Statistics and Its Interface Volume 13 (2020) 347–359

Multi-dimensional classification with semiparametric mixture model

Anqi Yin and Ao Yuan*

Compared to non-model based classification methods, the model based classification has the advantage of classification together with regression analysis, and is the interest of our investigation. For robustness, we propose and study a semiparametric mixture model, in which each sub-density is only assumed unimodal. The semiparametric maximum likelihood estimate is used to estimate the parametric and nonparametric components. Then the Bayesian classification rule is used to classify the subjects according to the model. Large sample properties of the estimates are investigated, simulation studies are conducted to evaluate the finite sample performance of the proposed model, and then the method is applied to analyze a real data.

AMS 2000 subject classifications: Primary 62H30; secondary 62J99.
Keywords and phrases: Classification, Mixture model, Maximum likelihood estimate, Semiparametric model.

1. INTRODUCTION

Classification is an area of extensive studies, in statistics and many other fields in engineering. Statistical methods for classification are extensive, they can be parametric, nonparametric or semiparametric. Parametric models are appealing because they are easy to implement for inferences, and are efficient when the underlying distribution is correctly specified. But they may lead to biased results when the model is misspecified, and result in incorrect classification results. It is known [38, 50] that, if the model is incorrectly specified, the estimation can be biased. In classification the correct model specification is particularly important, as the commonly used classification rules, such as the Bayesian classification rule, is based on density ratio.

Also, full nonparametric methods often do not work well for relatively high dimensional data. When the data dimension is relatively high, the commonly used nonparametric estimation, such as kernel density estimator often behaves not well, and any fully pre-specified model, such as the multivariate normal model, is more or less deviated form the true one. The multivariate copula models [59, 28] may still be unsatisfactory. Although regression parameters can be estimated efficiently with nonparametric model, but the location parameter, which is crucial in classification, cannot be estimated in nonparametric model.

For parametric methods, Celeux and Govaert [15] studied an EM algorithm for classification; Campbell et al. [13], McLachlan and Peel [44], Dasgupta and Raftery [21] and Fraley and Raftery [25] studied model based classification; Bartlett, Jordan, McAuliffe [1] proposed convexity and risk bounds of classification; Lin [41] studied loss of classification; Scott and Nowak [57], Han, Chen, Sun [33] and Rigollet and Tong [55] studied Neyman-Pearson classification; Tsybakov [64] and Zhang [73] studied statistical behavior of classification methods, Boucheron, Bousquet and Lugosi [10], Fraley and Raftery [25] and Fung [27] provide a comprehensive review.

Nonparametric methods, on the other hand, are robust because they do not make distributional assumption, but they are less efficient than the parametric methods when the latter are correctly specified or nearly so. Popular nonparametric methods include the k-means clustering [43], which is formulated by minimizing the within cluster distortion measure and maximizing the intra cluster distance, and is often used for exploratory clustering analysis. The hierarchical clustering methods iteratively merge (or split clusters according to some criteria until all data becomes one cluster (or until some stopping criterion is meet to prevent further splitting) to form a hierarchical tree. The tree is cut at a place to obtain clusters [72]. The popular supporting vector machine (SVM) [67, 9] is a minimax classifier. Given a training sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i$ is multi-dimensional data and y_i is the indicator $y_i = 1$ if $\mathbf{x}_i$ is from class A and $y_i = -1$ for class B. In this method, one needs to specify a set of known functions $\{\phi_1(\cdot), ..., \phi_k(\cdot)\}$ and some weights $w_1, ..., w_k$ to be determined. Then define the decision function $D_w(\mathbf{x}) = \sum_{j=1}^{k} w_j \phi_j(\mathbf{x}) + b$ for some given b. It classifies $\mathbf{x} \in A$ if $D_w(\mathbf{x}) > 0$. The optimal weights $\mathbf{w}^* = (w_1^*, ..., w_k^*)$ are obtained from the training sample by

$$\mathbf{w}^* = \arg \max_{\mathbf{w}:||\mathbf{w}||=1,} \min_{1 \le i \le n} \{y_i D_w(\mathbf{x}_i)\}.$$

This weight maximizes the closest distance M of all data points to the hyperplane $D_w(\mathbf{x}) = M$, and makes the optimal separation of the two clusters. This method is easy to use, but like many other similar methods, it does not provide

*Corresponding author.

regression coefficients estimation, and so the relationship between response and the covariates is unclear. The support vector network [19] is similar. Here the $\phi_j(\cdot)$'s and k need to be subjectively chosen. The random decision forest [35, 11] is another well known classification algorithm. Given a training sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i$ is covariate vector and y_i is the response, it first splits the $\mathbf{x}_i$'s into m trees using the classification tree algorithm. Then for a given sets of weight $w_j(\cdot,\cdot)$, it predict the response of a new point $\mathbf{x}'$ by

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m} w_j(\mathbf{x}_i, \mathbf{x}')y_i.$$

Often $w_j(\mathbf{x}_i, \mathbf{x}') = 1/m$ if $\mathbf{x}_i$ is one of the m points in the same leaf as $\mathbf{x}'$, and zero otherwise. Here the choice of weights and the number of trees m is still subjective.

Kernel method is also applied to nonparametric clustering approach to estimate the data distribution [26, 18, 14] and a mean-shift algorithm is inoked to iteratively transform and group the data.

Here our goal is to model the regression relationship between the response and covariates and classify the subjects into two groups based on the model. Since the membership of each subject is unknown, the model is a mixture. For robustness, we specify the common density of all group to be unimodal, otherwise unspecified. The assumption of unimodality is for indentifiablity of nonparametric mixture model. The unimodal model is closely related to monotonic model. Qin et al. [54] applied isotonic regression to predict genetic risk under monotone distributions. The unimodal density is piecewise monotone and our model is a mixture. Yuan, Zhou, Tan [71] studied a similar model for subgroup analysis, in which the subgroup density is required to be unimodal and symmetric around zero. The symmetry condition is rather restrictive in application, here we relax this condition.

Mixture models, parametric, nonparametric and semiparametric, are studied extensively. Titterington [61] considered minimal distance nonparametric estimation in mixture model. Titterington et al. [62] introduced statistical applications of finite mixture model. A general form is $G(x) = \int F(x,\theta)dH(\theta)$ (as in [42]), where $F(\cdot,\theta)$ is a known parametric disitribution and $H(\cdot)$ an unknown distribution. Qin [52, 53] studied goodness of fit tests based on semiparametric mixture model, and empirical likelihood ratio confidence interval for mixing proportion. McLachlan and Peel [45] provide an over view in this field. Hall and Zhou [32] considered nonparametric estimation of the model $F(x) = \pi\prod_{j=1}^{k} F_{j1}(x_j) + (1-\pi)\prod_{j=1}^{k} F_{j2}(x_j)$ ($k \geq 3$ for identifiability), where $x = (x_1, ..., x_k)$, F_{jl}'s are unknown distribution functions, π is unknown proportion. Hunter et al. [39] considered mixture models with symmetric distributions. Cruz-Medina and Hettmansperger [20] considered nonparametric estimation in semiparametric mixture model. Bordes et al. [7] studies models of the form $G(x) = \lambda F(x-\mu_1) + (1-\lambda)F(x-\mu_2)$, with unknown distribution $F(\cdot)$ and unknown parameters (μ_1, μ_2, λ). Pu and Arias-Castro [51] considered semiparametric estimation of symmetric mixture models with monotone and log-concave densities. Yuan and He [70] proposed semiparametric mixture model for this problem, in which the sub-densities of the clusters are modelled nonparametric via shape constraints. Using kernel density estimators, their method successfully handles the lower dimensional case.

The semiparametric maximum likelihood estimate (MLE) is used to estimate the model parameter and the unknown density. Then we use the commonly used Bayesian classification rule to classify the subjects according to this semiparametric model. Simulation studies are conducted to evaluate the performance of the proposed method, and the method is used to analyse a real data set. In Section 2 we describe the proposed method, in Section 3 we study the asymptotic behavior of the estimates, in Section 4 we present some simulation results and apply our method on a real dataset; lastly some concluding remarks are given at the end of the paper. The relevant technical proofs are given in the Appendix.

2. THE PROPOSED METHOD

The observed data are $D_n = \{(y_i, \mathbf{x}_i) : i = 1, ..., n\}$, where $y_i \in R$ is the response and $\mathbf{x}_i \in R^d$ is the covariates of the i-subject. Each subject belongs to one of k sub-groups, but the sub-group label is unknown, and our goal is to classify each subject to the most likely subgroup. We focus on the case of $k = 2$ subgroups. For this, let δ_i be the latent indicator of subject i belonging to group one/two ($\delta_i = 1/0$). Conditioning on the covariates, the model for the response can be specified as

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \delta_i\alpha + \epsilon_i,$$

where $\boldsymbol{\beta} \in R^d$ is the regression parameters (unknown), α is the extra effect for subgroup one, and ϵ_i is the residual error accounts for the departure of the above linear relationship specification.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha, \lambda)'$, we specify the likelihood for the observed data as

(1)
$$L(\boldsymbol{\theta}|D_n) = \prod_{i=1}^{n}\Big(\lambda g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \alpha) + (1-\lambda)g(y_i - \boldsymbol{\beta}'\mathbf{x}_i)\Big), \quad g \in \mathcal{G}$$

$\mathcal{G}$ be the collection of unimodal density functions with model at 0, and satisfies the following conditions

$$\lim_{y\to-\infty}\frac{g(y-\alpha)}{g(y)} = \lim_{y\to\infty}\frac{g(y)}{g(y-\alpha)} = 0, \quad \forall\ \alpha > 0.$$

The first condition is also used in modal regression problems, see Chen et al. [17]. The second condition is adopted from Hohmann and Holzman [36] for model identifiability, see

below. This condition is satisfied by most commonly used models.

