Statistical Science
2019, Vol. 34, No. 3, 391–404
https://doi.org/10.1214/19-STS698

An Overview of Semiparametric Extensions of Finite Mixture Models

Sijia Xiang, Weixin Yao and Guangren Yang

Abstract. Finite mixture models have offered a very important tool for exploring complex data structures in many scientific areas, such as economics, epidemiology and finance. Semiparametric mixture models, which were introduced into traditional finite mixture models in the past decade, have brought forth exciting developments in their methodologies, theories, and applications. In this article, we not only provide a selective overview of the newly-developed semiparametric mixture models, but also discuss their estimation methodologies, theoretical properties if applicable, and some open questions. Recent developments are also discussed.

Key words and phrases: EM algorithm, mixture models, mixture regression models, semiparametric mixture models.

1. WHY SEMIPARAMETRIC MIXTURE MODELS?

Parametric mixture models are popularly used since they are easy to interpret, quick to estimate and have well-studied theoretical properties. However, as any other parametric statistical inference, parametric mixture models are based on strong model assumptions, such as linearity and normality. Some of the assumptions are unrealistic in practice. In addition, model misspecification could be disastrous in parametric mixture models and might lead to misleading results and inferences. Please refer to Pommeret and Vandekerkhove (2018) for the advantage of a semiparametric method in testing a parametric assumption on the unknown component of the two-component mixture model with one known component.

As a result, many semiparametric mixture models are proposed to relax assumptions of fully parametric mixture models. Bordes, Mottelet and Vandekerkhove (2006), Bordes, Chauveau and Vandekerkhove (2007) and Hunter, Wang and Hettmansperger (2007), among others, studied a two-component mixture of locations model where the component density is only assumed to be symmetric. Chang and Walther (2007) studied a mixture of log-concave distributions for clustering. This model includes most standard parametric families, but suffers from non-identifiability. In addition, a two-component mixture of locations model with a known component has been extensively studied during the past decade by Bordes, Delmas and Vandekerkhove (2006), Bordes and Vandekerkhove (2010), Patra and Sen (2016), Hohmann and Holzmann (2013), Xiang, Yao and Wu (2014), Ma and Yao (2015), Huang et al. (2018b), and so on.

In addition, many semiparametric finite mixtures of regressions (FMR) models were proposed in the last decade. By allowing the mixing proportions to be dependent on a covariate, Young and Hunter (2010) and Huang and Yao (2012) studied semiparametric mixture of regressions models with varying proportions. Huang, Li and Wang (2013) and Xiang and Yao (2018) relaxed the parametric assumptions on the mean functions and/or variances to accommodate for complicated data structures. However, due to the application of kernel regression in the estimation procedure, the models were not suitable for data with high-dimensional predictors. Hunter and Young (2012) studied a FMR model where linearity was still assumed within each component, but the error terms were modeled fully

Sijia Xiang is Associate Professor, School of Data Sciences, Zhejiang University of Finance & Economics, Hangzhou, Zhejiang, 310018, P.R. China (e-mail: sjxiang@zufe.edu.cn). Weixin Yao is Associate Professor, Department of Statistics, University of California, Riverside, California, USA (e-mail: weixin.yao@ucr.edu). Guangren Yang is Associate Professor, Department of Statistics, School of Economics, Jinan University, Guangzhou, 510632, China (e-mail: tygr@jnu.edu.cn).

nonparametrically. However, since the degrees of freedom of the aforementioned models are difficult to define, the selection of the number of components remains an issue.

The need for semiparametric mixture models also comes from practice. In order to detect differentially expressed genes under two or more conditions in microarray data, Bordes, Delmas and Vandekerkhove (2006) proposed a semiparametric two-component mixture model (2.14) in which one component was known. Practically, a real-valued test statistic was calculated for each gene. Under the null hypothesis, each test statistic should have a known distribution F_0 and an unknown distribution F. The collected sample should come from a two-component mixture model with F_0 and F as its component distributions. Semiparametric mixture models are also needed in economics. Since the scatter plot of HPI change and GDP growth shows different patterns in different macroeconomic cycles, and since the patterns are clearly not linear, Huang, Li and Wang (2013) proposed a semiparametric mixture of regressions model (3.9). To model Return on Equity (ROE), Huang et al. (2018b) studied a special two-component model (2.14) to account for the fact that the earnings included in ROE is comprised of real earnings and manipulated earnings. In their model, the known component F_0 is assumed to be Pareto.

Both theoretically and practically, many semiparametric mixture models have been developed and demonstrated to have superior performance during the last few years. In Section 2, we will present a systematic overview of semiparametric mixture of *location* models, and in Section 3, we will discuss semiparametric mixture of *regression* models. For consistency purposes, we use the same notation system throughout the article, which might not be the same as the original articles. We conclude the article with a discussion section.

2. MIXTURE OF LOCATIONS

2.1 Introduction

Consider a C-component mixture model

$$g(\mathbf{x}) = \sum_{c=1}^{C} \pi_c f_c(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \tag{2.1}$$

where f_c's are unknown component densities and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)^\top$ is a vector of unknown mixture proportions satisfying $\pi_c > 0$ for all c and $\sum_{c=1}^{C} \pi_c = 1$. When C is unknown, the selection of C could cause the convergence rate of the maximum likelihood estimator (MLE) to vary, and therefore, is a crucial topic. See, for example, Leroux (1992), Dacunha-Castelle and Gassiat (1999), and Lemdani and Pons (1999) for more discussion on this topic in the parametric setup.

When the unknown component densities are modeled nonparametrically, (2.1) is referred to as a semiparametric mixture model by Bordes, Chauveau and Vandekerkhove (2007) and Benaglia, Chauveau and Hunter (2009). Finite mixture models with nonparametric components are very flexible, but are generally not identifiable without additional restrictions. Hall and Zhou (2003) showed, under some technical conditions, the identifiability of model (2.1) when $C = 2$, $d \geq 3$, and $f_c(\mathbf{x})$ is expressed as a product of d component-specific marginal densities of $\mathbf{x}$.

2.2 $d = 1$, Semiparametric Location-Shifted Mixture Model

When $d = 1$, researchers imposed shape restrictions, such as symmetry, on f_c's. Let $f_c(x) = f(x - \mu_c)$ and f be symmetric about the origin. Then mixture model (2.1) becomes

$$g(x) = \sum_{c=1}^{C} \pi_c f(x - \mu_c), \quad x \in \mathbb{R}. \tag{2.2}$$

Denote $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_C, \mu_1, \ldots, \mu_C)^\top$. Theoretical studies by Bordes, Mottelet and Vandekerkhove (2006) and Hunter, Wang and Hettmansperger (2007) showed that (2.2) is identifiable for $C \leq 3$ under some conditions. Specifically, when $C = 2, \pi \notin \{0, 1/2, 1\}$ and $\mu_1 \neq \mu_2$, identifiability holds for the following two-component location-shifted mixture model:

$$\begin{aligned} g(x) &= \pi f(x - \mu_1) \\ &\quad + (1 - \pi) f(x - \mu_2), \quad x \in \mathbb{R}. \end{aligned} \tag{2.3}$$

Assuming the component distributions to be symmetric, Bordes, Mottelet and Vandekerkhove (2006) proposed a cumulative distribution function (CDF) based M-estimation method to estimate the Euclidean and functional parts separately, and proved that their estimators are $n^{-1/4+\alpha}$ a.s. consistent for all $\alpha > 0$. To be more specific, let $F(\cdot)$ and $G(\cdot)$ be the CDFs of $f(\cdot)$ and $g(\cdot)$, respectively. Define $A_\theta = \pi \tau_{\mu_1} + (1 - \pi)\tau_{\mu_2}$ (von Neumann, 1931) with τ_μ ($\mu \in \mathbb{R}$) being an invertible operator from L_1 to L_1. Then the CDF version of (2.3) is equivalent to $G = A_\theta F$. Let S_r be a symmetry operator defined by $S_r\{F(\cdot)\} = 1 - F(-\cdot)$. Then the condition $G = A_\theta S_r A_\theta^{-1} G$ happens if and

only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ denotes the true value of the unknown Euclidean parameter. This is in line with the identifiability result listed above.

Define the following divergence function

$$K(\boldsymbol{\theta}) = K(\boldsymbol{\theta}; G) = \int_{\mathbb{R}} \{G_\theta(X) - G(x)\}^2 \, dG(x),$$

where $G_\theta = A_\theta S_r A_\theta^{-1} G$. Bordes, Mottelet and Vandekerkhove (2006) proposed a minimum contrast estimator for $\boldsymbol{\theta}$, defined by $\arg\min_{\theta\in\Theta} K(\boldsymbol{\theta}; \hat{G}_n)$, where Θ is a compact parametric space and $\hat{G}_n$ is the empirical CDF of the sample $(X_1, \ldots, X_n)$ drawn from G_{θ_0}. F is then estimated by $\hat{F}_n = \frac{1}{2}(I + S_r) A_{\hat{\theta}_n}^{-1} \hat{G}_n$, where I is the identity operator, and $I + S_r$ is imposed to guarantee the symmetry of F.

