in this section are given in the Appendix.

4.1 The Univariate Symmetric Location Model

Both Bordes et al. (2006) and Hunter et al. (2007) showed that, for univariate

$$X \sim \sum_{j=1}^{m} \lambda_j f(x - \mu_j), \tag{20}$$

where each λ_j is positive, all μ_j are distinct, $\sum_j \lambda_j = 1$, and f is some density function on $\mathbb{R}$ that is symmetric about zero, the parameters $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and f are uniquely identifiable when $m = 2$ (up to label-switching) from the density of X as long as $\lambda_1 \neq 1/2$. Furthermore, Hunter et al. (2007) showed that for $m = 3$, the parameters are uniquely identifiable except when $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ take values in a particular set of Lebesgue measure zero, conjecturing that a similar result may be shown for general m. We will assume here that f is absolutely continuous with respect to Lebesgue measure, though this assumption is not necessary for the above identifiability results to hold.

Although both Bordes et al. (2006) and Hunter et al. (2007) proposed methods for estimating the parameters in (20) given a simple random sample $x_1, \ldots, x_n$ distributed according to (20), these methods were inefficient

and were not easily generalizable beyond the case $m = 2$. Later, Bordes et al. (2007) proposed a stochastic EM-like estimation algorithm that is easily generalizable to any m; however, this algorithm does not possess the descent property of a typical EM algorithm. Here, we discuss an estimation algorithm that does guarantee a descent property.

Given a bandwidth h, together with initial parameter values $\boldsymbol{\theta}^0 = (\mathbf{f}^0, \boldsymbol{\lambda}^0, \boldsymbol{\mu}^0)$, iterate the following steps for $t = 0, 1, \ldots$:

- **Majorization step:** Define, for each i and j,

$$w_{ij}^t = \frac{\lambda_j^t \mathcal{N}_h f^t(x_i - \mu_j^t)}{\sum_{a=1}^m \lambda_a \mathcal{N}_h f^t(x_i - \mu_a^t)} \tag{21}$$

- **Minimization step, part 1:** Set, for $j = 1, \ldots, m$,

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n w_{ij}^t \tag{22}$$

- **Minimization step, part 2:** For any $u \in \mathbb{R}$, let

$$\begin{aligned} f^{t+1}(u) &= \frac{1}{2nh\lambda_j^{t+1}} \sum_{j=1}^m \sum_{i=1}^n w_{ij}^t \left[K\left(\frac{x_i - \mu_j^t - u}{h}\right) \right. \\ &\qquad \left. + K\left(\frac{x_i - \mu_j^t + u}{h}\right) \right]. \end{aligned} \tag{23}$$

- **Minimization step, part 3:** For $j = 1, \ldots, m$, let

$$\mu_j^{t+1} = \arg\max_{\mu} \int \sum_{i=1}^n w_{ij}^t K\left(\frac{x_i - u}{h}\right) \log f^{t+1}(u - \mu)\, du. \tag{24}$$

Equation (23) assures that $f(u) = f(-u)$, which is required due to the symmetry assumption. This algorithm guarantees that $\ell_h(\boldsymbol{\theta}^t)$ of Equation (11) is nonincreasing in t, where in this model we may express this

objective function in the form

$$\ell_h(\boldsymbol{\theta}^t) \quad = \quad -\sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j^t [\mathcal{N}_h f^t](x_i - \mu_j^t). \tag{25}$$

In other words, this algorithm has a provable descent property. However, the "minimization" step in this algorithm is slightly misnamed, since parts 1 through 3 do not result in a global minimization of the majorizing function. Instead, as verified in the Appendix, part 2 minimizes *only* as a function of f, while holding $\boldsymbol{\mu}$ fixed at $\boldsymbol{\mu}^t$. Then part 3 minimizes as a function of $\boldsymbol{\mu}$, while holding f fixed at f^{t+1}. Thus, each of these parts results in a lowering of the value of the majorizing function, which in turn guarantees a decrease in $\ell_h(\boldsymbol{\theta})$. It is a small drawback that the maximization of Equation (24) must be accomplished numerically, but since this is merely a one-dimensional maximization for each j, it can easily be accomplished as long as the integral in Equation (24) is inexpensive to calculate for a given μ.

One could modify the above algorithm by alternating between iterations that implement only parts 1 and 2 and iterations that implement only parts 1 and 3 of the maximization step. Because this idea holds part of the parameter vector fixed at each iteration and optimizes only with respect to the rest of the parameters, it produces something that might be called an MCM (majorization-conditional maximization) algorithm, analogous to the ECM (expectation conditional maximization) algorithm of Meng and Rubin (1993).

