Ann Inst Stat Math (2018) 70:489–521
https://doi.org/10.1007/s10463-017-0602-4

Robust variable selection for finite mixture regression models

Qingguo Tang[1] · **R. J. Karunamuni**[2]

Received: 18 May 2016 / Revised: 29 September 2016 / Published online: 25 February 2017

Abstract Finite mixture regression (FMR) models are frequently used in statistical modeling, often with many covariates with low significance. Variable selection techniques can be employed to identify the covariates with little influence on the response. The problem of variable selection in FMR models is studied here. Penalized likelihood-based approaches are sensitive to data contamination, and their efficiency may be significantly reduced when the model is slightly misspecified. We propose a new robust variable selection procedure for FMR models. The proposed method is based on minimum-distance techniques, which seem to have some automatic robustness to model misspecification. We show that the proposed estimator has the variable selection consistency and oracle property. The finite-sample breakdown point of the estimator is established to demonstrate its robustness. We examine small-sample and robustness properties of the estimator using a Monte Carlo study. We also analyze a real data set.

Keywords Finite mixture regression models · Variable selection · Minimum-distance methods

✉ R. J. Karunamuni
r.j.karunamuni@ualberta.ca

[1] School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China

[2] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada

1 Introduction

We consider the problem of simultaneous estimation and variable selection in finite mixture regression models. Let Y be a univariate response variable and let $\mathbf{X} \in \mathbb{R}^p$ be a vector of covariates that affect the outcome Y. Then, (Y, X) is said to have a finite mixture regression (FMR) model of order J if the conditional density of Y given $X = x$ is of the form

$$f_{\boldsymbol{\theta}}(y|\mathbf{x}) = \sum_{j=1}^{J} \alpha_j g(y; \varpi\left(\mathbf{x}^T \boldsymbol{\beta}_j\right), \gamma_j), \tag{1}$$

where g is a density function with respect to a σ-finite measure ν, ϖ is a link function, and $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T, \gamma_1, \ldots, \gamma_J)^T$ is the parameter vector with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^T$ and $\alpha_j > 0,\ j = 1, \ldots, J,\ \sum_{j=1}^{J} \alpha_j = 1$. See Titterington et al. (1985), Hennig (2000), Khalili and Chen (2007), Städler et al. (2010), Khalili and Lin (2013) and the references therein for more details on FMR models. Model (1) can be generalized to allow the α_j values to be functions of x. The density function g can take many parametric forms, including negative-binomial, normal and Poisson. In some FMR models, the dispersion parameters, γ_j, are assumed to be equal.

For a given design matrix $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$, model (1) is said to be *identifiable* if for any two parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$,

$$\sum_{j=1}^{J} \alpha_j g(y, \varpi\left(\mathbf{x}_i^T \boldsymbol{\beta}_j\right), \gamma_j) = \sum_{j=1}^{J^*} \alpha_j^* g(y, \varpi(\mathbf{x}_i^T \boldsymbol{\beta}_j^*), \gamma_j^*)$$

for each $i = 1, \ldots, n$ and all possible values of y, implies that $J = J^*$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$; see, e.g., Titterington et al. (1985), Hennig (2000) and Khalili and Chen (2007). In the above definition, one interprets $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ upto a permutation. The identifiability is potentially a serious problem when dealing with FMR models. The identifiability of an FMR model depends on several factors, such as component densities $g(y, \varpi(\mathbf{x}^T \boldsymbol{\beta}_j), \gamma_j)$, the maximum possible order J and the design matrix $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$. Many finite mixture models, including finite mixture of binomial, multinomial, normal and Poisson distributions, are identifiable. For more details on this issue, see Hennig (2000).

It is well known that finite mixture models provide a mathematical basis for the statistical modeling of a wide variety of random situations. Because of their flexibility, these models have received increasing interest over the years from both the practical and theoretical points of view. They have been applied in many fields including economics (Khalili and Chen 2007), neural networks (Bishop 1995) and machine learning (Jiang and Tanner 1999). A comprehensive account of the literature, theory, and applications of modeling via finite mixture models is given in the monographs of Titterington et al. (1985) and McLachlan and Peel (2000). For software implementations in FMR models, see Leisch (2004).

A penalty (or regularization) function generally facilitates variable selection in regression models. Various penalty functions have been used in the literature: the bridge regression (Frank and Friedman 1993), LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), adaptive LASSO (Zou 2006), the elastic net (Zou and Hastie 2005), the adaptive elastic net (Zou and Zhang 2009), and MCP (Zhang 2010) are well known. In particular, Khalili and Chen (2007), Khalili (2010), Khalili et al. (2011), Khalili and Lin (2013), and Städler et al. (2010) have investigated the variable selection problem for FMR models with versions of the above penalty functions. Generally speaking, in these articles the estimation of $\boldsymbol{\theta}$ is carried out by maximizing the penalized log-likelihood function

$$l_n(\boldsymbol{\theta}) - \mathcal{P}_n(\boldsymbol{\theta}),$$

where $l_n(\boldsymbol{\theta})$ is the (conditional) log-likelihood function based on a sample of size n from model (1), and $\mathcal{P}_n(\boldsymbol{\theta})$ is a penalty function. The resulting estimators have been shown to have nice properties in most cases, including the variable selection consistency and oracle property of Fan and Li (2001). However, these estimators can be highly unstable if the model is not completely correct, and they are not robust if the data are slightly contaminated. In general, a minor instability in the model can have severe consequences for finite mixture models (Karlis and Xekalaki 2001; Markatou 2000; Lu et al. 2003; Tang and Karunamuni 2013). For example, in the case of parametric univariate finite normal mixtures with all parameters unknown, the likelihood can grow without bound if one of the means coincides with a data point and the corresponding variance is allowed to go to zero; see e.g., McLachlan and Peel (2000).

In this paper, we propose a different approach: we replace the log-likelihood function $l_n(\boldsymbol{\theta})$ with a distance measure in order to develop a robust estimator. The rationale for this idea is that minimum-distance methods have a degree of automatic robustness to model misspecification (Donoho and Liu 1988). We examine the problem of simultaneous estimation and variable selection in a FMR model of the form (1) with the squared Hellinger distance as the measure of adequacy. Specifically, the proposed estimator of $\boldsymbol{\theta}$ is constructed by minimizing

$$\left\| f_{\boldsymbol{\theta},n}^{1/2} - \hat{f}_n^{1/2} \right\|^2 + \mathcal{P}_n(\boldsymbol{\theta}), \tag{2}$$

where $\mathcal{P}_n(\boldsymbol{\theta})$ is a penalty function, $f_{\boldsymbol{\theta},n}(y)$ and $\hat{f}_n(y)$ denote a consistent estimator and a nonparametric density estimator, respectively, of the *semiparametric mixture density* of Y and $\|.\|$ denotes the L_2-norm. We show that the proposed estimator has the variable selection consistency and oracle property of Fan and Li (2001). We also develop a feasible and effective algorithm for variable selection and parameter estimation. Further, we establish the global robustness properties of the proposed estimator by finding its *finite-sample breakdown point*. Recall that the breakdown point of an estimator is the proportion of incorrect observations (i.e., arbitrary values) it can handle before giving an arbitrarily large result (Donoho 1982; Donoho and Huber 1983).