It is known that mixture of nonparametric densities in the general case is not identifiable. Borders et al. [6] showed that if the components densities are symmetric and equal up to a shift parameter, plus a few conditions then the nonparametric mixture is identifiable. They estimate their model parameters and the unknown density by moment method or kernel method, etc.; while we estimate the model parameters by semiparametric MLE. Hohmann and Holzman [36] studied identifiability of nonparametric mixture model in more general form below

$$F(y|x) = (1 - \lambda(x))F_0(y) + \lambda(x)F_1(y).$$

They showed that the above model is nonparametrically identifiable, if

$$\lim_{y\to-\infty} F_1(y)/F_0(y) = 0$$

and

$$\lim_{y\to\infty} (1 - F_0(y))/(1 - F_1(y)) = 0.$$

By L'Hospital's rule, in terms of density functions f_1 and f_0, the above is equivalent to

$$\lim_{y\to-\infty} f_1(y)/f_0(y) = 0 \quad \text{and} \quad \lim_{y\to\infty} f_0(y)/f_1(y) = 0.$$

The above condition is satisfied for many commonly used distributions, such as the normal distributions with different means.

Here in our case, if we set $y = y_i - \boldsymbol{\beta}'\mathbf{x}_i - \alpha$ or $= y_i - \boldsymbol{\beta}'\mathbf{x}_i$, then model (1) is a special case of that in Hohmann and Holzman [36] and the identifiability condition can be meet.

We estimate $(\boldsymbol{\theta}, g)$ by $(\hat{\boldsymbol{\theta}}_n, \hat{g}_n)$,

$$(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) = \arg \max_{(\boldsymbol{\theta},g)\in(\boldsymbol{\Theta},\mathcal{G})} L(\boldsymbol{\theta}|D_n).$$

However, direct maximization of the above likelihood is not easy, especially for g, and a common alternative is to estimate them via the 'augmented' data model, which assume either the latent status δ_i's be observed, together with the EM-algorithm [22]. However, we'll see that none of these two augmented data model is easy to work with. So we will adopt the 'complete data' model, which assumes the δ_i's are observed, and this method works for our case. Below we describe the 'complete data' model and use this model for our estimation.

The "Complete data" model. Let $\mathbf{z}_i = (y_i, \mathbf{x}_i, \delta_i)$ $(i = 1, 2, ..., n)$ be the "complete data". Under this "complete data", the model is

$$(2)\quad f(y, \delta|\mathbf{x}) = \big[\lambda g(y - \boldsymbol{\beta}'\mathbf{x} - \alpha)\big]^{\delta}\big[(1 - \lambda)g(y - \boldsymbol{\beta}'\mathbf{x})\big]^{1-\delta}$$

and the "complete data" log-likelihood for the augmented data $\{(y_i, \mathbf{x}_i, \delta_i) : i = 1, ..., n\}$ is

(3)

$$\ell_n(\boldsymbol{\theta}, g) = \sum_{i=1}^{n} \Big(\delta_i \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \alpha) + (1 - \delta_i) \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \delta_i \log\lambda + (1 - \delta_i) \log(1 - \lambda)\Big).$$

The true parameter $(\boldsymbol{\theta}_0, g_0)$ are estimated by the MLE

$$(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) = \arg \max_{(\boldsymbol{\theta},g)\in(\boldsymbol{\Theta},\mathcal{G})} \ell_n(\boldsymbol{\theta}, g).$$

Our interest here is the joint maxima $(\hat{\boldsymbol{\theta}}_n, \hat{g}_n)$, so we use the following iterative maximization. The detailed algorithm for $\hat{g}_n$ part is non-trivial, we first give a general description of the algorithm, justify its convergence property, then give detailed description latter. For a starting value $\boldsymbol{\theta}^{(0)}$ of $\boldsymbol{\theta}$, find $g^{(1)}(\cdot) \in \mathcal{G}$ as the maxima of $\ell_n(\boldsymbol{\theta}^{(0)}, g)$, then fix $g^{(1)}$, find $\boldsymbol{\theta}^{(g)} \in \boldsymbol{\Theta}$ as the maxima of $\ell_n(\boldsymbol{\theta}, g^{(1)})$, and so on... until convergence of the sequence $\{(\boldsymbol{\theta}^{(r)}, g^{(r)})\}$.

It is known that the sequence $\{(\boldsymbol{\theta}^{(r)}, g^{(r)})\}$ increasing the likelihood at each iteration, and will converge to at least some local maxima of $\ell_n(\boldsymbol{\theta}, g)$. In fact, the increasing likelihood property is obvious, as for all integer r,

$$\ell_n(\boldsymbol{\theta}^{(r+1)}, g^{(r+1)}) \geq \ell_n(\boldsymbol{\theta}^{(r)}, g^{(r+1)}) \geq \ell_n(\boldsymbol{\theta}^{(r)}, g^{(r)}).$$

A formal justification of the convergence of the above iterative algorithm is a case of the block coordinate descent methods in [3].

The computation of $\boldsymbol{\theta}^{(r)}$ can be realized by the well known EM-algorithm [22]. It is known that there is no guarantee for the EM algorithm convergence to the MLE, and generally it converges to a local maxima ([68]; Theorem 3). If the underlying model has the concave property, then EM algorithm converges to the global MLE. On the other hand, it may converges to some local maxima. Thus in application, one needs to apply the EM algorithm with different starting values, to get possible different local stationary points, and compare the log-likelihood at these stationary points to find the global maxima.

Our algorithm is a semiparametric version of EM algorithm, see also [60], chap. 2 for bio-medical applications of this algorithm. The semiparametric and nonparametric EM algorithm were used in a large number of literatures, such as in [46, 12, 34, 20], and see the argument there for the convergence of such algorithm (pp. 67–68). Chen, Zhang and Davidian [16] applied the EM algorithm to a semiparametric random effects model, Borders, Chauveau and Vandekerknove [8] applied the EM algorithm to a semiparametric mixture model, using simulation studies to justify the convergence of the algorithm.

However, since the δ_i's are unobserved, we cannot estimate $(\boldsymbol{\theta}_0, g_0)$ directly based on (3). Instead, we use the

EM-algorithm [22]. In this model, given starting value $(\boldsymbol{\theta}^{(0)}, g^{(0)})$, compute the next step estimate $(\boldsymbol{\theta}^{(1)}, g^{(1)}), \ldots$. Generally, at the r-th step, let

$$Q_n\big(\boldsymbol{\theta}, g\big|\boldsymbol{\theta}^{(r)}, g^{(r)}\big) = E_{\boldsymbol{\delta}}\big[\ell_n(\boldsymbol{\theta}, g)\big|\mathbf{y}^n, \mathbf{x}^n, \boldsymbol{\theta}^{(r)}, g^{(r)}\big], \tag{4}$$

where the expectation is with respect to $\boldsymbol{\delta}$, and as if the true data is generated from parameter $(\boldsymbol{\theta}^{(r)}, g^{(r)})$.

Specifically, at each iteration r, we compute the following

i) compute the $\delta_i^{(r)}$'s as given in the Appendix, computation of (4).
ii) compute $g^{(r+1)}$ (see below).
iii) compute $\boldsymbol{\theta}^{(r+1)} = (\boldsymbol{\beta}^{'(r+1)}, \alpha^{(r+1)}, \lambda^{(r+1)})'$ as

$$\begin{aligned}\boldsymbol{\theta}^{(r+1)} &= \arg\max_{\boldsymbol{\theta}} Q_n\big(\boldsymbol{\theta}, g^{(r+1)}\big|\boldsymbol{\theta}^{(r)}, g^{(r)}\big)\\ &= \sup_{(\boldsymbol{\beta},\alpha,\lambda)} \sum_{i=1}^{n} \Big[\delta_i^{(r)} \log g^{(r+1)}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \alpha)\\ &\quad + (1-\delta_i^{(r)}) \log g^{(r+1)}(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \delta_i^{(r)} \log\lambda\\ &\quad + (1-\delta_i^{(r)}) \log(1-\lambda)\Big].\end{aligned}$$

In principle,

$$\begin{aligned}&g^{(r+1)}(\cdot)\\ &= \arg\max_{g\in\mathcal{G}} \sum_{i=1}^{n} \Big(\delta_i^{(r)} \log g(\epsilon_{1i}^{(r)}) + (1-\delta_i^{(r)}) \log g\big(\epsilon_{0i}^{(r)}\big)\Big),\end{aligned} \tag{5}$$

where $\epsilon_{0i}^{(r)} = y_i - \boldsymbol{\beta}^{(r)'}\mathbf{x}_i$ and $\epsilon_{1i}^{(r)} = y_i - \boldsymbol{\beta}^{(r)'}\mathbf{x}_i - \alpha^{(r)}$. We combine the $\epsilon_{0i}^{(r)}$'s and ϵ_{1i}'s, arrange them in increasing order, and denote them as $\{\epsilon_i^{(r)} : i = 1, ..., N\}$ with $N = 2n$, so that the technique of isotonic regression can be used to compute $g^{(r+1)}$. In particular, the R-code PAVA (pull adjacent violator algorithm, [4]) can be used to solve $g^{(r+1)}$. See below.