4.2 The location-scale model

It is not difficult to restrict model (2) somewhat while still retaining the essential nonparametric character of the estimation: We may assume that the various univariate density functions in Equation (2) have the same shape, not assumed to follow any parametric form, but that they differ from one another in a parametric way. There are various ways in which this may be accomplished. For example, Qin and Leung (2006) propose an "exponential tilt" idea in which the ratio of one component's density functions to another's has a specific parametric form, namely, $\log[f_{2k}(x)/f_{1k}(x)]$ is a quadratic function of x for each k. (They consider only the case $m = 2$ and $r = 3$.)

By contrast, we assume here, as in Benaglia et al. (2009a), that

$$f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f_j \left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right) \tag{26}$$

for unknown parameters $(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, f_j)$, $j = 1, \ldots, m$, we are assuming that the coordinates within each individual have the same shape of distribution (depending on the individual's mixture component) but may differ by a location and scale factor. One may restrict the model of Equation (26) even further by assuming that all $\boldsymbol{\mu}_j$ or all $\boldsymbol{\sigma}_j$ are the same, in which case we have either a scale-only or a location-only model, respectively. Alternatively, we may assume that

$$f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f_\ell \left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right), \tag{27}$$

in which case the individual differences, i.e., the mixture components, only account for differences up to a location and scale parameter, but otherwise the distributions of different blocks of coordinates do not relate to one another in any way. Equation (26) differs from Equation (27) by only a single

subscript on the density f, yet the interpretations of the two models are quite different.

As a special case of both (26) and (27), if all f_{jk} are assumed to have the same shape, then we may require that

$$f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f\left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}}\right) \tag{28}$$

for a single unspecified density function $f(\cdot)$.

Because f_j in equation (26) is completely unspecified, the location and scale parameters may be absorbed into f_j, so the parameters are not uniquely identiable even if each $f_{j\ell}$ is known. Therefore, one may assume some additional constraints on the $\mu_{j\ell}$ and $\sigma_{j\ell}$, such as $\sum_\ell \mu_{j\ell} = 0$ and $\sum_\ell \sigma_{j\ell} = 1$. In practice, however, it is typically not necessary to enforce these constraints. Similar arguments can be made for the parameters in equations (27) and (28).

Employing the block structure of Equation (12) instead of the less general Equation (2), we may modify the algorithm of Section 3.3. Equations (14) and (15) remain unchanged, but we must modify Equation (16) to either

$$f_j^{t+1}(u) = \frac{1}{nrh_j\lambda_j^{t+1}} \sum_{k=1}^{r}\sum_{i=1}^{n} w_{ij}^t K\left(\frac{u - x_{ik} + \mu_{jb_k}^t}{h_j\sigma_{jb_k}^t}\right) \tag{29}$$

or

$$f_\ell^{t+1}(u) = \frac{1}{nh_\ell\lambda_j^{t+1}C_\ell} \sum_{k=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^t I_{\{b_k=\ell\}} K\left(\frac{u - x_{ik} + \mu_{jb_k}^t}{h_\ell\sigma_{jb_k}^t}\right), \tag{30}$$

where $C_\ell = \sum_{k=1}^r I_{\{b_k=\ell\}}$, depending upon whether we take Equation (26) or Equation (27) as our assumption. In addition, the updates to the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ parameters would take place in a separate part of the minimization

step, as in Equation (24). The resulting algorithm would be similar to the one described in Section 20: It is not an MM algorithm exactly, but it is very similar and most importantly it guarantees a decrease in the desired objective function (13).

5 Numerical examples

5.1 A synthetic example

To illustrate the iterative and block- and component-specific bandwidths, we choose first a simulated example with heavy-tailed distributions and different scales among the coordinates. The model is multivariate with $r = 5$ repeated measures grouped into $B = 2$ blocks of sizes 3 and 2 ($b_1 = b_2 = b_3 = 1$ and $b_4 = b_5 = 2$) and $m = 2$ components. Block 1 corresponds to a mixture of two noncentral Student t distributions, $t(2, 0)$ and $t(10, 4)$, where the first parameter is the number of degrees of freedom and the second is the noncentrality. Block 2 corresponds to a mixture of Beta distributions, $\mathcal{B}(1, 1)$ (which is actually the uniform distribution over $[0, 1]$) and $\mathcal{B}(1, 5)$. The first component weight is $\lambda_1 = 0.4$. For this example, in which the coordinate densities are on different scales, it is obvious that the bandwidth should depend on the blocks and components.

A simple run of the original npMSL algorithm results in a single fixed bandwith $h = 0.527$, while a run with the above scheme gives the following (final) bandwidth matrix:

	component 1	component 2
block 1	0.316	0.343
block 2	0.089	0.038

The estimates of the component and block densities are shown in Figure 1.

In that figure, we see that the original npMSL algorithm gives a nearly identical estimate of the densities in the first block, but the estimates are dramatically different in the second block.