Bordes, Chauveau and Vandekerkhove (2007) pointed out that the direct estimator of f in Bordes, Mottelet and Vandekerkhove (2006) is generally not a probability density function (PDF) and that the numerical calculation was time consuming. On the other hand, Bordes, Chauveau and Vandekerkhove (2007) proposed to estimate f in (2.2) by $f_h(x) = \frac{1}{2n} \sum_{i=1}^n \sum_{c=1}^C p_{ic} \{K_h(x - x_i + \mu_c) + K_h(x + x_i - \mu_c)\}$, obtained in an EM context, where p_{ic} is the probability that x_i comes from component c, $K_h = K(x/h)/h$ and $K(\cdot)$ is a zero-symmetric kernel density function. A generalization of the EM algorithm for model (2.2) was proposed. However, obtaining the asymptotic behavior of these estimators remains a challenge.

Define $d_n(\boldsymbol{\theta}; \hat{G}_n) = \mathcal{D}[\sum_{c=1}^C \pi_c \hat{G}_n(x + \mu_c), \sum_{c=1}^C \pi_c \{1 - \hat{G}_n(x - \mu_c)\}]$, where $\mathcal{D}\{G_1, G_2\}$ is some measure of distance between distributions G_1 and G_2. Hunter, Wang and Hettmansperger (2007) proposed to estimate $\boldsymbol{\theta}$ by minimizing $d_n(\boldsymbol{\theta}; \hat{G}_n)$. They proved that for $C = 2$ or 3, under some technical conditions, the estimator of the Euclidean parameter is asymptotically normally distributed at the $\sqrt{n}$-rate, which is faster than the estimator proposed by Bordes, Mottelet and Vandekerkhove (2006). Balabdaoui (2017) formally proved the existence of such an estimator and established its asymptotic distribution.

Butucea and Vandekerkhove (2014) applied the Fourier analysis to invert the mixture operator, and related the symmetry of f to the fact that its Fourier transform has no imaginary part. Define $f^*(u) = \int_{\mathbb{R}} e^{ixu} f(x)\, dx$ as the Fourier transform of $f(x)$, and denote $M(\boldsymbol{\theta}, u) = \pi e^{iu\mu_1} + (1-\pi) e^{iu\mu_2}$. Then, model (2.2) implies

$$\begin{aligned} g^*(u) &= \{\pi e^{iu\mu_1} + (1-\pi) e^{iu\mu_2}\} f^*(u) \\ &= M(\boldsymbol{\theta}, u) f^*(u). \end{aligned} \tag{2.4}$$

The symmetry of f implies $\operatorname{Im}\{g^*(u)/M(\boldsymbol{\theta}, u)\} = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. By building a contrast function based on the characteristic function (2.4), the parameter $\boldsymbol{\theta}$ is then estimated by "$\arg\min_{\theta\in\Theta} S_n(\theta)$," where $S_n(\theta)$ is an estimator of the contrast $S(\theta) = \int_{\mathbb{R}} \{g^*(u)/M(\theta, u)\}^2 \, dW(u)$, and W is a Lebesgue absolutely continuous probability measure supported by $\mathbb{R}$. Under simpler conditions than Hunter, Wang and Hettmansperger (2007), Butucea and Vandekerkhove (2014) proved the central limit theorem of the estimators, and showed the minimax rate for estimating f was $n^{-2\beta/(2\beta+1)}$ for some $\beta > 1/2$. The authors argued that the estimators and the convergence results could be extended to the $C \geq 3$ cases when identifiability assumptions are satisfied.

Chee and Wang (2013) proposed a semiparametric MLE approach to estimate the Euclidean parameters. Specifically, they suggested to model the unknown density f of (2.3) as

$$\begin{aligned} &\tilde{f}_h(x; Q) \\ &\quad = \frac{1}{2} \int \{K_h(x - \sigma) + K_h(x + \sigma)\} \, dQ(\sigma), \end{aligned} \tag{2.5}$$

a generalization of the kernel-based method, where Q is a mixing distribution completely unspecified. Even with fixed $\boldsymbol{\theta}$, the estimation of Q is not a simple task since it is an optimization problem over an infinite dimensional space. As shown by Lindsay (1983), the NPMLE of Q is discrete with finite support points no more than the number of distinct observations. Let

$$\hat{Q}_n = \sum_{j=1}^m w_j \delta_{\sigma_j} \tag{2.6}$$

be a discrete estimator of Q which has mass at σ_j with probability w_j for $j = 1, \ldots, m$. Replacing Q by $\hat{Q}_n$, (2.5) becomes

$$\begin{aligned} &\tilde{f}_h(x; \mathbf{w}, \boldsymbol{\sigma}) \\ &\quad = \frac{1}{2} \sum_{j=1}^m w_j \{K_h(x - \sigma_j) + K_h(x + \sigma_j)\}, \end{aligned} \tag{2.7}$$

where $\mathbf{w} = (w_1, \ldots, w_m)^\top$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$. Note that the estimator of f in Bordes, Chauveau and Vandekerkhove (2007) is actually a special case of (2.7). Replacing the unknown density f by (2.7), model (2.2) becomes

$$\tilde{g}_h(x; \boldsymbol{\theta}, \mathbf{w}, \boldsymbol{\sigma}) = \sum_{c=1}^C \pi_c \tilde{f}_h(x - \mu_c; \mathbf{w}, \boldsymbol{\sigma}). \tag{2.8}$$

The estimation of $\boldsymbol{\theta}$ and f now becomes the estimation of $\boldsymbol{\theta}$, $\mathbf{w}$, $\boldsymbol{\sigma}$ and m. The log-likelihood of (2.8) is then maximized by algorithms of Wang (2010).

Xiang, Yao and Seo (2016) studied a method that is somewhat similar to Chee and Wang (2013). Instead of (2.5), they assumed the unknown density f to be

$$\check{f}(x; Q) = \int_{\mathbb{R}^+} \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right) dQ(\sigma), \tag{2.9}$$

where $\phi(x)$ is the standard normal density, and Q is also estimated by (2.6). Then, the authors proposed to iteratively update $\boldsymbol{\theta}$ and Q in turn until convergence. With fixed $\boldsymbol{\theta}$, other variables $\mathbf{w}$, $\boldsymbol{\sigma}$ and m are estimated through a gradient based algorithm. At a given $\hat{Q}_n$, $\boldsymbol{\theta}$ is updated by a regular EM algorithm. Xiang, Yao and Seo (2016) argued that (2.9) includes a rich class of continuous distributions, and the resulting estimators are robust against outliers. In addition, this method avoids the tedious work of the selection of the tuning parameters.

Wu, Yao and Xiang (2017) proposed to estimate (2.3) by minimizing a profile Hellinger distance between the assumed semiparametric two-component location-shifted mixture model and a nonparametric kernel density estimator.

2.3 With Shape Constraints

Nonparametric shape constraints are becoming increasingly popular in semiparametric mixture models.

Chang and Walther (2007) proposed mixtures of log-concave distributions for clustering. Examples of log-concave densities include normal, Laplace, logistic, as well as gamma and beta with certain parameter constraints. Since such distributions are not restricted to any parametric assumptions, the corresponding estimation results will not suffer from model misspecification. In addition, the estimation of log-concave distributions does not involve any tuning parameters. Chang and Walther (2007) assumed that each component in (2.1) is log-concave, that is, $\log f_c(x)$ is a concave function. Note that the corresponding log-likelihood is a concave function, and so the existence of a MLE is guaranteed. The computation algorithm consists of two parts. The first part computes the MLE of a Gaussian mixture through an EM algorithm. Define $\hat{\pi}_c$ and $\hat{f}_c$ as the MLEs, and

$$p_{ic} = \frac{\hat{\pi}_c \hat{f}_c(X_i)}{\sum_{c'=1}^{C} \hat{\pi}_{c'} \hat{f}_{c'}(X_i)} \tag{2.10}$$

as the classification probability of the ith observation belonging to the cth component. The second part of the algorithm also involves an EM algorithm. In the E-step, (2.10) is calculated with $\hat{f}_c(\cdot)$ replaced by the log-concave MLE $\tilde{f}_c(\cdot)$. The computation for $\tilde{\pi}_c$ in the M-step is still $\tilde{\pi}_c = \sum_{i=1}^{n} p_{ic}/n$, and p_{ic} is used as weights for X_i when the log-concave MLE $\tilde{f}_c$ was computed using the methods developed in Walther (2002) and Rufibach (2007). Their simulation study shows that only five iterations are required in the second part of the algorithm. The model is then extended to the *multivariate* situation. Assume $(N_1, \ldots, N_d)$ to be a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ, and $F_1, \ldots, F_d$ be CDFs of arbitrary univariate log-concave distributions. Then, within a component, observations $(X_{i1}, \ldots, X_{id})^\top \in \mathbb{R}^d$ are assumed to have density $(F_1^{-1}\Phi(N_1), \ldots, F_d^{-1}\Phi(N_d))$, where Φ denotes the CDF of the standard normal distribution. The joint density for the cth component is then defined as

$$f_c(x_1, \ldots, x_d) = \phi_{\mathbf{0},\Sigma}\{\Phi^{-1}F_1(x_1), \ldots, \Phi^{-1}F_d(x_d)\} \times \prod_{j=1}^{d} \frac{f_j(x_j)}{\phi_{\mathbf{0},\mathbf{I}}\{\Phi^{-1}F_j(x_j)\}},$$

where $\phi_{\boldsymbol{\mu},\Sigma}$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance Σ. The resulting EM algorithm is quite similar to the univariate case, and thus is omitted here. However, without symmetry, a main drawback of such a model is that it suffers from non-identifiability.