Other distance measures such as the L_1-norm, L_2-norm, and the Chi-squared distance can be used instead of the Hellinger distance in (2). However, the Hellinger distance has some practical and theoretical advantages in nonpenalized estimation problems. Specifically, minimum Hellinger distance (MHD) estimators for parametric models achieve efficiency at the model, and they simultaneously have excellent robustness properties in the presence of outliers and/or model misspecification (Beran 1977, 1978). Furthermore, Lindsay (1994) has shown that the maximum likelihood and MHD estimators are members of a larger class of efficient estimators with various second-order efficiency properties. MHD estimators have been developed in the literature for various setups and models, including for some parametric mixture models and semiparametric models. The literature is too extensive to give a complete listing here. A discussion of recent developments and some important references can be found in the articles of Cutler and Cordero-Braña (1996); Karunamuni and Wu (2011), Wu and Karunamuni (2012, 2015), Wu et al. (2010), and Tang and Karunamuni (2013).

This paper is organized as follows. Section 2 develops the proposed variable selection and estimation procedure and studies its asymptotic properties. Section 3 studies the robustness properties of the proposed penalized procedures, and Sect. 4 presents a computational algorithm. Monte Carlo studies, real-data examples, and concluding remarks are given in Sects. 5, 6, and 7, respectively.

2 Robust variable selection

We consider a FMR model in which $Y \in \mathbb{R}$ is a univariate continuous response variable, $\mathbf{X} \in \mathbb{R}^p$ is a continuous random covariate vector, and the conditional density of Y given $\mathbf{X} = \mathbf{x}$ as

$$f_{\boldsymbol{\theta}}(y|\mathbf{x}) = \sum_{j=1}^{J} \alpha_j g(y, \mathbf{x}^T \boldsymbol{\beta}_j, \gamma_j), \tag{3}$$

where $\alpha_j > 0, j = 1, \ldots, J$, $\sum_{j=1}^{J} \alpha_j = 1$, $\mathbf{x} = (x_1, \ldots, x_p)^T$, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^T$ are $p \times 1$ parameter vectors, the γ_j are parameters, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T, \gamma_1, \ldots, \gamma_J)^T$, and the function $g(y, z, u)$ satisfies $g(y, z, u) \geq 0$ and $\int g(y, z, u)\mathrm{d}y = 1$. We also assume that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^{J(p+2)-1}$. We shall assume throughout that the FMR model (3) is identifiable. Also assume that J is fixed and that $\boldsymbol{\theta}$ is the parameter of interest.

Let $f_{\boldsymbol{\theta},\eta}(y)$ denote the marginal mixture density function of Y given by

$$f_{\boldsymbol{\theta},\eta}(y) = \int f_{\boldsymbol{\theta}}(y|\mathbf{x})\mathrm{d}\eta(\mathbf{x}) = \sum_{j=1}^{J} \alpha_j \int g(y, \mathbf{x}^T \boldsymbol{\beta}_j, \gamma_j)\mathrm{d}\eta(\mathbf{x}), \tag{4}$$

where $\eta(\mathbf{x})$ denotes an unknown *mixing distribution* of $\mathbf{X}$. Model (4) is a *semiparametric mixture model*. Under fairly general conditions, including the identifiability of $f_{\boldsymbol{\theta}}(y|\mathbf{x})$, the semiparametric mixture model $f_{\boldsymbol{\theta},\eta}(y)$ is identifiable, and the maximum likelihood estimators of $(\boldsymbol{\theta}, \eta)$ are well established; see, e.g., Vaart (1996) and the references therein. In what follows, we assume that the model $f_{\boldsymbol{\theta},\eta}(y)$ identifiable in the sense that $\| f_{\boldsymbol{\theta}_1,\eta_1}^{1/2} - f_{\boldsymbol{\theta}_2,\eta_2}^{1/2} \|^2 = 0$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ and $\eta_1 = \eta_2$. For semiparametric estimation of $\boldsymbol{\theta}$, a natural approach is to replace η by a consistent estimator, e.g., the empirical distribution function of the $\mathbf{X}_i$'s, and then deals with the resulting marginal density of Y

$$f_{\boldsymbol{\theta},n}(y) = \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y|\mathbf{X}_i) = \sum_{j=1}^{J}\sum_{i=1}^{n}\frac{\alpha_j}{n} g(y, \mathbf{X}_i^T\boldsymbol{\beta}_j, \gamma_j), \tag{5}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from $\eta(\mathbf{x})$. This is the approach taken in this paper to obtain a robust estimator of $\boldsymbol{\theta}$. For a given n design points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the identifiability of model (5) follows from that of model (3).

Assume that we have a random sample $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ with the conditional density of Y_i given $\mathbf{X}_i = \mathbf{x}_i$ and the marginal distribution of $\mathbf{X}_i$ are $f_{\boldsymbol{\theta}}(y|\mathbf{x})$ given by (3) and $\eta(\mathbf{x})$, respectively, $i = 1, \ldots, n$. Let $\hat{f}_n(y)$ denote a kernel-type density estimator based on $Y_1, \ldots, Y_n$:

$$\hat{f}_n(y) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right), \tag{6}$$

where $K(\cdot)$ is a given kernel function and h_n is a bandwidth such that $h_n \to 0$ as $n \to \infty$. Let $\boldsymbol{\theta}_0$ denote the true parameter value of $\boldsymbol{\theta}$. Then, an MHD estimator of $\boldsymbol{\theta}_0$ can be defined by $\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \| f_{\boldsymbol{\theta},n}^{1/2} - \hat{f}_n^{1/2}\|^2$.

In the variable selection problem, it is assumed that some components of $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$, are equal to zero. The goal is to identify and estimate the subset model. It has been argued that folded concave penalties are preferable to convex penalties such as the L_1-penalty in terms of both model-estimation accuracy and variable selection consistency (Lv and Fan 2009; Fan and Lv 2011). Let $p_{\lambda_{nj}}(|t|) = p_{a,\lambda_{nj}}(|t|)$ be general folded concave penalty functions defined on $t \in (-\infty, +\infty)$ satisfying

(a) $p_{\lambda_{nj}}(t)$ are increasing and concave in $t \in [0, +\infty)$;

(b) $p_{\lambda_{nj}}(t)$ are differentiable in $t \in (0, +\infty)$ with $p'_{\lambda_{nj}}(0) := p'_{\lambda_{nj}}(0+) \geq a_1\lambda_{nj}$, $p'_{\lambda_{nj}}(t) \geq a_1\lambda_{nj}$ for $t \in (0, a_2\lambda_{nj}]$, $p'_{\lambda_{nj}}(t) \leq a_3\lambda_{nj}$ for $t \in [0, +\infty)$, and $p'_{\lambda_{nj}}(t) = 0$ for $t \in [a\lambda_{nj}, +\infty)$ with a prespecified constant $a > a_2$, where a_1, a_2 and a_3 are fixed positive constants.

The above family of general folded concave penalties contains several popular penalties, including the SCAD penalty (Fan and Li 2001) and the MCP penalty (Zhang 2010).