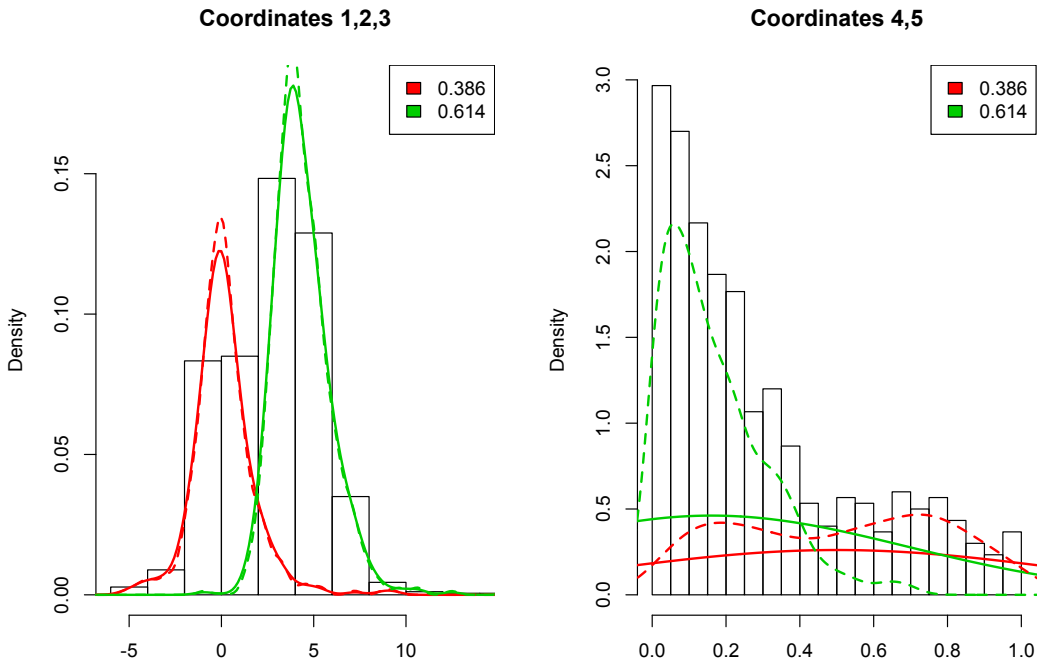

Figure 1: *The two colors designate the two components; the solid and dashed lines are the fixed-single-bandwidth and adaptive-multiple bandwidth npMSL solutions, respectively.*

Remark: The choice of the Gaussian kernel K in Figure 1 may explain the "leaking" of mass seen at the edges of the second block's density estimates. Though choice of kernel function is not generally very influential, a different choice such as a triangle density might prevent such leakage. Studying such a boundary correction could be the subject of future work.

5.2 Water-level dataset

As an illustration of the adaptive block- and component-wise bandwidth approach, we consider a benchmark dataset which has previously been analyzed by Hettmansperger and Thomas (2000) and Elmore et al. (2004) with a conditionally i.i.d. (independent and identically distributed) assumption, and more recently by Benaglia et al. (2009a) and Levine et al. (2010) under the same assumptions we make here. This experiment involves $n = 405$ children aged 11 to 16 years subjected to a written test as initially described by Thomas et al. (1993). In this test, each child is presented with eight rectangular drawings of a vessel on a sheet of paper, each tilted to one of $r = 8$ clock-hour orientations: 11, 4, 2, 7, 10, 5, 1, and 8 o'clock, in order of presentation to the subjects. The children's task was to draw a line representing the surface of still liquid in the closed, tilted vessel in each picture. The acute angle, in degrees, formed between the horizontal and this line was measured for each response, the associated sign being the sign of the slope of the line. The water-level dataset is available in the mixtools package (Young et al., 2009; Benaglia et al., 2009b).

As in Benaglia et al. (2009a) and Levine et al. (2010), it seems reasonable to weaken the conditionally i.i.d. assumption, assuming instead that only opposite clock-face orientations lead to conditionally independent and identically distributed responses, so that the eight coordinates may be organized into four blocks of two each, which is model (12) with B=4. According to the ordering of the clock-hour orientations, we thus define $\mathbf{b} = (4, 3, 2, 1, 3, 4, 1, 2)$. For instance, we see that $b_4 = b_7 = 1$, which means

block 1 relates to coordinates 4 and 7, corresponding to clock orientations 1:00 and 7:00.