It is known that $L(\boldsymbol{\theta}^{(r+1)}, g^{(r+1)}|D_n) \geq L(\boldsymbol{\theta}^{(r)}, g^{(r)}|D_n)$ for all r, and under suitable conditions, as $r \to \infty$,

$$(\boldsymbol{\theta}^{(r)}, g^{(r)}) \to (\hat{\boldsymbol{\theta}}_n, \hat{g}_n).$$

The iteration continues until a convergence criterion is meet for the $(\boldsymbol{\theta}^{(r)}, g^{(r)})$'s.

Computation of $g^{(r+1)}$. The computation of the NPMLE $g^{(r+1)}$ at each iteration is non-trivial, and needs more attention. Below we using the isotonic regression technique to compute it. Denote $\hat{g}$ for $g_n^{(r)}$ and similarly for $\hat{\delta}_i$, $\hat{\epsilon}_i$, etc. Suppose that $\hat{\epsilon}_i = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$ $(i = 1, ..., n)$. We arrange $\hat{\delta}_i$ $(i = 1, ..., n)$ in increasing order. Calculate $n\hat{\lambda}$ and set it as integer. Let $\hat{\epsilon}_{0i} = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$ $(i = 1, ..., n - n\hat{\lambda})$, $\hat{\epsilon}_{1i} = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\alpha}$ $(i = n - n\hat{\lambda} + 1, ..., n)$, combine the $\hat{\epsilon}_{0i}$'s and $\hat{\epsilon}_{1i}$'s, and arrange them as $\hat{\epsilon}_i$ $(i = 1, ..., n)$. Then

$$\hat{g}_n(\cdot) = \arg\max_{g\in\mathcal{G}} \sum_{i=1}^{N} \log g\big(\hat{\epsilon}_i\big).$$

Let

$$G_n(t) = \sum_{i=1}^{N} \frac{1}{n} I(\hat{\epsilon}_i \leq t)$$

be the weighted empirical distribution function of the $\hat{\epsilon}_i$'s, $G_n^-(\cdot)$ be the greatest convex minorant of $G_n(\cdot)$ on R^-, and $G_n^+(\cdot)$ be its least concave majorant on R^+. Modifying the argument in [56], pp. 332–334, we have

Lemma. *On R^-, $\hat{g}_n(\cdot)$ is the right derivative (slope) of $G_n^-(\cdot)$; and on R^+, $\hat{g}_n(\cdot)$ is the left derivative (slope) of $G_n^+(\cdot)$.*

For computation of $\hat{g}_n(\cdot)$, let $c_i = \hat{\epsilon}_i - \hat{\epsilon}_{i-1}$, $h_i = 1/[nc_i]$ and $w_i = nc_i$. Then by Theorem 1.5.1 in [56], p. 31, $\hat{g}_n$ is the following isotonic regression solution

$$\hat{g}_n = \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{n} w_i\big(h_i - g_i\big)^2.$$

Classification. Initially we considered two commonly used classification rules, the Bayesian rule and the Neyman-Pearson rule. The former rule is for the case the two groups to be classified have the same status of preference, the latter is for the case one group has more preference than the other. However, in our simulation studies, the Neyman-Pearson rule does not work well for the simulated rule, so we adopt the Bayesian rule below.

For each subject $i = 1, 2, ..., n$, the probability of subject i belonging to group 1 is

$$\begin{aligned}&P(\delta_i = 1|y_i, \mathbf{x}_i, h_n, \hat{\boldsymbol{\theta}})\\ &= \frac{\hat{\lambda}\hat{g}_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\alpha})}{\hat{\lambda}\hat{g}_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\alpha}) + (1-\hat{\lambda})\hat{g}_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)}.\end{aligned}$$

With the Bayesian rule, we classify this subject to the subgroup S_1 corresponds to $\delta_i = 1$, if $P(\delta_i = 1|y_i, \mathbf{x}_i, \hat{g}_n, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}}) > 1/2$ or

$$\hat{\lambda}\hat{g}_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\alpha}) > (1-\hat{\lambda})\hat{g}_n(\mathbf{y}_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i), \tag{6}$$

otherwise classify the i-th subject to subgroup S_0 corresponds to $\delta_i = 0$. From the above we see that a good classification requires both $\hat{g}_n$ and $\hat{\boldsymbol{\beta}}$ to be accurately estimated. For parametric model, if g is incorrectly specified, the classification error can be serious, and a semiparametric model is much safer.

3. ASYMPTOTIC PROPERTIES

The classification rule is good only if the parameters in the model can be estimated well. We see from the above that the classification only depends on the estimated parameters $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$. Now we study the asymptotic properties of the these parameter estimators in the model. As the model and that in [71] are similar, so are the conditions, results and proofs.

The classification is consistent only if the corresponding parameter estimation is consistent. We assume the condition for model identifiability as in [36], this condition is very easy to meet. We list the following conditions.

(C1). $\mathbf{X}$ has bounded support.
(C2). $\boldsymbol{\Theta}$ is bounded.
(C3). For all g in a neighborhood of $g_0 \in \mathcal{G}$, $g(\cdot)$ has derivative $\dot{g}(\cdot)$ and $\dot{g}(\cdot)/g(\cdot) \in L_1(P)$.
(C4). $\mathcal{G}$ is bounded.
(C5). g_0 is second order differentiable.
(C6). The derivative $\dot{g}_0(t) \neq 0$.
(C7). $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = o_p(n^{-1/3})$.

Conditions (C1)–(C2) and (C4) are reasonable and practical for most applications. Condition (C3) is for the proof of strong consistency of the semiparametric MLE, it will be true if $\dot{g}$ is integrable and the ratio g_0/g is bounded. (C5) is a commonly assumed condition to get asymptotic distribution of $\hat{\boldsymbol{\beta}}$, here it is used to get its convergence rate. (C6) is used to get asymptotic distribution of $\hat{g}$. The same condition is used to get the corresponding results in other literatures mentioned below. (C7) is a technical assumption used to derive the asymptotic distribution of $\hat{g}$, see the comment after Theorem 2.

Theorem 1. *Assume (C1)–(C4), then*

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s} \boldsymbol{\theta}_0. \qquad \sup_x |\hat{g}_n(x) - g_0(x)| \xrightarrow{a.s} 0.$$

As pointed out by a number of researchers [37] Section 3.2.2; [47] Sections 2–3 and [31], the $\sqrt{n}$-consistency of the MLE $\hat{\beta}$ is an open question. The reason is that $\hat{\beta}$ is bundled with the nonparametric MLE $\hat{g}(\cdot)$, which is a non-smooth piecewise step function. Below, as in [71], we give the convergence rate of $\hat{\boldsymbol{\beta}}$ and $\hat{g}$, with $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|$ being the Euclidean distance between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and $||\hat{g}_n - g_0||$ being any commonly used distance between two functions $\hat{g}$ and g_0.

Theorem 2. *Assume (C1)–(C6), then*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\hat{g} - g_0\| = O_p(n^{-1/3}).$$

Even though currently there is no proof of the sharper convergence rate given in (C8), numerical studies by the above mentioned authors suggest this rate. So we regard (C8) is reasonable. Denote $\xrightarrow{D}$ for convergence in distribution. Let $\mathbb{B}(\cdot)$ be the two-sided Browning motion originating from zero: a mean zero Gaussian process on R with $\mathbb{B}(0) = 0$, and $E\big(\mathbb{B}(s) - \mathbb{B}(h)\big)^2 = |s - h|$ for all $s, h \in R$.