Let $\mathcal{F}$ be the set of all densities with respect to the Lebesgue measure on the real line. We define a penalized MHD functional $T : \mathcal{F} \to \mathbb{R}^{J(p+2)-1}$ by

$$T(\phi) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\| f_{\boldsymbol{\theta},\eta}^{1/2} - \phi^{1/2}\right\|^2 + 2\sum_{j=1}^{J}\alpha_j^{1/2}\sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{0jk}|)|\beta_{jk}|\right\}, \tag{7}$$

where $\boldsymbol{\beta}_{0j} = (\beta_{0j1}, \ldots, \beta_{0jp})^T$, $j = 1, \ldots, J$, denote vectors of true parameter values and $f_{\boldsymbol{\theta},\eta}(y)$ is given by (4). Since $\eta(\mathbf{x})$ is unknown, $f_{\boldsymbol{\theta},\eta}(y)$ is unknown. Thus, by replacing $f_{\boldsymbol{\theta},\eta}(y)$ with a consistent estimator (5), we define an estimated penalized MHD functional $\hat{T} : \mathcal{F} \to \mathbb{R}^{J(p+2)-1}$ by

$$\hat{T}(\phi) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left\| f_{\boldsymbol{\theta},n}^{1/2} - \phi^{1/2} \right\|^2 + 2\sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{p} p'_{\lambda_{nj}} \left(|\beta_{jk}^{(0)}| \right) |\beta_{jk}| \right\}, \tag{8}$$

where $\boldsymbol{\beta}_j^{(0)} = (\beta_{j1}^{(0)}, \ldots, \beta_{jp}^{(0)})^T$, $j = 1, \ldots, J$, is obtained from $\boldsymbol{\theta}^{(0)} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \| f_{\boldsymbol{\theta},n}^{1/2} - \hat{f}_n^{1/2} \|^2$, an initial robust estimator of $\boldsymbol{\theta}$. Then, the proposed penalized MHD estimator of $\boldsymbol{\theta}_0$ (the true parameter value of $\boldsymbol{\theta}$) is defined by

$$\hat{\boldsymbol{\theta}} = \hat{T}(\hat{f}_n) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left\| f_{\boldsymbol{\theta},n}^{1/2} - \hat{f}_n^{1/2} \right\|^2 + 2\sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{p} p'_{\lambda_{nj}} \left(|\beta_{jk}^{(0)}| \right) |\beta_{jk}| \right\}. \tag{9}$$

Since $\| f_{\boldsymbol{\theta},n}^{1/2} - \hat{f}_n^{1/2} \|^2 = 2 - 2\int f_{\boldsymbol{\theta},n}^{1/2}(y) \hat{f}_n^{1/2}(y)\mathrm{d}y$, from (9) we have

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \int f_{\boldsymbol{\theta},n}^{1/2}(y) \hat{f}_n^{1/2}(y)\mathrm{d}y - \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{p} p'_{\lambda_{nj}} \left(|\beta_{jk}^{(0)}| \right) |\beta_{jk}| \right\}. \tag{10}$$

The functional $T(.)$ defined by (7) is said to be *essentially unique* if $f_{\boldsymbol{\theta},\eta}(y)$ is nondegenerate for any $\boldsymbol{\theta} \in T(\phi)$, and any other element of $T(\phi)$ can be obtained from $\boldsymbol{\theta} \in T(\phi)$ by permuting the labels of the components. The next theorem gives continuity of the functional $\hat{T}$ with respect to Hausdorff metric (Pollard 1981). The Hausdorff metric, say $\check{\delta}$, is defined for compact subsets A, B by $\check{\delta}(A, B) < \varepsilon$ if and only if every point of A is within a Euclidean distance ε of at least one point of B, and vice versa.

Theorem 1 *For each η, assume that the class $\{f_{\boldsymbol{\theta},\eta}(y) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is identifiable, where $f_{\boldsymbol{\theta},\eta}(y)$ is given by (4). Further assume that the function $g(y, z, u)$ in (4) is continuous in (z, u) for almost every y. Then, we have*

(i) *For every $\phi \in \mathcal{F}$, there exists $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\theta}$ minimizes the function inside (7). For every $\phi \in \mathcal{F}$, there exists $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\theta}$ minimizes the function inside (8).*

(ii) *If $T(\phi)$ and $\hat{T}(\phi)$ are essentially unique, then $\hat{T}(\phi_n) \to T(\phi)$ in the Hausdorff metric for any sequences $\{\phi_n\}_{n\geq 1}$ and $\{f_{\boldsymbol{\theta},n}\}_{n\geq 1}$ such that $\|\phi_n^{1/2} - \phi^{1/2}\| \to 0$ and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \| f_{\boldsymbol{\theta},n}^{1/2} - f_{\boldsymbol{\theta},\eta}^{1/2} \| \to 0$ as $n \to \infty$.*

(iii) *$\boldsymbol{\theta}_0 \in T(f_{\boldsymbol{\theta}_0,\eta})$ and $T(f_{\boldsymbol{\theta}_0,\eta})$ is essentially unique.*

The proofs of (i) and (ii) of Theorem 1 are similar to those of (1) and (2), respectively, of Theorem 2.1 in Tang and Karunamuni (2013). Let $D_0(\boldsymbol{\theta}, \eta) = \| f_{\boldsymbol{\theta},\eta}^{1/2} - f_{\boldsymbol{\theta}_0,\eta}^{1/2} \|^2 + 2\sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{0jk}|)|\beta_{jk}|$. Part (iii) in Theorem 1 follows from the fact that $D_0(\boldsymbol{\theta}_0, \eta) = 0$ and that $f_{\boldsymbol{\theta},\eta}$ is nondegenerate and the class $\{f_{\boldsymbol{\theta},\eta}(y) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is identifiable.

Theorem 3.1 of Tang and Karunamuni (2013) shows that $\|\hat{f}_n^{1/2} - f_{\boldsymbol{\theta}_0,\eta}^{1/2}\| \to_P 0$, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|f_{\boldsymbol{\theta},n}^{1/2} - f_{\boldsymbol{\theta},\eta}^{1/2}\| \to_P 0$. Then, by (ii) of Theorem 1, we have the next theorem.

Theorem 2 *Under the assumptions of Theorem 1, suppose that condition* (C3) *defined in "Appendix" holds and* $\int \sup_{t\in\Theta} |\frac{\partial f_t(y)}{\partial t}|dy < +\infty$ *in* (C1) *and*

$$\int_{|y|>L}\int_{\|x\|\le\tilde{L}} |x_r|\breve{g}_z(y,\mathbf{x})d\eta(\mathbf{x})dy \to 0, \quad \int_{|y|>L}\int_{\|x\|\le\tilde{L}} \breve{g}_u(y,\mathbf{x})d\eta(\mathbf{x})dy \to 0$$

in (C2). *If* $h_n \to 0$ *and* $nh_n \to \infty$, *then* $\hat{\boldsymbol{\theta}} \to_P \boldsymbol{\theta}_0$ *as* $n \to \infty$.

Next we show that the penalized MHD estimator defined by (9) has the oracle property. Without loss of generality, let $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j1}^T, \boldsymbol{\beta}_{j2}^T)^T$, where $\boldsymbol{\beta}_{j1} \in \mathbb{R}^{d_j}$ and $\boldsymbol{\beta}_{j2} \in \mathbb{R}^{p-d_j}$. The vector of true parameters is denoted by $\boldsymbol{\beta}_{0j} = (\beta_{0j1},\ldots,\beta_{0jp})^T = (\boldsymbol{\beta}_{0j1}^T, \boldsymbol{\beta}_{0j2}^T)^T$ with each element of $\boldsymbol{\beta}_{0j1}$ being nonzero and $\boldsymbol{\beta}_{0j2} = \mathbf{0}$. Generally, we split the vector of true parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^T, \boldsymbol{\theta}_{02}^T)^T$ such that $\boldsymbol{\theta}_{02} = \mathbf{0}$; that is, $\boldsymbol{\theta}_{02}$ consists of all $\boldsymbol{\beta}_{0j2}$, $j = 1,\ldots,J$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. For convenience of notation, we write $f_{\boldsymbol{\theta}}(y)$ for $f_{\boldsymbol{\theta},\eta}(y)$ defined by (4) in what follows.

