Statistics Surveys
Vol. 9 (2015) 1–31
ISSN: 1935-7516
DOI: 10.1214/15-SS108

Semi-parametric estimation for conditional independence multivariate finite mixture models

Didier Chauveau

Université d'Orléans, CNRS, MAPMO, UMR 7349, Orléans, France
e-mail: didier.chauveau@univ-orleans.fr

David R. Hunter

Department of Statistics, Pennsylvania State University, University Park PA 16802, USA
e-mail: dhunter@stat.psu.edu

and

Michael Levine

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
e-mail: mlevins@purdue.edu

Abstract: The conditional independence assumption for nonparametric multivariate finite mixture models, a weaker form of the well-known conditional independence assumption for random effects models for longitudinal data, is the subject of an increasing number of theoretical and algorithmic developments in the statistical literature. After presenting a survey of this literature, including an in-depth discussion of the all-important 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 in three or more dimensions. It possesses a descent property and can be easily adapted to situations where the data are 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 give a bandwidth selection procedure for our algorithm. Finally, we demonstrate the effectiveness of our algorithm using a simulation study and two psychometric datasets.

MSC 2010 subject classifications: Primary 62G05; secondary 62G07.
Keywords and phrases: Kernel density estimation, MM algorithms.

Received December 2013.

1. Introduction

The analysis of longitudinal data generally involves multivariate observations for each subject in which the correlation among observations for a given subject must be taken into account. A common method for modeling this situation is the so-called "conditional-independence model" [29], 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:

$$f_j(\mathbf{x}) = \prod_{k=1}^{r} f_{jk}(x_k). \tag{1.1}$$

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

$$\mathbf{X}_i \ \sim \ g_{\boldsymbol{\theta}}(\mathbf{x}_i) \ = \ \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}), \tag{1.2}$$

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

$$\sum_{j=1}^{m} \lambda_j = 1 \quad \text{and each } \lambda_j \geq 0. \tag{1.3}$$

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 $1 \leq k \leq r$.

One attractive feature of Model (1.2) is the fact that the multivariate density functions for each mixture component may be studied via their univariate marginal densities, since the conditional independence implies that each f_j factors into the product of its marginals. This aids in the interpretation of results, as we shall see in Section 5, since it is typically easier to analyze univariate density functions than multivariate ones. We emphasize, however, that Model (1.2) does *not* imply that the repeated measurements are independent; the dependence is captured by component membership.

The focus of the remainder of this article will be model (1.2) in the case when the dimension r is at least 3, for reasons that will become apparent in Section 2. First, however, we discuss additional existing literature on the general topic of nonparametric mixture models.

The case of univariate nonparametric mixtures, though not the cornerstone of this manuscript, deserves some additional remarks. When $r = 1$, Equation (1.2) becomes

$$g_{\boldsymbol{\theta}}(x_i) \;=\; \sum_{j=1}^{m} \lambda_j f_j(x_i),$$

and it is not hard to see that the parameters on the right hand side are not uniquely determined if the left hand side is known. Therefore, some additional constraints are necessary in the univariate case. To this end, Bordes et al. [8] and Hunter et al. [25] simultaneously propose a location-shift semiparametric mixture model

$$g_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{m} \lambda_j f(x - \mu_j)$$

for some symmetric (about zero), but otherwise completely unspecified, density f and unknown scalar location parameters μ_j. These authors prove identifiability of these parameters for $m \leq 3$ except in certain special cases (see Section 4.1). Bordes et al. [8] use the so-called minimum contrast method to estimate all of the parameters. The authors show that the mixing weight estimators converge to the true weights at the $\sqrt{n}$ rate if the location parameters are known; if they are not known, then the entire Euclidean parameter is $n^{1/4-\alpha}$ consistent for an arbitrarily small $\alpha > 0$. The proposed method is interesting but difficult to generalize to the case of more than two component densities.

Later, Bordes et al. [7] proposed a stochastic EM-like estimation algorithm that is easy to generalize but does not possess the descent property of a genuine EM algorithm. We discuss in Section 4.1 an alternative algorithm that does guarantee the descent property.

Another version of this problem is that of the mixture of the two distributions where only one of them is unknown. Such a model can be written as

$$g(x) = \lambda f(x) + (1 - \lambda)\phi(x), \tag{1.4}$$

where the probability density ϕ is known while f is unknown. This model has been considered in Robin et al. [38] together with its application to local false discovery rate (FDR) estimation. The other context where this model appears

is that of contamination, where the distribution ϕ for which reasonable assumptions can be made is contaminated by an arbitrary distribution f. Robin et al. [38] proposed an algorithm that converges to the unique solution of the problem but no large-sample analysis of the problem has been undertaken. A significant step forward in that regard was made in Bordes and Vandekerkhove [9], who considered model (1.4) with an additional assumption of the symmetry of the unknown density f and an additional location parameter and obtained identifiability together with a joint central limit theorem for the estimators of all the components. An EM-like algorithm for FDR estimation using this model is presented in Chauveau et al. [11]. Hohmann [23] considered two additional models, both consisting of just two components as well. One of them has an unknown component f with an unknown location parameter μ and a known component f_0 with an unknown location parameter ν; the other is the generalization of the previous one that also adds a scale parameter to the known component f_0.

The common thread in all of the nonparametric mixture models discussed heretofore is the assumption of a fixed, known number of components and no parametric assumptions on the component densities. An entirely different type of nonparametric mixture model is studied by Lindsay [32]. In this treatment, the mixture model is given by

$$g_Q(\mathbf{x}) = \int f(\mathbf{x};\boldsymbol{\theta}) dQ(\boldsymbol{\theta}),$$

where the mixing distribution $Q(\boldsymbol{\theta})$ is completely unspecified but the component densities $f(\mathbf{x};\boldsymbol{\theta})$ are known to come from a parametric family indexed by θ. In model (1.2), by contrast, $Q(\cdot)$ is assumed to be supported on m points whereas the $f(\mathbf{x})$ are fully general except for the conditional independence assumption. Throughout the rest of this article, our interpretation of "nonparametric mixture model" is that of Equation (1.2) rather than of Lindsay [32].

The multivariate finite-mixture version of the conditional independence assumption (1.2) has appeared in a growing body of literature on non- and semiparametric multivariate mixture models. Hettmansperger and Thomas [22] introduced a more restrictive version of (1.2) 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 [15] and Cruz-Medina and Hettmansperger [12]. Hall and Zhou [21] considered (1.2) in its full generality, establishing some rudimentary results concerning the identifiability of the parameters in this model. Other articles [16, 20] explored this identifiability question further, until Allman et al. [1] established the fundamental result that we elucidate fully in Section 2. Benaglia et al. [4] proposed an estimation algorithm for (1.2), which was later modified and put on more solid theoretical ground by Levine et al. [31], 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. [31], and in Section 5, we summarize numerical tests of the extended algorithm.

2. Identifiability

Without some further restrictions, the parameters of the finite nonparametric mixture model (1.2) are not identifiable. (NB: We avoid referring to the model itself as identifiable or not—a model is merely a set of probability distributions and it may in general be parameterized in multiple ways, some of which may be identifiable and others not.) Since the lack of identifiability results probably delayed the development of estimation and inference for these models, it is a good idea to recall steps that led to the complete description of the finite conditionally independent nonparametric density mixtures.