Theorem 3. *Assume (C1)–(C5) and (C7), then*

$$n^{1/3}(\hat{g}_n(t) - g_0(t)) \xrightarrow{D} \Big(4|\dot{g}_0(t)|g_0(t)\Big)^{1/3} \arg\max_{h \in R}\{\mathbb{B}(h) - h^2\}.$$

4. SIMULATION STUDY AND APPLICATION

4.1 Simulation studies

We simulate $n = 500$ i.i.d. data with response y_i and with several different dimensions of covariates $\mathbf{x}_i = (x_{i1}, ..., x_{i2})$, $\mathbf{x}_i = (x_{i1}, ..., x_{i4})$, $\mathbf{x}_i = (x_{i1}, ..., x_{i7})$ and $\mathbf{x}_i = (x_{i1}, ..., x_{i10})$. We first generate the covariates, sample the $\mathbf{x}_i$'s from different dimensional normal distribution with mean vector $\boldsymbol{\mu} = (3.1, 1.8)'$, $\boldsymbol{\mu} = (3.1, 1.8, -0.5, 1.2)'$, $\boldsymbol{\mu} = (3.1, 1.8, -0.5, 1.2, -2.3, 4.5, 2.4)'$ and $\boldsymbol{\mu} = (3.1, 1.8, -0.5, 1.2, -2.3, \ 4.5, 2.4, -3.0, 1.7, 2.1)'$ and some covariance matrix Γ, Then we generate the response data, which, given the covariates, are from the mixture $\lambda_0 g_1 + (1 - \lambda_0) g_2$. The y_i's are generated as

$$y_i = \boldsymbol{\beta}_0' \mathbf{x}_i + \delta_i \alpha_0 + \epsilon_i, \quad (i = 1, ..., n).$$

The distribution of the ϵ_i's is a mixture. With probability λ_0, $\epsilon_i \sim g_1(\cdot)$, and with probability $(1 - \lambda_0)$, $\epsilon_i \sim g_2(\cdot)$. We consider 2 cases for ϵ_i. In case 1, $g_1(\cdot)$ and $g_2(\cdot)$ are both skew normal distributions, and in case 2, $g_1(\cdot)$ and $g_2(\cdot)$ are both normal distributions.

Then with the simulated data $(y_i, \mathbf{x}_i)$'s, we first fit model (3) – actually model (4) treating the δ_i's as missing data via the EM-algorithm. In particular, we set the starting values as $\boldsymbol{\theta}^{(0)}$ as the MLE of $\boldsymbol{\theta}$ under standard normal distribution. Then compute $(g^{(r)}, \boldsymbol{\theta}^{(r)}, \delta_i^{(r)})$'s by the EM-algorithm. Convergence of the algorithm can be accessed by the relative error criterion, with a given ρ (here we choose $\rho = 10^{-4}$ to achieve high accuracy) or 100 iterations,

$$\frac{||\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}||}{||\boldsymbol{\theta}^{(r)}||} \leq \rho.$$

When the above criterion is meet at the $(r+1)$-th iteration, the EM algorithm is stopped, and $\boldsymbol{\theta}^{(r+1)}$ is treated as the profile MLE $\hat{\boldsymbol{\theta}}_n$.

Below in Table 1, Table 2, Table 3 and Table 4, we show estimation results with different choices of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha, \lambda)'$ based on the the following models under sample size $n = 500$. SD is the estimated standard deviation of the estimator. Table 1. Semiparametric model and normal mixture model when the true distribution is normal mixture. We see that the estimation results from the normal model is slightly better than those from the semiparametric model. This is expected as the normal mixture model is the true one generating the observed data. Table 2. Semiparametric model and skew normal mixture model when the true distribution is normal mixture. In this case, the skew normal mixture

Table 1. Parameter estimation under semiparametric model and normal mixture model when the true distribution is normal mixture. $n = 500$.

	α	λ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
1												
TRUE	3.110	0.800	1.920	4.130								
Semiparametric	3.099	0.766	1.881	4.125								
(SD)	(0.105)	(0.025)	(0.068)	(0.083)								
Mixture normal	3.102	0.801	1.925	4.129								
(SD)	(0.104)	(0.021)	(0.043)	(0.069)								
2												
TRUE	3.140	0.800	1.920	2.380	-1.750	0.580						
Semiparametric	3.135	0.766	1.885	2.392	-1.731	0.584						
(SD)	(0.116)	(0.029)	(0.086)	(0.093)	(0.067)	(0.030)						
Mixture normal	3.147	0.801	1.903	2.400	-1.755	0.582						
(SD)	(0.143)	(0.027)	(0.086)	(0.093)	(0.056)	(0.026)						
3												
TRUE	2.250	0.600	1.640	0.980	-2.340	1.170						
Semiparametric	2.262	0.560	1.606	0.965	-2.304	1.166						
(SD)	(0.112)	(0.030)	(0.086)	(0.103)	(0.081)	(0.033)						
Mixture normal	2.263	0.611	1.634	0.976	-2.339	1.167						
(SD)	(0.152)	(0.081)	(0.093)	(0.098)	(0.070)	(0.031)						
4												
TRUE	3.270	0.800	0.640	1.590	-2.530	2.620	-1.560	0.870	2.340			
Semiparametric	3.255	0.776	0.648	1.567	-2.512	2.612	-1.543	0.873	2.336			
(SD)	(0.130)	(0.025)	(0.152)	(0.177)	(0.125)	(0.146)	(0.122)	(0.061)	(0.059)			
Mixture normal	3.217	0.792	0.654	1.568	-2.513	2.615	-1.574	0.877	2.344			
(SD)	(0.208)	(0.031)	(0.156)	(0.182)	(0.126)	(0.146)	(0.121)	(0.066)	(0.059)			
5												
TRUE	2.460	0.700	1.880	1.390	-2.460	1.970	-1.340	1.910	0.790			
Semiparametric	2.473	0.646	1.841	1.450	-2.508	2.016	-1.307	1.893	0.789			
(SD)	(0.118)	(0.033)	(0.173)	(0.197)	(0.149)	(0.168)	(0.122)	(0.072)	(0.077)			
Mixture normal	2.458	0.687	1.845	1.452	-2.511	2.018	-1.332	1.897	0.792			
(SD)	(0.209)	(0.084)	(0.174)	(0.194)	(0.146)	(0.171)	(0.136)	(0.072)	(0.081)			
6												
TRUE	3.390	0.800	1.670	0.910	-1.640	3.140	-1.690	1.780	1.650	-2.380	1.630	2.310
Semiparametric	3.386	0.999	1.679	0.896	-1.630	3.138	-1.694	1.781	1.654	-2.376	1.622	2.315
(SD)	(0.119)	(0.001)	(0.213)	(0.184)	(0.207)	(0.181)	(0.088)	(0.063)	(0.114)	(0.117)	(0.068)	(0.040)
Mixture normal	3.399	0.802	1.677	0.899	-1.629	3.139	-1.694	1.779	1.652	-2.375	1.623	2.315
(SD)	(0.167)	(0.030)	(0.221)	(0.185)	(0.210)	(0.186)	(0.089)	(0.069)	(0.116)	(0.120)	(0.070)	(0.042)
7												
TRUE	2.570	0.600	2.330	1.780	-2.880	0.730	-1.240	2.140	1.570	-1.260	2.160	1.390
Semiparametric	2.561	0.998	2.333	1.785	-2.870	0.724	-1.246	2.137	1.571	-1.265	2.166	1.392
(SD)	(0.109)	(0.003)	(0.217)	(0.215)	(0.248)	(0.223)	(0.083)	(0.073)	(0.119)	(0.119)	(0.074)	(0.048)
Mixture normal	2.586	0.599	2.332	1.782	-2.869	0.724	-1.243	2.130	1.570	-1.264	2.168	1.392
(SD)	(0.123)	(0.070)	(0.227)	(0.220)	(0.255)	(0.228)	(0.080)	(0.074)	(0.123)	(0.120)	(0.079)	(0.050)

model is still correctly specified as the normal model is a special case of the skew normal. The semiparametric model has comparable or slightly better overall performance than the skew normal mixture, reflecting the flexibility of the model. Table 3. Semiparametric model and normal mixture model when the true distribution is skew normal mixture, we see that results from the semiparametric model is slightly better. Table 4. Semiparametric model and skew normal mixture model when the true distribution is skew normal mixture. In this case, as expected, the skew normal mixture model has slightly overall better performance, as it is the true model generating the data.

The estimated density $\hat{g}_n(\cdot)$, the corresponding normal density and the true density functions are shown in Figure 1 and Figure 2, for some selected data sets with different dimension of covariates. It is seen that $\hat{g}_n$ is much more closer to the true density than the normal density. This is important in classification as the commonly used Bayesian classification rule is based on density ratio, better density estimate implies more accurate classification.

Then we use the Bayesian rule to classify each of the $\mathbf{y}_i$'s, and below in Table 5 we report the overall classification error for the two groups based on the proposed semiparametric model, and compare with the commonly used mixture nor-

Table 2. Parameter estimation under semiparametric model and skew normal mixture model when the true distribution is normal mixture. $n = 500$.