Theorem 3 *Let* $V_1,\ldots,V_n$ *be a random sample from the joint distribution of* $(Y,\mathbf{X})$ *that satisfies the regularity conditions* (C1)–(C6) *in "Appendix". Let* $p_{\lambda_{nj}}(\cdot)$ *be general folded concave penalty functions satisfying assumptions* (a) *and* (b) *defined above. If* $\lambda_{nj} \to 0$ *and* $\sqrt{n}\lambda_{nj} \to \infty$ *as* $n \to \infty$, *then the penalized MHD estimator* $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T)^T$ *defined by (9) satisfies*

(1) *Sparsity:* $P(\hat{\boldsymbol{\theta}}_2 = \mathbf{0}) \to 1$.
(2) *Asymptotic normality:*

$$n^{1/2}[H_1(\boldsymbol{\theta}_{01})(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01}) - A_{n1}(\boldsymbol{\theta}_{01})] \to_d N(\mathbf{0}, \Sigma_1(\boldsymbol{\theta}_{01})), \tag{11}$$

where $H_1(\boldsymbol{\theta}_{01}) = -\int \ddot{S}_{\boldsymbol{\theta}_{01}}(y) f_{\boldsymbol{\theta}_0}^{1/2}(y)dy$, $\ddot{S}_{\boldsymbol{\theta}_{01}}(y) = \frac{\partial^2 S_{\boldsymbol{\theta}}(y)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^T}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, $S_{\boldsymbol{\theta}}(y) = f_{\boldsymbol{\theta}}^{1/2}(y)$,

$$\Sigma_1(\boldsymbol{\theta}_{01}) = \int \psi_{\boldsymbol{\theta}_{01}}(y)\psi_{\boldsymbol{\theta}_{01}}^T(y)dy - \int \left[\int \psi_{\boldsymbol{\theta}_{01}}(y) f_{\boldsymbol{\theta}_0}(y|x)dy\right]\left[\int \psi_{\boldsymbol{\theta}_{01}}(y) f_{\boldsymbol{\theta}_0}(y|x)dy\right]^T d\eta(x),$$

$A_{n1}(\boldsymbol{\theta}_{01}) = \frac{1}{2}\mu_2 h_n^2 \int \psi_{\boldsymbol{\theta}_{01}}(y) f_{\boldsymbol{\theta}_0}''(y)dy$ *with* $\mu_2 = \int v^2 K(v)dv$,

$$\psi_{\boldsymbol{\theta}_{01}}(y) = \dot{S}_{\boldsymbol{\theta}_{01}}(y)/[2 f_{\boldsymbol{\theta}_0}^{1/2}(y)],$$

$\dot{S}_{\boldsymbol{\theta}_1}(y) = \frac{\partial S_{\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}_1}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, $f''_{\boldsymbol{\theta}}(y) = \partial^2 f_{\boldsymbol{\theta}}(y)/\partial y^2$. Further, if $h_n = b_0 n^{-\gamma}$ for some $\gamma \in (1/4, 1/2)$ and some positive constant b_0, then we have

$$n^{1/2}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01}) \to_d N(\mathbf{0}, H_1^{-1}(\boldsymbol{\theta}_{01})\Sigma_1(\boldsymbol{\theta}_{01})H_1^{-1}(\boldsymbol{\theta}_{01})).$$

The maximization of (10) involves a nonlinear weighted L_1 regularization. To simplify computations, denote $M(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta},n}^{1/2}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y$. Then, $M(\boldsymbol{\theta})$ can be locally approximated by

$$M(\boldsymbol{\theta}) \approx M(\boldsymbol{\theta}^{(0)}) + \nabla M(\boldsymbol{\theta}^{(0)})^T(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^T \nabla^2 M(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}).$$

Using $\nabla M(\boldsymbol{\theta}^{(0)}) = \mathbf{0}$, we then solve the penalized minimization problem

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^T[-\nabla^2 M(\boldsymbol{\theta}^{(0)})](\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) + \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}| \right\}. \tag{12}$$

Let $H(\boldsymbol{\theta}_0) = -\int \ddot{S}_{\boldsymbol{\theta}_0}(y) f_{\boldsymbol{\theta}_0}^{1/2}(y)\mathrm{d}y$, where $\ddot{S}_{\boldsymbol{\theta}}(y) = \frac{\partial^2 S_{\boldsymbol{\theta}}(y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$. By arguments similar to those used in the proof of (22) in “Appendix”, it follows that $-\nabla^2 M(\boldsymbol{\theta}^{(0)}) \to_P H(\boldsymbol{\theta}_0)$. Following an argument similar to the proof of Theorem 5 of Zou and Li (2008) and using Theorem 3.3 of Tang and Karunamuni (2013), we have the next theorem.

Theorem 4 *Under the assumptions of Theorem 3, the estimator $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1^T, \tilde{\boldsymbol{\theta}}_2^T)^T$ defined by (12) satisfies the conclusions of Theorem 3.*

Remark 1 Under the assumptions of Theorem 1, suppose that condition (C3) defined in “Appendix” holds. If $\sum_{n=1}^{\infty} \exp(-\varpi n h_n^2) < \infty$ for any $\varpi > 0$ and $\int\int \breve{g}_u(y, \mathbf{x})\mathrm{d}\eta(\mathbf{x})\mathrm{d}y < +\infty$, $\int\int |x_r|\breve{g}_z(y, \mathbf{x})\mathrm{d}\eta(\mathbf{x})\mathrm{d}y < +\infty$, then the weak consistency result given in Theorem 2 can be strengthened to the strong consistency of $\hat{\boldsymbol{\theta}}$. Moreover, we have that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|f_{\boldsymbol{\theta},n}^{1/2} - f_{\boldsymbol{\theta},\eta}^{1/2}\|^2 \xrightarrow{a.s.} 0$ and $\boldsymbol{\beta}_j^{(0)} \xrightarrow{a.s.} \boldsymbol{\beta}_{0j}$ for $j = 1, \ldots, J$. By Devroye and Wagner (1979), it follows that $\|\hat{f}_n - f_{\boldsymbol{\theta}_0,\eta}\|_1 \xrightarrow{a.s.} 0$. Since $\|\hat{f}_n^{1/2} - f_{\boldsymbol{\theta}_0,\eta}^{1/2}\|^2 \le \|\hat{f}_n - f_{\boldsymbol{\theta}_0,\eta}\|_1$, we have $\|\hat{f}_n^{1/2} - f_{\boldsymbol{\theta}_0,\eta}^{1/2}\|^2 \xrightarrow{a.s.} 0$. Hence, from (7) and (9) we see that $\hat{\boldsymbol{\theta}}$ is Fisher consistent.

Remark 2 In (9), when $\alpha_j^{1/2}$ is replaced by α_j in the penalty terms, the conclusions of Theorems 1–4 still hold. For computational convenience, we have chosen $\alpha_j^{1/2}$ instead of α_j in the penalty terms; see computational algorithm in Sect. 4 for more details.