Hu, Wu and Yao (2016) proposed a log-concave maximum likelihood estimator (LCMLE) to estimate the mixture densities and provided its theoretical properties. It is assumed that $(X_1, \ldots, X_n)$ are independent d-dimensional random variables whose mixture distribution belongs to

$$\mathcal{G}_\eta = \left\{ g : g(x) = \sum_{c=1}^{C} \pi_c \exp\{\phi_c(x)\} \right\}, \tag{2.11}$$

where $\phi = (\phi_1, \ldots, \phi_C) \in \Phi_\eta$ and Φ_η is a constrained parameter space defined by $\Phi_\eta = \{(\phi_1, \ldots, \phi_C) : \phi_c$ is concave, $|S(\phi)| \geq \eta > 0$ for some $\eta \in (0, 1]\}$. Here, $M_c(\phi) = \max_{x\in\mathbb{R}^d}\{\phi_c(x)\}$, $M_{(1)}(\phi) = \min_c\{M_c(\phi)\}$, $M_{(C)}(\phi) = \max_c\{M_c(\phi)\}$, and $S(\phi) = M_{(1)}(\phi)/M_{(C)}(\phi)$. The LCMLE is then defined as

$$g_n = \arg\max_{g\in\mathcal{G}_\eta} \int \log(g)\, dQ_n,$$

where Q_n is the empirical distribution of $(X_1, \ldots, X_n)$. Hu, Wu and Yao (2016) proved the existence of the LCMLE for the log-concave mixture models and the consistency of the estimated mixture density.

Balabdaoui and Doss (2018) discussed the estimation and inference for mixtures of log-concave distributions assuming symmetry. Under some technical conditions, if the location and mixing estimators are $\sqrt{n}$-consistent, then the nonparametric log-concave MLE converges to the true symmetric density at the (usual) $n^{-2/5}$-rate in the L_1 distance.

Al Mohamad and Boumahdaf (2018) considered a semiparametric two-component mixture model where one component is parametric and the other is from a distribution family with linear constraints. A new estimation method is proposed which incorporates a prior linear information about the distribution of the unknown component and is based on ϕ-divergences. When the proportion of the parametric component is very low and the moment constraints hold, this method shows better performance than existing methods.

2.4 $d > 1$

When multivariate covariates $\mathbf{x} \in \mathbb{R}^d$ $(d > 1)$ are considered, a common restriction placed on f_c is that each joint density f_c is equal to the product of its marginal densities. In other words, the coordinates of the $\mathbf{x}$ vector are independent, conditional on the subpopulation or component from which $\mathbf{x}$ is drawn. Therefore, model (2.1) becomes

$$g(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \prod_{j=1}^{d} f_{cj}(x_j). \tag{2.12}$$

Hall and Zhou (2003) showed that when $C = 2$ and $d > 2$, identifiability of model (2.12) can typically be achieved. Allman, Matias and Rhodes (2009) proved that if the density functions $f_{1j}, \ldots, f_{Cj}$ are linearly independent except possibly on a set of Lebesgue measure zero, the parameters in (2.12) are identifiable whenever $d > 2$.

Benaglia, Chauveau and Hunter (2009) considered a more general case of (2.12) by assuming that the coordinates of $\mathbf{x}$ are conditionally independent and that there be blocks of coordinates with identical density. They proposed an EM-like estimation method. If all the blocks are of size 1, such as the setting in model (2.12), then the coordinates in $\mathbf{x}_i$ are conditionally independent but with different distributions. If there only exists one block, then the coordinates are not only conditionally independent but also identically distributed, that is, $f_{c1}(\cdot) = \cdots = f_{cd}(\cdot)$. Let b_j denote the block to which the jth coordinate belongs, where $1 \le b_j \le B$, and B is the total number of such blocks. Then model (2.1) becomes

$$g(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \prod_{j=1}^{d} f_{cb_j}(x_j). \tag{2.13}$$

At the tth iteration $(t = 1, 2, \ldots)$, in the E-step, the "posterior" probabilities of component inclusion $p_{ic}^{(t)}$, conditional on the current estimators, are calculated in the same manner as in any regular EM algorithms. In the M-step, the algorithm updates the mixing proportion by $\pi_c^{(t+1)} = n^{-1} \sum_{i=1}^{n} p_{ic}^{(t)}$, and the density as

$$f_{cl}^{(t+1)}(u) = \frac{1}{nhC_l\pi_c^{(t+1)}} \times \sum_{j=1}^{d} \sum_{i=1}^{n} p_{ic}^{(t)} I\{b_j = l\} K\left(\frac{u - x_{ij}}{h}\right),$$

for $c = 1, \ldots, C, l = 1, \ldots, B$, where $C_l = \sum_{j=1}^{d} I\{b_j = l\}$ is the number of coordinates in the lth block. However, the authors did not discuss the theoretical properties of the estimators or of the algorithm.

To improve the work of Benaglia, Chauveau and Hunter (2009), Levine, Hunter and Chauveau (2011) introduced a smoothed log-likelihood function which replaces the component density function $f_c(x)$ with a nonlinear smoother $\mathcal{N} f_c(\mathbf{x}) = \exp \int K_h^d(\mathbf{x} - \mathbf{u}) \times \log f_c(\mathbf{u})\, d\mathbf{u}$, where $K_h^d(\mathbf{u}) = h^{-d} \prod_{j=1}^{d} K^d(u_j/h)$, $K^d(u) = \prod_{j=1}^{d} K(u_j)$, $\mathbf{u} = (u_1, \ldots, u_d)^\top$. Then the new EM algorithm is proved to have the monotonicity property similar to the manner of an maximization-minimization (MM) algorithm.

Chauveau, Hunter and Levinez (2015) extended an algorithm to estimate the parameters in nonparametric multivariate finite mixture models assuming conditional independence. Similar to the work of Benaglia, Chauveau and Hunter (2009), Chauveau, Hunter and Levinez (2015) also assumed that conditionally independently and identically distributed coordinates belong to the same block. The algorithm is quite similar to the one of Benaglia, Chauveau and Hunter (2009), except that they used different bandwidths for each component and block. Applying the smoothed log-likelihood, similarly to what Levine, Hunter and Chauveau (2011) did, Chauveau, Hunter and Levinez (2015) proved that the algorithm attains the ascent property of a typical EM algorithm. Additionally, due to the algorithm's good properties and ease of calculation, the authors further extended it to the univariate model (2.2).

2.5 $d = 1$, $C = 2$ with a Known Component

Consider the following two-component mixture model:

$$g(x) = (1-\pi) f_0(x) + \pi f(x-\mu), \quad x \in \mathbb{R}, \tag{2.14}$$

where f_0 is a known PDF, $f \in \mathscr{F}$, where $\mathscr{F} = \{f : f \geq 0, \int f(x)\,dx = 1 \text{ and } f(-x) = f(x)\}$, and $\boldsymbol{\theta} = (\pi, \mu)^\top$ are the unknown parameters. Model (2.14) is motivated by the detection of differentially expressed genes under two or more conditions in microarray data analysis (Bordes, Delmas and Vandekerkhove, 2006) and the sequential clustering algorithm (Song, Nicolae and Song, 2010), and is also commonly used in contamination problems in astronomy and biology, among other fields (Patra and Sen, 2016). It is an extension of the classical two-component mixture models in the sense that the unknown component f is not restricted to any distribution families, but assumed only to be symmetric. In the parametric setup, this model is sometimes referred to as a contamination model.