	α	λ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	γ_1	γ_2
1														
TRUE	3.110	0.800	1.920	4.130									1.500	1.800
Semiparametric	3.127	0.765	1.884	4.115										
(SD)	(0.114)	(0.025)	(0.075)	(0.071)										
Skew normal	3.131	0.803	1.919	4.124									0.993	1.002
(SD)	(0.117)	(0.023)	(0.058)	(0.069)									(0.109)	(0.229)
2														
TRUE	3.140	0.800	1.920	2.380	-1.750	0.580							1.500	1.800
Semiparametric	3.141	0.767	1.899	2.370	-1.718	0.579								
(SD)	(0.117)	(0.027)	(0.082)	(0.095)	(0.068)	(0.027)								
Skew normal	3.102	0.796	1.936	2.371	-1.748	0.578							1.008	1.029
(SD)	(0.237)	(0.045)	(0.120)	(0.103)	(0.069)	(0.027)							(0.162)	(0.630)
3														
TRUE	2.250	0.600	1.640	0.980	-2.340	1.170							1.700	1.300
Semiparametric	2.246	0.555	1.590	0.979	-2.300	1.164								
(SD)	(0.105)	(0.035)	(0.109)	(0.116)	(0.103)	(0.038)								
Skew normal	2.130	0.610	1.650	0.988	-2.349	1.169							1.421	1.126
(SD)	(0.313)	(0.149)	(0.152)	(0.114)	(0.082)	(0.039)							(1.122)	(0.893)
4														
TRUE	3.270	0.800	0.640	1.590	-2.530	2.620	-1.560	0.870	2.340				3.600	2.900
Semiparametric	3.266	0.770	0.619	1.614	-2.544	2.640	-1.545	0.861	2.341					
(SD)	(0.121)	(0.025)	(0.177)	(0.209)	(0.149)	(0.181)	(0.117)	(0.069)	(0.063)					
Skew normal	3.133	0.773	0.639	1.601	-2.537	2.639	-1.592	0.873	2.350				1.359	1.030
(SD)	(0.345)	(0.065)	(0.194)	(0.235)	(0.170)	(0.192)	(0.151)	(0.081)	(0.077)				(1.077)	(0.810)
5														
TRUE	2.460	0.700	1.880	1.390	-2.460	1.970	-1.340	1.910	0.790				2.400	2.700
Semiparametric	2.465	0.651	1.861	1.391	-2.463	1.987	-1.325	1.908	0.786					
(SD)	(0.122)	(0.036)	(0.185)	(0.218)	(0.163)	(0.186)	(0.115)	(0.070)	(0.069)					
Skew normal	2.328	0.700	1.853	1.403	-2.468	2.003	-1.372	1.909	0.797				1.202	1.135
(SD)	(0.399)	(0.133)	(0.193)	(0.228)	(0.168)	(0.189)	(0.143)	(0.078)	(0.075)				(0.931)	(0.951)
6														
TRUE	3.390	0.800	1.670	0.910	-1.640	3.140	-1.690	1.780	1.650	-2.380	1.630	2.310	2.700	3.800
Semiparametric	3.382	0.999	1.689	0.897	-1.616	3.122	-1.690	1.782	1.642	-2.377	1.634	2.312		
(SD)	(0.155)	(0.002)	(0.209)	(0.177)	(0.200)	(0.205)	(0.074)	(0.068)	(0.107	(0.105)	(0.070	(0.042)		
Skew normal	3.148	0.774	1.739	0.902	-1.587	3.070	-1.720	1.824	1.625	-2.362	1.632	2.311	2.311	0.985
(SD)	(0.455)	(0.063)	(0.285)	(0.184)	(0.236)	(0.275)	(0.094)	(0.111)	(0.123)	(0.116)	(0.078)	(0.048)	(0.048)	(0.090)
7														
TRUE	2.570	0.600	2.330	1.780	-2.880	0.730	-1.240	2.140	1.570	-1.260	2.160	1.390	2.500	3.300
Semiparametric	2.548	0.998	2.309	1.806	-2.905	0.738	-1.233	2.134	1.573	-1.272	2.170	1.391		
(SD)	(0.104	(0.002	(0.208	(0.191	(0.210	(0.217	(0.098	(0.071	(0.120	(0.129	(0.081	(0.051		
Skew normal	2.298	0.591	2.320	1.817	-2.906	0.730	-1.245	2.151	1.568	-1.275	2.172	1.391	1.391	1.042
(SD)	(0.383)	(0.169)	(0.233)	(0.199)	(0.222)	(0.240)	(0.123)	(0.108)	(0.135)	(0.145)	(0.100)	(0.062)	(0.062)	(0.263)

mal model, mixture skew normal, K-means, support vector machine (SVM), and the classification errors based on the real likelihood ratio values (denoted as Exact), under sample size $n = 500$. The seven data sets generated from the mixture skew normal, as in Tables 3 and Table 4, are used for the classification. We see that the mixture skew normal model has the smallest overall classification error, this is expected as the data are generated from this distribution; the semiparametric method has apparent smaller classification errors than the other models for most data sets, except data sets 6 and 7.

4.2 Real data analysis

We analyse the data DATATOP (Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism) [49]. It is sponsored by NIH (The National Institutes of Health), and a multicenter randomized controlled clinical trial for studying the early Parkinson's disease treatment. The DATATOP trial was conducted at 28 US and Canadian sites from September 1987 to November 1989. About 800 patients with the early stages of untreated Parkinson's disease were enrolled in the trial and were randomly assigned to one of four

Table 3. Parameter estimation under semiparametric model and normal mixture model when the true distribution is skew normal mixture. $n = 500$.

	α	λ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
1												
TRUE	3.110	0.800	1.920	4.130								
Semiparametric	3.191	0.808	1.872	4.121								
(SD)	(0.092)	(0.028)	(0.053)	(0.055)								
Mixture normal	3.195	0.831	1.873	4.122								
(SD)	(0.092)	(0.019)	(0.044)	(0.058)								
2												
TRUE	3.140	0.800	1.920	2.380	-1.750	0.580						
Semiparametric	3.290	0.811	1.843	2.399	-1.720	0.575						
(SD)	(0.102)	(0.027)	(0.072)	(0.086)	(0.056)	(0.022)						
Mixture normal	3.497	0.865	1.755	2.413	-1.689	0.566						
(SD)	(0.095)	(0.018)	(0.063)	(0.086)	(0.052)	(0.020)						
3												
TRUE	2.250	0.600	1.640	0.980	-2.340	1.170						
Semiparametric	2.219	0.559	1.605	0.977	-2.318	1.178						
(SD)	(0.107)	(0.039)	(0.101)	(0.118)	(0.089)	(0.037)						
Mixture normal	2.424	0.816	1.434	1.011	-2.268	1.154						
(SD)	(0.130)	(0.049)	(0.073)	(0.098)	(0.061)	(0.033)						
4												
TRUE	3.270	0.800	0.640	1.590	-2.530	2.620	-1.560	0.870	2.340			
Semiparametric	3.444	0.813	0.620	1.609	-2.547	2.623	-1.498	0.861	2.321			
(SD)	(0.107)	(0.023)	(0.161)	(0.198)	(0.139)	(0.161)	(0.101)	(0.069)	(0.063)			
Mixture normal	3.617	0.856	0.596	1.623	-2.566	2.627	-1.450	0.849	2.306			
(SD)	(0.109)	(0.018)	(0.168)	(0.203)	(0.140)	(0.166)	(0.100)	(0.068)	(0.061)			
5												
TRUE	2.460	0.700	1.880	1.390	-2.460	1.970	-1.340	1.910	0.790			
Semiparametric	2.460	0.659	1.878	1.403	-2.469	1.977	-1.302	1.896	0.786			
(SD)	(0.114)	(0.031)	(0.169)	(0.201)	(0.147)	(0.170)	(0.113)	(0.073)	(0.062)			
Mixture normal	2.803	0.845	1.846	1.416	-2.500	1.962	-1.187	1.876	0.750			
(SD)	(0.100)	(0.029)	(0.175)	(0.195)	(0.143)	(0.170)	(0.111)	(0.067)	(0.063)			
6												
TRUE	3.390	0.800	1.670	0.910	-1.640	3.140	-1.690	1.780	1.650	-2.380	1.630	2.310
Semiparametric	3.554	0.999	1.629	0.925	-1.670	3.177	-1.672	1.754	1.654	-2.375	1.619	2.316
(SD)	(0.104)	(0.001)	(0.194)	(0.161)	(0.181)	(0.183)	(0.075)	(0.060)	(0.097)	(0.102)	(0.064)	(0.038)
Mixture normal	3.709	0.852	1.578	0.927	-1.699	3.221	-1.649	1.719	1.663	-2.383	1.619	2.318
(SD)	(0.099)	(0.018)	(0.194)	(0.151)	(0.171)	(0.178)	(0.069)	(0.061)	(0.105)	(0.107)	(0.068)	(0.040)
7												
TRUE	2.570	0.600	2.330	1.780	-2.880	0.730	-1.240	2.140	1.570	-1.260	2.160	1.390
Semiparametric	2.541	0.999	2.352	1.775	-2.874	0.726	-1.243	2.135	1.564	-1.248	2.160	1.389
(SD)	(0.108)	(0.002)	(0.223)	(0.203)	(0.228)	(0.218)	(0.105)	(0.077)	(0.114)	(0.125)	(0.085)	(0.051)
Mixture normal	2.747	0.781	2.234	1.786	-2.941	0.835	-1.190	2.048	1.585	-1.262	2.161	1.393
(SD)	(0.104)	(0.041)	(0.210)	(0.164)	(0.209)	(0.212)	(0.085)	(0.061)	(0.107)	(0.114)	(0.074)	(0.045)

treatment groups: (1) active deprenyl, (2) active tocopherol, (3) active deprenyl and tocopherol, and (4) placebo. The development of disability requiring the onset of levodopa therapy was the primary endpoint in the DATATOP trial. The Parkinson Study Group reported the results that deprenyl (10 mg per day) slowed the disease progression measured by the total Unified Parkinson's Disease Rating Scale (UPDRS), its subscales about motor, mentation and activities of daily living. Moreover, some covariates such as baseline age, years of education and gender could be potential confounder, that are associated with the deprenyl treatment and the disease progression.