Remark 3 If $\mathbf{X}$ is a discrete random covariate vector with possible values $\tilde{\mathbf{x}}_{\boldsymbol{k}}, \boldsymbol{k} \in \mathbb{N}^p$ for $\mathbb{N} = \{1, 2, \ldots\}$ and the probability function of $\mathbf{X}$ is $\eta(\mathbf{x})$, then the marginal density function of Y is $f_{\boldsymbol{\theta}}(y) = \sum_{\boldsymbol{k}} f_{\boldsymbol{\theta}}(y|\tilde{\mathbf{x}}_{\boldsymbol{k}})\eta(\tilde{\mathbf{x}}_{\boldsymbol{k}})$. Under some conditions similar to (C1)–(C6) in "Appendix", the conclusions of Theorems 1–4 then still hold. For example, in condition (C2), $\int_{|y|>L}\int_{\|x\|\le\tilde{L}}\breve{g}_u(y,\mathbf{x})\mathrm{d}\eta(\mathbf{x})\mathrm{d}y \to 0$ can be replaced by $\int_{|y|>L}\sum_{\|\tilde{\mathbf{x}}_{\boldsymbol{k}}\|\le\tilde{L}}\breve{g}_u(y,\tilde{\mathbf{x}}_{\boldsymbol{k}})\eta(\tilde{\mathbf{x}}_{\boldsymbol{k}})\mathrm{d}y \to 0$.

Remark 4 For fixed design $\mathbf{x}$, we regard $\mathbf{X}$ as a discrete uniform random vector with probability function $\eta(\mathbf{x}_i) = P\{\mathbf{X} = \mathbf{x}_i\} = 1/n$ for $i = 1, \ldots, n$. In this case, the marginal density function of Y becomes $f_{\boldsymbol{\theta},n}(y) = \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y|\mathbf{x}_i)$, and the parameter $\boldsymbol{\theta}$ can be still estimated by (9).

Remark 5 It is important to note here that the robust penalized MHD estimator $\hat{\boldsymbol{\theta}}$ defined by (9) is a semiparametric estimator, obtained by replacing the infinite-dimensional nuisance parameter η by a consistent estimator, the empirical distribution function of the $\mathbf{X}_i$'s in this case. The price one pays for this procedure may be a slight reduction in efficiency of some components of $\boldsymbol{\theta}$. To further examine this point on efficiency of $\hat{\boldsymbol{\theta}}_1$, we consider the following example.

Set $J = 1$ in (3) and suppose $\beta_{01} \ne 0$, $\beta_{02} \ne 0$, $\beta_{03} = \ldots = \beta_{0p} = 0$. Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ and $\boldsymbol{\theta}_{01} = (\beta_{01}, \beta_{02})^T$. Assume that $X_1 \equiv 1$ and $X_2 \sim N(0, \sigma^2)$, where $\sigma > 0$ is a known constant. We further assume that $f_{\boldsymbol{\theta}_{01}}(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\gamma}\exp\{-\frac{[y-(\beta_{01}+\beta_{02}x_2)]^2}{2\gamma^2}\}$, where γ is a known parameter. Then, the marginal density function of Y is $f_{\boldsymbol{\theta}_{01}}(y) = \frac{1}{\sqrt{2\pi(\sigma^2\beta_{02}^2+\gamma^2)}}\exp\{-\frac{(y-\beta_{01})^2}{2(\sigma^2\beta_{02}^2+\gamma^2)}\}$. Under the reduced model when all zero effects are removed, the Fisher information is $I_1(\boldsymbol{\theta}_{01}) = diag(1/\gamma^2, \sigma^2/\gamma^2)$ and $I_1^{-1}(\boldsymbol{\theta}_{01}) = \gamma^2 diag(1, 1/\sigma^2)$. It is easy to prove that $H_1(\boldsymbol{\theta}_{01}) = \frac{1}{4(\sigma^2\beta_{02}^2+\gamma^2)^2}diag(\sigma^2\beta_{02}^2+\gamma^2, 2\sigma^4\beta_{02}^2)$ and

$$\Sigma_1(\boldsymbol{\theta}_{01}) = \frac{\gamma^2}{16(\sigma^2\beta_{02}^2+\gamma^2)^2}diag(1, 2\sigma^4\beta_{02}^2(2\sigma^2\beta_{02}^2+\gamma^2)/(\sigma^2\beta_{02}^2+\gamma^2)^2).$$

Hence

$$H_1^{-1}(\boldsymbol{\theta}_{01})\Sigma_1(\boldsymbol{\theta}_{01})H_1^{-1}(\boldsymbol{\theta}_{01}) = \gamma^2 diag\left(1, \frac{1}{\sigma^2}[1+\gamma^2/(2\sigma^2\beta_{02}^2)]\right).$$

From Wang et al. (2007), the asymptotic covariance matrix of least-absolute deviation (LAD) lasso estimator of $\boldsymbol{\theta}_{01}$ is $\frac{\pi}{2n}I_1^{-1}(\boldsymbol{\theta}_{01})$. Let $\varepsilon = Y - (\beta_{01} + \beta_{02}X_2)$. If ε is independent of X_2, then by Theorem 1 of Wang et al. (2013), the asymptotic covariance matrix of exponential squared loss (ESL) estimator of $\boldsymbol{\theta}_{01}$ is $L(\gamma_0)I_1^{-1}(\boldsymbol{\theta}_{01})/n$ with $L(\gamma_0) = \left[\frac{\gamma_0}{2\gamma^2+\gamma_0}(2 - \frac{\gamma_0}{2\gamma^2+\gamma_0})\right]^{-3/2}$. Note that $L(\gamma_0)$ decreases as γ_0 increases and tends 1 as $\gamma_0 \to +\infty$. By comparing, we find that the asymptotic variance of the penalized MHD estimator of β_{01} is equal to that of the maximum penalized likelihood estimator of β_{01}, but it is less than that of the LAD-lasso and ESL estimators of β_{01}. On the other hand, the asymptotic variance of the penalized MHD estimator of β_{02} is larger than that of the maximum penalized likelihood estimator of β_{02}. When $\gamma^2/\sigma^2\beta_{02}^2 < (\pi - 2)$, that is $Var(\varepsilon) < (\pi - 2)Var(\beta_{02}X_2)$, the asymptotic variance of the penalized MHD estimator of β_{02} is less than that of the LAD-lasso estimator of β_{02}. For the ESL method, larger γ_0 can improve the efficiency of the estimator; however, it reduces the breakdown point of the estimator.

An alternative approach that may be used here is to first assume the density of $\mathbf{X}$ exists, call it η, and then replace the unknown density η with a suitable estimator in the joint density $f_\theta(y, \mathbf{x}) = f_\theta(y|\mathbf{x})\eta(\mathbf{x})$ of $(Y, \mathbf{X})$, and finally minimize the Hellinger distance between the corresponding expression and a nonparametric density estimator $\hat{f}_n(y, \mathbf{x})$ of the density of $(Y, \mathbf{X})$. This method compares well with the likelihood method of Khalili and Chen (2007) while retaining good robustness properties. However, since the vector $(Y, \mathbf{X})$ is $\mathbb{R}^{p+1}$-valued, the above approach may face difficult issues related to high-dimensional construction of a nonparametric density estimator $\hat{f}_n(y, \mathbf{x})$, as the curse of dimensionality affects the rates of convergence of $\hat{f}_n(y, \mathbf{x})$, and which, in turn, may restrict its application. Thus, we have implemented a semi-parametric approach in this paper for the construction of our estimator. The proposed estimator is easy to implement in practice where the asymptotic efficiency issues are less important, but the robustness is more desirable.

