SEMIPARAMETRIC MIXTURE REGRESSION UNDER A SYMMETRIC UNIMODAL ERROR DISTRIBUTION

June 28, 2018

Abstract: Semiparametric mixture regression models are often used in clustering and survival analysis problems. However, as of now there have been no provably efficient estimators for them. In this paper, we propose a special case of a semiparametric mixture location-shift family mixture model. For model identifiability, each error subdistribution is assumed to be symmetric, unimodal and same up to a shift. The semiparametric maximum likelihood estimator is shown to be strongly consistent, its parametric component to be asymptotically efficient and its nonparametric component to have, pointwise and after being multiplied by the cube root of the sample size, a weak limit of a Chernoff random variable times a constant. Simulation studies support the proposed estimation method.

Key words and phrases: Isotonic regression, linear regression, maximum likelihood estimation, mixture model, symmetric density, unimodal density.

1 Introduction

The mixture regression model is commonly used in clustering, classification, personalized precision medicine, subgroup analysis, and survival analysis. In this paper, we propose and study a mixture linear regression model, in which the error sub-distribution is assumed unimodal.

Let $\boldsymbol{D}_n = \{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ be the observed data from n individuals, where the y_i's are the responses and $\boldsymbol{x}_i$'s are the covariates, independent and identically distributed as $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^d$. Let the data be generated from k subgroups, and consider the corresponding linear regression mixture model

$$y = \boldsymbol{\beta}_0^T \boldsymbol{x} + \boldsymbol{\alpha}_0^T \boldsymbol{\gamma} + e, \quad E(e \mid \boldsymbol{x}) = 0,$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ is unobserved and is a group membership indicator: one of its components equals to 1 and the other components equal to 0. We wish to estimate $\boldsymbol{\alpha}_0 = (\alpha_{10}, \ldots, \alpha_{k0})^T$, a group effects vector, and $\boldsymbol{\beta}_0 \in \mathbb{R}^d$, the regression coefficients vector which explains the linear relationship between the response y and covariates $\boldsymbol{x}$. The error term unexplained by the linear relationship is represented by e.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$. The commonly used method for inferring $\boldsymbol{\theta}$, then, is to specify some parametric model for the conditional density of the data given the covariates,

$$g(y - \boldsymbol{\beta}^T \boldsymbol{x} - \boldsymbol{\alpha}^T \boldsymbol{\gamma} \mid \boldsymbol{x}, \eta) = \sum_{j=1}^{k} \delta_j(\boldsymbol{x}) g_j(y - \boldsymbol{\beta}^T \boldsymbol{x} - \alpha_j \mid \eta),$$

where the $\delta_j(\cdot)$'s are given functions, the $g_j(\cdot \mid \eta)$'s are some parametric density functions and η is the model parameter. Then $\boldsymbol{\theta}$ could be estimated by the maximum likelihood estimator under this model. However, when the underlying error distribution is deviated far from the subjectively specified model, the performance of the estimates can be questionable or even inconsistent. To address this issue, numerous researchers proposed various semiparametric methods, the most general method being to replace the parametric $g_j(\cdot \mid \eta)$'s by nonparametric densities $f_j(\cdot - \alpha_j)$ to get

$$g(y - \boldsymbol{\beta}^T \boldsymbol{x} - \boldsymbol{\alpha}^T \boldsymbol{\gamma} \mid \boldsymbol{x}) = \sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f_j(y - \boldsymbol{\beta}^T \boldsymbol{x} - \alpha_j).$$

Recently, semiparametric mixture models of the above form have been used extensively, such as in Mallapragada et al. (2010), Huang and Yao (2012), Zhu and Hunter (2015) and the references in Zhu and Hunter (2015). Butucea et al. (2017) and Huang et al. (2013) considered similar models.

However, it is known that a semiparametric mixture model in this general form is non-identifiable. Identifiability of finite-component semiparametric mixture models is a basic problem and has generated lots of literature, for example, Titterington et al. (1985), Hennig (2000) and Frühwirth-Schnatter (2006). For the location-shift family mixture model

$$g(y) = \sum_{j=1}^{k} \delta_j f(y - \alpha_j), \tag{0}$$

which satisfies $f_j(\cdot) = f(\cdot - \alpha_j)$ for all j, the identifiability condition is relatively easy to meet. For the case $k = 2$, if $f(\cdot)$ is symmetric and $\delta_j \notin \{0, 1/2, 1\}$, then model (0) is identifiable (Bordes et al., 2006, Theorem 2.1). For $k = 3$, if $\delta_j \neq 0$ and $(\alpha_2 - \alpha_1)/(\alpha_3 - \alpha_2) \notin \{1/3, 1/2, 1, 2, 3\}$, then model (0) is identifiable (Hunter et al., 2007, Corollary 1). Since for $k = 2, 3$, the parameter set in which the k-component mixture is non-identifiable has Lebesgue measure zero, Hunter et al. (2007) conjectured that for any finite k, the k-component mixture model (0) is almost surely identifiable.

For the special class of Pólya frequency functions (of infinite order) (Schoenberg, 1951), which are log-concave and hence unimodal, if $E(f(Y)) = 0$ then model (0) is identifiable (Balabdaoui and Butucea, 2014), even (surprisingly) if the number of components k is treated as a parameter. Bordes et al. (2006) and Hunter et al. (2007) proposed distance-based estimators for $(\boldsymbol{\theta}, f)$, but their estimators are not semiparametric efficient, in the sense that they do not converge to a normal distribution with asymptotic variance equal to the semiparametric information lower bound.

In Section 2 we introduce our model and use isotonic regression techniques to derive an expression for the infinite-dimensional component of the estimator, in terms of the finite-dimensional component. In Section 3 we present our main results on the asymptotic behavior of the estimator, including strong consistency of both the infinite-dimensional and finite-dimensional components of the estimator, semiparametric efficiency of the finite-dimensional component, and a weak limit theorem for the infinite-dimensional component. In Section 4 we describe a simulation study and its results. In Section 5 we summarize our results and make a few concluding remarks. In the Appendix we provide proofs of the identifiability of the model, of the form of the infinite-dimensional component, and of the asymptotic results, as well as the table and the two figures.

2 The method

Let us now turn to our problem and method. Define a density function on $\mathbb{R}$ to be unimodal at M if it is non-decreasing on $(-\infty, M]$ and non-increasing on $[M, \infty)$. Let

$$\mathcal{U} = \big\{f(\cdot): \ f \text{ is a density function on } \mathbb{R}\text{, unimodal at and symmetric around 0}\big\}.$$

Let

$$\mathcal{U} \supseteq \mathcal{F} \supseteq \mathcal{U} \cap \big\{f: \ f \text{ is a step function}\big\}.$$

Conditioning on the covariates $\boldsymbol{x}$, we consider the more general setup of a k-component regression mixture

$$y = \boldsymbol{\beta}^T \boldsymbol{x} + e, \quad g(e \mid \boldsymbol{x}) = \sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f(e - \alpha_j), \quad (\boldsymbol{\theta}, f) \in \boldsymbol{\Theta} \times \mathcal{F}, \tag{1}$$

where $\delta_j(\boldsymbol{x})$ is the mixing proportion for $f(\cdot - \alpha_j)$, and satisfies

$$\delta_j(\boldsymbol{x}) > 0 \ \text{ and } \ \delta_j(\boldsymbol{x}) = E(\gamma_j \mid \boldsymbol{x}), \quad \text{for each } 1 \leq j \leq k.$$

This implies $\sum_{j=1}^k \delta_j(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$. Also, $E(Y \mid \boldsymbol{x}) = \mu(\boldsymbol{x}, \boldsymbol{\beta}_0)$, where $\mu(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{x} + \sum_{j=1}^k \alpha_j \delta_j(\boldsymbol{x})$.

Here we have assumed that the $\delta_j(\cdot)$'s depend on $\boldsymbol{x}$ because in applications, such as personalized precision medicine studies and subgroup analysis, the membership proportions are determined by the subject's profile. Though the $\delta_j(\cdot)$'s can be estimated (for example, via logistic regression), to allow us to focus our attention on our goal of estimating $\boldsymbol{\theta}$ and f, we also assume these to be known.

Proposition 1. *Under the identifiability condition for model (0), model (1) is identifiable.*

2.1 The complete-data likelihood

Since estimation under model (1) is difficult due to the additive mixture, a convenient way is to formulate (1) as a multiplicative mixture. Under the independent and identically distributed complete data $\boldsymbol{D}_n^c = \{(y_i, \boldsymbol{x}_i, \boldsymbol{\gamma}_i) : i = 1, \ldots, n\}$, conditioning on the $\boldsymbol{x}_i$'s, the likelihood can be formulated as

$$L(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n^c) = \prod_{i=1}^{n} \prod_{j=1}^{k} \Big(\delta_j(\boldsymbol{x}_i) f(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i - \alpha_j)\Big)^{\gamma_{ij}}$$

with corresponding log-likelihood, ignoring a term without the parameters of interest,

$$l(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n^c) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \log f(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i - \alpha_j).$$

Let $l(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n)$ be the log-likelihood function for the original data. Denote $(\boldsymbol{\theta}_0, f_0)$ to be the true parameters generating the observed data. We estimate $(\boldsymbol{\theta}_0, f_0)$ by the semiparametric maximum likelihood estimator $(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)$,

$$(\hat{\boldsymbol{\theta}}_n, \hat{f}_n) = \underset{(\boldsymbol{\theta}, f) \in \boldsymbol{\Theta} \times \mathcal{F}}{\operatorname{arg\,max}} \; l(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n). \tag{2}$$

Let $\boldsymbol{\theta}^{(0)}$ and $f^{(0)}$ be values which make the Expectation-Maximization algorithm converge to the maximum likelihood estimate. Their existence is a difficult problem and may not be guaranteed. Nevertheless, we shall assume that they exist. Then let

$$Q(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n, \boldsymbol{\theta}^{(r)}, f^{(r)}) = E(l(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n^c) \mid \boldsymbol{D}_n, \boldsymbol{\theta}^{(r)}, f^{(r)}), \quad r = 0, 1, \ldots.$$

Thus,

$$Q(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n, \boldsymbol{\theta}^{(r)}, f^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij}^{(r)} \log f(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i - \alpha_j),$$

where

$$\gamma_{ij}^{(r)} = \frac{\delta_j(\boldsymbol{x}_i) f^{(r)}(y_i - {\boldsymbol{\beta}^{(r)}}^T \boldsymbol{x}_i - \alpha_j^{(r)})}{\sum_{j=1}^{k} \delta_j(\boldsymbol{x}_i) f^{(r)}(y_i - {\boldsymbol{\beta}^{(r)}}^T \boldsymbol{x}_i - \alpha_j^{(r)})}.$$

Let

$$(\boldsymbol{\theta}^{(r+1)}, f^{(r+1)}) = \underset{(\boldsymbol{\theta}, f) \in \boldsymbol{\Theta} \times \mathcal{F}}{\operatorname{arg\,max}} \; Q(\boldsymbol{\theta}, f \mid \boldsymbol{D}_n, \boldsymbol{\theta}^{(r)}, f^{(r)}). \tag{3}$$