Some of the first results in this direction were obtained by Hall and Zhou [21]. This publication was, essentially, a trailblazer in that it was the first to consider a mixture of nonparametric components without any training data available. As the authors themselves noticed, "Very little is known of the potential for consistent nonparametric inference in mixtures without training data and that problem motivates the present paper". The particular motivation in that case was an attempt to estimate the distributions of medical test results when disease status was unavailable. Considering only a special case of $m = 2$ components, Hall and Zhou [21] also suggested an estimation approach that was somewhat awkward and difficult to implement; a more efficient approach was suggested later by Hall et al. [20], who also noticed what came to be known as the "curse of dimensionality in the reverse"—for a given number of components m, there is a lower bound r_m that the dimensionality of observations must exceed for the model to be identifiable. Hall et al. [20] showed that, if $r \geq r_m$ and $m \geq 2$, both the component densities and mixing weights can be estimated $\sqrt{n}$-consistently.

Another important step on the road to the full-fledged description of identifiability conditions was the paper by Kasahara and Shimotsu [26] concerning dynamic discrete choice models in economics. It takes a somewhat different approach to the identifiability than the above literature. One reason for that is that the result of Hall et al. [20] has limited applicability in that context since the number r_m, as characterized in Hall and Zhou [21], is typically too large for economic applications. Also, most models used in economics have covariates whose inclusion must be accounted for and that are not considered in Hall et al. [20]. Finally, the conditional independence, while reasonable in many applications, is sometimes difficult to justify in dynamic discrete choice models for panel data that are commonly considered in economics. Kasahara and Shimotsu [26] considered a number of finite mixture models used in applied work in economics and, under different assumptions on Markov property, stationarity and type-invariance in transition process, obtained sufficient conditions for their identifiability.

The fundamental result concerning identifiability of finite mixtures of nonparametric, conditionally independent measure products is due to Allman et al. [1]. It is based on an algebraic result of Kruskal [27, 28] 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 three 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 a 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

$$[A_1,A_2,A_3]=\sum_{j=1}^{m}\mathbf{a}_j^1\bigotimes\mathbf{a}_j^2\bigotimes\mathbf{a}_j^3. \tag{2.1}$$

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, $1\le j\le m$, 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 $\mathbf{p}_{jk}=P(X_k=\cdot\,|\,Z=j)$. Choose one of the three matrices (say, A_1) and define $\tilde{A}_1=\operatorname{diag}(\boldsymbol{\lambda})A_1$, 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 (2.1).

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 2.1. *Let $I_k=rank_K A_k$. If*

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

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 18, 36]. It was understood long ago that finite mixtures of Bernoulli products are not identifiable in a strict sense [see 19]; however, these mixtures are known to be well behaved in practice with respect to statistical parameter inference [see, for example, 10]. Allman et al. [1] explained this seeming contradiction by providing exact sufficient conditions for *generic* identifiability of these mixtures, up to 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. [1].

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. [1] 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 2.1.

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, $1 \le j \le m$. 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

$$\mathcal{P} = \sum_{j=1}^{m} \lambda_j \mu_j = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} \mu_j^k. \tag{2.2}$$

This model implies that the r variates are, yet again, independent conditional 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. [1].

Theorem 2.2. *Let* $\mathcal{P}$ *be a mixture of nonparametric measure products as defined in (2.2) and, for every variate* $k \in \{1, \ldots, r\}$*, assume 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. Parameter estimation via EM-like algorithms

We do not attempt here to document the full history of the development of algorithms for estimating the model parameters of Equation (1.2) and the various special cases of that model that have been considered in the literature, since many if not most of the theoretical articles cited in Sections 1 and 2 introduce estimation algorithms of one sort or another. However, we do trace one particular line of algorithmic development, a sequence of algorithms of increasing flexibility and theoretical grounding that culminates in the novel contributions of the current article. This line of algorithms combines the well-known techniques of kernel density estimation and EM algorithms for (parametric) finite mixture models.

Bordes et al. [7] was the first article to introduce an EM-like algorithm that incorporates a kernel density estimator of the unknown component densities. The idea of that article, which specifically considered the particular univariate nonparametric finite mixture model described in Section 4.1, is that each observation in the dataset may be randomly assigned, at every iteration of the algorithm, to one of the mixture components. This assignment is based on the probabilities of component membership for each observation, which are a natural byproduct of a finite-mixture EM algorithm, resulting in a stochastic algorithm. Benaglia et al. [4] extends this algorithm to the case of Equation (1.2) in its full generality and refines it by replacing the random assignment of observations to components by a deterministic algorithm that assigns observations fractionally to the various components according to their component membership probabilities. This EM-like algorithm, which was published even before the identifiability result of Allman et al. [1], was by far the most efficient estimation algorithm in the literature for Equation (1.2) that existed at the time. Yet it lacked one crucial feature of any true EM algorithm: It did not possess any provable descent property. This flaw was remedied by Levine et al. [31], who modified the algorithm slightly and put it on more solid theoretical footing by proving a descent property. It is this modified algorithm that we discuss here, along with some novel extensions. First, however, it is necessary to introduce some notation.

3.1. Notational conventions

Let Ω be a compact subset of $\mathbb{R}^r$ 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

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

This $\mathcal{N}_{\mathbf{h}}$ operator is strictly concave [13, 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 (1.1). Letting H denote the $m \times r$ bandwidth matrix $(\mathbf{h}_1^\top, \ldots, \mathbf{h}_m^\top)^\top$, we may extend $\mathcal{S}$ to $\mathcal{F}$ by defining $\mathcal{S}_H \mathbf{f} = (\mathcal{S}_{\mathbf{h}_1} f_1, \ldots, \mathcal{S}_{\mathbf{h}_m} f_m)^\top$. Similarly, we let $\mathcal{N}_H \mathbf{f} = (\mathcal{N}_{\mathbf{h}_1} f_1, \ldots, \mathcal{N}_{\mathbf{h}_m} f_m)^\top$.

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 (1.2), and

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

The introduction of the bandwidth matrix H above is an extension of the notation used in Levine et al. [31], where each component and coordinate is assumed to use the same scalar bandwidth h.

3.2. The descent property

Let $g(\mathbf{x})$ now represent a known target density function. Following Levine et al. [31], we define the functional

$$\ell_H(\boldsymbol{\theta}, g) = \int_{\Omega} g(\mathbf{x}) \log \frac{g(\mathbf{x})}{[\mathcal{M}_{\boldsymbol{\lambda}} \mathcal{N}_H \mathbf{f}](\mathbf{x})}\, d\mathbf{x}, \tag{3.1}$$

which can be viewed as a penalized Kullback-Leibler distance between $g(\mathbf{x})$ and $(\mathcal{M}_{\boldsymbol{\lambda}} N_H \mathbf{f})(\mathbf{x})$. Suppose we wish to find parameter values $\boldsymbol{\theta}$ that minimize $\ell_H(\boldsymbol{\theta}, g)$. Letting $\boldsymbol{\theta}^0 = (\boldsymbol{\lambda}^0, \mathbf{f}^0)$ denote the starting values we use in our iterative algorithm, define

$$\hat{f}_{jk}(u) \;=\; \alpha_{jk} \int K_{h_{jk}}(x_k - u) g(\mathbf{x}) w_j^0(\mathbf{x})\, d\mathbf{x}, \tag{3.2}$$

where

$$w_j^0(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\lambda_j^0 \mathcal{N}_{\mathbf{h}_j} f_j^0(\mathbf{x})}{\mathcal{M}_{\boldsymbol{\lambda}^0} \mathcal{N}_H \mathbf{f}^0(\mathbf{x})}, \tag{3.3}$$

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

$$\hat{\lambda}_j = \frac{\int g(\mathbf{x}) w_j^0(\mathbf{x})\, d\mathbf{x}}{\sum_{a=1}^m \int g(\mathbf{x}) w_a^0(\mathbf{x})\, d\mathbf{x}} = \int g(\mathbf{x}) w_j^0(\mathbf{x})\, d\mathbf{x}. \tag{3.4}$$

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

$$\ell_H(\hat{\boldsymbol{\theta}}, g) \;\leq\; \ell_H(\boldsymbol{\theta}^0, g). \tag{3.5}$$

This fact may be verified by proving that the algorithm defined by iterating Equations (3.2) and (3.4) is an example of a so-called MM algorithm, which stands for majorization-minimization algorithm [24], a class of algorithms that generalizes the well-known EM algorithms. The proof is nearly identical to a result of Levine et al. [31] except for the presence of the different bandwidth values H. The analogous result for sample data, in which $g(\mathbf{x})\,d\mathbf{x}$ is replaced by the empirical distribution $d\tilde{G}_n(\mathbf{x})$, is discussed in the next section and proved in Appendix A.