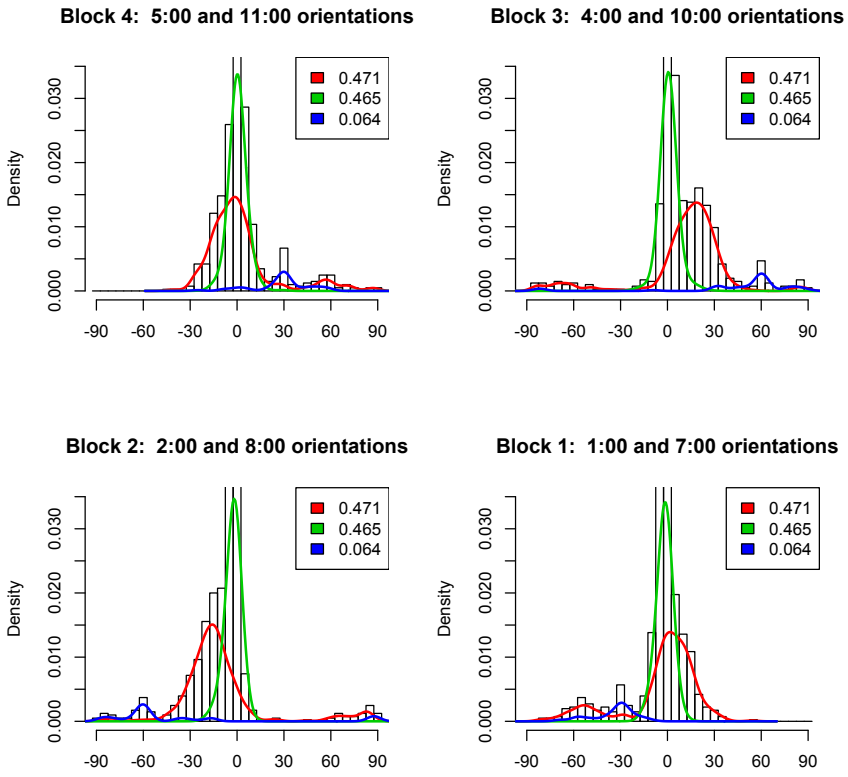

Figure 2: *The water-level data analyzed using the npMSL algorithm with* $m = 3$ *mixture components and a fixed bandwidth* $h = 4$.

We first consider here the $m = 3$-component model as studied in Levine et al. (2010) to compare the npMSL with fixed bandwidth against the adaptive strategy. Figure 2 gives the solution returned by the npMSL algorithm with a fixed bandwidth preset to $h = 4$, as in Benaglia et al. (2009a) and Levine et al. (2010). This value has been chosen by trial an error by these

authors, instead of allowing the algorithm compute a fixed bandwidth value using Silverman's rule as in (18). However, using that rule would result in a fixed bandwidth value of $h = 1.47$, and correspondingly more jagged component densities, but qualitatively the same overall solution. The interpretation of this solution is that component 2 (green) represents the 46.5% of the subjects who know how to do the task—the "competent group"—whereas component 3 (blue) represents the 6.4% of the subjects who always draw the line parallel to the vessel bottom. The first component (red, with 47%) is perhaps the most interesting: These subjects in the "slightly confused group" appear to perform the task nearly correctly for the more vertically oriented vessels (1, 5, 7, and 11 o'clock) but tend to allow the water level to slant somewhat with the vessel itself when the vessel is tipped to a more horizontal orientation.

Figure 3 gives the solution returned by the npMSL algorithm with the adaptive bandwidth given by (19). The corresponding bandwidth matrix is displayed in Table 1, which shows that the bandwith differences are mostly between components.

Table 1: Adaptive bandwidths per block and components for the Water level data, at the npMSL last iteration.

	component 1	component 2	component 3
block 1	12.17	1.46	0.975
block 2	14.0	2.74	2.276
block 3	19.19	2.55	2.276
block 4	12.36	1.28	1.63