Let $h(\boldsymbol{x})$ be the density function of $\boldsymbol{x}$, and define the joint probability distribution of y and $\boldsymbol{x}$ by

$$p(y, \boldsymbol{x} \mid \boldsymbol{\theta}, f) = g(y \mid \boldsymbol{x}, \boldsymbol{\theta}, f) h(\boldsymbol{x}).$$

Let P be the probability measure corresponding to $p(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0, f_0)$. Let the semimetric on $\boldsymbol{\Theta} \times \mathcal{F}$ be

$$d((\boldsymbol{\theta}_1, f_1), (\boldsymbol{\theta}_2, f_2)) = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|f_1 - f_2\|_2,$$

where $| \cdot |$ is the Euclidean norm. Also, assume that $\boldsymbol{\Theta}$ is closed and $\mathcal{F}$ is closed in $\mathcal{L}_2(P)$. ($\mathcal{F} \subseteq \mathcal{L}_2(P)$ since every function in $\mathcal{F}$ is bounded.) Under some certain conditions, the pair $(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)$ defined in equation (2) is equivalent to

$$(\hat{\boldsymbol{\theta}}_n, \hat{f}_n) = \lim_{r \to \infty} (\boldsymbol{\theta}^{(r)}, f^{(r)}).$$

Maximization with respect to the infinite-dimensional component $f(\cdot)$ is much harder than maximization with respect to the finite-dimensional component $\boldsymbol{\theta}$. Conceptually, when $\boldsymbol{\theta}^{(r)}$ is given, for each i and j, let

$$e_{ji}^{(r)} = y_i - {\boldsymbol{\beta}^{(r)}}^T \boldsymbol{x}_i - \alpha_j^{(r)}$$

and let $e_{j0}^{(r)} = 0$. Reindex the $e_{ji}^{(r)}$'s as $e_i^{(r)} (i = 1, \ldots, nk)$ and rearrange indices so that the absolute values of $e_i^{(r)}$ are in nondecreasing order over i. Then, let $e_0^{(r)} = 0$. For each $i = 1, \ldots, nk$, let $\gamma_i^{(r)}$

be the $\gamma_j^{(r)}$ that corresponds to $e_{ji}^{(r)}$. Then, as in Robertson et al. (1988, p. 326), the nonparametric maximum likelihood estimator of a unimodal density on $\mathbb{R}$ is a step function that is càdlàg on $\mathbb{R}^-$ and càglàd on $\mathbb{R}^+$, where here and hereafter we follow the convention $\mathbb{R}^- = (-\infty, 0)$ and $\mathbb{R}^+ = (0, \infty)$. (Here, càdlàg means "right-continuous with left limits" and càglàd means "left-continuous with right limits".) For each $i = 0, \ldots, nk$, let $f_i = f^{(r)}(e_i^{(r)})$. Let $\{f_i\}$ be the density function obtained from the f_i's. That is,

$$\{f_i\}(t) = \sum_{i=0}^{nk} I(-|e_i^{(r-1)}| < |t| \leq |e_i^{(r)}|) f_i.$$

Let $f_{ji} = f(e_{ji}^{(r)})$. Let $\{f_{ji}\}$ be the density function obtained from the f_{ji}'s. Then (3) can be rewritten as

$$f^{(r+1)}(\cdot) = \operatorname*{arg\,max}_{\{f_{ji}\} \in \mathcal{F}} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij}^{(r)} \log f_{ji}. \tag{4}$$

As shown in the Appendix, an explicit form for the maximization in (4) can be derived by using isotonic regression methods and taking a limit.

Let $I(\cdot)$ be the indicator function. Let

$$\hat{\gamma}_{ij} = \lim_{r \to \infty} \gamma_{ij}^{(r)},$$

$$\hat{e}_{ji} = y_i - \hat{\boldsymbol{\beta}}_n^T \boldsymbol{x}_i - \hat{\alpha}_{nj}$$

and

$$F_n(t) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{\hat{\gamma}_{ij}}{n} \frac{I(\hat{e}_{ji} \leq t) + I(\hat{e}_{ji} > -t)}{2}. \tag{5}$$

Let $\hat{F}_n^-(\cdot)$ be the greatest convex minorant of $F_n(\cdot)$ on $\mathbb{R}^-$, and $\hat{F}_n^+(\cdot)$ be the least concave majorant of $F_n(\cdot)$ on $\mathbb{R}^+$.

Lemma 1. *With probability 1, (i). $\hat{f}_n(t)$ is the right derivative of $\hat{F}_n^-(t)$, for each $t \in \mathbb{R}^-$; and (ii). $\hat{f}_n(t)$ is the left derivative of $\hat{F}_n^+(t)$, for each $t \in \mathbb{R}^+$.*

This lemma gives an expression for $\hat{f}_n$ in terms of $\hat{\boldsymbol{\theta}}_n$.

3 Theoretical results

Rewrite the density function $g(e \mid \boldsymbol{x})$ as $g(y \mid \boldsymbol{x}, \boldsymbol{\theta}, f)$. Since $h(\boldsymbol{x})$ is constant with respect to $\boldsymbol{\theta}$ and f, we can define our log-likelihood function to be

$$l(\boldsymbol{\theta}, f \mid y, \boldsymbol{x}) = \log g(y \mid \boldsymbol{x}, \boldsymbol{\theta}, f).$$

We need the following regularity conditions:

Condition 1. The support of $g(y \mid \boldsymbol{x}, \boldsymbol{\theta}, f)$ is $\mathbb{R}$, for all $\boldsymbol{\theta}$, f and $\boldsymbol{x}$ such that $h(\boldsymbol{x}) > 0$.

Condition 2. For P-almost every y and $\boldsymbol{x}$, $l(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})$ is Lipschitz in $(\boldsymbol{\theta}, f)$.

Condition 3. Θ is bounded.

Condition 4. $\mathcal{F}$ is uniformly bounded.

We show the convergence of the estimators based on the following theorems:

Theorem 1. *If Conditions 1 – 4 hold, then*

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \to 0 \textit{ almost surely}, \quad \textit{and} \quad \|\hat{f}_n - f_0\|_2 \to 0 \textit{ almost surely}.$$

Let $\dot{f}_0(\cdot)$ be the derivative of $f_0(\cdot)$. Let $\mathcal{L}_2^0(P)$ be the subspace of $\mathcal{L}_2(P)$ consisting of functions from $\mathbb{R} \times \mathbb{R}^d$ to $\mathbb{R}$ with zero mean and finite variance. Let

$$l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x}) = \partial l(\boldsymbol{\theta}, f \mid y, \boldsymbol{x}) / \partial \boldsymbol{\theta}.$$

For any function h from $\mathbb{R}^{d+1}$ to $\mathbb{R}$ and any functional j, define the operator $\partial/\partial f$ by

$$(\partial/\partial f) j(f)[h] = (\partial j((1 + \lambda h) f)/\partial \lambda)\big|_{\lambda=0},$$

and define

$$l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h] = \partial l(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})/(\partial f)[h] = \partial l(\boldsymbol{\theta}_0, (1 + \lambda h) f_0 \mid y, \boldsymbol{x})/\partial \lambda\big|_{\lambda=0}.$$

Let

$$l_{\boldsymbol{\theta}, f}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})[g] = \partial l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, f + \lambda g \mid y, \boldsymbol{x})/\partial \lambda \|_{\lambda=0},$$

and

$$l_{f, \boldsymbol{\theta}}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x}) = \partial l_f(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})/\partial \boldsymbol{\theta}.$$

Let

$$l_{\boldsymbol{\theta}, \boldsymbol{\theta}}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x}) = \partial^2 l(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})/\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}$$

and

$$l_{f,f}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})[g_1, g_2] = \partial l_f(\boldsymbol{\theta}, f + \lambda g_2 \mid y, \boldsymbol{x})[g_1]/\partial \lambda|_{\lambda=0}.$$

For any measure μ and function h, define $\mu(h) = \int h d\mu$. Since this is a linear functional, we can (and will) abbreviate this as μh.

Let $\tilde{l}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})$ be the efficient score function for $\boldsymbol{\theta}$ at $(\boldsymbol{\theta}_0, f_0)$, and let

$$\tilde{I}(\boldsymbol{\theta}_0 \mid f_0) = E((\tilde{l}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}))^{\otimes 2}).$$

Let $\mathbb{B}(\cdot)$ be the two-sided Brownian motion originating from zero: a mean zero Gaussian process on $\mathbb{R}$ with $\mathbb{B}(0) = 0$, and

$$E\big(\mathbb{B}(s) - \mathbb{B}(h)\big)^2 = |s - h|$$

for all $s, h \in \mathbb{R}$. Let $\tilde{\gamma}_j = \lim_{n \to \infty} \hat{\gamma}_j$.

Condition 5. For every $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, $d^2((\boldsymbol{\theta}, f), (\boldsymbol{\theta}_0, f_0)) \lesssim P(l(\boldsymbol{\theta}_0, f_0) - l(\boldsymbol{\theta}, f))$.

Condition 6. $\tilde{I}(\boldsymbol{\theta}_0 \mid f_0)$ is positive definite.

Condition 7. $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})$ exists.

Condition 8. For all $h \in \mathcal{L}_2^0(P)$, $Pl_f(\boldsymbol{\theta}_0, f_0)[h] = \partial Pl(\boldsymbol{\theta}_0, f_0)[h]/\partial f$.

Condition 9. For all h in the inverse image $l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})^{-1}(\mathcal{L}_2^0(P))^{d+k}$, $\mid l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h] \mid_2 < \infty$.

Condition 10. $\partial Pl(\boldsymbol{\theta}_0, f_0)/\partial \boldsymbol{\theta} = Pl_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)$.

Condition 11. For P-almost every y and $\boldsymbol{x}$, $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})$ is Lipschitz in $(\boldsymbol{\theta}, f)$.

Condition 12. $(Pl_{\boldsymbol{\theta},f})(\boldsymbol{\theta}_0, f_0)[g] = (\partial^2 Pl(\boldsymbol{\theta}_0, f_0)/\partial\boldsymbol{\theta}\partial f)[g]$, for each function g from $\mathbb{R}^{d+1}$ to $\mathbb{R}$.

Condition 13. $(Pl_{f,f})(\boldsymbol{\theta}_0, f_0)[g_1, g_2] = (\partial^2 Pl(\boldsymbol{\theta}_0, f_0)/\partial f^2)[g_1, g_2]$, for all functions g_1 and g_2 from $\mathbb{R}^{d+1}$ to $\mathbb{R}$.

Condition 14. $(Pl_{\boldsymbol{\theta},\boldsymbol{\theta}})(\boldsymbol{\theta}_0, f_0) = \partial^2 Pl(\boldsymbol{\theta}_0, f_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$.

Condition 15. $(Pl_{f,\boldsymbol{\theta}})(\boldsymbol{\theta}_0, f_0)[g] = (\partial^2 Pl(\boldsymbol{\theta}_0, f_0)/\partial f\partial\boldsymbol{\theta}^T)[g]$, for each function g from $\mathbb{R}^{d+1}$ to $\mathbb{R}$.

Condition 16. $\dot{f}_0(t)$ exists and $\dot{f}_0(t) \neq 0$.

Condition 17. The support of $\boldsymbol{x}$ is bounded.

Condition 18. For all y and $\boldsymbol{x}$, $l_{\boldsymbol{\theta},\boldsymbol{\theta}}(\boldsymbol{\theta}, f_0 \mid y, \boldsymbol{x})$ exists in a neighborhood about $\boldsymbol{\theta}_0$.