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 (1.2), or one may instead take the view that the truth is not contained in our model class and that the goal of estimation is to find in some sense a "best" vector $\boldsymbol{\theta}$ to approximate the truth by a density of the form (1.2). 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 parameters, a discrete version of (3.1) is

$$\begin{aligned} \ell_H(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \int \log \frac{1}{[\mathcal{M}_{\boldsymbol{\lambda}}\mathcal{N}_H\mathbf{f}](\mathbf{x})}\, d\tilde{G}_n(\mathbf{x}) \\ &= -\sum_{i=1}^{n} \log\left\{[\mathcal{M}_{\boldsymbol{\lambda}}\mathcal{N}_H\mathbf{f}](\mathbf{x}_i)\right\}. \end{aligned} \tag{3.6}$$

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 (3.1). In its new form, Equation (3.6), 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. [31]: In that algorithm, it is possible to fix certain subsets of the coordinates in the $\mathbf{x}$ vectors to be identically distributed, in addition to being conditionally independent. This idea, first introduced by Benaglia et al. [4], generalizes the assumptions of earlier work such as Hettmansperger and Thomas [22] and Elmore et al. [17], who assumed that the coordinates are conditionally independent *and* identically distributed, as well as Hall and Zhou [21], who assumed only conditional independence as in Model (1.2). The assumption that

some coordinates have identical marginal densities, if realistic, allows for more precise estimation because data on these coordinates may be pooled. As we illustrate in Section 5, some applications naturally suggest certain subsets of coordinates that may reasonably be assumed to be identically distributed.

To formalize these ideas, we say that groups 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

$$g_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_{ik}). \tag{3.7}$$

A simplification is possible when $b_k = k$ for all k, whereby (3.7) becomes (1.2). Assuming model (3.7) and letting $h_{j\ell}$ be the bandwidth used in the jth component and the ℓth block, the objective function of Equation (3.6) may be written

$$\ell_H(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j \exp\left\{ \sum_{k=1}^{r} \int K_{h_{jb_k}}(x_{ik} - u) \log f_{jb_k}(u)\, du \right\}. \tag{3.8}$$

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

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

$$w_{ij}^t = \frac{\lambda_j^t \mathcal{N}_{\mathbf{h}_j} f_j^t(\mathbf{x}_i)}{\sum_{a=1}^{m} \lambda_a^t \mathcal{N}_{\mathbf{h}_a} f_a^t(\mathbf{x}_i)} = \frac{\lambda_j^t \prod_{k=1}^{r} \mathcal{N}_{h_{jb_k}} f_{jb_k}^t(x_{ik})}{\sum_{a=1}^{m} \lambda_a^t \prod_{k=1}^{r} \mathcal{N}_{h_{ab_k}} f_{ab_k}^t(x_{ik})}. \tag{3.9}$$

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

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^t \tag{3.10}$$

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

$$f_{j\ell}^{t+1}(u) = \frac{1}{n h_{j\ell} \lambda_j^{t+1} C_\ell} \sum_{k=1}^{r} \sum_{i=1}^{n} w_{ij}^t I_{\{b_k = \ell\}} K\left(\frac{u - x_{ik}}{h_{j\ell}} \right), \tag{3.11}$$

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 (3.11) need not be the same as those in the E-step (3.9). However, in order to prove that our new algorithm retains the desirable descent property, we require an analogue of Equation (3.2), 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. [31] to show that $\ell_H(\boldsymbol{\theta}^t)$ is nonincreasing in t using the algorithm in this section. In other words, equations (3.9) through (3.11) ensure the descent property

$$\ell_H(\boldsymbol{\theta}^{t+1}) \leq \ell_H(\boldsymbol{\theta}^t). \tag{3.12}$$

3.4. Bandwidth selection

As discussed in Benaglia et al. [4], the selection of a bandwidth in a mixture setting like (3.7) can be an intricate problem, and there are several reasons that using a single, fixed bandwidth or set of bandwidths throughout an iterative algorithm, as is implicit in Equations (3.9) through (3.11), is not always appropriate. An iterative bandwidth scheme adapting the well-known rule of Silverman [39, p. 46] is proposed in Benaglia et al. [5] for a similar algorithm. Briefly, it amounts to replacing, in Silverman's rule

$$h = 0.9 \min\left\{\mathrm{SD}, \frac{\mathrm{IQR}}{1.34}\right\} n^{-1/5} \tag{3.13}$$

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 to the MSL algorithm and gives estimated bandwidths at the $(t+1)$th iteration,

$$h_{j\ell}^{t+1} = 0.9 \min\left\{\sigma_{j\ell}^{t+1}, \frac{IQR_{j\ell}^{t+1}}{1.34}\right\} (nC_\ell \lambda_j^{t+1})^{-1/5}, \tag{3.14}$$

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. [5], but using here the w_{ij}^t to weight the data.

Since the adaptive bandwidth selection strategy of Equation (3.14) was originally tailored for estimation of Gaussian distributions, it may not be the best choice in our nonparametric setting. Another strategy frequently used in the literature is the cross validation (CV) approach. As explained in Eggermont and LaRiccia [14, Section 7.2], the "leave-one-out" method for cross-validation in the case of a univariate sample $y_1, \ldots, y_n$ minimizes the function

$$\mathrm{CV}(h) = \left\|\hat{f}^{nh}\right\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_{(i)}^{nh}(y_i),$$

where f^{nh} is the kernel density estimator using bandwidth h and

$$f_{(i)}^{nh}(y) = \frac{1}{n-1}\sum_{j\neq i} K_h(y - y_i). \tag{3.15}$$

In other words, the leave-one-out CV method basically relies on $n+1$ kernel density estimators. It is therefore straightforward, though computationally intensive, to derive bandwidths $h_{j\ell}^t$ using $n+1$ weighted KDEs as in Equation (3.11) for each component j, block ℓ, and iteration t. A computationally faster approach arises if we partition the individuals among the components

at each iteration, then apply the fast `bw.ucv()` function available in R. We may achieve this partitioning either stochastically, by taking $\mathbf{z}_i^t$ to be a single random draw from a multinomial distribution with probability vector $\mathbf{w}_i^t$, or deterministically, by setting

$$z_{ij_0}^t = 1, \quad \text{where} \quad j_0 = \underset{j=1,\ldots,m}{\arg\max} \left\{ w_{ij}^t \right\}$$

and $z_{ij}^t = 0$ for $j \neq j_0$. We refer to the latter strategy as the maximum a posteriori (MAP) method. For well-separated mixtures and for individuals belonging clearly to one component, the MAP approach should result in $\mathbf{z}^t$ very close to the weights $\mathbf{w}^t$, so the "adaptive MAP CV" bandwidth should be close to the weighted CV bandwidth, but computationally simpler and faster. We implement the leave-one-out MAP CV strategy in our experiments in Section 5.1. We have also experimented with a "leave-k-out" CV strategy in which $\lfloor n/k \rfloor$ different subsets of size $n - k$ are used to produce kernel density estimates in place of Equation (3.15), but the final results are similar and so we omit the details here. For all CV approaches, including both leave-one-out and leave-k-out for $k > 1$, it is straightforward to derive an adaptive method for bandwidth estimation based on weighted kernel density estimates.