In this study, the UPDRS was treated as response data, and three key movement dysfunction measures from UPDRS subscales (motor, mentation and activities of daily living) as well as four covariates information, such as baseline age, treatment (whether receive deprenyl), gender and years of education, were treated as covariates. We aimed to use the data to construct a semiparametric model, then use the Bayesian rule to classify the 800 patients into one of two subgroups. It is a longitudinal data and each patient had 6 or 9 repeated measurements. Therefore, we calculated the mean value for activities of daily living, motor and mentation. Also, one patient had no UPDRS, we excluded

Table 4. Parameter estimation under semiparametric model and skew normal mixture model when the true distribution is skew normal mixture. $n = 500$.

	α	λ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	γ_1	γ_2
1														
TRUE	3.110	0.800	1.920	4.130									2.500	2.500
Semiparametric	3.104	0.802	1.890	4.121										
(SD)	(0.084)	(0.030)	(0.053)	(0.052)										
Skew normal	3.109	0.800	1.919	4.130									2.685	3.105
(SD)	(0.083)	(0.022)	(0.030)	(0.044)									(0.538)	(1.033)
2														
TRUE	3.140	0.800	1.920	2.380	-1.750	0.580							2.500	2.500
Semiparametric	3.142	0.801	1.907	2.372	-1.725	0.578								
(SD)	(0.072)	(0.024)	(0.055)	(0.063)	(0.047)	(0.018)								
Skew normal	3.144	0.799	1.924	2.376	-1.744	0.579							2.902	3.139
(SD)	(0.072)	(0.017)	(0.050)	(0.059)	(0.035)	(0.016)							(0.803)	(1.070)
3														
TRUE	2.250	0.600	1.640	0.980	-2.340	1.170							3.400	1.800
Semiparametric	2.247	0.581	1.597	0.970	-2.304	1.164								
(SD)	(0.079)	(0.038)	(0.076)	(0.081)	(0.053)	(0.023)								
Skew normal	2.241	0.601	1.643	0.980	-2.342	1.169							4.253	1.907
(SD)	(0.080)	(0.026)	(0.054)	(0.076)	(0.041)	(0.020)							(0.997)	(0.292)
4														
TRUE	3.270	0.800	0.640	1.590	-2.530	2.620	-1.560	0.870	2.340				1.800	1.300
Semiparametric	3.269	0.797	0.621	1.609	-2.537	2.627	-1.556	0.857	2.351					
(SD)	(0.109)	(0.022)	(0.134)	(0.184)	(0.141)	(0.139)	(0.087)	(0.053)	(0.047)					
Skew normal	3.269	0.803	0.623	1.610	-2.538	2.626	-1.569	0.858	2.353				1.962	1.371
(SD)	(0.109)	(0.020)	(0.134)	(0.184)	(0.140)	(0.139)	(0.083)	(0.052)	(0.047)				(0.343)	(0.277)
5														
TRUE	2.460	0.700	1.880	1.390	-2.460	1.970	-1.340	1.910	0.790				1.500	1.800
Semiparametric	2.474	0.663	1.870	1.383	-2.459	1.980	-1.314	1.915	0.779					
(SD)	(0.095)	(0.029)	(0.172)	(0.213)	(0.146)	(0.165)	(0.114)	(0.070)	(0.066)					
Skew normal	2.471	0.705	1.871	1.384	-2.463	1.983	-1.337	1.919	0.781				1.770	2.199
(SD)	(0.094)	(0.029)	(0.173)	(0.212)	(0.145)	(0.166)	(0.101)	(0.072)	(0.064)				(0.830)	(0.836)
6														
TRUE	3.390	0.800	1.670	0.910	-1.640	3.140	-1.690	1.780	1.650	-2.380	1.630	2.310	1.500	1.600
Semiparametric	3.320	0.999	1.668	0.942	-1.659	3.143	-1.703	1.794	1.639	-2.367	1.624	2.312		
(SD)	(0.344)	(0.001)	(0.213)	(0.162)	(0.199)	(0.201)	(0.079)	(0.081)	(0.104)	(0.108)	(0.066)	(0.040)		
Skew normal	3.321	0.792	1.668	0.941	-1.657	3.143	-1.702	1.792	1.637	-2.366	1.625	2.311	1.709	2.232
(SD)	(0.345)	(0.043)	(0.213)	(0.160)	(0.198)	(0.200)	(0.080)	(0.082)	(0.104)	(0.107)	(0.065)	(0.040)	(0.740)	(1.375)
7														
TRUE	2.570	0.600	2.330	1.780	-2.880	0.730	-1.240	2.140	1.570	-1.260	2.160	1.390	1.700	1.800
Semiparametric	2.528	0.999	2.297	1.826	-2.928	0.766	-1.240	2.124	1.575	-1.262	2.159	1.390		
(SD)	(0.314)	(0.002)	(0.201)	(0.159)	(0.193)	(0.197)	(0.092)	(0.084)	(0.098)	(0.099)	(0.069)	(0.041)		
Skew normal	2.535	0.605	2.295	1.824	-2.927	0.768	-1.238	2.122	1.576	-1.264	2.160	1.391	2.718	2.007
(SD)	(0.314)	(0.092)	(0.203)	(0.157)	(0.191)	(0.196)	(0.087)	(0.081)	(0.098)	(0.098)	(0.068)	(0.040)	(1.627)	(1.002)

Table 5. Classification error ($n = 500$*).*

Data set	Exact	Semiparametric	Mixture normal	Skew normal	K-means	SVM
1	0.028	0.016	0.024	0.012	0.512	0.056
2	0.028	0.016	0.052	0.018	0.488	0.064
3	0.288	0.164	0.194	0.138	0.492	0.176
4	0.006	0.012	0.040	0.006	0.538	0.184
5	0.122	0.118	0.158	0.116	0.530	0.326
6	0.068	0.480	0.064	0.020	0.522	0.216
7	0.178	0.398	0.210	0.318	0.548	0.370

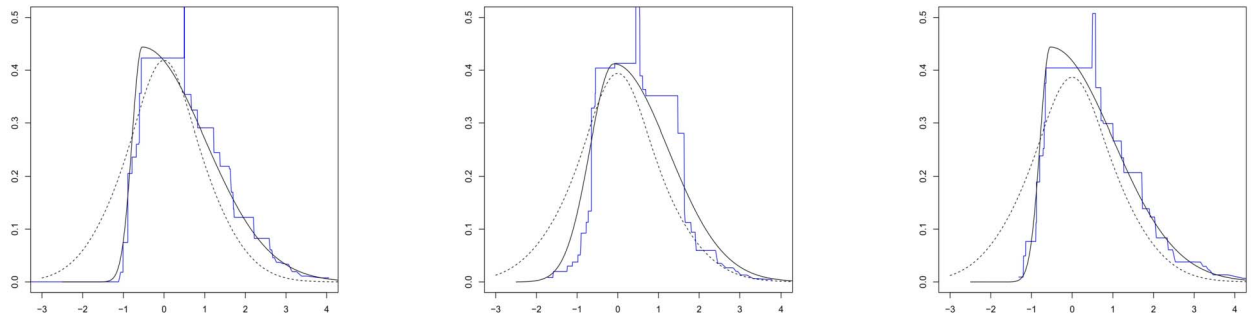

Figure 1. Left panel: True density (4-dimensional covariates, Skew normal with skewness parameter 2.5, lambda=0.8, solid black line), estimated density (solid blue line) and normal density (dotted line). Middle panel: True density (4-dimensional covariates, Skew normal with skewness parameter 1.5, lambda=0.6, solid black line), estimated density (solid blue line) and normal density (dotted line). Right panel: True density (7-dimensional covariates, Skew normal with skewness parameter 2.5, lambda=0.8, solid black line), estimated density (solid blue line) and normal density (dotted line).

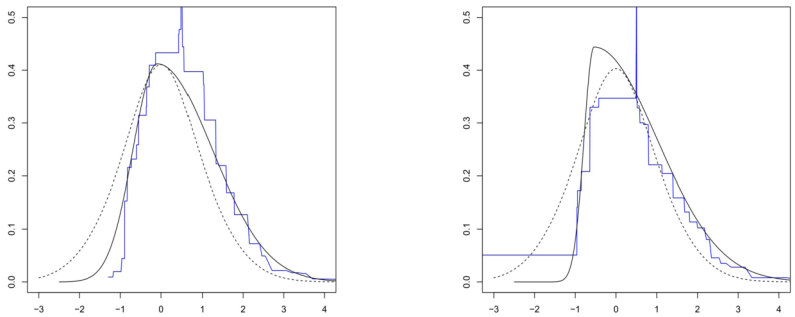

Figure 2. Left panel: True density (7-dimensional covariates, Skew normal with skewness parameter 1.5, lambda=0.7, solid black line), estimated density (solid blue line) and normal density (dotted line). Right panel: True density (10-dimensional covariates, Skew normal with skewness parameter 2.5, lambda=0.8, solid black line), estimated density (solid blue line) and normal density (dotted line).

this observation and therefore we considered 799 patients at last.