To prove Theorem 2, we first prove the following result:

Lemma 2. *If Conditions 1 – 5 hold, then* $d((\hat{\boldsymbol{\theta}}_n, \hat{f}_n), (\boldsymbol{\theta}_0, f_0)) = O_p(n^{-1/3})$.

Theorem 2. *If Conditions 1 – 15 hold, then*

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \rightarrow N(\mathbf{0}, \tilde{I}^{-1}(\boldsymbol{\theta}_0 \mid f_0)) \text{ in distribution.}$$

Theorem 3. *If Conditions 1 – 4 and Conditions 16 – 18 hold, then*

$$n^{1/3}\left(\hat{f}_n(t) - f_0(t)\right) \rightarrow \left(2|\dot{f}_0(t)|f_0(t)\left(\sum_{j=1}^{k}\Gamma_j^{1/2}\right)^2\right)^{1/3} \arg\max_{h\in\mathbb{R}}(\mathbb{B}(h) - h^2) \text{ in distribution,}$$

where $\Gamma_j = E(\tilde{\gamma}_j^2)$.

4 Simulation results

We simulate $n = 1000$ independent and identically distributed data with 1-dimensional response y_i's and with covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})^T$. We first generate the covariates, sample the $\boldsymbol{x}_i$'s from the 5-dimensional normal distribution with mean vector $\boldsymbol{\mu} = (3.1, 1.8, -0.5, 0.6, 1.5)^T$ and some given covariance matrix Γ, say

$$\Gamma^{1/2} = \begin{pmatrix} 1.13 & -0.27 & 0.55 & 0.82 & 0.47 \\ & 1.34 & -0.14 & 0.57 & 0.34 \\ & & 1.52 & 0.57 & -0.53 \\ & & & 0.72 & -0.40 \\ & & & & 0.96 \end{pmatrix}.$$

Set $\boldsymbol{\beta}_0 = (1.2, -2.1, 0.6, 1.5, 0.8)^T$. Set $k = 2$, $\boldsymbol{\alpha}_0 = (0.15, 0.87)^T$, $\boldsymbol{\zeta} = (1.1, 0.2, -0.4, 0.6, 0.8)^T$, and

$$\delta_1(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\zeta}^T\boldsymbol{x})}{1 + \exp(\boldsymbol{\zeta}^T\boldsymbol{x})}, \quad \delta_2(\boldsymbol{x}) = \frac{1}{1 + \exp(\boldsymbol{\zeta}^T\boldsymbol{x})}.$$

Then conditional on the covariate $\boldsymbol{x}_i$, we sample the error ϵ_i. Sampling ϵ_i from the additive mixture

$$\delta(\boldsymbol{x}_i)f(\cdot - \alpha_{0,1}) + \delta_2(\boldsymbol{x})f(\cdot - \alpha_{0,2})$$

model is straightforward. We set $f(\cdot)$ to be the Student's t-distribution with 5 degrees of freedom. Then we generate the response data, given the covariates, as $y_i = \boldsymbol{\beta}_0^T \boldsymbol{x}_i + \alpha_{0,1} + \epsilon_i$ if ϵ_i is from $f(\cdot - \alpha_{0,1})$ and $y_i = \boldsymbol{\beta}_0^T \boldsymbol{x}_i + \alpha_{0,2} + \epsilon_i$ if ϵ_i is from $f(\cdot - \alpha_{0,2})$.

We used the nonparametric maximum likelihood estimator to estimate $\hat{f}_n$ under unimodal constraint, it is uniquely obtained as a formulation of the isotonic regression problem (see proof of Lemma 1). The pool adjacent violators algorithm (Best and Chakravarti, 1990) is a convenient computational tool to perform such order restricted maximization or minimization, and is available in R. Mair et al. (2009) gives a review of the algorithm history and computational aspects. To use the pool adjacent violators algorithm, we use the following iterative procedure. Given starting values $\boldsymbol{\theta}^{(0)}$ (we can first set $f(\cdot) \sim N(0,1)$ and use the maximum likelihood estimate from this model as $\boldsymbol{\theta}^{(0)}$) , set

$$\hat{\epsilon}_{1,i} = y_i - \boldsymbol{\beta}_i^{(0)T} \boldsymbol{x}_i - \alpha_1^{(0)} \quad \text{and} \quad \hat{\epsilon}_{2,i} = y_i - \boldsymbol{\beta}_i^{(0)T} \boldsymbol{x}_i - \alpha_2^{(0)}, \quad (i = 1, \ldots, n),$$

then use the pool adjacent violators algorithm to estimate $\hat{f}^{(1)}(\cdot)$. Plugging in $\hat{f}^{(1)}(\cdot)$ into (1), use maximum likelihood to estimate $\hat{\theta}^{(1)}$. Then iterate to get $\hat{f}^{(r)}(\cdot)$ and $\hat{\boldsymbol{\theta}}^{(r)}$ until the convergence criterion is met and the final values are essentially equal to $\hat{f}_n(\cdot)$ and $\hat{\boldsymbol{\theta}}_n$.

The data was sampled from 3 different generating parameter values. The first time, the e_i's were sampled from our mixture model with f_0 being the Laplace distribution with location parameter 0 and scale parameter 1, and true parameters $\boldsymbol{\beta}_0 = (1.2, -2.1, 0.6, 1.5, 0.8)^T$ and $\boldsymbol{\alpha}_0 = (0.15, 1.25)^T$, and starting values $f^{(0)} = N(0,1)$ and $\boldsymbol{\beta}^{(0)} = (1.3, -2.2, 0.7, 1.4, 0.7)^T$ and $\boldsymbol{\alpha}^{(0)} = (0.21, 1.19)^T$. For the second data set f_0 was fixed to be the p-generalized normal distribution (also known as the exponential power distribution or Subbotin distribution) with location parameter $\mu = 0$, shape parameter $\beta = 6$, and scale parameter α equal to the default in the R package `pgnorm`. The same true Euclidean parameters and same starting values were used. For the third data set the p-generalized normal was used as the true distribution but set $\boldsymbol{\beta}_0 = (0.8, 1.5, -1.3, 0.7, 2.1)^T$ and $\boldsymbol{\alpha}_0 = (0, 2.2)^T$, and the starting values were $\boldsymbol{\beta}^{(0)} = (0.9, 1.6, -1.2, 0.8, 2.0)^T$ and $\boldsymbol{\alpha}^{(0)} = (0.06, 2.25)^T$. The results are given in Tables 1 and 2.

Under both profile and normal densities, the estimates of $\boldsymbol{\beta}$ were very accurate. It can be seen from the table that the estimates' mean values (computed over 200 trials with n=1000 for the first set, over 100 trials with n=100 for the second set, and 10 trials with n=100 for the third set) are at most .13 off from the true parameter values. However, estimating $\boldsymbol{\alpha}$ is difficult, probably due to the unobserved memberships; for example, the average value of $\hat{\alpha}_{n,1}$ is off by 0.50 for the normal estimate in data set 1, and by a whopping 0.83 for the that in data set 3. Similarly, the average value of $\hat{\alpha}_{n,2}$ is off by 0.32 for the profile estimate in data set 2 and by 0.99 for the profile estimate in data set 3.

Generally, the estimate under the profile likelihood tended to have the greatest likelihood, then the true parameter values under the true distribution, then the estimate under the normal distribution. This suggests that the estimates under the profile likelihood actually were the maximum likelihood estimates; the code actually optimized the parameters correctly. But we do not know why the estimates for $\boldsymbol{\alpha}$ be so consistently off from the true values.

On the other hand, the estimates for the density function were only fairly accurate. The actual regression was performed using R's package `Iso`'s `ufit` function. It seemed to return different results than those from running `Iso`'s `pava` twice, on the negative side and on the positive side. Usually, `ufit` would perform better than `pava`. One interesting thing to note is that oftentimes the estimated density function would have a large spike at and around 0. This is not precluded by Theorem 1, as we are only guaranteed (with probability 1) convergence of $\hat{f}_n$ to f in the L_2

seminorm and not, for example, in the supremum norm. In the future, this problem could potentially be eliminated with interpolation.

We compare $\hat{\boldsymbol{\theta}}_n$ with the maximum likelihood estimator $\bar{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ under the normal model in Tables 1 and 2, and the estimated $\hat{f}_n$ and the normal density in Figures 1 and 2. These tables and figures are at the end of the Appendix.

5 Discussion

Work on this project has raised three important questions. First, it has already been shown that for a location shift family of mixtures, unimodality and symmetry is a sufficient condition for identifiability (with the exception of a set of measure zero of mixing proportions). Is symmetry a necessary condition? We conjecture that unimodality alone is sufficient. Then, there is the problem of determining the efficient score. We have tried but have not been able to find a formula for the efficient score under this likelihood. Finally, there is the problem of estimating the mixing proportions when they are unknown. It would be interesting to see how estimating the mixing proportions would affect the results we derived in Theorems 1, 2 and 3.

A Appendix

A.1 Proof of Lemma 1

We draw inspiration from Robertson et al. (1988, p. 332-334). Reindex the $\pm e_{ji}^{(r)}$'s as $\{e_i^{(r)} : i = 1, \ldots, N; N = 2kn\}$ such that the $e_i^{(r)}$'s are in increasing order, and write $f_i = f(e_i^{(r)})$. Let w be the integer such that $e_w^{(r)} < 0 < e_{w+1}^{(r)}$, $c_i = e_{i+1}^{(r)} - e_i^{(r)}$ (if $i = 1, \ldots, w-1$), $-e_i^{(r)}$ (if $i = w$), $e_i^{(r)}$ (if $i = w+1$), $e_i^{(r)} - e_{i-1}^{(r)}$ (if $i = w+2, \ldots, N$). With probability 1, $c_i \neq 0$, $i = 1, \ldots, N$, and r exists (upon which $r = kn$). It follows from the argument by Robertson et al. (1988, p. 326) that $f_n^{(r+1)}(\cdot)$ is a step function, cádlág on $(-\infty, 0]$ and cáglád on $[0, \infty)$, that equals zero on $(-\infty, e_1^{(r)}) \cup (e_N^{(r)}, \infty)$. Hence, the constraint $\int \{f_{ji}\}(t)dt = 1$ in (4) is written as

$$\sum_{i=1}^{w-1} f_i \cdot (e_{i+1}^{(r)} - e_i^{(r)}) + f_w \cdot (0 - e_w^{(r)}) + f_{w+1} \cdot (e_{w+1}^{(r)} - 0) + \sum_{i=w+2}^{N} f_i \cdot (e_i^{(r)} - e_{i-1}^{(r)}) = 1. \qquad (A.1)$$

To simplify (4), let $g_i = \gamma^{(r)}(\boldsymbol{x}_i)/(Nc_i)$, $\gamma^{(r)}(\boldsymbol{x}_i)$ be the $\gamma_j^{(r)}(\boldsymbol{x}_i)$ corresponding to $e_{ji}^{(r)}$, and $w_i = Nc_i$. Now we can rewrite (4) as

$$f_n^{(r+1)}(\cdot) = \operatorname*{arg\,max}_{\{f_i\} \in \mathcal{F}} \sum_{i=1}^{N} g_i w_i \log(f_i),$$

however, written in this form will lead to simplification using results in isotonic regression.