The major difference between our MSL algorithm and the Benaglia et al. [4] algorithm or some modified version that uses a different bandwidth updating scheme such as a cross-validation algorithm is that the former 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. We do not tackle this difficult question 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, we find that a slightly different scheme works well: Simply run the first stage, in which bandwidths are allowed to change, until our convergence criterion is satisfied. (The result is, in fact, the same for the two schemes, since the second stage makes no changes because the convergence criterion is already satisfied.)

4. Extensions of the estimation algorithm

Here, we discuss two novel extensions of the basic idea of the algorithm of Section 3.3 to situations related to model (3.7). 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

In this chapter we denote $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)'$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$. Both Bordes et al. [8] and Hunter et al. [25] showed that, for univariate

$$X \sim \sum_{j=1}^{m} \lambda_j f(x - \mu_j), \tag{4.1}$$

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. [25] showed that for $m = 3$, the parameters are uniquely identifiable except when $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ take values in a particular set of Lebesgue measure zero, conjecturing that a similar result may be shown for general m.

Although both Bordes et al. [8] and Hunter et al. [25] propose methods for estimating the parameters in (4.1) given a simple random sample $x_1, \ldots, x_n$ distributed according to (4.1), these methods are inefficient and not easily generalizable beyond the case $m = 2$. Bordes et al. [7] propose a stochastic EM-like estimation algorithm that is easily generalizable to any m and that works well in practice; 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 and initial parameter values $\boldsymbol{\theta}^0 = (f^0, \boldsymbol{\lambda}^0, \boldsymbol{\mu}^0)$, iterate the following steps for $t = 0, 1, \ldots$:

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

$$w_{ij}^t = \frac{\lambda_j^t \mathcal{N}_h f^t(x_i - \mu_j^t)}{\sum_{a=1}^m \lambda_a \mathcal{N}_h f^t(x_i - \mu_a^t)} \tag{4.2}$$

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

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^t \tag{4.3}$$

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

$$f^{t+1}(u) = \frac{1}{2nh\lambda_j^{t+1}} \sum_{j=1}^{m} \sum_{i=1}^{n} w_{ij}^t \left[K\left(\frac{x_i - \mu_j^t - u}{h}\right) + K\left(\frac{x_i - \mu_j^t + u}{h}\right) \right]. \tag{4.4}$$

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

$$\mu_j^{t+1} = \arg\max_{\mu} \int \sum_{i=1}^{n} w_{ij}^t K\left(\frac{x_i - u}{h}\right) \log f^{t+1}(u - \mu)\, du. \tag{4.5}$$

Equation (4.4) 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 (3.6) is nonincreasing in t, where in this model we may express this objective function in the form

$$\ell_h(\boldsymbol{\theta}^t) \quad = \quad -\sum_{i=1}^{n} \log \sum_{j=1}^{m} \lambda_j^t [\mathcal{N}_h f^t](x_i - \mu_j^t). \tag{4.6}$$

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 (4.5) 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 (4.5) is inexpensive to calculate for a given μ. In practice, there is a tradeoff to be made here between accuracy of the integral and speed of implementing the algorithm.

Although we do not implement this variant in the code we test in Section (5), one could modify the 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 [33].

4.2. The location-scale model

It is not difficult to restrict model (3.7) somewhat while still retaining the essential nonparametric character of the estimation: We may assume that the various univariate density functions in Equation (3.7) 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, the idea of "exponential tilting" assumes that $\log[f_{j\ell}(x)/f_{1\ell}(x)]$ follows a specified parametric form in x for all components $j > 1$ and blocks ℓ. This idea in which the log-ratio of densities is linear in x was introduced in the mixture setting by Anderson [2], then extended by Leung and Qin [30], among others. These authors all employ the least restrictive block structure in Equation (3.7), i.e., $b_k = k$.

By contrast, we assume here, as in Benaglia et al. [4], that

$$f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f_j \left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right) \tag{4.7}$$

for unknown parameters $(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j, f_j)$, $j = 1, \ldots, m$. We thus assume 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 (4.7) 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

$$f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f_\ell \left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right), \tag{4.8}$$

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. Although Equation (4.7) differs from Equation (4.8) by only a single subscript on the density f, the interpretations of the two models are quite different.

As a special case of both (4.7) and (4.8), if all f_{jk} are assumed to have the same shape, then we may require that

$$f_{j\ell}(x) = \frac{1}{\sigma_{j\ell}} f \left(\frac{x - \mu_{j\ell}}{\sigma_{j\ell}} \right) \tag{4.9}$$

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

Because f_j in equation (4.7) is completely unspecified, the location and scale parameters may be absorbed into f_j, so the parameters are not uniquely identifiable 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, the algorithm may work even if these constraints are not enforced. Similar arguments can be made for the parameters in equations (4.8) and (4.9).

We may modify the algorithm of Section 3.3 to incorporate the block structure of Equation (3.7). Equations (3.9) and (3.10) remain unchanged, but we must modify Equation (3.11) to either

$$f_j^{t+1}(u) = \frac{1}{nrh_j\lambda_j^{t+1}} \sum_{k=1}^{r} \sum_{i=1}^{n} w_{ij}^t K \left(\frac{u - x_{ik} + \mu_{jb_k}^t}{h_j \sigma_{jb_k}^t} \right) \tag{4.10}$$

or

$$f_\ell^{t+1}(u) = \frac{1}{nh_\ell\lambda_j^{t+1}C_\ell} \sum_{k=1}^{r} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^t I_{\{b_k=\ell\}} K \left(\frac{u - x_{ik} + \mu_{jb_k}^t}{h_\ell \sigma_{jb_k}^t} \right), \tag{4.11}$$

where $C_\ell = \sum_{k=1}^{r} I_{\{b_k=\ell\}}$, depending upon whether we take Equation (4.7) or Equation (4.8) 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 (4.5). The resulting algorithm would be similar to the one described in Section 4.1: It is not an MM algorithm exactly, but it is very similar and most importantly it guarantees a decrease in the desired objective function (3.8).

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 non-central 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 non-centrality. 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 $\lambda_1 = 0.4$. This example, in which the coordinate densities are on different scales, is specifically designed so that the bandwidth should depend on the blocks and components.

A simple run of the original MSL algorithm on a sample of size $n = 300$ results in a single fixed bandwidth $h = 0.511$, while a run with the adaptive Silverman's rule gives the following (final) bandwidth matrix:

	component 1	component 2
block 1	0.330	0.332
block 2	0.085	0.037

A run of the same sample using the adaptive MAP CV strategy gives very similar results:

	component 1	component 2
block 1	0.360	0.350
block 2	0.036	0.0077

The estimates of the component and block densities using Silverman's rule are shown in Figure 1, where we see nearly identical estimates of the densities in the first block, whereas the two algorithms produce dramatically different estimates in the second block. The reason for this is clear in this example: A fixed bandwidth estimate of $h \approx 0.5$ is over-estimated for both blocks, but dramatically so for the second block, in which the support of the mixture is $[0, 1]$ and therefore a much smaller bandwidth is appropriate. On the other hand, adjusting the single bandwidth downward to fit the second block will result in more variable estimates in the first block, and in any event it is not obvious how to redesign the automatic bandwidth selector if we constrain h to be the same for every block and component.