Bordes, Delmas and Vandekerkhove (2006) showed the identifiability of model (2.14) when f has third-order moment and is zero-symmetric. Similar to Bordes, Mottelet and Vandekerkhove (2006), the inversion of the CDF of model (2.14) leads to

$$F(x) = \frac{1}{\pi}\{G(x+\mu) - (1-\pi)F_0(x+\mu)\}, \tag{2.15}$$

where F_0 is the CDF of f_0. Define

$$H_1(x;\mu,m,G) = \frac{\mu}{m}G(x+\mu) + \frac{m-\mu}{m}F_0(x+\mu),$$

$$H_2(x;\mu,m,G) = 1 - \frac{\mu}{m}G(\mu-x) + \frac{\mu-m}{m}F_0(\mu-x),$$

where m is the first-order moment of G. Then, by the symmetry of F, the estimator of μ is defined as $\hat{\mu}_n = \arg\min_\mu d\{H_1(\cdot;\mu,\hat{m}_n,\hat{G}_n), H_2(\cdot;\mu,\hat{m}_n,\hat{G}_n)\}$, where d is the L_q distance, and $\hat{G}_n$ and $\hat{m}_n$ are the empirical versions of G and m, derived from a sample of size n. Then, $\hat{\pi}_n = \hat{m}_n/\hat{\mu}_n$. However, the estimator was shown to be numerically unstable and the theoretical properties were not shown.

Similar to Bordes, Delmas and Vandekerkhove (2006), Bordes and Vandekerkhove (2010) also considered (2.15), and defined

$$H_1(x;\boldsymbol{\theta},G) = \frac{1}{\pi}G(x+\mu) + \frac{1-\pi}{\pi}F_0(x+\mu),$$

$$H_2(x;\boldsymbol{\theta},G) = 1 - \frac{1}{\pi}G(\mu-x) + \frac{1-\pi}{\pi}F_0(\mu-x).$$

However, Bordes and Vandekerkhove (2010) considered

$$d(\boldsymbol{\theta}) = \int_{\mathbb{R}} H^2(x;\boldsymbol{\theta},G)\,dG(x),$$

where $H(x;\boldsymbol{\theta},G) = H_1(x;\boldsymbol{\theta},G) - H_2(x;\boldsymbol{\theta},G)$. In order to estimate $\boldsymbol{\theta}$ by a differentiable optimization routine, another empirical version of d is defined as

$$d_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n H^2(X_i;\boldsymbol{\theta},\tilde{G}_n),$$

where $\tilde{G}_n(x) = \int_{-\infty}^x \hat{g}_n(t)\,dt$ is a smoothed version of $\hat{G}_n$ and $\hat{g}_n(x) = \frac{1}{nh}\sum_{i=1}^n K(\frac{x-X_i}{h})$. By these improvements, the authors showed the asymptotic normality of the estimators.

Maiboroda and Sugakova (2011) considered a generalized estimating equations (GEE) method to estimate the Euclidean parameters of model (2.14). Let $(X_1,\ldots,X_n)$ be a sample generated from (2.14), and z, z_0 and δ be three random variables such that $z \sim f$, $z_0 \sim f_0$ and $\delta \sim \mathrm{B}(1,\pi)$. Then, $X_i \sim \delta(z+\mu) + (1-\delta)z_0$. Denote $h_j (j = 1, 2)$ as two odd functions, and let $H_j(\mu) = \mathrm{E}h_j(z_0 - \mu)$ for any $\mu \in \mathbb{R}$. It is easy to see that $\mathrm{E}h_j(X_i - \mu) = \pi\mathrm{E}h_j(z) + (1-\pi)H_j(\mu) = (1-\pi)H_j(\mu)$ where the second equality is derived directly by the oddness of h_i and the symmetry of f. Motivated by this, the authors proposed the following unbiased estimating equations for the estimation of $\boldsymbol{\theta}$:

$$\begin{cases} \hat{h}_1(\mu) - (1-\pi)H_1(\mu) = 0, \\ \hat{h}_2(\mu) - (1-\pi)H_2(\mu) = 0, \end{cases}$$

where $\hat{h}_j(\mu) = n^{-1}\sum_{i=1}^n h_j(X_i - \mu)$. Maiboroda and Sugakova (2011) proved, under mild conditions, the consistency and the asymptotic normality of their estimators.

Patra and Sen (2016) studied model (2.14) without assuming the symmetry of f. The article uses ideas from shape restricted function estimation and develops "tuning parameter free" estimators that are easy to implement and have good finite sample performance. Consider the first estimator of the unknown CDF F,

$$\hat{F}(x;\pi) = \frac{\hat{G}_n(x) - (1-\pi)F_0(x)}{\pi},$$

where $\hat{G}_n$ is the empirical CDF. This estimator is easy to calculate but is not guaranteed to satisfy the conditions of a distribution function: lying between 0 and 1 and nondecreasing. Therefore, a second estimator of F is proposed as $\tilde{F}(x;\pi)$, which is the minimizer of $\frac{1}{n}\sum_{i=1}^n\{W(X_i) - \hat{F}(X_i;\pi)\}^2$ over all distribution

functions W. Since the two estimators indeed all depend on π, the authors suggested estimating π by

$$\hat{\pi} = \inf\left\{ p \in (0,1) : p d_n\{\hat{F}(x;p), \tilde{F}(x;p)\} \le \frac{c_n}{\sqrt{n}} \right\},$$

where c_n is a sequence of constants and d_n stands for the L_2 distance. It is shown that the estimating procedure is consistent for a broad range of c_n. In addition, the "elbow" of $pd_n(\hat{F}(x;p), \tilde{F}(x;p))$, that is, the point that has the maximum curvature, is a good estimator of π, and is free of tuning. Once an estimator of π is decided, say $\hat{\pi}_n$, then it is natural to estimate F by $\tilde{F}(\cdot;\hat{\pi}_n)$.

There are some generalizations of model (2.14). For example, Hohmann and Holzmann (2013) studied

$$g(x) = (1-\pi) f_0(x-\nu) + \pi f(x-\mu), \quad x \in \mathbb{R},$$

where ν is another nonnull location parameter. They showed identifiability under assumptions made on the tails of the characteristic function for the true underlying mixture. The authors applied methodologies quite similar to Bordes and Vandekerkhove (2010), and constructed asymptotically normally distributed estimators.

Xiang, Yao and Wu (2014) and Ma and Yao (2015) studied another transformation of model (2.14), assuming f_0 to be known with an unknown parameter. That is,

$$g(x;\boldsymbol{\theta}, f) = (1-\pi) f_0(x;\xi) + \pi f(x-\mu), \quad x \in \mathbb{R}, \tag{2.16}$$

where ξ is an unknown parameter and $\boldsymbol{\theta} = (\pi, \mu, \xi)^\top$. Ma and Yao (2015) studied the identifiability conditions of model (2.16) and proposed a general class of estimation equations based estimators, who have a nice connection to the most efficient estimator in the sense of semiparametric efficiency. The estimator of Xiang, Yao and Wu (2014) is based on the minimum profile Hellinger distance. Define the Hellinger distance between two functions g_1, g_2 as

$$d_H(g_1, g_2) = \|g_1^{1/2} - g_2^{1/2}\|,$$

where $\|\cdot\|$ denotes the L_2-norm. It is a natural idea to estimate $\boldsymbol{\theta}$ and f by minimizing $d_H\{g(\cdot;\boldsymbol{\theta}, f), \hat{g}_n\}$ over $\boldsymbol{\theta} \in \Theta$ and $f \in \mathcal{F}$, where $\hat{g}_n$ is a nonparametric kernel density estimator of the data. Note that this optimization problem involves both the parametric part $\boldsymbol{\theta}$ and the nonparametric part f. Therefore, the authors suggest to apply the profile idea to implement the calculation. First, for any $\boldsymbol{\theta}$, define $f(\boldsymbol{\theta}, \hat{g}_n) = \arg\min_{l \in \mathcal{F}} d_H\{g(\cdot;\boldsymbol{\theta}, l), \hat{g}_n\}$. Then the minimum profile Hellinger distance estimator of $\boldsymbol{\theta}$ is defined as $\hat{\boldsymbol{\theta}}_H = \arg\min_{\boldsymbol{\theta} \in \Theta} d_H[g\{\cdot;\boldsymbol{\theta}, f(\boldsymbol{\theta}, \hat{g}_n)\}, \hat{g}_n]$. The algorithm works by iterating between updating the parameter $\boldsymbol{\theta}$ and updating the nonparametric function f. Xiang, Yao and Wu (2014) further showed the asymptotic normality of the estimator.