For $u \in \mathbb{R}^+$, let $\Phi(u) = u \log u$, $\phi(u) = 1 + \log u$, and

$$\Delta_\Phi(u, v) = \Phi(u) - \Phi(v) - (u - v)\phi(v) = u \log u - u \log v - (u - v).$$

Note $\Phi(\cdot)$ is convex on $\mathbb{R}^+$. Since $\gamma^{(r)}(\boldsymbol{x}_i) \log \gamma^{(r)}(\boldsymbol{x}_i)$ is a constant with respect to f_i, $\arg\min_S(-\cdot) = \arg\max_S(\cdot)$ and for each $\{f_i\} \in \mathcal{F}$, $\sum_{i=1}^{N} f_i c_i = 1$, we have

$$\begin{aligned}\operatorname*{arg\,max}_{\{f_i\} \in \mathcal{F}} \sum_{i=1}^{N} g_i w_i \log(f_i) &= \operatorname*{arg\,min}_{\{f_i\} \in \mathcal{F}} \sum_{i=1}^{N} \Big(\gamma^{(r)}(\boldsymbol{x}_i) \log \gamma^{(r)}(\boldsymbol{x}_i) - \gamma^{(r)}(\boldsymbol{x}_i) \log f_i - \big(\gamma^{(r)}(\boldsymbol{x}_i) - f_i\big)\Big) c_i \\ &= \operatorname*{arg\,min}_{\{f_i\} \in \mathcal{F}} \sum_{i=1}^{N} \Delta_\Phi\big(\gamma^{(r)}(\boldsymbol{x}_i), f_i\big) c_i = \operatorname*{arg\,min}_{\{f_i\} \in \mathcal{F}} \sum_{i=1}^{N} \Delta_\Phi\big(g_i, f_i\big) w_i.\end{aligned}$$

We have that $\Phi(u)$ is a convex function and $\phi(u) = d\Phi(u)/du$, so the solution to the above minimization is equivalent to

$$\operatorname*{arg\,min}_{\{f_i\}\in\mathcal{F}} \sum_{i=1}^{N} w_i\big(g_i - f_i\big)^2$$

(Robertson et al., 1988, p. 31, Theorem 1.5.1). Let $W_i = \sum_{j=1}^{i} w_j = Ne_i^{(r)}$ and $G_i = \sum_{j=1}^{i} w_j g_j = \sum_{j=1}^{i} \gamma^{(r)}(\boldsymbol{x}_j)/2$. Hence, on $\mathbb{R}^-$, $f_n^{(r)}(\cdot)$ is the right derivative of the greatest convex minorant of the sum diagram of

$$\{(W_i/N, G_i/N) : i = 1, \ldots, N\},$$

and on $\mathbb{R}^+$, $f_n^{(r)}(\cdot)$ is the left derivative of the least concave majorant of the sum diagram (Robertson et al., 1988, p. 7-8, Theorem 1.2.1; p. 332-334). Let $I(\cdot)$ be the indicator function, and let

$$\begin{aligned} F_n^{(r+1)}(t) &= \sum_{i=1}^{N} \frac{\gamma^{(r)}(\boldsymbol{x}_i)}{2n}\big(I(e_i^{(r)} \le t)I(t \ge 0) + I(e_i^{(r)} < t)I(t < 0)\big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \frac{\gamma_j^{(r)}(\boldsymbol{x}_i)}{n} \frac{I(e_{ji}^{(r)} \le t) + I(e_{ji}^{(r)} > -t)}{2} \end{aligned}$$

be a modified weighted empirical distribution function of the $e_i^{(r)}$'s. Note that $F_n(t)$ has the same greatest convex minorant on $\mathbb{R}^-$ and least concave majorant on $\mathbb{R}^+$ as the sum diagram. Thus, for each r, $f_n^{(r+1)}$ is equal to the right derivative of the least concave majorant of $F_n^{(r+1)}(t)$ on $\mathbb{R}^-$ and the left derivative of the greatest convex minorant of $F_n^{(r)}(t)$ on $\mathbb{R}^+$. Thus, taking the limit as r goes to infinity of $f^{(r)}$, we find $\hat{f}_n$ is equal to the right derivative of the least concave majorant of $\hat{F}_n(t)$ on $\mathbb{R}^-$ and the left derivative of the greatest convex minorant of $\hat{F}_n(t)$ on $\mathbb{R}^+$.

A.2 Proof of Proposition 1

Assume that

$$\sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f(y - \boldsymbol{\beta}^T\boldsymbol{x} - \alpha_j) \equiv \sum_{j=1}^{k} \delta_j(\boldsymbol{x})\tilde{f}(y - \tilde{\boldsymbol{\beta}}^T\boldsymbol{x} - \tilde{\alpha}_j), \quad \text{for each } (\boldsymbol{\theta}, f), (\tilde{\boldsymbol{\theta}}, \tilde{f}) \in \boldsymbol{\Theta}\times\mathcal{F}; y \in \mathbb{R}; \boldsymbol{x} \in \mathbb{R}^d.$$

We want to show $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}, \tilde{f}) = (\boldsymbol{\beta}, \boldsymbol{\alpha}, f)$. First, take $\boldsymbol{x} = \boldsymbol{0}$. Then the above reduces to

$$\sum_{j=1}^{k} \delta_j(\boldsymbol{0}) f(y - \alpha_j) \equiv \sum_{j=1}^{k} \delta_j(\boldsymbol{0})\tilde{f}(y - \tilde{\alpha}_j),$$

and by the identifiability of model (0) we have $(\tilde{\boldsymbol{\alpha}}, \tilde{f}) = (\boldsymbol{\alpha}, f)$.

Now we need to show $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$, or equivalently, $\tilde{\beta}_i = \beta_i, i = 1, \ldots, d$. Without loss of generality, we only show $\tilde{\beta}_1 = \beta_1$. Set $\boldsymbol{x} = (x_1, 0, \ldots, 0)^T$. Then the above equation becomes

$$\sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f(y - \beta_1 x_1 - \alpha_j) \equiv \sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f(y - \tilde{\beta}_1 x_1 - \alpha_j).$$

Without loss of generality, assume $\alpha_1 < \cdots < \alpha_k$. We show that if $\beta_1 > \tilde{\beta}_1$, then this equality cannot hold. In fact, since $f(\cdot)$ is monotone increasing on $\mathbb{R}^-$, we can choose $y \in \mathbb{R}$ and $x_1 < 0$ such that $y - \beta_1 x_1 - \alpha_1 < 0$, so $y - \beta_1 x_1 - \alpha_j < 0$ and $y - \tilde{\beta}_1 x_1 - \alpha_j < 0$ for all j. We can choose x_1 such that $f(y - \beta_1 x_1 - \alpha_j) < f(y - \tilde{\beta}_1 x_1 - \alpha_j)$ for at least one j. Thus the condition $y - \tilde{\beta}_1 x_1 - \alpha_j < y - \beta_1 x_1 - \alpha_j$ for all j implies $f(y - \tilde{\beta}_1 x_1 - \alpha_j) \leq f(y - \beta_1 x_1 - \alpha_j)$ for all j, and strict inequality holds for at least one j. Then

$$\sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f(y - \beta_1 x_1 - \alpha_j) > \sum_{j=1}^{k} \delta_j(\boldsymbol{x}) f(y - \tilde{\beta}_1 x_1 - \alpha_j),$$

contradiction. The case $\beta_1 < \tilde{\beta}_1$ similarly yields a contradiction. Thus $\tilde{\beta}_1 = \beta_1$.

A.3 Proof of Theorem 1 (Sketch)

We only give a sketch of the proof, due to lack of space. Let $P(y, \boldsymbol{x})$ be the probability measure of $p(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0, f_0)$. Define

$$m(y, \boldsymbol{x} \mid \boldsymbol{\theta}, f) = \log \frac{p(y, \boldsymbol{x} \mid \boldsymbol{\theta}, f)}{p(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0, f_0)}.$$

By Condition 1, the numerator of the above expression will always be zero whenever the denominator is zero, so m is well-defined. We use the notation

$$Pm(\boldsymbol{\theta}, f) = \int m(y, \boldsymbol{x} \mid \boldsymbol{\theta}, f) dP(y, \boldsymbol{x}),$$

the true mean of m; and

$$P_n m(\boldsymbol{\theta}, f) = n^{-1} \sum_{i=1}^{n} m(y_i, \boldsymbol{x}_i \mid \boldsymbol{\theta}, f),$$

the empirical mean of m from the data $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$.

$Pm(\boldsymbol{\theta}, f)$ is the negative Kullback-Leibler divergence of $p(y, \boldsymbol{x} \mid \boldsymbol{\theta}, f)$ from $p(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0, f_0)$, and so is always nonpositive, attaining its maximum value of 0 whenever $p(y, \boldsymbol{x} \mid \boldsymbol{\theta}, f) = p(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0, f_0)$ almost everywhere.

Now use Conditions Conditions 2 – 4 and Theorems 2.7.5 and 2.4.1 in van der Vaart and Wellner (1996) to show $\mathcal{M}$ is a Glivenko-Cantelli class with respect to P. By Theorem 5.8 in (van der Vaart, 2002, p. 386),

$$d((\hat{\boldsymbol{\theta}}_n, \hat{f}_n), (\boldsymbol{\theta}_0, f_0)) \to 0 \text{ almost surely.}$$

A.4 Proof of Lemma 2 (Sketch)

Use Lemma 3.4.2 in van der Vaart and Wellner (1996) and Theorem 3.2.5 in van der Vaart and Wellner (1996) to conclude

$$n^{1/3} d((\hat{\boldsymbol{\theta}}_n, \hat{f}_n), (\boldsymbol{\theta}_0, f_0)) = O_p(1).$$

A.5 Proof of Theorem 2

Let Λ be the nuisance tangent space for this model. Condition 6 implies that $\tilde{l}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})$ is square-integrable, and hence that $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})$ and $\Pi(l(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) \mid \Lambda)$ are square-integrable. For any $r \in \mathbb{N}$ and any function $\boldsymbol{h}(\cdot) = (h_1(\cdot), \ldots, h_r(\cdot))^T$ from $\mathbb{R}^{d+1}$ to $\mathbb{R}^r$, define

$$l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}] = (l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h_1], \ldots, l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h_r])^T.$$

Consider the set of functions of the form $Bl_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}]$ with $B \in \text{Hom}(\mathbb{R}^{d+k}, \mathbb{R}^r)$ (the set of $(d+k) \times r$ real matrices), $h \in (\mathcal{L}_2^0(P))^r$ and $r \in \mathbb{N}$. With Conditions 8 and 9, this space is a subset of $(\mathcal{L}_2^0(P))^{d+k}$. The closure in $(\mathcal{L}_2(P))^{d+k}$ of this space is equal to Λ, by definition of nuisance tangent space. Clearly, Λ is symmetric, in that if $(h_1, \ldots, h_{d+k})^T$ is in Λ then for any permutation

σ, so is $(\sigma(h_1), \ldots, \sigma(h_{d+k}))^T$. By Condition 6 and Condition 10, $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) \in (\mathcal{L}_2^0(P))^{d+k}$. By the Hilbert projection theorem, there exists an element, say λ^*, of Λ, such that

$$\text{for each } \lambda \in \Lambda, E((l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) - \lambda^*(y, \boldsymbol{x}))^T \lambda(y, \boldsymbol{x})) = 0.$$

There exists a sequence $B_q l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}_q]$, with $B_q \in \mathbb{R}^{r_q}$ and $\boldsymbol{h}_q \in (\mathcal{L}_2^0(P))^{r_q}$, for some $r_q \in \mathbb{N}$, such that

$$\lambda^* = \lim_{q \to \infty} B_q l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}_q].$$

By inspection, $l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h]$ is continuous with respect to h and linear with respect to h. Therefore $l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}]$ must be continuous with respect to $\boldsymbol{h}$, so the inverse image of any closed set is closed. Therefore, $l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})^{-1}(\mathcal{L}_2^0(P))^{d+k}$ is closed.