Table 6 shows the results from the semiparametric model, mixture normal model, skew normal mixture and SVM. We see that with the semiparametric model, about 38.9% patients were classified into group 1, while with the mixture normal model and the skew normal mixture model, nearly all of the patients were classified into group 1. The regression coefficients estimates $\boldsymbol{\beta}$ and the estimated treatment effect α are similar for the three models. The SVM method does not have regression coefficient estimates, so only its MSEs are reported. The k-means does not provide predicted value and is not show here. As for real data the true distribution is unknown, different methods cannot be assessed according to their estimates. We used 10-fold cross validation method to compare their mean square error (MSE) for the response data. And the MSEs for the semiparametric model and skew normal mixture model are apparently smaller.

Concluding remarks. In this study, we proposed a semiparametric mixture model, in which the regression relationship between the response and covariates are specified parametric, and the group density function is specified nonparametric, only assume to be unimodal. The Bayesian classification rule is used to classify the subjects according to this semiparametric model. The parameters are estimated via semiparametric MLE. Nonparametric methods like SVM and k-means clustering are robust, but they do not consider the relationship between response and covariates, which is important in some applications. With the semiparametric model, the model parameters are estimated and the relationship between response and covariates are also characterized by the regression coefficients. Based on our simulation studies, for data with low or moderate dimensions, the proposed model works better than the commonly used normal mixture model, but it does not work well for data with covariates dimension greater than 10. This phenomenon is quite common for most models. For data with high dimensional covariates, dimension reduction method such as LASSO can be used to select the significant components. Based on our simulation studies, the semiparametric model estimates the residual error density much better than the normal mixture model, this makes better classification as the commonly used Bayesian classification is based on density ratio.

Table 6. *Parameter estimation and classification results (real data).*

	α	λ	β_1	β_2	β_3	β_4	β_5	β_6	β_7	Prop.	MSE
Semiparametric	4.145	0.394	0.045	-0.343	-0.165	0.090	-0.031	1.990	1.234	0.389	2.960
Normal mixture	4.998	1.000	0.034	-0.228	-0.063	0.083	-0.058	1.883	1.236	1.000	6.524
Skew normal mixture	3.273	0.475	0.044	-0.308	-0.172	0.108	-0.032	1.898	1.253	0.999	2.819
SVM	–	–	–	–	–	–	–	–	–	–	8.989

*Prop.: classification proportion of group 1 according to different models

As pointed by the Associate Editor, generally the regression coefficients in the two subgroups can be different, with $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, or even the sub-densities in the two groups can be different with $g_1(\cdot)$ and $g_2(\cdot)$. But this will double the number of regression parameters to be estimated and reduce the efficiency. In fact, in most investigations in subgroup analysis, common regression coefficients are assumed, such as in [58, 24, 69]. According to our limited simulation studies, incorporating different regression parameters for the two groups does not improve the results. The main reason may be due to the double of regression parameters requires much larger data sample size. However, investigating this more general setup can be our future stuty topic.

APPENDIX

Computation of (5). We have

$$\begin{aligned}\delta_i^{(r)} &:= P(\delta_i = 1|\mathbf{y}^n, \mathbf{x}^n, g^{(r)}, \boldsymbol{\theta}^{(r)}) \\ &= 1 - P(\delta_i = 0|\mathbf{y}^n, \mathbf{x}^n, g^{(r)}, \boldsymbol{\theta}^{(r)}) = P(\delta_i = 1|y_i, \mathbf{x}_i, g^{(r)}, \boldsymbol{\theta}^{(r)}) \\ &= \frac{\lambda^{(r)} g^{(r)}(y_i - \boldsymbol{\beta}'^{(r)}\mathbf{x}_i - \alpha^{(r)})}{\lambda^{(r)} g^{(r)}(y_i - \boldsymbol{\beta}'^{(r)}\mathbf{x}_i - \alpha^{(r)}) + (1-\lambda^{(r)}) g^{(r)}(y_i - \boldsymbol{\beta}'^{(r)}\mathbf{x}_i)}.\end{aligned}$$

Then get

(7)

$$\begin{aligned}Q_n\big(\boldsymbol{\theta}, g\big|\boldsymbol{\theta}^{(r)}, g^{(r)}\big) = \sum_{i=1}^{n} \Big[&\delta_i^{(r)} \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \alpha) \\ &+ (1-\delta_i^{(r)}) \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \delta_i^{(r)} \log\gamma \\ &+ (1-\delta_i^{(r)}) \log(1-\gamma) \Big].\end{aligned}$$

Proof of the Lemma. Without loss of generality we assume the $\hat{\epsilon}_i$'s are arranged in increasing order. Let $c_i = \hat{\epsilon}_i - \hat{\epsilon}_{i-1}$, r be the integer such that $\hat{\epsilon}_r < 0 < \hat{\epsilon}_{r+1}$. Denote $g_i = g(\hat{\epsilon}_i)$. Since $\hat{g}_n(\cdot)$ is a step function, and takes zero on $(-\infty, \hat{\epsilon}_1] \cup [\hat{\epsilon}_n, \infty)$, the constraint $\int g(t)dt = 1$ in is written as

(8)
$$\begin{aligned}&\sum_{j=1}^{r-1} g_j(\hat{\epsilon}_j - \hat{\epsilon}_{j-1}) + g_r(0 - \hat{\epsilon}_r) + g_{r+1}(\hat{\epsilon}_{r+1} - 0) \\ &\quad + \sum_{j=r+2}^{n} g_j(\hat{\epsilon}_j - \hat{\epsilon}_{j-1}) = 1.\end{aligned}$$

See the estimation of unimodal density in Section 7.3 in [56], pp. 332–334.

By Example 1.5.7 in [56], pp. 38–39, the $\hat{g}_n$ in the maximization in (7) is

$$\hat{g}(\cdot) = \arg\max_{g\in\mathcal{G}} \sum_{i=1}^{n} h_i w_i \log(g_i),$$

where $h_i = \hat{\lambda}_i)/[Nc_i]$ and $w_i = Nc_i$. The above is the same as (7), however, written in this form will lead to simplification using results in isotonic regression.

Let $\Phi(u) = u\log u\ u \in R^+$ and $\Delta_\Phi(u,v) = \Phi(u) - \Phi(v) - (u-v)\Phi(v) = u\log u - u\log v - (u-v)$. Then $\Phi(\cdot)$ is convex on R^+. Note that the right hand side of (7) is the same as

$$\begin{aligned}&= \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{N} \Big(\hat{\lambda}_i \log(\hat{\lambda}_i) - \hat{\lambda}_i \log g_i - \big(\hat{\lambda}_i\big)(-g_i)\Big) c_i, \\ &\quad \text{subject to (8)} \\ &= \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{N} \Delta_\Phi\big(\hat{\lambda}_i, g_i\big) c_i, \quad \text{subject to (8)} \\ &= \arg\max_{g\in\mathcal{G}} \sum_{i=1}^{N} \Delta_\Phi\big(h_i, g_i\big) w_i.\end{aligned}$$

By Theorem 1.5.1 in [56], p. 31, the above minimization is the same as the following isotonic regression solution

$$\arg\min_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i\big(h_i - g_i\big)^2.$$

Let $W_i = \sum_{j=1}^{i} w_j = N\hat{\epsilon}_i$ and $G_i = \sum_{j=1}^{i} w_j h_j = \sum_{j=1}^{i} \hat{\lambda}_i$. By Theorem 1.2.1 in [56], pp. 7–8, and the description of estimation of unimodal density in Section 7.3 in [56], pp. 332–334, on R^-, $\hat{g}_n(\cdot)$ is the right derivative (slope) of the greatest convex minorant of the sum diagram of $\{(W_i, G_i) : i = 1, ..., n\}$, and on R^+, $\hat{g}_n(\cdot)$ is the left derivative (slope) of the least concave majorant of the sum diagram. Note that in terms of slopes of the greatest convex minorant (the least concave majorant), the sum diagram of $\{(W_i, G_i) : i = 1, ..., n\}$ and that of $\{(\hat{\epsilon}_i, G_i/N) : i = 1, ..., n\}$ are the same. So, let $I(\cdot)$ be the indicator function,

$$G_n(t) = \sum_{i=1}^{n} \frac{\hat{\lambda}_i}{N} I(\hat{\epsilon}_i \le t)$$

be the weighted empirical function of the $\hat{\epsilon}_i$'s, $H_n^-(\cdot)$ be the greatest convex minorant of $G_n(\cdot)$ on R^-, and $G_n^+(\cdot)$ be its least concave majorant on R^+, then on R^-, $\hat{g}_n(\cdot)$ is the right derivative (slope) of $G_n^-(\cdot)$; and on R^+, $\hat{g}_n(\cdot)$ is the left derivative (slope) of $G_n^+(\cdot)$. □

Proofs of Theorems 1–3 are similar to those in [71], are omitted, and will be provided upon request.