3 Robustness properties

In this section, we establish some robustness properties of the proposed penalized MHD estimator by finding its *finite sample breakdown point*. Roughly speaking, this is the proportion of incorrect observations an estimator can handle before giving an arbitrarily large result (Donoho 1982; Donoho and Huber 1983). This is a global measurement of robustness in terms of resistance to outliers. Following Donoho (1982), we define a measure of discrepancy between parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{j=1}^{J} \left(\left\| \boldsymbol{\beta}_{j1} - \boldsymbol{\beta}_{j2} \right\|^2 + \frac{\gamma_{j1}}{\gamma_{j2}} + \frac{\gamma_{j2}}{\gamma_{j1}} \right).$$

Suppose $\boldsymbol{V}_n = (V_1, \ldots, V_n)$ is a data set of size n from the distribution of $(Y, \mathbf{X})$, and suppose $\boldsymbol{V}_{n_1} = (V_{n+1}, \ldots, V_{n+n_1})$ denotes any (contaminated) data set of size n_1. An estimator $\hat{\boldsymbol{\theta}}$ is said to break down if $d(\hat{\boldsymbol{\theta}}(\boldsymbol{V}_n \cup \boldsymbol{V}_{n_1}), \hat{\boldsymbol{\theta}}(\boldsymbol{V}_n)) = \infty$ for an appropriate choice of $V_{n+1}, \ldots, V_{n+n_1}$. Let n_1^* denote the smallest number of contaminating points for which $\hat{\boldsymbol{\theta}}$ breaks down. Then, the finite-sample breakdown point $b^*(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_n)$ of $\hat{\boldsymbol{\theta}}$ at $\boldsymbol{V}_n$ is defined as $n_1^*/(n + n_1^*)$. Define

$$\delta^* = \liminf_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \left\| f_{\boldsymbol{\theta}_1}^{1/2} - f_{\boldsymbol{\theta}_2}^{1/2} \right\|^2,$$

where the limit is taken as $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \to \infty$. Let $\hat{f}_n(h_n)$ be a kernel density estimator of $f_{\boldsymbol{\theta}_0}$ with bandwidth h_n and given some i.i.d. data $(Y_1, \ldots, Y_n)$. Let $\hat{\boldsymbol{\theta}}(h_n)$ be the estimator of $\boldsymbol{\theta}_0$ defined by (9) based on $\hat{f}_n(h_n)$. Let $f_n(\hat{\boldsymbol{\theta}}(h_n)) = f_{\hat{\boldsymbol{\theta}}(h_n),n}$. Then, we have the following result on the breakdown point of $\hat{\boldsymbol{\theta}}$.

Theorem 5 *Assume that the parameter space is bounded. Then, the breakdown point of* $\hat{\boldsymbol{\theta}}(h_n)$ *satisfies*

$$b^*(\hat{\boldsymbol{\theta}}(h_n), \boldsymbol{V}_n) \geq (\delta^*/4 - a_{n,n_1} - \varepsilon_{n,n_1}/4)/(2 - a_{n,n_1}), \tag{13}$$

where $a_{n,n_1} = \left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_n^{1/2}(h_{n+n_1}) \right\|^2$,

$$\begin{aligned}\varepsilon_{n,n_1} = {} & \left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) \right\|^2 + 2\left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 \\ & + 3\left\| f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 + 4P^*_{n,n_1}\end{aligned}$$

and

$$\begin{aligned}P^*_{n,n_1} = {} & \sum_{j=1}^{J} \hat{\alpha}_{j,n}^{1/2} \sum_{k=1}^{p} p'_{\lambda_{(n+n_1)j}}(|\beta^{(0)}_{jk,n+n_1}|)|\hat{\beta}_{jk,n}| \\ & - \sum_{j=1}^{J} \hat{\alpha}_{j,n+n_1}^{1/2} \sum_{k=1}^{p} p'_{\lambda_{(n+n_1)j}}(|\beta^{(0)}_{jk,n+n_1}|)|\hat{\beta}_{jk,n+n_1}|.\end{aligned}$$

Since $\left\| \hat{f}_n^{1/2}(h_{n+n_1}) - f_{\boldsymbol{\theta}_0}^{1/2} \right\|^2 \leq \left\| \hat{f}_n(h_{n+n_1}) - f_{\boldsymbol{\theta}_0} \right\|_1$ and by Devroye and Wagner (1979), $\|\hat{f}_n(h_{n+n_1}) - f_{\boldsymbol{\theta}_0}\|_1 \to_P 0$ as $h_{n+n_1} \to 0$ and $nh_{n+n_1} \to \infty$. Hence, $\left\| \hat{f}_n^{1/2}(h_{n+n_1}) - f_{\boldsymbol{\theta}_0}^{1/2} \right\|^2 \to_P 0$. By Theorem 2, $\hat{\boldsymbol{\theta}}(h_n) \to_P \boldsymbol{\theta}_0$, and then by the dominated convergence theorem, it follows that $\left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - f_{\boldsymbol{\theta}_0}^{1/2} \right\|^2 \leq \|f(\hat{\boldsymbol{\theta}}(h_n)) - f_{\boldsymbol{\theta}_0}\|_1 \to_P 0$. Hence, $a_{n,n_1} \to_P 0$ as $h_{n+n_1} \to 0$ and $nh_{n+n_1} \to \infty$. Further, we have

$$\begin{aligned}& \left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1}) \right\|^2 \\ & \quad + 2\sum_{j=1}^{J} \hat{\alpha}_{j,n+n_1}^{1/2} \sum_{k=1}^{p} p'_{\lambda_{(n+n_1)j}}(|\beta^{(0)}_{jk,n+n_1}|)|\hat{\beta}_{jk,n+n_1}| \\ & \leq \left\| f_{n+n_1}^{1/2}(\boldsymbol{\theta}^{(0)}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1}) \right\|^2 \\ & \quad + 2\sum_{j=1}^{J} {\alpha^{(0)}_{j,n+n_1}}^{1/2} \sum_{k=1}^{p} p'_{\lambda_{(n+n_1)j}}(|\beta^{(0)}_{jk,n+n_1}|)|\beta^{(0)}_{jk,n+n_1}|\end{aligned}$$

and $\left\| f_{n+n_1}^{1/2}(\boldsymbol{\theta}^{(0)}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1}) \right\|^2 \leq \left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1}) \right\|^2$. Hence, it follows that

$$\begin{aligned}&\sum_{j=1}^{J}\hat{\alpha}_{j,n+n_1}^{1/2}\sum_{k=1}^{p}p'_{\lambda_{(n+n_1)j}}(|\beta_{jk,n+n_1}^{(0)}|)|\hat{\beta}_{jk,n+n_1}|\\&\quad\le\sum_{j=1}^{J}\alpha_{j,n+n_1}^{(0)}{}^{1/2}\sum_{k=1}^{p}p'_{\lambda_{(n+n_1)j}}(|\beta_{jk,n+n_1}^{(0)}|)|\beta_{jk,n+n_1}^{(0)}|\le C_0\max_{1\le j\le J}\lambda_{(n+n_1)j},\end{aligned}$$

where C_0 is a positive constant. Therefore, $|P_{n,n_1}^{*}|\le C_1\max_{1\le j\le J}\lambda_{(n+n_1)j}$, where C_1 is a positive constant. Using Theorem 3.1 of Tang and Karunamuni (2013), we have $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|f_n^{1/2}(\boldsymbol{\theta})-f^{1/2}(\boldsymbol{\theta})\right\|^2\to_P 0$. Thus, $\varepsilon_{n,n_1}\to_P 0$ as $n\to\infty$. Therefore, with probability tending to 1,

$$b^{*}(\hat{\boldsymbol{\theta}}(h_n),\boldsymbol{V}_n)\ge\delta^{*}/8. \tag{14}$$