We have

$$\lim_{q \to \infty} (\sum_{i=1}^{r_q} (B_q)_{1i} l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h_{qi}], \ldots, \sum_{i=1}^{r_q} (B_q)_{d+k,i} l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h_{qi}])^T =$$

$$l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\lim_{q \to \infty} B_q \boldsymbol{h}_q].$$

Let $\boldsymbol{h}^* = \lim_{q \to \infty} B_q \boldsymbol{h}_q$. Then $l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}^*] = \lambda^*(y, \boldsymbol{x})$.

Therefore, $\lim_{q \to \infty} B_q \boldsymbol{h}_q \in (\mathcal{L}_2^0(P))^r$. Therefore, $l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h] \in (\mathcal{L}_2^0(P))^{d+k}$. By symmetry, we have

$$E((l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) - \lambda^*(y, \boldsymbol{x})) l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[h]) = \mathbf{0}$$

for all $h \in \mathcal{L}_2^0(P)$. Thus, $\Pi(l(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x} \mid \Lambda) = l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}^*]$, so

$$\tilde{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) = l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) - l_f(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\boldsymbol{h}^*].$$

Let

$$\mathcal{M}_1 = \{l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}, f \in \mathcal{F}\}$$

and

$$\mathcal{M}_2 = \{l_f(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})[\boldsymbol{h}^*] : \boldsymbol{\theta} \in \boldsymbol{\Theta}, f \in \mathcal{F}\}.$$

From Condition 3 we know $\boldsymbol{\Theta}$ is Donsker, and from Condition 4 we know $\mathcal{F}$ is Donsker, as it is a set of unimodal functions. Similarly as the entropy computation of $\mathcal{M}$ in the proof of Theorem 1, we can show that

$$N_{[\,]}(\epsilon, \mathcal{M}_j, \mathcal{L}_2(P)) = O\big(\exp(C/\epsilon)\big)$$

for some $0 < C < \infty$, and so $\tilde{J}_{[\,]}(1, \mathcal{M}_j, \mathcal{L}_2(P)) < \infty$ $(j = 1, 2)$. Hence, from Theorem 6.8 in van der Vaart (2002, p. 401) we have that $\mathcal{M}_1$ and $\mathcal{M}_2$ are Donsker. Thus from Corollary 2.3.12 (van der Vaart and Wellner, 1996, p. 115), we have

$$\lim_{\epsilon \downarrow 0} \lim_{\delta \downarrow 0} \limsup_{n \to \infty} P^*(\sup_{f \in \mathcal{M}_{1,\delta}} |n^{1/2}(P_n - P)f| > \epsilon) = 0$$

and

$$\lim_{\epsilon \downarrow 0} \lim_{\delta \downarrow 0} \limsup_{n \to \infty} P^*(\sup_{f \in \mathcal{M}_{2,\delta}} |n^{1/2}(P_n - P)f| > \epsilon) = 0,$$

where

$$\mathcal{M}_{1,\delta} = \{f - g : f, g \in \mathcal{M}_1, \rho_P(f - g) < \delta\},$$

$$\mathcal{M}_{2,\delta} = \{f - g : f, g \in \mathcal{M}_2, \rho_P(f - g) < \delta\}$$

and $\rho_P(f) = (P(f - Pf)^2)^{1/2}$. For any function h from $\mathbb{R} \times \mathbb{R}^d$ to $\mathbb{R}$, let $Ph = E(h(y, \boldsymbol{x}))$ and $P_n h = n^{-1} \sum_{i=1}^n h(y_i, \boldsymbol{x}_i)$, where $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ are the data. Since by Theorem 1 $d((\hat{\boldsymbol{\theta}}_n, f_n), (\boldsymbol{\theta}_0, f_0)) \to 0$ almost surely, and since by inspection for all $(y, \boldsymbol{x})$, $l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})$ and $l_f(\boldsymbol{\theta}, f \mid y, \boldsymbol{x})$ are continuous in $(\boldsymbol{\theta}, f)$, we have

$$n^{1/2}(P_n - P)(l_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \hat{f}_n) - l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)) = o_p(1) \tag{A.1}$$

and

$$n^{1/2}(P_n - P)(l_f(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)[\boldsymbol{h}^*] - l_f(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*]) = o_p(1). \tag{A.2}$$

Since $(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)$ maximizes $l(\boldsymbol{\theta}, f \mid \mathbf{D}_n)$, and since $P_n l(\boldsymbol{\theta}, f) \propto l(\boldsymbol{\theta}, f \mid \mathbf{D}_n)$, and since by Condition 10 $l_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \hat{f}_n \mid y, \boldsymbol{x})$ and $l_f(\hat{\boldsymbol{\theta}}_n, \hat{f}_n \mid y, \boldsymbol{x})[\boldsymbol{h}^*]$ exist and by the homogeneity of the differential operator, we have $P_n l_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \hat{f}_n) = \mathbf{0}$ and $P_n l_f(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)[\boldsymbol{h}^*] = \mathbf{0}$. Also, by Condition 10 and Condition 8 we have $P l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0) = \mathbf{0}$ and $P l_f(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*] = \mathbf{0}$. Hence

$$n^{1/2}(P l_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \hat{f}_n) + P_n l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)) = o_p(1)$$

and

$$n^{1/2}(P l_f(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)[\boldsymbol{h}^*] + P_n l_f(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*]) = o_p(1).$$

From Theorem 1, with probability 1, for large enough n, $(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)$ will be in the disks of convergence for the Taylor series of $P l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)$ and $P l_f(\boldsymbol{\theta}_0, f_0)$. Note also that $P(a_n = o_p(1)) = 1$ implies $a_n = o_p(1)$, for any sequence (a_n). From expanding $P l_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)$ and using Lemma 2 we find

$$P l_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n, \hat{f}_n \mid y, \boldsymbol{x}) - P l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x}) - P l_{\boldsymbol{\theta},\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - P l_{\boldsymbol{\theta},f}(\boldsymbol{\theta}_0, f_0 \mid y, \boldsymbol{x})[\hat{f}_n - f_0]$$

$$= O_p(d^2((\hat{\boldsymbol{\theta}}_n, \hat{f}_n), (\boldsymbol{\theta}_0, f_0))) = O_p(n^{-2/3}).$$

Hence, adding $n^{-1/2}$ times (A.1) we have

$$-P_n l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0) - P l_{\boldsymbol{\theta},\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - P l_{\boldsymbol{\theta},f}(\boldsymbol{\theta}_0, f_0)[\hat{f}_n - f_0] = o_p(n^{-1/2}). \tag{A.3}$$

Similarly, from expanding $P l_f(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)[\boldsymbol{h}^*]$ we find

$$P l_f(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)[\boldsymbol{h}^*] - P l_f(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*] - P l_{f,\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - P l_{f,f}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*, \hat{f}_n - f_0] =$$

$$O_p(d^2((\hat{\boldsymbol{\theta}}_n, \hat{f}_n), (\boldsymbol{\theta}_0, f_0))) = O_p(n^{-2/3}).$$

Hence, adding $n^{-1/2}$ times (A.2), we have

$$-P_n l_f(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*] - P l_{f,\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*](\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - P l_{f,f}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*, \hat{f}_n - f_0] = o_p(n^{-1/2}). \tag{A.4}$$

Given Condition 12 we have

$$-P l_{\boldsymbol{\theta},f}(\boldsymbol{\theta}_0, f_0)[g] = P\big(l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0) l_f(\boldsymbol{\theta}_0, f_0)[g]\big) \text{ for all } g,$$

and given Condition 13 we have

$$-P\big(l_{f,f}(\boldsymbol{\theta}_0, f_0)[g_1, g_2]\big) = P\big(l_f(\boldsymbol{\theta}_0, f_0)[g_1] l_f(\boldsymbol{\theta}_0, F_0)[g_2]\big) \text{ for all } g_1, g_2.$$

Thus,

$$P l_{\boldsymbol{\theta},f}(\boldsymbol{\theta}_0, f_0)[\hat{f}_n - f_0] - P l_{f,f}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*, \hat{f}_n - f_0]$$

$$= Pl_{\boldsymbol{\theta},f}(\boldsymbol{\theta}_0, f_0)[\hat{f}_n - f_0] - Pl_{f,f}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*, \hat{f}_n - f_0] - Pl_{\boldsymbol{\theta},f}(\boldsymbol{\theta}_0, f_0)[f_0 - f_0] + Pl_{f,f}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*, f_0 - f_0]$$

$$= O_p(d^3((\hat{\boldsymbol{\theta}}_n, \hat{f}_n), (\boldsymbol{\theta}_0, f_0))) = O_p(n^{-1}).$$

From Conditions 14 and 15,

$$P\big(l_{\boldsymbol{\theta},\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0) - l_{f,\boldsymbol{\theta}}(\boldsymbol{\theta}_0, f_0)[\boldsymbol{h}^*]\big) = -\tilde{I}(\boldsymbol{\theta}_0 \mid f_0).$$

Hence, subtracting (A.3) from (A.4), multiplying both sides by $n^{1/2}$ and left-multiplying by $\tilde{I}^{-1}(\boldsymbol{\theta}_0, f_0)$, we have

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \tilde{I}^{-1}(\boldsymbol{\theta}_0, f_0) n^{1/2} P_n \tilde{l}(\boldsymbol{\theta}_0, f_0) + o_p(1),$$

which gives the desired result.