The qualitative interpretation appears simpler here, since the competent

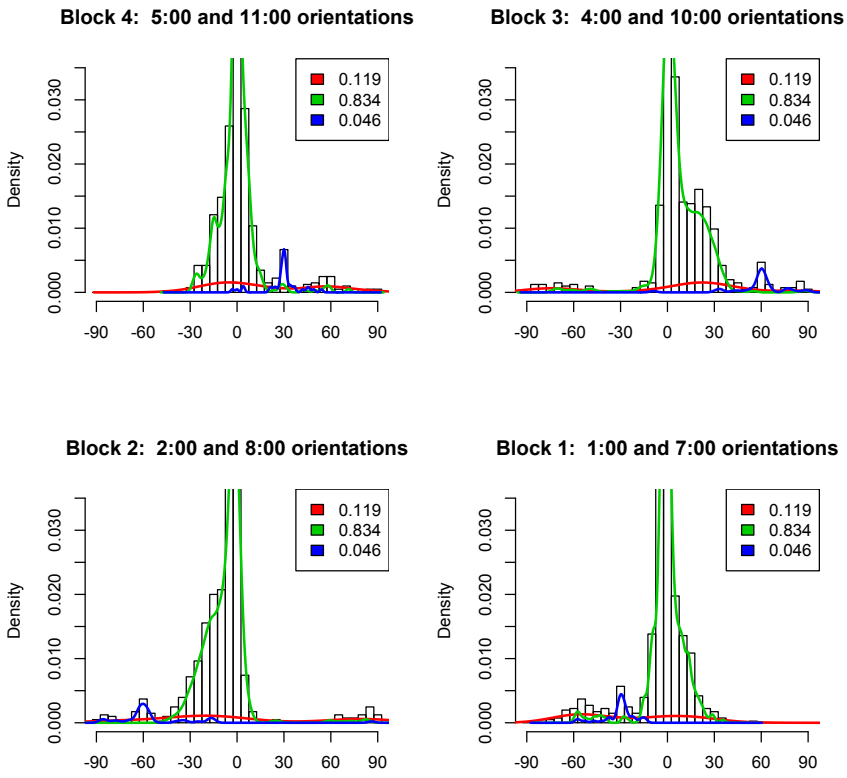

Figure 3: *The water-level data analyzed using the npMSL algorithm with* $m = 3$ *mixture components and adaptive bandwidths given in Table 1.*

group now represents 83% of the subjects (but seems to encompass most of the previous slightly confused group), while the group of subjects who always draw the line parallel to the vessel bottom lowers to 4.6%, with more clear peaks on ± 30 and ± 60 due to this component smaller bandwidths. An interesting fact is also that the first (red) component is far less important (12%) and appears to retain qualities of the previous slightly confused group but also includes some even stranger behavior that is close to uniform, or

"totally guessing." Hence in this example, allowing bandwidth to change adaptively results in a very different qualitative interpretation.

However, if we fit a $m = 4$ components model with the npMSL algorithm and adaptive bandwidth strategy, we identify all four previously mentioned groups. A typical result is in Fig. 4, and the final bandwidth matrix is omitted for brevity. The competent group represents again about 45% of the subjects, and is distinct from the 43% slightly confused group. The group who always draw the line parallel to the vessel bottom drops to 3.7% which is more in accordance with the result from Fig.3, and distinct from the 7% totally guessing group.

5.3 A psychometric data example

The data in this section come from a large-scale psychometrics study exploring cognitive task performances for children with specific language impairments, presented in Miller et al. (2001). Response (or Reaction) Times (RT) with which the children respond to a range of tasks are recorded in milliseconds. We focus in particular on one experiment that Miller et al. (2001) call a "simple RT task": The child is instructed to strike a key as quickly as possible in response to a visual signal, which itself is emitted after a delay following the word "ready" said by the child. There are 8 trials for each of three time delays of 1, 2 and 5 seconds. Tasks are mixed into a much longer sequence of trials so that the child does not know exactly what the next task would be, so that independence of the repeated measures for each child may reasonably be assumed. This dataset with $n = 82$ subjects and $r = 24$ coordinates is available in the mixtools package (Young et al., 2009;

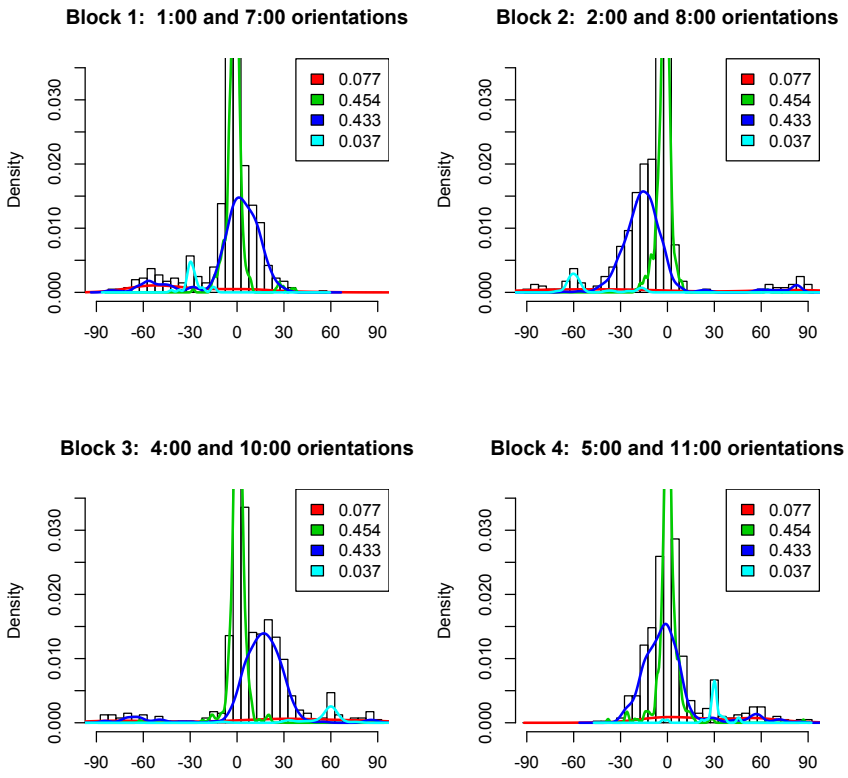

Figure 4: *The water-level data analyzed using the npMSL algorithm with* $m = 4$ *mixture components and adaptive bandwidths strategy.*

Benaglia et al., 2009b) for the R statistical software environment (R Development Core Team, 2010), and is loaded by the `data(RTdata2)` command.