Now assume that the conditional densities $g(y,\mathbf{x}^T\boldsymbol{\beta}_j,\gamma_j)$ belong to a normal family. If $d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\to\infty$, then there exists some integer l, $1\le l\le J$, such that $\left\|\boldsymbol{\beta}_{l1}-\boldsymbol{\beta}_{l2}\right\|^2+\frac{\gamma_{l1}}{\gamma_{l2}}+\frac{\gamma_{l2}}{\gamma_{l1}}\to\infty$. Hence $\left\|\boldsymbol{\beta}_{l1}\right\|+\gamma_{l1}^{-1}\to\infty$ or $\left\|\boldsymbol{\beta}_{l2}\right\|+\gamma_{l2}^{-1}\to\infty$. Suppose $\left\|\boldsymbol{\beta}_{l1}\right\|+\gamma_{l1}^{-1}\to\infty$. Then

$$\lim_{\|\boldsymbol{\beta}_{l1}\|+\gamma_{l1}^{-1}\to\infty}f_{\boldsymbol{\theta}_1}(y)=\sum_{j\ne l}\alpha_j\int g(y,\mathbf{x}^T\boldsymbol{\beta}_j,\gamma_j)\mathrm{d}\eta(x)=:f_{-l,\boldsymbol{\theta}_1}(y).$$

It follows that

$$\begin{aligned}\lim_{\|\boldsymbol{\beta}_{l1}\|+\gamma_{l1}^{-1}\to\infty}\int f_{\boldsymbol{\theta}_1}^{1/2}(y)f_{\boldsymbol{\theta}_2}^{1/2}(y)\mathrm{d}y&=\int f_{-l,\boldsymbol{\theta}_1}^{1/2}(y)f_{\boldsymbol{\theta}_2}^{1/2}(y)\mathrm{d}y\\&\le\left(\int f_{-l,\boldsymbol{\theta}_1}(y)\mathrm{d}y\int f_{\boldsymbol{\theta}_2}(y)\mathrm{d}y\right)^{1/2}\\&=(1-\alpha_l)^{1/2}.\end{aligned}$$

Therefore,

$$\delta^{*}\ge 2\min_j[1-(1-\alpha_j)^{1/2}]. \tag{15}$$

When $J=1$ in model (3), i.e., $\alpha_1=1$ and $\alpha_j=0$ for $j\ge 2$, we have $\delta^{*}=2$.

We note from (14) and (15) that if α_j is small for some j, then the breakdown point of the estimator $\hat{\boldsymbol{\theta}}(h_n)$ is small. This can be interpreted as follows: when $J>1$ and α_j is small, then data from the j-th component are about $[\alpha_j n]$ among the data set $\boldsymbol{V}_n=(V_1,\ldots,V_n)$, where $[\alpha_j n]$ denotes the integer part of $\alpha_j n$. Further, $\boldsymbol{V}_{n_1}=(V_{n+1},\ldots,V_{n+n_1})$ is also the contaminated data of the estimators $\hat{\boldsymbol{\beta}}_j$ and $\hat{\gamma}_j$ of j-th component. Note that small $n_1/(n+n_1)$ can make $n_1/([n\alpha_j]+n_1)$ large, and large $n_1/([n\alpha_j]+n_1)$ can result in breakdown of the estimators $\hat{\boldsymbol{\beta}}_j$ and $\hat{\gamma}_j$. When the estimators $\hat{\boldsymbol{\beta}}_j$ and $\hat{\gamma}_j$ break down, the estimator $\hat{\boldsymbol{\theta}}(h_n)$ also breaks down.

Remark 6 A positive value for the breakdown point indicates that the estimator is robust against data contamination. Assume that the density estimates $\hat{f}_n^{1/2}(h_{n+n_1})$ and $\hat{f}_{n_1}^{1/2}(h_{n+n_1})$ based on $\boldsymbol{V}_n$ and $\boldsymbol{V}_{n_1}$, respectively, have disjoint supports. Then, the lower bound $(\delta^*/4 - a_{n,n_1} - \varepsilon_{n,n_1}/4)/(2 - a_{n,n_1})$ of (13) can be replaced by $1 - [(2 - (\delta^* - \varepsilon_{n,n_1})/4)/(2 - a_{n,n_1})]^2$. Since $a_{n,n_1} \to_P 0$ and $\varepsilon_{n,n_1} \to_P 0$, we have

$$b^*(\hat{\boldsymbol{\theta}}(h_n), \boldsymbol{V}_n) \geq 1 - (1 - \delta^*/8)^2 \geq 1 - [3/4 + \max_j (1 - \alpha_j)^{1/2}]^2,$$

with probability tending to one. When $J = 1$ in model (3), we obtain $b^*(\hat{\boldsymbol{\theta}}(h_n), \boldsymbol{V}_n) \geq 7/16 = 0.4375$, which is close to the maximum breakdown point value 0.5.

Remark 7 When $J = 1$ in model (3), Maronna (1976) indicated that the nonpenalized M-estimator of $\boldsymbol{\theta}_0$ has a breakdown point with an upper bound of $1/(p+1)$. According to Devlin et al. (1981), the empirical breakdown point of M-estimators might in fact be much lower than the upper bound. By (14) and (15), we see that the MHD estimator of $\boldsymbol{\theta}_0$ has an asymptotic breakdown at least $1/4$, and it is at least $7/16$ under some cases, as indicated in Remark 6. Hence, for large p, the MHD estimator has better robustness properties than the M-estimator.

Remark 8 Donoho (1982) and Donoho and Huber (1983) introduced another notion of finite-sample robustness, namely the *replacement breakdown point,* based on replacing some with new data. For example, n_1 of the n values of V_n are changed arbitrarily, replacing them with new values and producing a new data set V_n^* of size n. Then, the breakdown point is defined as the smallest fraction of contaminating points for which the estimator breaks down. Toma (2008) studied the replacement breakdown point for some multivariate distributions. We believe similar results can be obtained here when using the replacement breakdown point as the measure of robustness.

Remark 9 The influence function approach is another useful method for evaluating the robustness properties of estimators; it describes the local stability of an estimator in the presence of an outlier. See Hampel et al. (1986) for the theory behind influence functions of estimators and for a characterization of robustness based on influence functions. For a fixed point $z \in \mathbb{R}$, let δ_z denote the uniform density on the interval $(z - \gamma,\ z + \gamma)$, where $\gamma > 0$ is very small, and let $f_{\boldsymbol{\theta},\eta,z,\varepsilon} = (1 - \varepsilon) f_{\boldsymbol{\theta},\eta} + \varepsilon\delta_z$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\varepsilon \in [0, 1)$, where $f_{\boldsymbol{\theta},\eta}$ is given by (4). The density function $f_{\boldsymbol{\theta},\eta,z}$ models an experiment where independent observations distributed according to $f_{\boldsymbol{\theta},\eta}$ are mixed with approximately $100\varepsilon\%$ gross errors located near z. We then define the influence function of estimators of type (9) as $\mathrm{IF}_\eta(z) = \lim_{\varepsilon\downarrow 0} \frac{1}{\varepsilon}\{T^(f_{\boldsymbol{\theta},\eta,z,\varepsilon}) - T(f_{\boldsymbol{\theta},\eta})\}$, assuming that the limit exists. This definition is consistent with the one defined in Shen (1995) for semiparametric models. It is clear that $\mathrm{IF}_\eta(z)$ mostly depends on the first term on right-hand side of (7) and appears to be of the form $\mathrm{IF}_\eta(z) = c_1 m_\eta(z) + c_2$, where c_1 and c_2 are constants depending on the penalty functions $p'_{\lambda_{nj}}$, and $m_\eta(z) = \lim_{\varepsilon\downarrow 0} \frac{1}{\varepsilon}\{T_1(f_{\boldsymbol{\theta},\eta,z,\varepsilon}) - T_1(f_{\boldsymbol{\theta},\eta})\}$ with functional $T_1 : \mathcal{F} \to \mathbb{R}^{J(p+2)-1}$ given by $T_1(\phi) = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \| f_{\boldsymbol{\theta},\eta}^{1/2} - \phi^{1/2}\|^2$. As observed in Beran (1977), $m_\eta(z)$ can be an unbounded function for many families, such as the normal location-scale family. Thus, the influence function $\mathrm{IF}_\eta(z)$ can be an unbounded function in some cases.