Assuming that f_0 follows a Pareto distribution with unknown parameter ξ, Huang et al. (2018b) proposed another special case of model (2.16). The identifiability is discussed, and a novel estimation method is studied using smoothed likelihood and profile-likelihood techniques. A smoothing kernel $K_{h,\mu}(x,t) = (2h)^{-1}[K\{(x-t)/h\} + K\{(2\mu - x - t)/h\}]$ is defined, which is μ-symmetric, and correspondingly a nonlinear smoothing operator for $f(\cdot)$ is defined as

$$\mathcal{N}_\mu f(x) = \exp\left\{ \int K_{h,\mu}(x,t) \log f(t)\, dt \right\}.$$

Replacing f by its nonlinear smoother, the smoothed log-likelihood of a data is then defined as

$$\ell(\mu, \pi, \xi, f) = \sum_{i=1}^{n} \log\{(1-\pi) f_0(X_i;\xi) + \pi \mathcal{N}_\mu f(X_i)\}.$$

The authors proposed an estimation method that separates μ from π, ξ and f. Given a known μ, or an estimator of it, denoted by μ_0, the maximum likelihood estimator of π, ξ and f can be calculated by maximizing $\ell(\mu_0, \pi, \xi, f)$ via an EM algorithm. Denote the estimators by $\hat{\pi}_\mu$, $\hat{\xi}_\mu$, and $\hat{f}_\mu(\cdot)$. Then, the estimator of μ is calculated through maximizing the profile likelihood $\hat{\ell}_p(\mu) = \ell(\mu, \hat{\pi}_\mu, \hat{\xi}_\mu, \hat{f}_\mu)$, using some advanced numerical methods.

Nguyen and Matias (2014) studied a special case of (2.14) when $f_0(\cdot) = 1$, and proved an impossibility result. They showed that the quadratic risk of any estimator of π does not have a parametric convergence rate when f is not 0 on any nonempty interval. This happens mainly because the Fisher information for the model is 0 when f is bounded away from 0 for all nonempty intervals. We conjecture that such results might also hold for model (2.2) and model (2.12), which could be an interesting topic for future work.

3. SEMIPARAMETRIC MIXTURE OF REGRESSIONS

3.1 Introduction

Assume $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is a random sample from $(\mathbf{x}, Y)$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top (p < n)$ is a

vector of predictors. The goal of a typical finite mixture of regressions (FMR) model is to describe the conditional distribution of $Y_i|\mathbf{x}_i$ using a mixture of linear regressions with assumed Gaussian errors. That is, let $\mathcal{C}$ be a latent class index random variable with $P(\mathcal{C}=c|\mathbf{x})=\pi_c$ for $c=1,\ldots,C$. Given $\mathcal{C}=c$, suppose that the response y depends on $\mathbf{x}$ in a linear way $y=\mathbf{x}^\top\boldsymbol{\beta}_c+\varepsilon_c$, where $\varepsilon_c\sim N(0,\sigma_c^2)$. Then the conditional distribution of Y given $\mathbf{x}$ is

$$Y|\mathbf{x}\sim\sum_{c=1}^{C}\pi_c\phi\big(y|\mathbf{x}^\top\boldsymbol{\beta}_c,\sigma_c^2\big), \tag{3.1}$$

where $\boldsymbol{\theta}=(\pi_1,\ldots,\pi_C,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_C,\sigma_1^2,\ldots,\sigma_C^2)^\top$ is the vector of parameters, $\phi(y|\mu,\sigma^2)$ is the normal density with mean μ and variance σ^2, $0\le\pi_c\le1$, and $\sum_{c=1}^{C}\pi_c=1$. See McLachlan and Peel (2000) for comprehensive discussions.

3.2 Mixture of Regression Models with Varying Proportions

In a parametric mixture of regressions model, the mixing proportions are assumed to be known and fixed as π_c, $c=1,\ldots,C$. However, if the covariates $\mathbf{x}$ contain some information about the relative weights, then model (3.1) is mistakenly specified and might provide misleading results. In the following paragraphs, several FMR models with varying proportions are discussed. The error density is assumed to be known throughout the section.

The first model is

$$Y|\mathbf{x}\sim\sum_{c=1}^{C}\pi_c(\mathbf{x})\phi\big(y|\mathbf{x}^\top\boldsymbol{\beta}_c,\sigma_c^2\big), \tag{3.2}$$

whose identifiability was discussed by Huang and Yao (2012) under some mild conditions. If $\pi_c(\mathbf{x})$ is modeled as a logistic function, then model (3.2) becomes the hierarchical mixtures of experts (HME; Jacobs, Peng and Tanner, 1997) in the neural network setting. Young and Hunter (2010), on the other hand, modeled $\pi_c(\mathbf{x})$ as

$$\pi_c(\mathbf{x}_i)=E[z_{ic}|\mathbf{x}_i], \tag{3.3}$$

where z_{ic} is a component indicator variable that is 1 if the ith observation is from the cth component, and 0 otherwise. Note that if one treats z_{ic} as a response, then (3.3) indicates nothing but a mean structure in a regression analysis. Therefore, Young and Hunter (2010) proposed to estimate $\pi_c(\mathbf{x}_i)$ by local polynomial regression (Fan and Gijbels, 1996) as

$$\begin{aligned}&\arg\min_{\boldsymbol{\alpha}}\sum_{l=1}^{n}K_{\mathbf{h}}(\mathbf{x}_i-\mathbf{x}_l)\\&\quad\times\left\{z_{ic}-\left(\alpha_0+\sum_{t=1}^{p}\alpha_t(x_{i,t}-x_{l,t})\right)\right\}^2,\end{aligned} \tag{3.4}$$

where $\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\ldots,\alpha_p)^\top$, and $K_{\mathbf{h}}(\mathbf{x}_i-\mathbf{x}_l)$ is a multivariate kernel density function. However, since z_{ic} is not known in reality, they proposed to run the EM algorithm for model (3.1) first, and then use the converged value of the classification probability, denoted by p_{ic}^{∞}, to replace z_{ic}. Given estimators of $\pi_c(\mathbf{x})$, $\boldsymbol{\beta}_c$ and σ_c can then be estimated through a regular EM algorithm. However, due to the "curse of dimensionality," the authors only did simulation study for the $p=1$ case, and argued that extra cautions should be given for high-dimensional predictors cases. Theoretical results were not discussed for this method.

Huang and Yao (2012), on the other hand, studied $\pi(\mathbf{x})$ fully nonparametrically, and proposed a one-step backfitting procedure to achieve the optimal convergence rates for both the regression parameters and the nonparametric functions of mixing proportions. They further derived the asymptotic bias and variance of the one-step estimator.

Allowing the response to come from other distribution families, model (3.2) can be extended to a mixture of GLMs with varying proportions (Wang, Yao and Huang, 2014):

$$Y|\mathbf{x}\sim\sum_{c=1}^{C}\pi_c(\mathbf{x})\,f_c(y|\mathbf{x},\boldsymbol{\theta}_c), \tag{3.5}$$

where f_c is a function of the exponential family whose mean is $\mu_c(\mathbf{x})=g_c^{-1}(\mathbf{x}^\top\boldsymbol{\beta}_c)$ and $g_c(\cdot)$ is a component-specific link function. For example, when a binomial response Y is considered, Cao and Yao (2012) studied a special case of (3.5), where both the component proportions and the success probabilities depend on the predictor nonparametrically. That is,

$$\begin{aligned}Y|_{X=x}\sim{}&\pi_1(x)\,\mathrm{Bin}(y;N,0)\\&+\pi_2(x)\,\mathrm{Bin}\big\{y;N,p(x)\big\},\end{aligned} \tag{3.6}$$

where $\pi_1(x)+\pi_2(x)=1$, and $\mathrm{Bin}(Y;N,p)$ denotes the probability mass function of a binomially distributed random variable Y with the number of trials N and success probability p. Note that the first component is a degenerate distribution with mass 1 on 0.

Therefore, model (3.6) has wide application in data with extra number of zeros. Cao and Yao (2012) successfully applied model (3.6) to a rain dataset from a global climate model and a historical rain dataset from Edmonton, Canada.

3.3 Nonparametric Errors

Traditional FMR models (3.1) are all based on the assumption of normally distributed errors. The estimation results might be biased or even misleading if this assumption is problematic. Different methods have been proposed to relax this assumption in the following article.

Hunter and Young (2012) studied a FMR model where linearity is still assumed within each component, but instead of normality, the error terms are modeled fully nonparametrically as $\varepsilon_i \sim g$. That is,

$$Y|\mathbf{x} \sim \sum_{c=1}^{C} \pi_c g\big(y - \mathbf{x}^\top \boldsymbol{\beta}_c\big). \tag{3.7}$$

Without loss of generality, g is assumed to have median 0. The identifiability of model (3.7) can be achieved whenever the regression planes are not parallel. If some further conditions are put on g, then (3.7) can still be identifiable even when the regression planes are parallel. Similar to Levine, Hunter and Chauveau (2011), Hunter and Young (2012) proposed an estimation procedure that maximizes the following smoothed log-likelihood:

$$\ell_s(\boldsymbol{\pi}, \boldsymbol{\beta}, g) = \sum_{i=1}^{n} \log\left\{\sum_{c=1}^{C} \pi_c \mathcal{N}_h g\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_c\big)\right\},$$

where $\mathcal{N}_h g = \exp \int h^{-1} K\{(x-u)/h\} \log g(u)\, du$ is a nonlinear smoother. The effectiveness of the new methods was demonstrated through numerical studies.