A.6 Proof of Theorem 3

Without loss of generality, assume $t \in \mathbb{R}^+$; the proof for $t \in \mathbb{R}^-$ is similar. Condition 16 excludes $t = 0$ and $f(t) = 0$. Recall $F_n(t)$ from (5), defining the process

$$\hat{S}_n(a) = \operatorname*{arg\,max}_s (F_n(s) - as), \quad a \in \mathbb{R}^+.$$

As shown in Lemma 1, $\hat{f}_n(t)$ is the left derivative of the least concave majorant of $F_n(t)$, so by the argument in Example 3.2.14 (van der Vaart and Wellner, 1996, p. 296-297),

$$\hat{f}_n(t) \le a \quad \text{if and only if} \quad \hat{S}_n(a) \le t, \quad \text{for all } t, a \in \mathbb{R}^+.$$

We are to evaluate the distribution function of $n^{1/3}(\hat{f}_n(t) - f_0(t))$. For $f_0(t) > 0$, the above relationship gives

$$\text{pr}\left(n^{1/3}(\hat{f}_n(t) - f_0(t)) \le x\right) = \text{pr}\left(\hat{S}_n(f_0(t) + xn^{-1/3}) - t \le 0\right),$$

for each $x \in \mathbb{R}$, for each n such that $f_0(t) + xn^{-1/3} \ge 0$. Substituting $s = t + hn^{-1/3}$ in the definition of $\hat{S}_n$, we have

$$\hat{S}_n\big(f_0(t) + xn^{-1/3}\big) - t = n^{-1/3} \operatorname*{arg\,max}_h \Big(F_n(t + hn^{-1/3}) - \big(f_0(t) + xn^{-1/3}\big)(t + hn^{-1/3})\Big)$$

$$= n^{-1/3} \operatorname*{arg\,max}_h \Big(F_n(t + hn^{-1/3}) - f_0(t)hn^{-1/3} - xhn^{-2/3}\Big)$$

$$= n^{-1/3} \operatorname*{arg\,max}_h \Big(n^{2/3} F_n(t + hn^{-1/3}) - f_0(t)hn^{1/3} - xh\Big)$$

$$= n^{-1/3} \operatorname*{arg\,max}_h \Big(n^{2/3}\big(F_n(t + hn^{-1/3}) - F_n(t)\big) - f_0(t)hn^{1/3} - xh\Big)$$

$$= n^{-1/3} \operatorname*{arg\,max}_h B_n(h),$$

where

$$B_n(h) = n^{2/3}\big(F_n(t + hn^{-1/3}) - F_n(t)\big) - f_0(t)hn^{1/3} - xh.$$

Thus,

$$\text{pr}\Big(n^{1/3}\big(\hat{f}_n(t) - f_0(t)\big) \le x\Big) = \text{pr}\Big(\operatorname*{arg\,max}_h B_n(h) \le 0\Big).$$

Below we only need to evaluate the asymptotic probability of the event $\{\text{arg max}_h B_n(h) \leq 0\}$. Let $F_0(\cdot)$ be the distribution function of $f_0(\cdot)$, and for all j and any function $g(\boldsymbol{x}, e_j)$ let $P_n g = \sum_{i=1}^{n} g(\boldsymbol{x}, e_j)$ and

$$Pg = \int g(\boldsymbol{x}, e_j) dP(\boldsymbol{x}, e_j) = \int g(\boldsymbol{x}, e_j) h(\boldsymbol{x}) f_0(e_j) d(\boldsymbol{x}, e_j).$$

We rewrite

$$B_n(h) = n^{2/3} \sum_{j=1}^{k} (P_n - P) \left(\tilde{\gamma}_j \frac{I(\hat{e}_j \leq t + hn^{-1/3}) - I(\hat{e}_j \leq t) + I(\hat{e}_j > -t - hn^{-1/3}) - I(\hat{e}_j > -t)}{2} \right)$$

$$+n^{2/3} \left(\sum_{j=1}^{k} P \left(\tilde{\gamma}_j \frac{I(\hat{e}_j \leq t + hn^{-1/3}) - I(\hat{e}_j \leq t) + I(\hat{e}_j > -t - hn^{-1/3}) - I(\hat{e}_j > -t)}{2} \right) \right.$$

$$\left. -f_0(t) h n^{-1/3} \right) - xh,$$

where $\hat{e}_j = y - \hat{\boldsymbol{\beta}}_n^T \boldsymbol{x} - \hat{\alpha}_{nj}$.

Let $e_j = y - \boldsymbol{\beta}_0^T \boldsymbol{x} - \alpha_{0j}$, and let

$$\xi_{nj} = e_j - \hat{e}_j = (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \boldsymbol{x} + (\hat{\alpha}_{nj} - \alpha_{0j}).$$

Since Conditions 17 and 18 hold and $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability holds by Theorem 1, it can be shown that

$$n^{1/2} \xi_{nj} = z_{n0} + n^{-1/2} z_{n1},$$

for some z_{n0}, z_{n1} with $E(z_{n0}) = 0$ and $E(z_{n1}) < \infty$. Hence $\xi_{nj} = O_p(n^{-1/2})$. Also,

$$E(\xi_{nj}) = n^{-1/2} E(z_{n0}) + n^{-1} E(z_{n1}) = O(n^{-1}).$$

Now we consider the first part of $B_n(h)$. Note that

$$I(\hat{e}_j \leq t) = I(e_j \leq t + \xi_{nj}) = I(e_j \leq t) + I(t < e_j \leq t + \xi_{nj}) - I(t + \xi_{nj} < e_j \leq t).$$

Note that

$$\text{var}(I(t < e_j \leq t + \xi_{nj})) = E(I(t < e_j \leq t + \xi_{nj})) - E^2(I(t < e_j \leq t + \xi_{nj})),$$

and

$$E(I(t < e_j \leq t + \xi_{nj})) = E(E(I(t < e_j \leq t + \xi_{nj}) \mid \xi_{nj})) = E(F_0(t + \xi_{nj}) - F_0(t))$$
$$= E(f_0(t)\xi_{nj} + O(\xi_{nj}^2)) = f_0(t) E(\xi_{nj}) + O(n^{-1}) = O(n^{-1}),$$

hence

$$\text{var}((P_n - P) I(t < e_j \leq t + \xi_{nj})) = \text{var}(P_n I(t < e_j \leq t + \xi_{nj}))$$
$$= \frac{1}{n^2} \text{var} \left(\sum_{i=1}^{n} I(t < e_{ji} \leq t + \xi_{nji}) \right) = \frac{1}{n} \text{var}(I(t < e_j \leq t + \xi_{nj}))$$
$$= \frac{1}{n} O(n^{-1}) = O(n^{-2}),$$

and hence

$$\text{var}(n^{2/3}(P_n - P)I(t < e_j \le t + \xi_{nj})) = O(n^{-2/3}).$$

So by Chebyshev's Inequality

$$\text{for each } \epsilon, n^{2/3}(P_n - P)I(t < e_j \le t + \xi_{nj}) = o_p(1).$$

Similar procedures show that overall,

$$n^{2/3}\sum_{j=1}^{k}(P_n - P)\left(\tilde{\gamma}_j \frac{I(\hat{e}_j \le t + hn^{-1/3}) - I(\hat{e}_j \le t) + I(\hat{e}_j > -t - hn^{-1/3}) - I(\hat{e}_j > -t)}{2}\right)$$

$$= n^{2/3}\sum_{j=1}^{k}(P_n - P)\left(\tilde{\gamma}_j \frac{I(e_j \le t + hn^{-1/3}) - I(e_j \le t) + I(e_j > -t - hn^{-1/3}) - I(e_j > -t)}{2}\right) + o_p(1).$$

For the second part, we have as a consequence of $e_j \perp \boldsymbol{x}$, (lengthy calculation)

$$n^{2/3}\sum_{j=1}^{k} P\left(\tilde{\gamma}_j \frac{I(\hat{e}_j \le t + hn^{-1/3}) - I(\hat{e}_j \le t) + I(\hat{e}_j > -t - hn^{-1/3}) - I(\hat{e}_j > -t)}{2}\right)$$

$$= n^{2/3}(F_0(t + hn^{-1/3}) - F_0(t) + O(n^{-1})).$$

So

$$B_n(h) = n^{2/3}\sum_{j=1}^{k}(P_n - P)\big[\tilde{\gamma}_j \frac{I(e_j \le t + hn^{-1/3}) - I(e_j \le t) + I(e_j > -t - hn^{-1/3}) - I(e_j > -t)}{2}\big]$$

$$+n^{2/3}\big(F_0(t + hn^{-1/3}) - F_0(t) - f_0(t)hn^{-1/3}\big) - xh + o_p(1)$$

$$= B_{1,n}(h) + B_{2,n}(h) - xh + o_p(1),$$

where

$$B_{1,n}(h) = n^{2/3}\sum_{j=1}^{k}(P_n - P)\big[\tilde{\gamma}_j \frac{I(e_j \le t + hn^{-1/3}) - I(e_j \le t) + I(e_j > -t - hn^{-1/3}) - I(e_j > -t)}{2}\big]$$

and

$$B_{2,n}(h) = n^{2/3}\big(F_0(t + hn^{-1/3}) - F_0(t) - f_0(t)hn^{-1/3}\big).$$

Note that

$$B_{2,n}(h) = n^{2/3}\int_t^{t+hn^{-1/3}} f_0(s) - f_0(t)ds \sim n^{2/3}\int_t^{t+hn^{-1/3}} \dot{f}_0(t)(s-t)ds = \frac{1}{2}\dot{f}_0(t)h^2.$$

Write

$$B_{1,n}(h) = n^{1/2}\sum_{j=1}^{k}(P_n - P)g_{j,n,h} = \sum_{j=1}^{k} B_{1,j,n}(h),$$

where

$$g_{j,n,h}(\boldsymbol{x}, e_j) = n^{1/6}\tilde{\gamma}_j \frac{I(e_j \le t + hn^{-1/3}) - I(e_j \le t) + I(e_j > -t - hn^{-1/3}) - I(e_j > -t)}{2}$$

and $B_{1,j,n}(h) = (P_n - P)g_{j,n,h}$.

Now we check the conditions for Theorem 2.11.23 (van der Vaart and Wellner, 1996, p.221). Let $\mathcal{G}_{j,n} = \{g_{j,n,h} : |h| \leq K\}$. Since $0 < \tilde{\gamma}_j \leq 1$, $\mathcal{G}_{j,n}$ has an envelope

$$G_{j,n}(e_j) = n^{1/6}(I(t < e_j \leq t + Kn^{-1/3}) - I(-t - Kn^{-1/3} < e_j \leq -t))/2,$$

as K is positive. We have

$$PG_{j,n}^2 = n^{1/3}P\left(\frac{I(-t \leq e_j \leq t + Kn^{-1/3}) + I(-t - Kn^{-1/3} \leq e_j \leq t)}{4}\right)$$

$$= \frac{n^{1/3}}{4}\left(\int_t^{t+Kn^{-1/3}} f_0(e_j)de_j + \int_{-t-Kn^{-1/3}}^{-t} f_0(e_j)de_j\right) = \frac{n^{1/3}}{2}(F_0(t + Kn^{-1/3}) - F_0(t))$$

$$= \frac{n^{1/3}}{2}(f_0(t)Kn^{-1/3} + O(n^{-2/3})) = O(1).$$

Note that $G_{j,n}(e_j) \leq n^{1/6}$, so for each $\eta > 0,\ I(G_{j,n}(e_j) > \eta n^{1/2}) = 0,$ for each $n \geq N$, where $N \geq 1/(2\eta)^3$, and so

$$P[G_{j,n}^2 I(G_{j,n} > \eta n^{1/2})] = O(0) \to 0.$$

Note

$$Pg_{j,n,h} = n^{1/6}E(\tilde{\gamma}_j)(F_0(t + hn^{-1/3}) - F_0(t)) = n^{1/6}E(\tilde{\gamma}_j)(f_0(t)hn^{-1/3} + O(n^{-2/3}))$$

$$= O(n^{-1/6}) \to 0,$$

and

$$Pg_{j,n,s}g_{j,n,h} = \begin{cases} \frac{\Gamma_j f_0(t)}{2}(s \wedge h) + O(n^{-1/3}) & \text{if } sh > 0 \text{ or } t = 0 \\ O(0) & \text{if } sh \leq 0 \text{ and } t \neq 0 \end{cases}$$

$$\to \begin{cases} \frac{\Gamma_j f_0(t)}{2}(s \wedge h) & \text{if } s, h > 0 \text{ or } t = 0 \\ 0 & \text{if } sh \leq 0 \text{ and } t \neq 0 \end{cases} = \frac{\Gamma_j f_0(t)}{2}\text{cov}\big(\mathbb{B}(h), \mathbb{B}(s)\big),$$

provided $t \neq 0$. Now that we have a formula for $Pg_{j,n,s}g_{j,n,h}$, it readily follows that

$$P(g_{j,n,s} - g_{j,n,h})^2 = Pg_{j,n,s}^2 - 2Pg_{j,n,s}g_{j,n,h} + 2Pg_{j,n,h}^2 = \frac{\Gamma_j f_0(t)}{2}|s - h| + O(n^{-1/3})$$

so for any $\delta_n \downarrow 0$,

$$\sup_{(s,h)\in[-K,K]^2:|s-h|<\delta_n} P(g_{j,n,s} - g_{j,n,h})^2 \sim \frac{\Gamma_j f_0(t)}{2}|s - h| < \frac{\Gamma_j f_0(t)}{2}\delta_n \to 0.$$