This experiment supports a model with $B = 3$ blocks of 8 coordinates each, each block corresponding to a delay between the "ready" sign and the stimulus. This data set is interesting because it illustrates the potential interest of the conditional independence model for multivariate data with a large number of coordinates and block structure suggested by scientific

considerations.

We ran the npMSL algorithm with fixed and adaptive bandwidth strategies. Results in Fig. 5 show that there is almost no difference between the two, which is not surprising because the component densities have similar scaling properties. However, one can see that the third block, which corresponds to the longer delay of 5 seconds, shows densities slightly shifted to the right. We find that no matter what the delay is, we can essentially describe the two groups as a "faster group" and a "slower group", where the former represents 72% of the subjects.

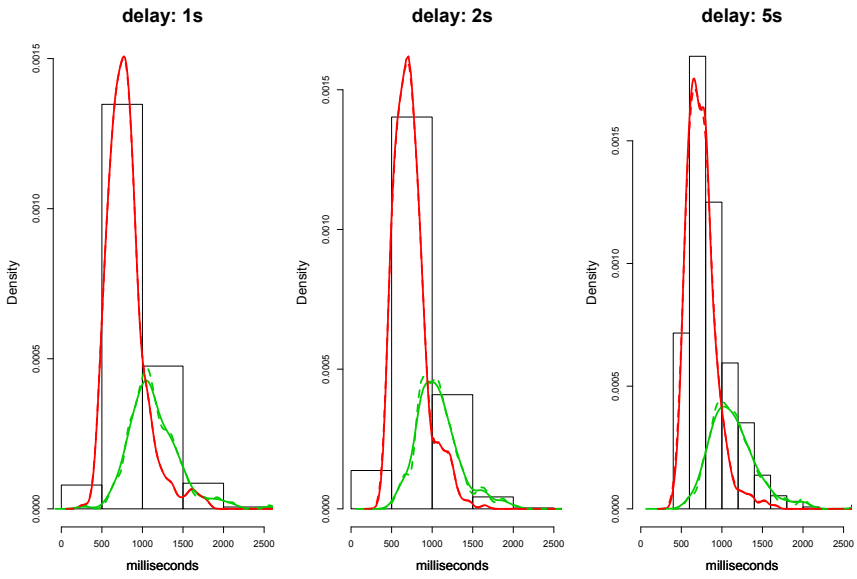

Figure 5: *Density estimates for the simple RT task with $B = 3$ blocks of 8 coordinates each, and $m = 2$ components: npMSL with fixed bandwidth (dashed line), and adaptive bandwidths (solid line). The component weights are* $(0.72, 0.28)$.

6 Discussion

This manuscript reviews the justification for the conditional independence assumption in multivariate finite mixture models and summarizes what is known about the identifiability of parameters in these models when no assumption is made about the parametric form of the component densities. In particular, we review the important results in Allman et al. (2009), who prove that conditional independence implies identifiability under weak assumptions as long as the multivariate observations have dimension at least three.

We review the npMSL algorithm of Levine et al. (2010) and introduce a method for selecting bandwidths, which is an important aspect of the practical implementation of this algorithm. In addition, we extend the idea of Levine et al. (2010) to the special cases of a univariate location mixture of symmetric components and a multivariate location-scale mixture. These special cases require a generalization of the notion of MM (majorization-minimization) algorithms since it is impossible to achieve a closed-form global minimization with respect to all parameters in the second "M" step. Finally, we give proofs of the descent properties of our algorithms when the bandwidths are held constant.

The important feature of the npMSL algorithm and the extension we introduce in the current article is that it is shown to minimize (at least locally) a particular objective function. This function may be considered a nonlinearly smoothed version of the nonparametric likelihood function. The fact that our estimators may be shown to optimize this function opens

the door for potential results on asymptotic properties of the algorithm, such as consistency and convergence rates. Such results appear much more difficult to establish for the similar npEM algorithm of Benaglia et al. (2009a, 2010) because that algorithm does not appear to optimize any type of a loglikelihood-like function despite its resemblance to an EM algorithm.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2010). Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger*.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51(11):5429–5443.

Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34(3):1204–1232.

Carreira-Perpiñán, M. Á. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12(1):141–152.

Cruz-Medina, I. R. and Hettmansperger, T. P. (2004). Nonparametric estimation in semi-parametric univariate mixture models. *J. Stat. Comput. Simul.*, 74(7):513–524.