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.

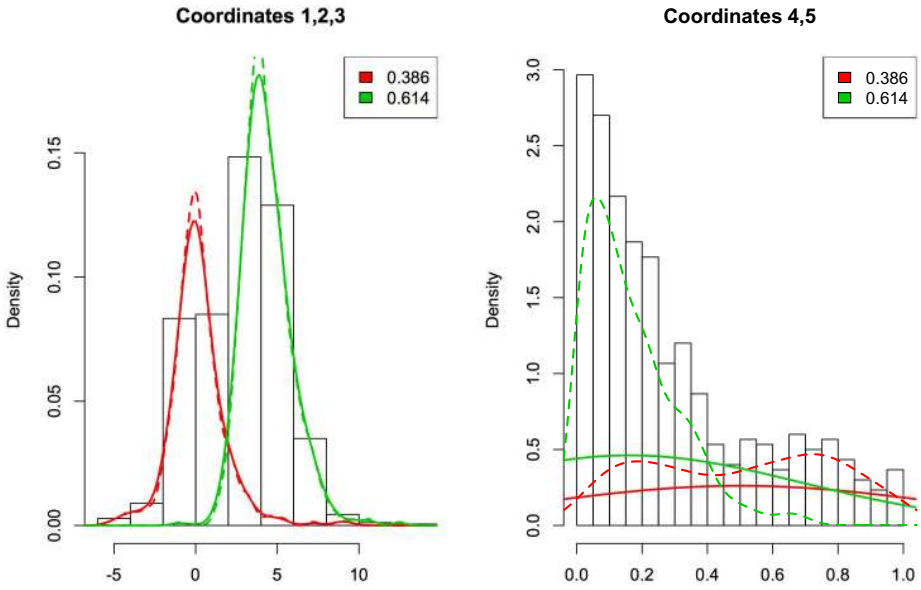

FIG 1. *Solid and dashed lines show fixed-single-bandwidth and adaptive-multiple bandwidth MSL solutions, respectively. The two components are displayed in different colors.*

5.2. The water-level dataset

To illustrate the adaptive block- and component-wise bandwidth approach, we consider a benchmark dataset previously analyzed by Hettmansperger and Thomas [22] and Elmore et al. [17] with a conditionally i.i.d. (independent and identically distributed) assumption, and more recently by Benaglia et al. [4] and Levine et al. [31] 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. [40]. 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 [41, 6].

As in Benaglia et al. [4] and Levine et al. [31], 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. Thus the eight coordinates may be organized into four blocks of two each, which is model (3.7) 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, corre-

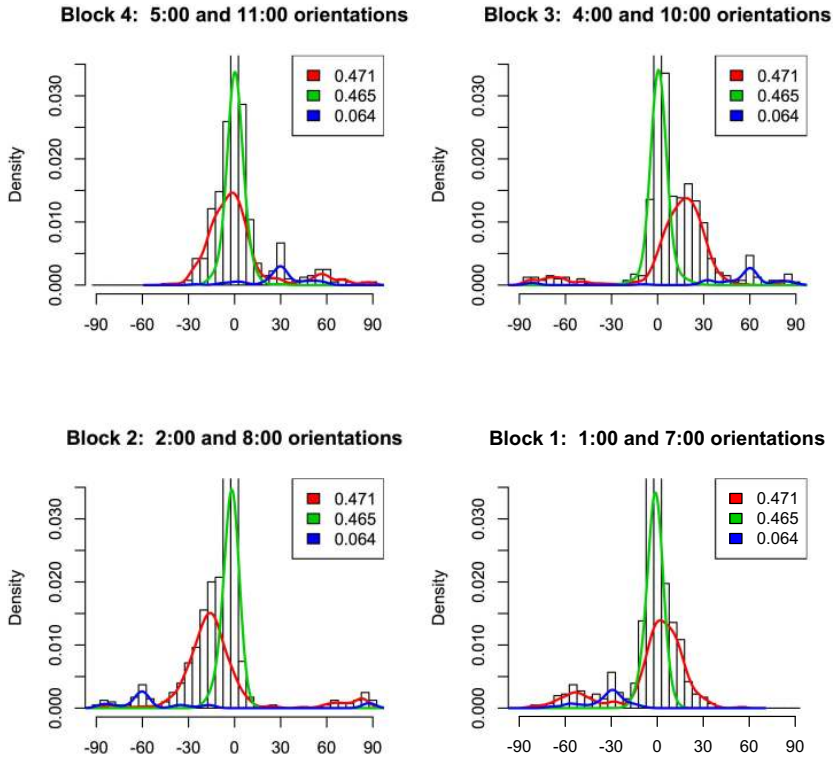

FIG 2. *The water-level data analyzed using the MSL algorithm with* $m = 3$ *mixture components and a fixed bandwidth* $h = 4$.

sponding to clock orientations 1:00 and 7:00. An alternative to this particular pairing of clock-hour orientiations is to consider each of the eight orientations as distinct. As we show in Appendix B, treating the dataset in this way provides evidence that the pairing approach is reasonable; thus, we proceed with the four-block assumption.

We first consider here the 3-component model as studied in Levine et al. [31] to compare the MSL with fixed bandwidth against the adaptive strategy. Figure 2 gives the solution returned by the MSL algorithm with a fixed bandwidth preset to $h = 4$, as in Benaglia et al. [4] and Levine et al. [31]. 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 (3.13). 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 con-

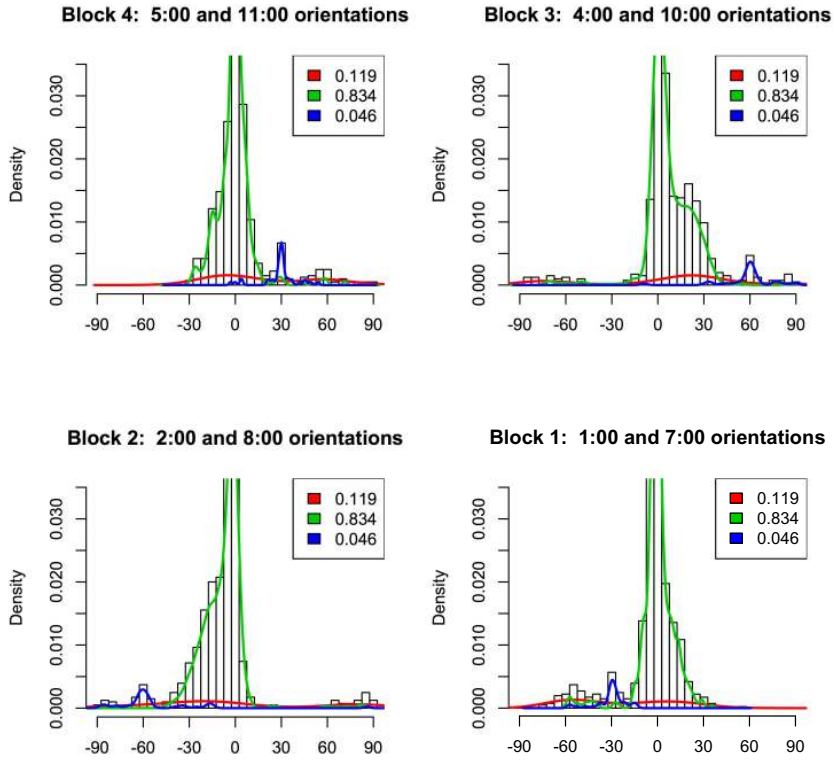

FIG 3. *The water-level data analyzed using the MSL algorithm with $m = 3$ mixture components and adaptive bandwidths given in Table 1.*

TABLE 1
Adaptive bandwidths per block and components for the Water level data, at the MSL 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

fused 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 MSL algorithm with the adaptive bandwidth given by (3.14). The corresponding bandwidth matrix is displayed in Table 1, which shows that the bandwidth differences are mostly between components.