For $\epsilon > 0$, for all

$$k \in \{-(K\epsilon^2)^{-1}, -\lfloor(K\epsilon^2)^{-1}\rfloor, \ldots, 0, \ldots, \lfloor(K\epsilon^2)^{-1}\rfloor, (K\epsilon^2)^{-1}\},$$

let $h_k = k\epsilon^2 K^2$ and let $g_k = g_{n,h_k}$. Then, for all $g_{j,n,h} \in \mathcal{G}_{j,n}$, there is k such that

$$g_{j,n,h_{k-1}} \leq g_{j,n,h} \leq g_{j,n,h_k},$$

and

$$\|g_{j,n,h_k}-g_{j,n,h_{k-1}}\|_2=\left(\int(g_{j,n,h_k}-g_{j,n,h_{k-1}})^2dP\right)^{1/2}\sim\left((h_k-h_{k-1})n^{-1/3}f_0(t)\right)^{1/2}$$
$$=\left(\epsilon^2K^2E(\tilde{\gamma}_j)f_0(t)\right)^{1/2}\sim\epsilon\|G_{j,n}\|_2,$$

that is, the set of functions

$$\{g_{n,k}:k=(K\epsilon^2)^{-1},-\lfloor(K\epsilon^2)^{-1}\rfloor,\ldots,0,\ldots,\lfloor(K\epsilon^2)^{-1}\rfloor,(K\epsilon^2)^{-1}\}$$

is an $(\epsilon\|G_{j,n}\|_2)$-bracketing cover of $\mathcal{G}_{j,n}$, with covering number

$$N_{[\,]}(\epsilon\|G_{j,n}\|_2,\mathcal{G}_{j,n},\mathcal{L}_2(P))=\lfloor(K\epsilon^2)^{-1}\rfloor+O(1)=O(\epsilon^{-2}).$$

Thus, for any $\delta_n\downarrow 0$,

$$\int_0^{\delta_n}(\log N_{[\,]}(\epsilon\|G_n\|_2,\mathcal{G}_{j,n},\mathcal{L}_2(P)))^{1/2}d\epsilon=O(1)\int_0^{\delta_n}(-2\log\epsilon)^{1/2}d\epsilon$$
$$=O(1)\int_{(-2\log\delta_n)^{1/2}}^{\infty}xe^{-x^2/2}dx=O(1)\delta_n\to 0.$$

By Theorem 2.11.23 (van der Vaart and Wellner, 1996, p. 221), $B_{1,j,n}(h)$ is asymptotically tight in $l^\infty[-K,K]$, where $l^\infty[-K,K]$ is the set of all bounded, real-valued functions on $[-K,K]$, and converges in distribution to $(\Gamma_j f_0(t)/2)^{1/2}\mathbb{B}(h)$. Therefore in generality we have

$$B_{1,n}(h)=\sum_{j=1}^k B_{1,j,n}(h)\Rightarrow\sum_{j=1}^k\left(\left(\frac{\Gamma_j f_0(t)}{2}\right)^{1/2}\mathbb{B}(h)\right)=\left(\left(\frac{f_0(t)}{2}\right)^{1/2}\sum_{j=1}^k\Gamma_j^{1/2}\right)\mathbb{B}(h)\text{ in distribution,}$$

where $\Rightarrow$ denote uniform convergence in $l^\infty[-K,K]$, so

$$\operatorname*{arg\,max}_{h\in[-K,K]}B_n(h)\to\operatorname*{arg\,max}_{h\in[-K,K]}\left(\mathbb{B}(h)\left(\frac{f_0(t)}{2}\right)^{1/2}\sum_{j=1}^k\Gamma_j^{1/2}+\frac{1}{2}\dot{f}_0(t)h^2-xh\right)\text{ in distribution.}$$

Let $\hat{h}_n=\text{arg max}_{h\in\mathbb{R}}B_{1,j,n}(h)$. As in p. 297 of van der Vaart and Wellner (1996), $\hat{h}_n$ is bounded in probability, and hence

$$\operatorname*{arg\,max}_{h\in\mathbb{R}}B_n(h)\to\operatorname*{arg\,max}_{h\in\mathbb{R}}\left(\mathbb{B}(h)\left(\frac{f_0(t)}{2}\right)^{1/2}\sum_{j=1}^k\Gamma_j^{1/2}+\frac{1}{2}\dot{f}_0(t)h^2-xh\right)\text{ in distribution.}$$

The right hand side can be rewritten, using Problem 3.2.5 (van der Vaart and Wellner, 1996, p.308), as

$$\left(\frac{2f_0(t)}{\dot{f}_0^2(t)}\left(\sum_{j=1}^k\Gamma_j^{1/2}\right)^2\right)^{1/3}\operatorname*{arg\,max}_h(\mathbb{B}(h)-h^2)+\frac{x}{\dot{f}_0(t)}.$$

Note $\dot{f}_0(t)<0$ on $\mathbb{R}^+$ given Condition 16, so

$$\text{pr}\left(n^{1/3}\big(\hat{f}_n(t)-f_0(t)\big)\le x\right)\to\text{pr}\left(\left(\frac{2f_0(t)}{\dot{f}_0^2(t)}(\sum_{j=1}^k\Gamma_j^{1/2})^2\right)^{1/3}\operatorname*{arg\,max}_h\left(\mathbb{B}(h)-h^2\right)\le-\frac{x}{\dot{f}_0(t)}\right)$$
$$=\text{pr}\left(\left(2|\dot{f}_0(t)|f_0(t)\left(\sum_{j=1}^k\Gamma_j^{1/2}\right)^2\right)^{1/3}\operatorname*{arg\,max}_h\left(\mathbb{B}(h)-h^2\right)\le x\right).$$

Table 1: **Estimates of θ_0 from simulated data ($f_0 =$ Laplace$(0,1)$)**

$\boldsymbol{\theta}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
$\boldsymbol{\theta}_0$	(1.2, -2.1, 0.6, 1.5, 0.8)	(0.15, 1.25)
$\hat{\boldsymbol{\theta}}_n$ (best)	(1.15, -2.10, 0.64, 1.51, 0.79)	(0.53, 1.12)
(mean)	(1.14, -2.12, 0.62, 1.49, 0.78)	(0.60, 0.96)
[sd]	[0.036, 0.030, 0.032, 0.079, 0.069]	[0.116, 0.151]
$\bar{\boldsymbol{\theta}}_n$ (best)	(1.21, -2.10, 0.57, 1.38, 0.70)	(0.50, 1.02)
(mean)	(1.13, -2.12, 0.62, 1.47, 0.76)	(0.65, 1.00)
[sd]	[0.035, 0.034, 0.032, 0.087, 0.077]	[0.162, 0.171]

Table 2: **Estimates of θ_0 from simulated data ($f_0 =$ pgnorm(6))**

$\boldsymbol{\theta}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
$\boldsymbol{\theta}_0$	(1.2, -2.1, 0.6, 1.5, 0.8)	(0.15, 1.25)
$\hat{\boldsymbol{\theta}}_n$ (best)	(1.17, -2.11, 0.62, 1.43, 0.75)	(0.59, 1.02)
(mean)	(1.15, -2.12, 0.61, 1.48, 0.76)	(0.58, 0.93)
[sd]	[0.019, 0.012, 0.014, 0.040, 0.041]	[0.077, 0.103]
$\bar{\boldsymbol{\theta}}_n$ (best)	(1.16, -2.11, 0.62, 1.43, 0.74)	(0.59, 1.02)
(mean)	(1.15, -2.12, 0.61, 1.46, 0.75)	(0.62, 0.96)
[sd]	[0.020, 0.010, 0.012, 0.045, 0.050]	[0.11, 0.12]
$\boldsymbol{\theta}_0$	(0.8, 1.5, -1.3, 0.7, 2.1)	(0, 2.2)
$\hat{\boldsymbol{\theta}}_n$ (best)	(0.62, 1.47, -1.18, 0.72, 2.10)	(0.75, 1.71)
(mean)	(0.67, 1.51, -1.24, 0.62, 2.01)	(0.73, 1.21)
[sd]	[0.053, 0.053, 0.063, 0.127, 0.107]	[0.292, 0.498]
$\bar{\boldsymbol{\theta}}_n$ (best)	(0.62, 1.50, -1.22, 0.67, 2.05)	(0.92, 1.78)
(mean)	(0.67, 1.51, -1.25, 0.59, 1.97)	(0.83, 1.38)
[sd]	[0.051, 0.057, 0.067, 0.147, 0.135]	[0.386, 0.428]

Figure 1: **Estimates of f_0 from simulated Laplace data** Below: $f = f_0 =$ Laplace$(0, 1)$ (dashed), Normal$(0, 1)$ (dotted), $\hat{f}_n$ (solid).