Eggermont, P. P. B. (1999). Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. *Applied Mathematics and Optimization*, 39(1):75–91.

Elmore, R. and Wang, S. (2003). Identifiability and estimation in finite mixture models with multinomial components. Technical report, Department of Statistics, Pennsylvania State University.

Elmore, R. T., Hall, P., and Neeman, A. (2005). An application of classical invariant theory to identifiability in nonparametric mixtures. *Annales de l'institut Fourier*, 55(1):1–28.

Elmore, R. T., Hettmansperger, T. P., and Thomas, H. (2004). Estimating component cumulative distribution functions in finite mixture models. *Communications in Statistics. Theory and Methods*, 33(9):2075–2086.

Glick, N. (1973). Sample-based multinomial classification. *Biometrics*, 29(2):241–256.

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, pages 542–548.

Hall, P., Neeman, A., Pakyari, R., and Elmore, R. T. (2005). Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678.

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

Hettmansperger, T. P. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society, Series B*, 62(4):811–825.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30–37.

Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293.

Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Levine, M., Hunter, D. R., and Chauveau, D. (2010). Smoothed likelihood for multivariate mixtures. Technical Report 10–04, Penn State Department of Statistics.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267.

Miller, C. A., Kail, R., and Leonard, L. B. (2001). Speed of processing in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 44:416–433.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2):103–134.

Qin, J. and Leung, D. H.-Y. (2006). Semiparametric analysis in conditionally independent multivariate mixture models. unpublished manuscript.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Thomas, H., Lohaus, A., and Brainerd, C. J. (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development*, 58(9).

Young, D. S., Benaglia, T., Chauveau, D., Elmore, R. T., Hettmansperger, T. P., Hunter, D. R., Thomas, H., and Xuan, F. (2009). mixtools: Tools for mixture models. R package version 0.3.3.

A Proofs of descent properties

Recall throughout this section that the parameter vector $\boldsymbol{\theta}$ consists of the mixing weights $\boldsymbol{\lambda}$ and the univariate densities $f_{j\ell}$, $1 \leq j \leq m$ and $1 \leq \ell \leq B$. For a given (fixed) $\boldsymbol{\theta}^t$, let the constants w_{ij}^t be defined as in Equation (14). We first state and prove two lemmas, each based on the following definition:

$$b_H^t(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^t \log\left\{\lambda_j[\mathcal{N}_{\mathbf{h}_j} f_j](\mathbf{x}_i)\right\}. \tag{31}$$

Lemma 1. *Let $\ell_H(\boldsymbol{\theta})$ be defined as in Equation (11). Then*

$$\ell_H(\boldsymbol{\theta}) - \ell_H(\boldsymbol{\theta}^t) \;\leq\; b_H^t(\boldsymbol{\theta}) - b_H^t(\boldsymbol{\theta}^t). \tag{32}$$

Proof:

$$\begin{aligned}
\ell_H(\boldsymbol{\theta}) - \ell_H(\boldsymbol{\theta}^t) &= -\sum_{i=1}^{n} \log \sum_{j=1}^{m} \frac{\lambda_j[\mathcal{N}_{\mathbf{h}_j} f_j](\mathbf{x}_i)}{[\mathcal{M}_{\boldsymbol{\lambda}^t}\mathcal{N}_H \mathbf{f}^t](\mathbf{x}_i)} \\
&= -\sum_{i=1}^{n} \log \sum_{j=1}^{m} w_{ij}^t \frac{\lambda_j[\mathcal{N}_{\mathbf{h}_j} f_j](\mathbf{x}_i)}{\lambda_j^t[\mathcal{N}_{\mathbf{h}_j} f_j^t](\mathbf{x}_i)} \\
&\leq -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^t \log \frac{\lambda_j[\mathcal{N}_{\mathbf{h}_j} f_j](\mathbf{x}_i)}{\lambda_j^t[\mathcal{N}_{\mathbf{h}_j} f_j^t](\mathbf{x}_i)} \\
&= b_H^t(\boldsymbol{\theta}) - b_H^t(\boldsymbol{\theta}^t),
\end{aligned}$$

where the inequality follows from the convexity of the negative logarithm function and the fact that $\sum_j w_{ij}^t = 1$ for each i. □

Remark: In the terminology of MM algorithms (see, for example, Hunter and Lange, 2004), the result of Lemma 1 means that $b_H^t(\boldsymbol{\theta})$ is said to *majorize* $\ell_H(\boldsymbol{\theta})$ at the point $\boldsymbol{\theta} = \boldsymbol{\theta}^t$.