The qualitative interpretation appears simpler here, since the competent group now represents 83% of the subjects (but seems to encompass most of

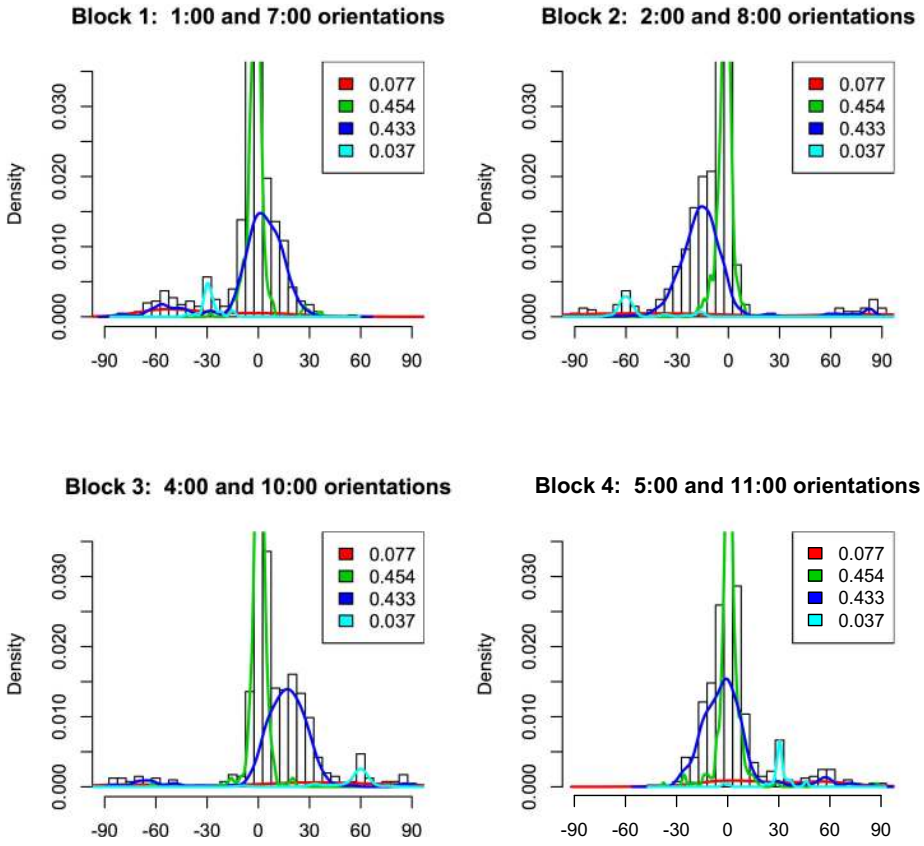

FIG 4. *The water-level data analyzed using the MSL algorithm with* $m = 4$ *mixture components and adaptive bandwidths strategy.*

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 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 4-component model with the MSL 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 again comprises 45% of the subjects, as distinct from the 43% slightly confused group. The group that always draws the line parallel to the vessel bottom drops to 3.7%, more in accordance with the result from Fig. 3, and distinct from the 7% totally guessing group.

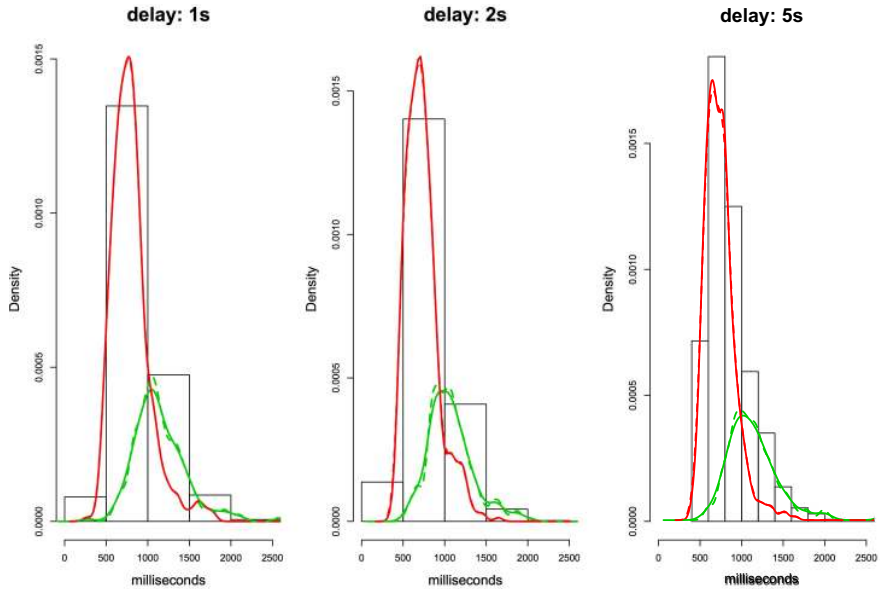

FIG 5. *Density estimates for the simple RT task with $B = 3$ blocks of 8 coordinates each, and $m = 2$ components: MSL with fixed bandwidth (dashed line), and adaptive bandwidths (solid line). The component weights are* $(0.72, 0.28)$.

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. [34]. 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. [34] 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" spoken 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 will be and 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 [41, 6] for the R statistical software environment [37], 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 MSL 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.

5.4. Measurements on abalone

Nash et al. [35] report measurements on a large dataset of $n = 4177$ abalone, which are a type of marine life akin to large snails that are often harvested for food. The dataset is publicly available at the Machine Learning Repository at the University of California, Irvine [3].

There are $m = 7$ continuous measurements made on each animal in the dataset, many of which (e.g., whole weight, shucked weight, and viscera weight) are highly correlated with one another. Since this correlation is likely to persist even within each mixture component, one might reasonably suspect that the conditional independence assumption is violated in this dataset. Thus, we use principal components (PC) analysis to find a linear transformation of the original dataset, then apply the modified MSL algorithm of Levine et al. [31], which assumes conditional independence and allows for different bandwidths for each component/coordinate combination, to the PC scores obtained from the PC analysis. Figure 6 plots the 4177 measurements on their original scale but using labelings derived from the analysis of the PC scores only. According to Bache and Lichman [3], it is of interest to use the seven measurements (as well as the animals' sex, which we ignore in our analysis) to try to predict the number of rings counted after cutting the shell through the cone, staining it, and observing the animal through a microscope. The number of rings is related to the age of the animal. In Figure 7, we see the ring distributions for each of the three components after classifying animals according to their highest posterior probability.

Our choice here of $m = 3$ is arbitrary and more detailed analyses might explore different numbers of components. We have successfully run our algorithm on this dataset with up to $m = 20$ components. One interesting facet of our analysis of these data is that it is the first dataset for which we have noticed a qualitative difference between the results of the EM-like algorithm of Benaglia et al. [4] and the smoothed version of the algorithm presented by Levine et al. [31]. For now, we cannot explain this difference, but its presence suggests a topic for future study.