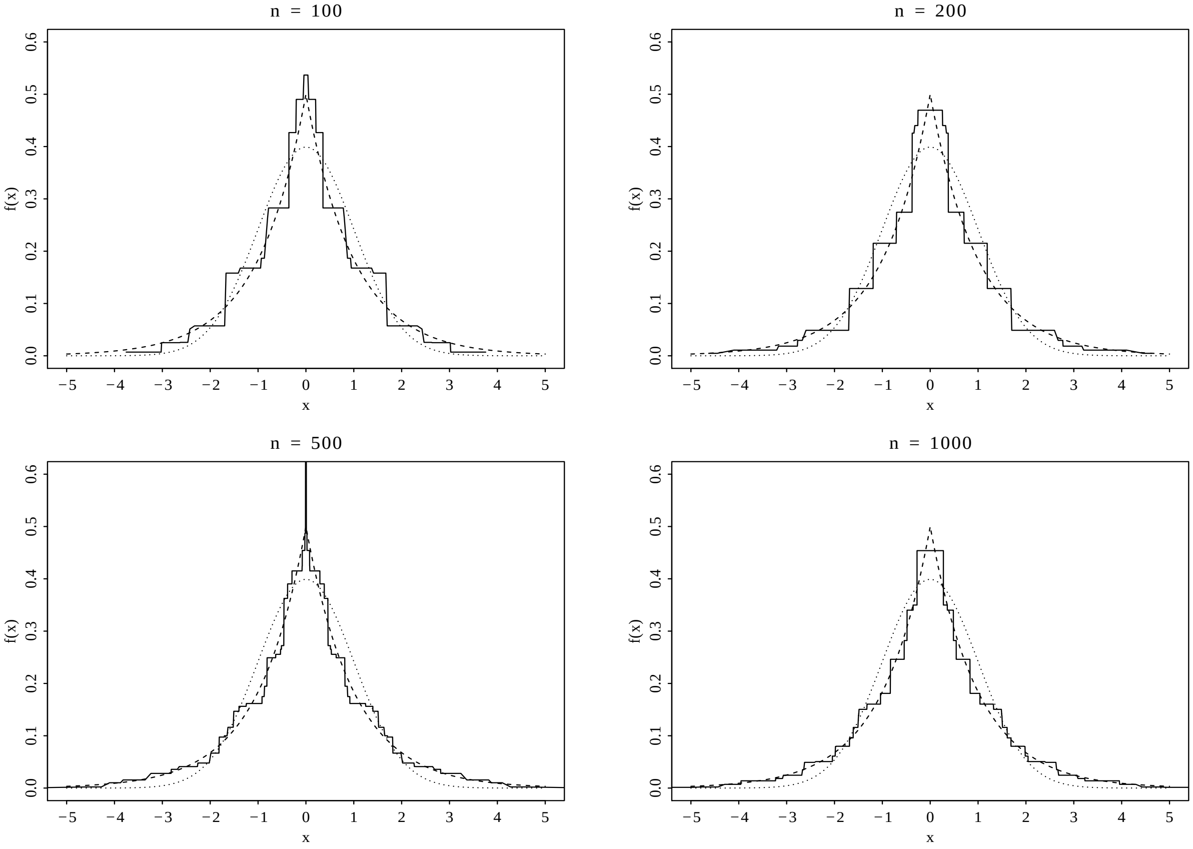
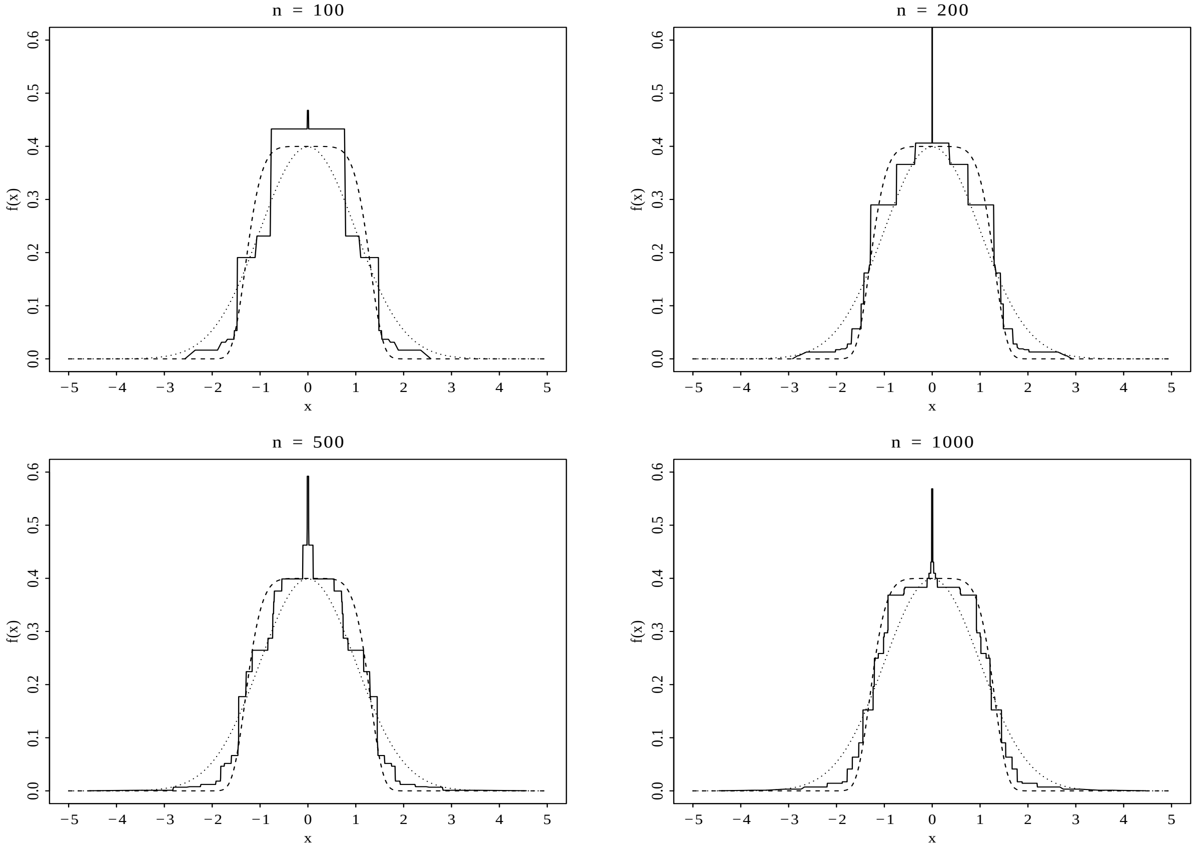


Figure 2: **Estimates of f_0 from simulated p-generalized normal data** Below: $f = f_0 =$ pgnorm(6) (dashed), Normal$(0, 1)$ (dotted), $\hat{f}_n$ (solid).



References

Atsedeweyn, A. A. and Srinivasa Rao, K. (2014) Linear regression model with new symmetric distributed errors. *Journal of Applied Statistics*, **41**, 364–381.

Balabdaoui, F. and Butucea, C. (2014) On location mixtures with pólya frequency components. *Statistics & Probability Letters*, **95**, 144–149.

Best, M. J. and Chakravarti, N. (1990) Active set algorithms for isotonic regression; a unifying framework. *Math. Programming*, **47**, 425–439. URL: http://dx.doi.org/10.1007/BF01580873.

Bianco, A. M., Garcia Ben, M. and Yohai, V. J. (2005) Robust estimation for linear regression with asymmetric errors. *Canad. J. Statist.*, **33**, 511–528. URL: http://dx.doi.org/10.1002/cjs.5550330404.

Bickel, P. J. and Fan, J. (1996) Some problems on the estimation of unimodal densities. *Statist. Sinica*, **6**, 23–45.

Birgé, L. (1989) The Grenander estimator: a nonasymptotic approach. *Ann. Statist.*, **17**, 1532–1549. URL: http://dx.doi.org/10.1214/aos/1176347380.

Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006) Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, **34**, 1204–1232. URL: http://dx.doi.org/10.1214/009053606000000353.

Box, G. E., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for experimenters: design, innovation, and discovery*, vol. 2. Wiley-Interscience New York.

Boyarshinov, V. and Magdon-Ismail, M. (2006) Linear time isotonic and unimodal regression in the L_1 and L_∞ norms. *J. Discrete Algorithms*, **4**, 676–691. URL: http://dx.doi.org/10.1016/j.jda.2005.07.001.

Brockett, P., Charnes, A. and Paick, K. (1995) Information-theoretic approach to unimodal density estimation. *IEEE Transactions on Information Theory*, **41**, 824–829.

Brunner, L. J. (1995) Bayesian linear regression with error terms that have symmetric unimodal densities. *J. Nonparametr. Statist.*, **4**, 335–348. URL: http://dx.doi.org/10.1080/10485259508832625.

Butucea, C., Tzoumpe, R. N., Vandekerkhove, P. et al. (2017) Semiparametric topographical mixture models with symmetric errors. *Bernoulli*, **23**, 825–862.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York.

Grenander, U. (1956) On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, **39**, 125–153 (1957).

Hall, P. and Kang, K.-H. (2005) Unimodal kernel density estimation by data sharpening. *Statist. Sinica*, **15**, 73–98.

Hanson, T. and Johnson, W. O. (2002) Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.*, **97**, 1020–1033. URL: http://dx.doi.org/10.1198/016214502388618843.

Hennig, C. (2000) Identifiability of models for clusterwise linear regression. *J. Classification*, **17**, 273–296. URL: http://dx.doi.org/10.1007/s003570000022.

Huang, M., Li, R. and Wang, S. (2013) Nonparametric mixture of regression models. *Journal of the American Statistical Association*, **108**, 929–941.

Huang, M. and Yao, W. (2012) Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, **107**, 711–724.

Hunter, D. R., Wang, S. and Hettmansperger, T. P. (2007) Inference for mixtures of symmetric distributions. *Ann. Statist.*, **35**, 224–251. URL: http://dx.doi.org/10.1214/009053606000001118.

Kiefer, J. and Wolfowitz, J. (1976) Asymptotically minimax estimation of concave and convex distribution functions. *Probability Theory and Related Fields*, **34**, 73–85.

Linton, O. and Xiao, Z. (2007) A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory*, **23**, 371–413. URL: http://dx.doi.org/10.1017/S026646660707017X.

Lo, S.-H. (1986) Estimation of a unimodal distribution function. *Ann. Statist.*, **14**, 1132–1138. URL: http://dx.doi.org/10.1214/aos/1176350054.

Mair, P., Hornik, K. and de Leeuw, J. (2009) Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, **32**, 1–24.

Mallapragada, P. K., Jin, R. and Jain, A. (2010) Non-parametric mixture models for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 334–343. Springer.

Prakasa Rao, B. L. S. (1969) Estimation of a unimodal density. *Sankhyā Ser. A*, **31**, 23–36.

— (1970) Estimation for distributions with monotone failure rate. *Ann. Math. Statist.*, **41**, 507–519.

Robertson, T. (1967) On estimating a density which is measurable with respect to a σ-lattice. *Ann. Math. Statist.*, **38**, 482–493.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988) *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

Schoenberg, I. (1951) On Pólya frequency functions I: the totally positive functions and their Laplace transforms. *Journal d'Analyse Mathématique*, **1**, 331–374.

Stout, Q. F. (2008) Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.*, **53**, 289–297. URL: http://dx.doi.org/10.1016/j.csda.2008.08.005.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Turnbull, B. C. and Ghosh, S. K. (2014) Unimodal density estimation using Bernstein polynomials. *Comput. Statist. Data Anal.*, **72**, 13–29. URL: http://dx.doi.org/10.1016/j.csda.2013.10.021.

van der Vaart, A. (2002) Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, vol. 1781 of *Lecture Notes in Math.*, 331–457. Springer, Berlin.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. URL: `http://dx.doi.org/10.1007/978-1-4757-2545-2`.

Wegman, E. J. (1970) Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.*, **41**, 457–471.

Yao, W. and Zhao, Z. (2013) Kernel density-based linear regression estimate. *Comm. Statist. Theory Methods*, **42**, 4499–4512. URL: `http://dx.doi.org/10.1080/03610926.2011.650269`.

Yuan, A. and de Gooijer, J. G. (2007) Semiparametric regression with kernel error model. *Scand. J. Statist.*, **34**, 841–869. URL: `http://dx.doi.org/10.1111/j.1467-9469.2006.00531.x`.

Zeckhauser, R. and Thompson, M. (1970) Linear regression with non-normal error terms. *The Review of Economics and Statistics*, 280–286.

Zhu, X. and Hunter, D. R. (2015) Clustering via finite nonparametric ica mixture models. *arXiv preprint arXiv:1510.08178*.