Received 31 May 2019

REFERENCES

[1] Bartlett, P. L., Jordan, M. I., McAuliffe, J. D. (2006). Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association* **101 (473)** 138–156. MR2268032

[2] Begun, J. M., Hall, W. J., Huang, W., Weller, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* **11** 432–452. MR0696057

[3] Bertsekas, D. P. (2016). *Nonlinear Programming*, 3rd edition. Athena Scientific. MR3587371

[4] Best, M. J., Chakravarti, N. (1990). Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming* **47** 425–439. MR1068274

[5] Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, Maryland. MR1245941

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

[7] Bordes, L., Delmas, C., Vandekerkhove, P. (2006b). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics* **33** 733–752. MR2300913

[8] Bordes, L., Chauveau, D., Vandekerknove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis* **51** 5429–5443. MR2370882

[9] Boser, B. E., Guyon, I. M., Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Worshop on Computational Learning Theory*, San Mateo, CA.

[10] Boucheron, S., Bousquet, O., Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.* **9** 323–375. MR2182250

[11] Breiman, L. (2001). Random forests. *Machine Learning* **45 (1)** 5–32. MR3874153

[12] Campbell, G. (1981). Nonparametric bivariate estimation with ranodmly censored data. *Biometrica* **68** 417–422. MR0626401

[13] Campbell, J. G., Fraley, C., Murtagh, F., Raftery, A. E. (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters* **18** 1539–1548.

[14] Carreira-Perpinan, M. A. (2006). Fast nonparametric clustering with Gaussian blurring mean-shift. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.

[15] Celeux, G., Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis* **14** 315–332. MR1192205

[16] Chen, J., Zhang, D., Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biometrics* **3 (3)** 347–360. MR2703312

[17] Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., Wasserman, L. (2016). Nonparametric modal regression. *Annals of Statistics* **44** 489–514. MR3476607

[18] Cheng, Y. (1995). Mean shift, mode seeking and clustering. *IEEE Trans. PAMI* **17** 790–799.

[19] Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20** 273–297.

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

[21] Dasgupta, A., Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **441** 294–302.

[22] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, **39** 1–38. MR0501537

[23] Diao, G., Yuan, A. (2018). A class of semiparametric cure models with current Status data. *Life Time Data Analysis* **25 (1)** 26–51. MR3896658

[24] Fan, A., Song, R., Lu, W. (2017). Change-plane analysis for subgroup detection and sample size calculation. *Journal of the American Statistical Association* **112** 769–778. MR3671769

[25] Fraley, C., Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97** 611–631. MR1951635

[26] Fukunaga, K., Hostetler, L. D. (1975). The estimation of the gradient of a density function, with application in pattern recognition. *IEEE Trans. PAMI* **25** 1499–1504. MR0388638

[27] Fung, G. (2004). A comprehensive overview of basic clustering algorithms, manuscript.

[28] Genest, C., Ghoudi, K., Rivest, L. P. (1995). semiparametric estimation procedure of dependence parameters in a multivariate families of distributions. *Biometrika* **82** 543–552. MR1366280

[29] Groeneboom, P. (1988). Brownian motion with a parabolic drift and Airy functions. *Probability Theory and Related Fields* **81** 79–109. MR0981568

[30] Groeneboom, P., Wellner, J. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhäuser Verlag, Basel. MR1180321

[31] Groeneboom, P., Hendrickx, K. (2018). Current status linear regression. *Annals of Statistics* **46 (4)** 1415–1444. MR3819105

[32] Hall, P., Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics* **31 (1)** 201–224. MR1962504

[33] Han, M., Chen, D., Sun, Z. (2008). Analysis to Neyman-Pearson classification with convex loss function. *Analysis in Theory and Applications* **24 (1)** 18–28. MR2422456

[34] Hanley, J. A., Parnes, M. N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics* **39** 129–139. MR0712744

[35] Ho, T. K. (1995). Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995, 278–282.

[36] Hohmann, D., Holzmann, H. (2013). Two-component mixtures with independent coordinates as conditional mixtures: nonparametric identification and estimation. *Electronic Journal of Statistics* **7** 859–880. MR3044502

[37] Huang, J., Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In D. Lin and T. Fleming (eds.) *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, pp. 123–169. New York: Springer-Verlag.

[38] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. MR0216620

[39] Hunter, D. R., Wang, S., Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *Annals of Statistics* **35** 224–251. MR2332275

[40] Kosorok, M. (2008). Bootstrapping the Grenander estimator. *IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* **1** 282–292. MR2462212

[41] Lin, Y. (2004). A Note on margin-based loss functions in classification. *Statistics and Probability Letters* **68** 73–82. MR2064687
[42] Lindsay, B. G., Lesperance, M. L. (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference* **47** 29–39. MR1360957
[43] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, eds. L. M. Cam and J. Neyman, Berkeley, CA: University of California Press: 281–297. MR0214227
[44] McLachlan, G. J., Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Lecture Notes in Computer Science* **1451** 658–666. MR1682742
[45] McLachlan, G. J., Peel, D. (2000). *Finite mixture models*. Wiley, New York. MR1789474
[46] Muñoz, A. (1980). Nonparametric estimation from censored bivariate observations Technical Report, Department of Statistics, Stanford University. MR2630946
[47] Murphy, S. A., van der Vaart, A. W., Wellner, J. A. (1999). Current status regression. *Mathematical Methods of Statistics***8 (3)** 407–425. MR1735473
[48] Olkin, I. (1987). A semiparametric approach to density estimation. *Journal of the American Statisticial Association* **82** 858–865. MR0909993
[49] Parkinson Study Group (1989). DATATOP: a multicenter controlled clinical trial in early Parkinson's disease. *Arch. Neurol.* **46** 1052–1060.
[50] Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimators. *Metrika* **14** 249–272.
[51] Pu, X., Arias-Castro, E. (2018). Semiparametric estimation of symmetric mixture models with monotone and log-concave densities. arXiv:1702.08897.
[52] Qin, J. (1998). Semiparametric likelihood based method for goodness of fit tests and estimation in upgraded mixture models. *Scandinavian Journal of Statistics* **25** 681–691. MR1666804
[53] Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics* **27** 1368–1384. MR1740107
[54] Qin, J., Garcia, T. P., Ma, Y., Tang, M.-X., Marder, K., Wang, Y. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. *Annals of Applied Statistics* **8** 1182–1208. MR3262550
[55] Rigollet, P., Tong, X. (2011). Neyman-Pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research* **12** 2831–2855. MR2854349
[56] Robertson, T., Wright, F. T., Dykstra, R. (1988). *Order Restricted Statistical Inference*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore. MR0961262
[57] Scott, C., Nowak, R. (2005). A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory* **51 (11)** 3806–3819. MR2239000
[58] Shen, J., He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association* **110** 303–312. MR3338504
[59] Sklar A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8** 229–231. MR0125600
[60] Tan, M., Tian, G.-L., Ng, K. W. (2009). *Bayesian Missing Data Problems: EM, Data Augmentation and Non-iterative Computation*, London and Boca Raton, Florida: Chapman and Hall/CRC. MR2562244
[61] Titterington, D. M. (1983). Minimum-distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistic Society*, Ser. B, **45** 37–46. MR0701074
[62] Titterington, D. M., Smith, A. F. M., Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley, Chichester. MR0838090
[63] Tsaitis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926
[64] Tsybakov, A. (2004). Optimal Aggregation of Classifiers in Statistical Learning. *Annals of Statistics* **32** 135–166. MR2051002
[65] van der Vaart, A. (2002). *Semiparametric Statistics*, Eds. Cachan, J. M. M., Groningen, F. T., Paris, B. T., Springer. MR1915446
[66] van der Vaart, A., Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer. MR1385671
[67] Vapnik, V. N., Chervonenkis, A. Ya. (1974). *The theory of pattern recognition*. Nauka, Moscow.
[68] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11** 95–103. MR0684867
[69] Yuan, A., Chen, X., Zhou, Y., Tan, M. T. (2018). Subgroup analysis with semiparametric models toward precision medicine. *Statistics in Medicine* **37** 1830–1845. MR3799843
[70] Yuan, A., He, W. (2008). Semi-parametric clustering method for microarray data analysis. *Journal of Bioinformatics and Computational Biology* **6** 261–282.
[71] Yuan, A., Zhou, Y., Tan, M. T. (2020). Sub-group analysis with nonparametric unimodal symmetric error distribution. *Communications in Statistics: Theory and Methods*, in press.
[72] Zhang, T., Ramakrishnan, R., Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *SIGMOD Record* **25** 103–114.
[73] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* **32** 56–85. MR2051001

Anqi Yin
Department of Biostatistics, Bioinformatics and Biomathematics
Georgetown University
Washington DC, 20057
USA
E-mail address: ay380@georgetown.edu

Ao Yuan
Department of Biostatistics, Bioinformatics and Biomathematics
Georgetown University
Washington DC, 20057
USA
E-mail address: ay312@georgetown.edu