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. [1], who prove that condi-

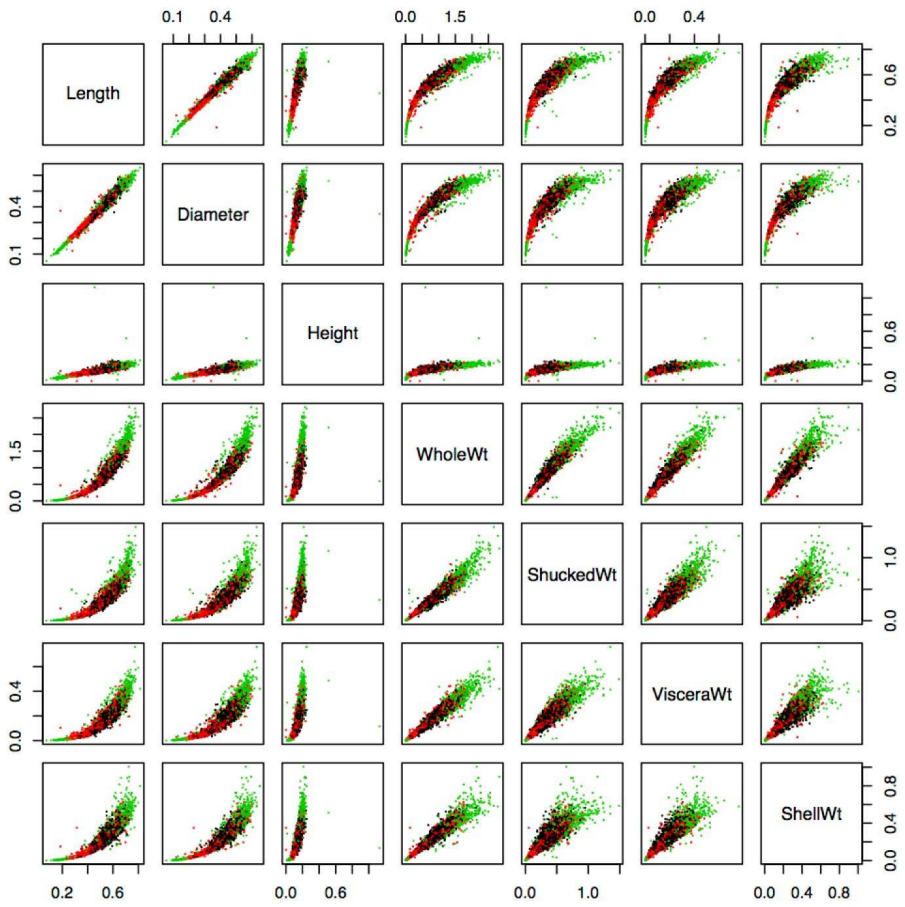

FIG 6. *Pairwise scatterplots for seven different measurements on a sample of 4177 abalone. Colors indicate the mixture component of highest posterior probability according to the model fit applied to the principal component scores. Estimates of the three population proportions are 0.415 for component 1 (black), 0.403 for component 2 (red), and 0.182 for component 3 (green).*

tional independence implies identifiability under weak assumptions as long as the multivariate observations have dimension at least three.

We review the MSL algorithm of Levine et al. [31] and introduce multiple methods for selecting bandwidths, an important aspect of the practical implementation of this algorithm. In addition, we extend the idea of Levine et al. [31] 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.

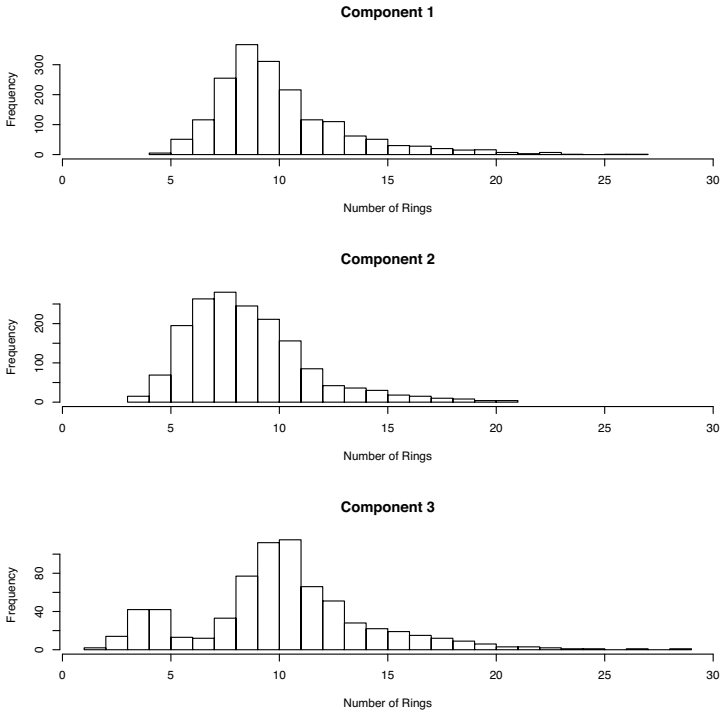

FIG 7. *Distribution of ring count for abalone classified into three groups according to their maximum posterior component probabilities.*

The important feature of the MSL 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 algorithm of Benaglia et al. [4, 5] because that algorithm does not appear to optimize any type of a loglikelihood-like function despite its resemblance to an EM algorithm.

Appendix 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 $f_{j\ell}$, $1 \le j \le m$ and $1 \le \ell \le B$. For a given (fixed) $\boldsymbol{\theta}^t$, let the constants w_{ij}^t be defined as in Equation (3.9). We first state and prove two lemmas, each based on the following definition:

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

Lemma A.1. *Let $\ell_H(\boldsymbol{\theta})$ be defined as in Equation (3.6). Then*

$$\ell_H(\boldsymbol{\theta}) - \ell_H(\boldsymbol{\theta}^t) \quad \leq \quad b_H^t(\boldsymbol{\theta}) - b_H^t(\boldsymbol{\theta}^t). \tag{A.2}$$

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, 24], the result of Lemma A.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 A.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 (3.10) and (3.11), 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 (3.10) gives the minimizer.

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

$$\begin{aligned}
b_H^t(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} w_{ij}^t \sum_{k=1}^{r} I_{\{b_k=\ell\}} \log \left\{ [N_{h_{j\ell}} f_{j\ell}](x_{ik}) \right\} \\
&\quad + \text{something not involving } f_{j\ell}.
\end{aligned} \tag{A.3}$$

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 (3.11). 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 (A.3) 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 A.1. *Let $\ell_H(\boldsymbol{\theta})$ be defined as in Equation (3.6). Then the algorithm given in steps (3.9) through (3.11) imply the descent property (3.12).*

Proof. Since Lemma A.2 implies that $b_H^t(\boldsymbol{\theta}^{t+1)} \leq b_H^t(\boldsymbol{\theta}^t)$, Lemma A.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 A.1. *For the univariate symmetric location model of Equation (4.1), the algorithm of Equations (4.2) through (4.5) guarantees $\ell_h(\boldsymbol{\theta}^{t+1}) \leq \ell_h(\boldsymbol{\theta}^t)$, where $\ell_h(\boldsymbol{\theta})$ is defined in Equation (4.6).*

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 A.1 and A.2 and Theorem A.1. Since Equation (3.6) is the same as Equation (4.6) for this special case, Lemmas A.1 and A.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 (A.1). 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

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

Using the same argument as in Lemma A.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 (4.3) and (4.4). Then, Equation (4.5) 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.