Ma et al. (2018) extended the identifiability result for model (3.7) by allowing different error densities for each component. They established the consistency and asymptotic normality of their estimators and of those by Hunter and Young (2012).

Hu, Yao and Wu (2017) assumed the error densities to be log-concave. That is, the model has the same form as (3.7), where $g_c(x) = \exp\{\phi_c(x)\}$ for some unknown concave function $\phi_c(x)$.

The three articles studied above all focus on the mean regressions. By regressing the conditional quantiles (such as median) on the covariates without any parametric assumptions, Wu and Yao (2016) studied a semiparametric mixture of quantile regressions model. Given $\mathcal{C} = c$,

$$Y = \mathbf{x}^\top \boldsymbol{\beta}_c(\tau) + \varepsilon_c(\tau), \tag{3.8}$$

where $\boldsymbol{\beta}_c(\tau) = (\beta_{0c}(\tau), \ldots, \beta_{pc}(\tau))^\top$ is the τth quantile regression coefficient for the cth component. The only assumption on the error density $g_c(\cdot)$ is that the τth quantile is zero. Model (3.8) is believed to be more robust than regular FMR model, and could reveal more detailed data structure. Wu and Yao (2016) proposed an EM-type algorithm which incorporates the kernel regression to estimate the parameters and error densities.

3.4 Semiparametric Mixtures of Nonparametric Regressions

In the traditional FMR model (3.1) and the models discussed above, linearity is always assumed in the mean functions. In the following, different models have been proposed to relax this assumption.

Motivated by a US house price index dataset, Huang, Li and Wang (2013) proposed the following model:

$$Y|_{X=x} \sim \sum_{c=1}^{C} \pi_c(x) N\big\{m_c(x), \sigma_c^2(x)\big\}, \quad x \in \mathbb{R}, \tag{3.9}$$

where $\pi_c(\cdot)$, $m_c(\cdot)$, and $\sigma_c^2(\cdot)$ are unknown but smooth functions, and $\sum_{c=1}^{C} \pi_c(\cdot) = 1$. Note that the errors are assumed to follow a normal distribution, and model (3.9) is still considered as a semiparametric mixture model. Since there are nonparametric functions, kernel regression is used in a modified EM algorithm. Specifically, like any regular EM algorithm, at $(t+1)$th iteration $(t = 1, 2, \ldots)$, a "posterior" probability is calculated and labeled as $p_{ic}^{(t+1)}$, based on current estimators. Then, at the M-step, to update the estimators, the following local objective function is maximized with respect to π_c, m_c and σ_c:

$$\sum_{i=1}^{n} \sum_{c=1}^{C} p_{ic}^{(t+1)} \big[\log \pi_c + \log \phi\big\{Y_i | m_c, \sigma_c^2\big\}\big] K_h(X_i - x).$$

Although model (3.9) is very flexible, it lacks efficiency. Taking both matters into account, Xiang and Yao (2018) suggested a new model by assuming the mixing proportions and variances to be constant. The model is defined as

$$Y|_{X=x} \sim \sum_{c=1}^{C} \pi_c \phi\big\{y | m_c(x), \sigma_c^2\big\}, \tag{3.10}$$

where $m_c(\cdot)$s are the unknown smooth functions. Due to the coexistence of both global and local parameters, model (3.10) is more difficult to estimate. An efficient one-step backfitting estimation procedure, similar to the ones discussed in Huang and Yao (2012) and Cao

and Yao (2012), was proposed. A generalized likelihood ratio test was also proposed to compare between model (3.9) and model (3.10), and was shown to have the Wilks type of phenomenon.

Similar to the issue discussed in the previous section, model (3.9) and model (3.10) are not suitable for data with high-dimensional predictors due to the application of kernel regression in the estimation procedure. As a result, Xiang and Yao (2017) studied a series of FMR models with single-index. First, replacing the one-dimensional covariate x in (3.9) by $\boldsymbol{\alpha}^\top \mathbf{x}$, a mixture of single-index models (MSIM) is defined as

$$Y|\mathbf{x} \sim \sum_{c=1}^{C} \pi_c(\boldsymbol{\alpha}^\top \mathbf{x}) \phi\{y|m_c(\boldsymbol{\alpha}^\top \mathbf{x}), \sigma_c^2(\boldsymbol{\alpha}^\top \mathbf{x})\}. \tag{3.11}$$

When $C = 1$, model (3.11) reduces to a single index model (Ichimura, 1993; Härdle, Hall and Ichimura, 1993). If $\mathbf{x}$ is a scalar, then model (3.11) reduces to model (3.9). Models with nonparametric means are flexible, but are difficult to estimate and interpret. Introducing single-index into the mixing proportions of model (3.2), Xiang and Yao (2017) proposed another model:

$$Y|\mathbf{x} \sim \sum_{c=1}^{C} \pi_c(\boldsymbol{\alpha}^\top \mathbf{x}) N(\mathbf{x}^\top \boldsymbol{\beta}_c, \sigma_c^2). \tag{3.12}$$

The global parameters $\boldsymbol{\alpha}, \boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_C^\top)^\top$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_C^2)^\top$ and the nonparametric functions $\boldsymbol{\pi}(\cdot) = (\pi_1(\cdot), \ldots, \pi_C(\cdot))^\top$ are estimated alternately by fixing the others.

3.5 Semiparametric Regression Models for Longitudinal/Functional Data

In this section, we introduce the applications of semiparametric mixture models to more complex data, such as longitudinal and functional data. Early works of such models can be found in Yao, Fu and Lee (2011). They extended traditional functional linear models to the framework of classical mixture regression models, and proposed a functional mixture regression model.

Rich in information, intensive longitudinal data (ILD) are becoming increasingly popular in behavioral sciences. However, since ILD are often heterogeneous and nonlinear, they are difficult to analyze. Dziak et al. (2015) proposed a mixture of time-varying effect models (MixTVCM), which incorporated time-varying effect model (TVEM) into a mixture model framework. Conditional on time-invariant subject-level covariates $s_1, \ldots, s_Q$, the probability that individual i comes from class c is

$$\pi_{ic} = P(\mathcal{C}_i = c) = \frac{\exp(\gamma_{0,c} + \sum_{q=1}^{Q} \gamma_{1qc} s_q)}{\sum_{t=1}^{k} \exp(\gamma_{0,t} + \sum_{q=1}^{Q} \gamma_{1qt} s_q)},$$

and within each component, the means are assumed to be the same as the TVEM model in Tan et al. (2012):

$$\begin{aligned}\mu_{ij} &= E(y_{ij}|\mathcal{C}_i = c) \\ &= \beta_{0c}(t_{ij}) + \beta_{10}(t_{ij}) x_{ij1} + \cdots + \beta_{pc}(t_{ij}) x_{ijP},\end{aligned}$$

where $x_1, \ldots, x_P$ are the observation-level covariates. The covariance structure of Y_{ij} is assumed to be of the form

$$\operatorname{cov}(y_{ij}, y_{ij'}) = \sigma_a^2 \rho^{|t_{ij} - t_{ij'}|} + \sigma_e^2,$$

where σ_a^2 and σ_e^2 denote the variances of subject-level and observation-level errors, respectively. Though nonparametric in means, MixTVEM is still considered as semiparametric since normality is assumed for the error distributions. In order to make the model identifiable, it is assumed that individuals are clustered into one and only one latent class. In the presence of mixture structure, the EM algorithm is used for estimation. The penalized B-spline is used to approximate $\beta(\cdot)$'s, where the penalization is considered to ensure a smooth and parsimonious shape.

To deal with inhomogeneous data collected at irregular, possibly subject-depending time points, which occurs when data are functional, Huang et al. (2014) proposed a new estimation procedure for the mixture of Gaussian processes. Conditional on $\mathcal{C} = c$, the model assumes

$$\begin{aligned}y_{ij} =& \mu_c(t_{ij}) + \sum_{q=1}^{\infty} \xi_{iqc} \nu_{qc}(t_{ij}) \\ &+ \varepsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, N_i,\end{aligned} \tag{3.13}$$

where ε_{ij}'s are i.i.d. and $N(0, \sigma^2)$ distributed, $\mu_c(t)$ is the mean of a Gaussian process with covariance function $G_c(s, t)$, and ξ_{iqc} and $\nu_{qc}(t)$ are the functional principal component (FPC) score and eigenfunctions of $G_c(s, t)$ (Karhunen-Loève theorem, Roger and Pol, 1991).