Lemma 2. *If $\boldsymbol{\theta}^{t+1} = (\boldsymbol{\lambda}^{t+1}, \mathbf{f}^{t+1})$, where λ_j^{t+1} and $f_{j\ell}^{t+1}$ are defined as in Equations (15) and (16), respectively, then $\boldsymbol{\theta}^{t+1}$ minimizes $b_H^t(\boldsymbol{\theta})$.*

Proof: As a function of $\boldsymbol{\lambda}$,

$$b_H^t(\boldsymbol{\theta}) = -\sum_{j=1}^{m} \log \lambda_j \left(\sum_{i=1}^{n} w_{ij}^t \right) + \text{something not involving } \boldsymbol{\lambda}.$$

Subject to the constraint $\sum_j \lambda_j = 1$, this is straightforward to minimize via a standard argument using a Lagrange multiplier. Since $\sum_i \sum_j w_{ij}^t = n$, Equation (15) gives the minimizer.

As a function of $f_{j\ell}$,

$$\begin{aligned} b_H^t(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} w_{ij}^t \sum_{k=1}^{r} I_{\{b_k=\ell\}} \log \left\{ [N_{h_{j\ell}} f_{j\ell}](x_{ik}) \right\} \\ &\quad + \text{something not involving } f_{j\ell}. \end{aligned} \tag{33}$$

The piece involving $f_{j\ell}$ may be rewritten

$$-\int \sum_{i=1}^{n} \sum_{k=1}^{r} w_{ij}^t I_{\{b_k=\ell\}} K_{h_{j\ell}}(x_{ik} - u) \log f_{j\ell}(u)\, du,$$

which is a constant times $-\int f_{j\ell}^{t+1}(u) \log f_{j\ell}(u)\, du$ if we define $f_{j\ell}^{t+1}$ as in Equation (16). However, this is merely the Kullback-Leibler divergence between $f_{j\ell}^{t+1}$ and $f_{j\ell}$ plus something not involving $f_{j\ell}$. We conclude that (33) is minimized for each j and ℓ by setting $f_{j\ell} = f_{j\ell}^{t+1}$. □

Putting the two lemmas together, we obtain the following:

Theorem 3. *Let $\ell_H(\boldsymbol{\theta})$ be defined as in Equation (11). Then the algorithm given in steps (14) through (16) imply the descent property (17).*

Proof: Since Lemma 2 implies in particular that $b_H^t(\boldsymbol{\theta}^{t+1)} \leq b_H^t(\boldsymbol{\theta}^t)$, Lemma 1 gives

$$\ell_H(\boldsymbol{\theta}^{t+1}) - \ell_H(\boldsymbol{\theta}^t) \;\leq\; b_H^t(\boldsymbol{\theta}^{t+1}) - b_H^t(\boldsymbol{\theta}^t) \;\leq\; 0.$$

□

Corollary 1. *Assuming Model (20), the algorithm described by Equations (21) through (24) guarantees that $\ell_h(\boldsymbol{\theta}^{t+1}) \leq \ell_h(\boldsymbol{\theta}^t)$, where $\ell_h(\boldsymbol{\theta})$ is defined in Equation (25).*

Proof: In this case, the observations $x_1, \ldots, x_n$ are not vector-valued (i.e., $r = 1$), so there is only a single block and we may drop the subscript ℓ wherever it occurs in Lemmas 1 and 2 and Theorem 3. Since Equation (11) is the same as Equation (25) for this special case, Lemmas 1 and 2 imply that the desired result holds whenever $b_h^t(\boldsymbol{\theta}^{t+1}) \leq b_h^t(\boldsymbol{\theta}^t)$, where $b_h^t(\boldsymbol{\theta})$ is the appropriately modified form of Equation (31). Using a simple change of variable together with the fact that $f(u) = f(-u)$, we may rewrite

$$\begin{aligned} \log\{[\mathcal{N}_h f_j](x_i)\} &= \int K_h(x_i - u) \log f(u - \mu_j)\, du \\ &= \frac{1}{2} \int [K_h(x_i - \mu_j - u) + K_h(x_i - \mu_j + u)] \log f(u)\, du. \end{aligned}$$

Thus, $b_h^t(\boldsymbol{\theta})$ becomes

$$-\int \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^t \left[K_h(x_i - \mu_j - u) + K_h(x_i - \mu_j + u)\right] \log f(u)\, du + \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^t \log \lambda_j.$$

Using the same argument as in Lemma 2, if $\boldsymbol{\mu}$ is fixed at $\boldsymbol{\mu}^t$, then $b_h^t(\boldsymbol{\theta})$ is minimized as a function of $\boldsymbol{\lambda}$ and f only by $\boldsymbol{\lambda}^{t+1}$ and f^{t+1} of Equations (22) and (23). Then, Equation (24) can only ensure a further decrease in the value of $b_h^t(\boldsymbol{\theta})$ when f is fixed at f^{t+1}. □

Remark: Similar reasoning to that used in the preceding proof, but without the extra step required because of the symmetry of f in that proof, demonstrates that the algorithms described in Section 4.2 also guarantee the descent properties as claimed in that section.