Appendix B: Justification of four-block assumption in water-level dataset

In the water-level example in Section 5.2, we pair the eight repeated measurements into four pairs of two clock angles each. As shown in Figure 8 for the

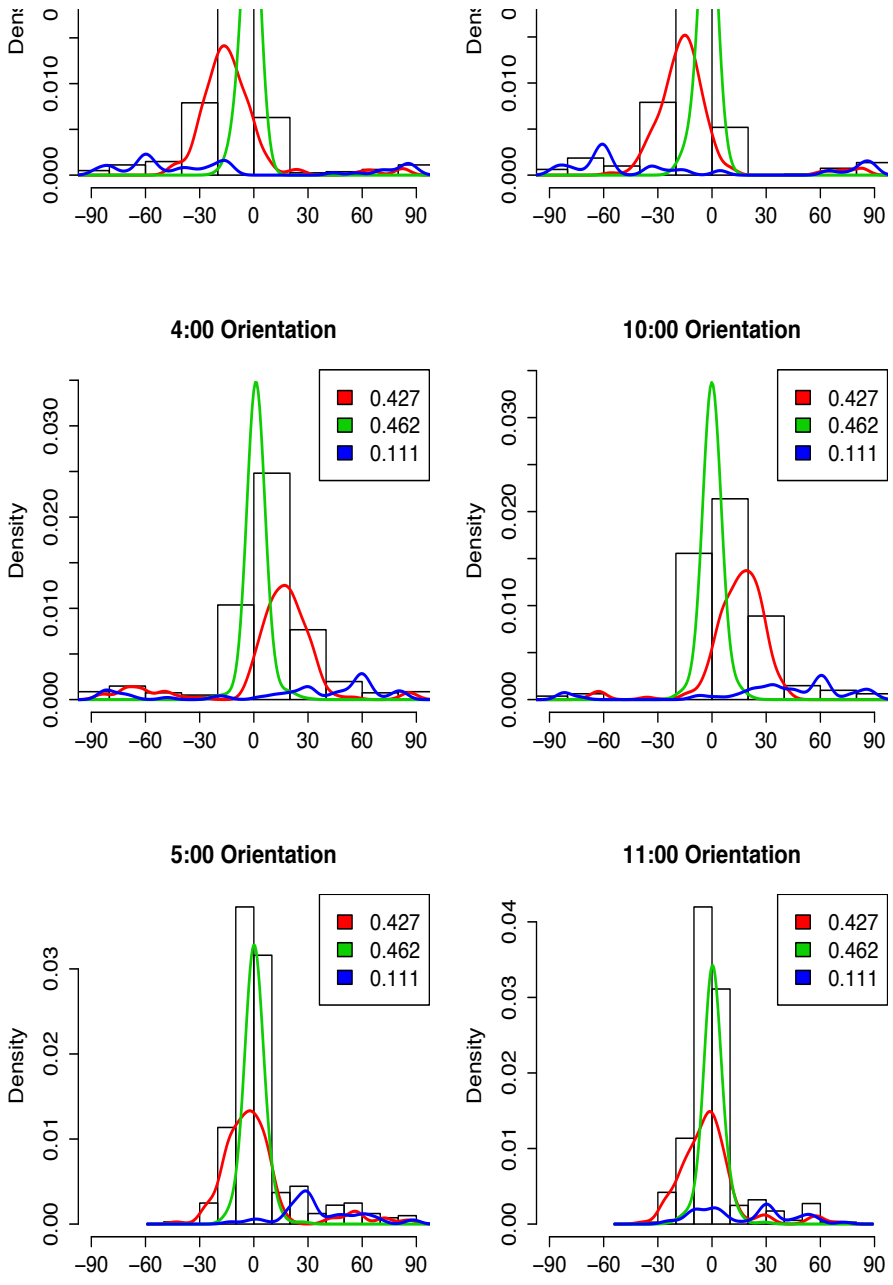

FIG 8. *This analysis of the water-level data uses the MSL algorithm with $m = 3$ assuming that each of the eight clock angle orientations follows its own distinct set of component densities. The orientations are paired so that the similarities in each of the four rows are obvious.*

three-component solution, this choice is justified by the fact that when we assume that each of the eight orientations follows a distinct set of component densities, the similarities between the pairs of orientations are clear. The results are similar when we consider the four-component solution, so we omit the corresponding plots for the four-component solution here.

Acknowledgements

We gratefully acknowledge the support of Award No. DMS–1209007 from the National Science Foundation.

References

[1] 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. MR2549554

[2] Anderson, J. (1979). Multivariate logistic compounds. *Biometrika*, 66(1):17–26. MR0529143

[3] Bache, K. and Lichman, M. (2013). University of California, Irvine machine learning repository. http://archive.ics.uci.edu/ml.

[4] 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.

[5] 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*.

[6] 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.

[7] 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. MR2370882

[8] Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34(3):1204–1232. MR2278356

[9] Bordes, L. and Vandekerkhove, P. (2010). Semiparametric two-component mixture model with a known component: An asymptotically normal estimator. *Mathematical Methods of Statistics*, 19(1):22–41. MR2682853

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

[11] Chauveau, D., Saby, N. P. A., Orton, T. G., Lemercier, B., Walter, C., and Arrouays, D. (2014). Large-scale simultaneous hypothesis

testing in monitoring carbon content from French soil database: A semi-parametric mixture approach. *Geoderma*, 219:117–124.

[12] Cruz-Medina, I. R. and Hettmansperger, T. P. (2004). Nonparametric estimation in semi-parametric univariate mixture models. *Journal of Statistical Computation and Simulation*, 74(7):513–524. MR2073229

[13] 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. MR1654562

[14] Eggermont, P. P. B. and LaRiccia, V. N. (2001). *Maximum Penalized Likelihood Estimation*. Springer, New York. MR1837879

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

[16] 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. MR2141286

[17] 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. MR2103063

[18] Glick, N. (1973). Sample-based multinomial classification. *Biometrics*, 29(2):241–256. MR0347009

[19] Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31:542–548. MR1274807

[20] Hall, P., Neeman, A., Pakyari, R., and Elmore, R. T. (2005). Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678. MR2202653

[21] Hall, P. and Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224. MR1962504

[22] 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. MR1796294

[23] Hohmann, D. (2010). *Identification and Estimation in Semiparametric Two-Component Mixtures*. PhD thesis, Philipps Universität Marburg.

[24] Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30–37. MR2055509

[25] Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251. MR2332275

[26] Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175. MR2477846

[27] 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. MR0488592

[28] 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. MR0444690

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

[30] Leung, D. and Qin, J. (2006). Semi-parametric inference in a bivariate (multivariate) mixture model. *Statistica Sinica*, 16(1):153. MR2256084

[31] Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 98:403–416. MR2806437

[32] Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–163. JSTOR.

[33] Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267. MR1243503

[34] 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.

[35] Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. (1994). The population biology of abalone (*Haliotis* species) in Tasmania. I. blacklip abalone (*H. rubra*) from the north coast and islands of Bass Strait. Technical report, Tasmania Sea Fisheries Division. Technical Report No. 48 (ISSN 1034-3288).

[36] 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.

[37] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[38] Robin, S., Bar-Hen, A., Daudin, J.-J., and Pierre, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–5493. MR2407654

[39] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London. MR0848134

[40] 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).

[41] 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.