To analyze heterogeneous functional data with functional covariates, given $\mathcal{C} = c$, Wang et al. (2016) proposed to model $\{y(t), t \in T\}$ in a functional-linear way:

$$y(t) = \mathbf{X}(t)^\top \boldsymbol{\beta}_c(t) + \varepsilon_c(t), \tag{3.14}$$

where $\mathbf{X}(t)$ is a random covariate process of dimension p, and $\boldsymbol{\beta}_c(t)$ is a smooth regression coefficient function of cth component. $\varepsilon_c(t)$ is a Gaussian process with

mean zero, independent of $\mathbf{X}(t)$, and is assumed to be of the form

$$\varepsilon_c(t) = \zeta_c(t) + e(t),$$

where $\zeta(t)$ denotes a trajectory process with covariance $\Gamma_c(s,t) = \operatorname{cov}\{\xi_c(s), \xi_c(t)\}$, and $e(t)$ is the measurement error with constant variance σ^2. For ease of notation, define $y_{ij} = y_i(t_{ij})$, $j = 1, \ldots, N_i$, and similarly define ε_{cij}, e_{ij}, etc. Similar to Huang et al. (2014), by the Karhunen-Loève theorem, model (3.14) can be represented as

$$y_{ij} = \mathbf{X}_i(t_{ij})^\top \boldsymbol{\beta}_c(t_{ij}) + \sum_{q=1}^{\infty} \xi_{iqc} v_{qc}(t_{ij}) + e_{ij}, \tag{3.15}$$

where $v_{qc}(\cdot)$'s are the eigenfunctions of $\Gamma_c(s,t)$ and λ_{qc}'s are the corresponding eigenvalues, and ξ_{iqc}'s are the uncorrelated FPC of $\zeta_c(t)$ satisfying $E(\xi_{iqc}) = 0$ and $\operatorname{var}(\xi_{iqc}) = \lambda_{qc}$. Ignoring the correlation structure, y_{ij} can be thought to be coming from the following mixture of Gaussian process:

$$y(t) \sim \sum_{c=1}^{C} \pi_c N\{\mathbf{X}(t)^\top \boldsymbol{\beta}_c(t), \sigma_c^{*2}(t)\},$$

where $\sigma_c^{*2}(t) = \Gamma_c(t,t) + \sigma^2$. Then, the parameters π_c, $\boldsymbol{\beta}_c(\cdot)$, and $\sigma_c^{*2}(\cdot)$ can be estimated by an EM-type algorithm, which is very close to the one discussed above in Huang et al. (2014).

3.6 Some Additional Topics

Additionally, we explore a few more interesting topics. For example, Vandekerkhove (2013) studied a two-component mixture of regressions model where the mixing proportion, slope, intercept and error distribution of one component is unknown while the other is known. The method proposed by Vandekerkhove (2013) performs well for datasets of reasonable sizes. However, the performance is not desirable as the sample size increases, since this method is based on the optimization of a contrast function of size $O(n^2)$. Bordes, Kojadinovic and Vandekerkhove (2013) also studied the same model as Vandekerkhove (2013), and proposed a new method-of-moments estimator whose computation order is of $O(n)$. Young (2014) extended the mixture of linear regression models to incorporate changepoints by assuming one or more of the components as piecewise linear. Such a model is a great combination of the traditional mixture of linear regression models and the standard changepoint regression model. Faicel (2016) proposed a new fully unsupervised algorithm to learn regression mixture models with unknown number of components. Unlike the standard EM for mixture of regressions, this method does not require accurate initialization. Montuelle and Le Pennec (2014) studied a mixture of Gaussian regressions model with logistic weights, and proposed to estimate the number of components and other parameters through a penalized maximum likelihood approach. Butucea, Ngueyep Tzoumpe and Vandekerkhove (2017) considered a nonlinear mixture of regression models with one known component. A local estimation procedure based on the symmetry of local noise is proposed to estimate the proportion and locations functions. Huang et al. (2018a) proposed a semiparametric hidden Markov model with nonparametric regression in which the mean and variance of emission model are unknown smooth functions. See also Gassiat, Cleynen and Robin (2016), de De Castro, Gassiat and Le Corff (2017), Gassiat, Rousseau and Vernet (2018), Gassiat and Rousseau (2016), and Dannemann, Holzmann and Leister (2014) for more discussion on nonparametric/semiparametric hidden Markov models, and Gassiat (2017) for a survey of mixtures of nonparametric components and hidden Markov models. Huang et al. (2018b) investigated the identifiability and statistical inference for mixture of varying coefficient models, in which each mixture component follows a varying coefficient model and the mixing proportions and dispersion parameters are unknown smooth functions.

4. DISCUSSION

This article summarizes several semiparametric extensions of parametric mixture of locations model and regressions model. Detailed model settings and corresponding estimation methods are presented. As we have seen, this field has received much attention, but there are still a great number of questions and issues remaining unaddressed. Choosing the number of components in mixture models is an important problem which has attracted much attention in statistical research. For parametric mixture models, some popular and simple approaches involve information criteria, such as AIC or BIC, and likelihood ratio tests. See McLachlan and Peel (2000), Chen, Chen and Kalbfleisch (2004) and Chen and Li (2009) for more details. For semiparametric mixture models, however, one main difficulty lies in the definition of model complexity. Huang, Li and Wang (2013) applied the degrees of freedom derived in Fan, Zhang and Zhang (2001), and proposed an information criterion approach for model selection. It is still an open and interesting topic waiting for other attempts. In addition, since many of the models we discussed are closely connected or even nested, testing

procedures are desired for model selection in addition to data-driven methods. For example, Pommeret and Vandekerkhove (2018) investigated a semiparametric testing approach to test whether the Gaussian assumption made by McLachlan, Bean and Jones (2006) on the unknown component of their false discovery type mixture model is correct or not.

Furthermore, in some of the articles that we reviewed, such as Bordes, Chauveau and Vandekerkhove (2007), Benaglia, Chauveau and Hunter (2009) and Hunter and Young (2012), only EM-type algorithms are proposed without rigorous theoretical justifications or asymptotic properties. Due to the application of kernel density estimators, those EM-type algorithms do not possess the ascent property of a standard EM algorithm. More research is required to investigate the convergence properties of the above EM-type algorithms and establish some theoretical properties about the semiparametric mixture estimators, such as the optimal convergence rate and semiparametric efficiency. We hope that this article could inspire researchers who are interested in this field to shine more light on the topic.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, the Associate Editor and three referees for numerous helpful comments during the preparation of the article. The authors would also like to thank Jian Guo and Edward Schuberg for proofreading the manuscript. Xiang's research is supported by the NSF of China Grant 11601477. Yao's research is supported by the NSF Grant DMS-1461677 and the Department of Energy with the award DE-EE0007328. Yang's research was supported by the National Natural Science Foundation of China Grant 11471086 and 11871173, the National Social Science Foundation of China Grant 16BTJ032, the Fundamental Research Funds for the Central Universities 15JNQ019, the Science and Technology Program of Guangzhou 2016201604030074 and the Science and Technology Planning Project of Guangdong 2016A050503033.

REFERENCES

AL MOHAMAD, D. and BOUMAHDAF, A. (2018). Semiparametric two-component mixture models when one component is defined through linear constraints. *IEEE Trans. Inform. Theory* **64** 795–830. MR3762592

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

BALABDAOUI, F. (2017). Revisiting the Hodges–Lehmann estimator in a location mixture model: Is asymptotic normality good enough? *Electron. J. Stat.* **11** 4563–4595. MR3724489

BALABDAOUI, F. and DOSS, C. R. (2018). Inference for a two-component mixture of symmetric distributions under log-concavity. *Bernoulli* **24** 1053–1071. MR3706787

BENAGLIA, T., CHAUVEAU, D. and HUNTER, D. R. (2009). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *J. Comput. Graph. Statist.* **18** 505–526. MR2749842

BORDES, L., CHAUVEAU, D. and VANDEKERKHOVE, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Comput. Statist. Data Anal.* **51** 5429–5443. MR2370882

BORDES, L., DELMAS, C. and VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Stat.* **33** 733–752. MR2300913

BORDES, L., KOJADINOVIC, I. and VANDEKERKHOVE, P. (2013). Semiparametric estimation of a two-component mixture of linear regressions in which one component is known. *Electron. J. Stat.* **7** 2603–2644. MR3121625

BORDES, L., MOTTELET, S. and VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. MR2278356

BORDES, L. and VANDEKERKHOVE, P. (2010). Semiparametric two-component mixture model with a known component: An asymptotically normal estimator. *Math. Methods Statist.* **19** 22–41. MR2682853

BUTUCEA, C., NGUEYEP TZOUMPE, R. and VANDEKERKHOVE, P. (2017). Semiparametric topographical mixture models with symmetric errors. *Bernoulli* **23** 825–862. MR3606752

BUTUCEA, C. and VANDEKERKHOVE, P. (2014). Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.* **41** 227–239. MR3181141

CAO, J. and YAO, W. (2012). Semiparametric mixture of binomial regression with a degenerate component. *Statist. Sinica* **22** 27–46. MR2933166

CHANG, G. T. and WALTHER, G. (2007). Clustering with mixtures of log-concave distributions. *Comput. Statist. Data Anal.* **51** 6242–6251. MR2408591

CHAUVEAU, D., HUNTER, D. R. and LEVINEZ, M. (2015). Estimation for conditional independence multivariate finite mixture models. *Stat. Surv.* **9** 1–31.

CHEE, C.-S. and WANG, Y. (2013). Estimation of finite mixtures with symmetric components. *Stat. Comput.* **23** 233–249. MR3016941

CHEN, H., CHEN, J. and KALBFLEISCH, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 95–115. MR2035761

CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Ann. Statist.* **37** 2523–2542. MR2543701

DACUNHA-CASTELLE, D. and GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209. MR1740115

DANNEMANN, J., HOLZMANN, H. and LEISTER, A. (2014). Semiparametric hidden Markov models: Identifiability and estimation. *Comput. Statist.* **6** 418–425.

DE CASTRO, Y., GASSIAT, É. and LE CORFF, S. (2017). Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Trans. Inform. Theory* **63** 4758–4777. MR3683535

DZIAK, J. J., LI, R., TAN, X., SHIFFMAN, S. and SHIYKO, M. P. (2015). Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychol. Methods* **20** 444–469.

FAICEL, C. (2016). Unsupervised learning of regression mixture models with unknown number of components. *J. Stat. Comput. Simul.* **86** 2308–2334. MR3502164

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

FAN, J., ZHANG, C. and ZHANG, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** 153–193. MR1833962

GASSIAT, E. (2017). Mixtures of nonparametric components and hidden Markov models. In *Handbook of Mixture Analysis* (S. Fruhwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 343–360. CRC Press, Boca Raton, FL. MR3889699

GASSIAT, E., CLEYNEN, A. and ROBIN, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.* **26** 61–71. MR3439359

GASSIAT, E. and ROUSSEAU, J. (2016). Nonparametric finite translation hidden Markov models and extensions. *Bernoulli* **22** 193–212. MR3449780

GASSIAT, E., ROUSSEAU, J. and VERNET, E. (2018). Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electron. J. Stat.* **12** 703–740. MR3769193

HALL, P. and ZHOU, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224. MR1962504

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

HOHMANN, D. and HOLZMANN, H. (2013). Semiparametric location mixtures with distinct components. *Statistics* **47** 348–362. MR3043704

HU, H., WU, Y. and YAO, W. (2016). Maximum likelihood estimation of the mixture of log-concave densities. *Comput. Statist. Data Anal.* **101** 137–147. MR3504841

HU, H., YAO, W. and WU, Y. (2017). The robust EM-type algorithms for log-concave mixtures of regression models. *Comput. Statist. Data Anal.* **111** 14–26. MR3630215

HUANG, M., LI, R. and WANG, S. (2013). Nonparametric mixture of regression models. *J. Amer. Statist. Assoc.* **108** 929–941. MR3174674

HUANG, M. and YAO, W. (2012). Mixture of regression models with varying mixing proportions: A semiparametric approach. *J. Amer. Statist. Assoc.* **107** 711–724. MR2980079

HUANG, M., LI, R., WANG, H. and YAO, W. (2014). Estimating mixture of Gaussian processes by kernel smoothing. *J. Bus. Econom. Statist.* **32** 259–270. MR3207838

HUANG, M., WANG, S., WANG, H. and JIN, T. (2018a). Maximum smoothed likelihood estimation for a class of semiparametric Pareto mixture densities. *Stat. Interface* **11** 31–40. MR3690796

HUANG, M., WANG, S., YAO, W. and CHEN, Y. (2018b). Statistical inference and applications of mixture of varying coefficient models. *Scand. J. Stat.* **45** 618–643. MR3858949

HUNTER, D. R., WANG, S. and HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275

HUNTER, D. R. and YOUNG, D. S. (2012). Semiparametric mixtures of regressions. *J. Nonparametr. Stat.* **24** 19–38. MR2885823

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

JACOBS, R. A., PENG, F. and TANNER, M. A. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Netw.* **10** 231–241.

LEMDANI, M. and PONS, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli* **5** 705–719. MR1704563

LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360. MR1186253

LEVINE, M., HUNTER, D. R. and CHAUVEAU, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403–416. MR2806437

LINDSAY, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94. MR0684866

MA, Y. and YAO, W. (2015). Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electron. J. Stat.* **9** 444–474. MR3326131

MA, Y., WANG, S., XU, L. and YAO, W. (2018). Semiparametric mixture regression with unspecified error distributions. Available at arXiv:1811.01117.

MAIBORODA, R. and SUGAKOVA, O. (2011). Generalized estimating equations for symmetric distributions observed with admixture. *Comm. Statist. Theory Methods* **40** 96–116. MR2747201

MCLACHLAN, G. J., BEAN, R. W. and JONES, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22** 1608–1615.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley Interscience, New York. MR1789474

MONTUELLE, L. and LE PENNEC, E. (2014). Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electron. J. Stat.* **8** 1661–1695. MR3263134

NGUYEN, V. H. and MATIAS, C. (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Stat.* **41** 1167–1194. MR3277044

PATRA, R. K. and SEN, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 869–893. MR3534354

POMMERET, D. and VANDEKERKHOVE, P. (2018). Semiparametric false discovery rate model Gaussianity test. Available at https://hal.archives-ouvertes.fr/hal-01868272.

ROGER, G. and POL, S. (1991). *Stochastic Finite Elements: A Special Approach*. Springer, Berlin.

RUFIBACH, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Stat. Comput. Simul.* **77** 561–574. MR2407642

SONG, S., NICOLAE, D. L. and SONG, J. (2010). Estimating the mixing proportion in a semiparametric mixture model. *Comput. Statist. Data Anal.* **54** 2276–2283. MR2720488

TAN, X., SHIYKO, M. P., LI, R., LI, Y. and DIERKER, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychol. Methods* **17** 61–77.

VANDEKERKHOVE, P. (2013). Estimation of a semiparametric mixture of regressions model. *J. Nonparametr. Stat.* **25** 181–208. MR3039977

VON NEUMANN, J. (1931). Die Eindeutigkeit der Schrödingerschen Operatoren. *Math. Ann.* **104** 570–578. MR1512685

WALTHER, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.* **97** 508–513. MR1941467

WANG, Y. (2010). Maximum likelihood computation for fitting semiparametric mixture models. *Stat. Comput.* **20** 75–86. MR2578078

WANG, S., YAO, W. and HUANG, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statist. Probab. Lett.* **93** 41–45. MR3244553

WANG, S., HUANG, M., WU, X. and YAO, W. (2016). Mixture of functional linear models and its application to CO_2-GDP functional data. *Comput. Statist. Data Anal.* **97** 1–15. MR3447032

WU, Q. and YAO, W. (2016). Mixtures of quantile regressions. *Comput. Statist. Data Anal.* **93** 162–176. MR3406203

WU, J., YAO, W. and XIANG, S. (2017). Computation of an efficient and robust estimator in a semiparametric mixture model. *J. Stat. Comput. Simul.* **87** 2128–2137. MR3656095

XIANG, S. and YAO, W. (2017). Semiparametric mixtures of regressions with single-index for model based clustering. Available at arXiv:1708.04142.

XIANG, S. and YAO, W. (2018). Semiparametric mixtures of nonparametric regressions. *Ann. Inst. Statist. Math.* **70** 131–154. MR3742821

XIANG, S., YAO, W. and SEO, B. (2016). Semiparametric mixture: Continuous scale mixture approach. *Comput. Statist. Data Anal.* **103** 413–425. MR3522641

XIANG, S., YAO, W. and WU, J. (2014). Minimum profile Hellinger distance estimation for a semiparametric mixture model. *Canad. J. Statist.* **42** 246–267. MR3208338

YAO, F., FU, Y. and LEE, T. C. M. (2011). Functional mixture regression. *Biostatistics* **12** 341–353.

YOUNG, D. S. (2014). Mixtures of regressions with changepoints. *Stat. Comput.* **24** 265–281. MR3165553

YOUNG, D. S. and HUNTER, D. R. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Comput. Statist. Data Anal.* **54** 2253–2266. MR2720486