4 Computational algorithm

In this section, we discuss the computational issues and propose an algorithm to compute the penalized MHD estimator defined by (9). First, we write

$$\begin{aligned}
&\int f_{\boldsymbol{\theta},n}^{1/2}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y-\sum_{j=1}^{J}\alpha_j^{1/2}\sum_{k=1}^{p}p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}| \\
&=\sum_{j=1}^{J}\sqrt{\alpha_j}\left[\int\sqrt{\alpha_j(y)g_{nj}(y,\boldsymbol{\beta}_j,\gamma_j)\hat{f}_n(y)}\mathrm{d}y-\sum_{k=1}^{p}p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}|\right] \\
&=:\sum_{j=1}^{J}\sqrt{\alpha_j}H_{nj}(\boldsymbol{\beta}_j,\gamma_j)=:G_n(\boldsymbol{\theta},\alpha(y)),
\end{aligned}$$

where $g_{nj}(y,\boldsymbol{\beta}_j,\gamma_j)=\frac{1}{n}\sum_{i=1}^{n}g(y,\mathbf{X}_i^T\boldsymbol{\beta}_j,\gamma_j)$, $\alpha(y)=(\alpha_1(y),\ldots,\alpha_J(y))^T$ and

$$\alpha_j(y)=\alpha_j g_{nj}(y,\boldsymbol{\beta}_j,\gamma_j)/\left(\sum_{l=1}^{J}\alpha_l g_{nl}(y,\boldsymbol{\beta}_l,\gamma_l)\right).$$

Treating $\alpha(y)$ as fixed and maximizing $G_n(\boldsymbol{\theta},\alpha(y))$ with respect to $\boldsymbol{\theta}$ subject to the constraint $\sum_{j=1}^{J}\alpha_j=1$, we obtain

$$(\hat{\boldsymbol{\beta}}_j,\hat{\gamma}_j)=\operatorname{argmax}_{(\boldsymbol{\beta}_j,\gamma_j)}H_{nj}(\boldsymbol{\beta}_j,\gamma_j),\ \hat{\alpha}_j=\frac{H_{nj}^2(\hat{\boldsymbol{\beta}}_j,\hat{\gamma}_j)}{\sum_{l=1}^{J}H_{nl}^2(\hat{\boldsymbol{\beta}}_l,\hat{\gamma}_l)}.$$

Treating $\boldsymbol{\theta}$ as fixed and maximizing $G_n(\boldsymbol{\theta},\alpha(y))$ with respect to $\alpha_j(y)$ subject to the constraint $\sum_{j=1}^{J}\alpha_j(y)=1$, we have

$$\alpha_j^*(y)=\frac{\alpha_j g_{nj}(y,\boldsymbol{\beta}_j,\gamma_j)}{\sum_{l=1}^{J}\alpha_l g_{nl}(y,\boldsymbol{\beta}_l,\gamma_l)}.$$

Then, the proposed algorithm can be summarized as follows:

Let $\boldsymbol{\theta}^{(0)}=(\alpha_1^{(0)},\ldots,\alpha_J^{(0)},\boldsymbol{\beta}_1^{(0)},\ldots,\boldsymbol{\beta}_J^{(0)},\gamma_1^{(0)},\ldots,\gamma_J^{(0)})$ be an initial estimator of $\boldsymbol{\theta}$.

Step 1. For $m=1,2,\ldots,$ compute

$$\alpha_j^{(m-1)}(y)=\frac{\alpha_j^{(m-1)}g_{nj}(y,\boldsymbol{\beta}_j^{(m-1)},\gamma_j^{(m-1)})}{\sum_{l=1}^{J}\alpha_l^{(m-1)}g_{nl}(y,\boldsymbol{\beta}_l^{(m-1)},\gamma_l^{(m-1)})}.$$

Step 2. Solve the following optimization problem to obtain $\boldsymbol{\beta}_j^{(m)}$, $\gamma_j^{(m)}$, and $\alpha_j^{(m)}$, $j = 1, \ldots, J$:

$$(\boldsymbol{\beta}_j^{(m)}, \gamma_j^{(m)}) = \operatorname{argmax}_{(\boldsymbol{\beta}_j,\gamma_j)} H_{nj}^{(m)}(\boldsymbol{\beta}_j, \gamma_j),\ \alpha_j^{(m)} = \frac{[H_{nj}^{(m)}(\boldsymbol{\beta}_j^{(m)},\gamma_j^{(m)})]^2}{\sum_{l=1}^{J}[H_{nl}^{(m)}(\boldsymbol{\beta}_l^{(m)},\gamma_l^{(m)})]^2},$$

where $H_{nj}^{(m)}(\boldsymbol{\beta}_j, \gamma_j) = \int \sqrt{\alpha_j^{(m-1)}(y) g_{nj}(y, \boldsymbol{\beta}_j, \gamma_j) \hat{f}_n(y)}\mathrm{d}y - \sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)$ $|\beta_{jk}|$.

Step 3. If $\|\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}\| < \varepsilon$ for a small prespecified threshold value ε, stop; otherwise, set $m = m + 1$ and return to step 1.

Since $H_{nj}^{(m)}(\boldsymbol{\beta}_j^{(m)}, \gamma_j^{(m)}) \geq H_{nj}^{(m)}(\boldsymbol{\beta}_j^{(m-1)}, \gamma_j^{(m-1)})$, it is easy to show that $G_n(\boldsymbol{\theta}^{(m)}, \alpha^{(m)}(y)) \geq G_n(\boldsymbol{\theta}^{(m-1)}, \alpha^{(m-1)}(y))$. Hence, $G_n(\boldsymbol{\theta}^{(m)}, \alpha^{(m)}(y))$ is increasing as a function of m. The algorithm consists of a sequence of weighted single-component MHD variable selection problems. The algorithm is similar to the EM algorithm which consists of a sequence of weighted maximum likelihood problems. However, it is difficult to find $\operatorname{argmax}_{(\boldsymbol{\beta}_j,\gamma_j)} H_{nj}^{(m)}(\boldsymbol{\beta}_j, \gamma_j)$. Thus, we make following adjustments. Set

$$R_{nj}(\boldsymbol{\beta}_j, \gamma_j) = \int \sqrt{\alpha_j(y) g_{nj}(y, \boldsymbol{\beta}_j, \gamma_j) \hat{f}_n(y)}\mathrm{d}y.$$

Let $(\boldsymbol{\beta}_j^{(0)}, \gamma_j^{(0)})$ be the maximizer of $R_{nj}(\boldsymbol{\beta}_j, \gamma_j)$. Set

$$w_{nj}^{(0)}(y) = \sqrt{\alpha_j(y) \hat{f}_n(y) / g_{nj}(y, \boldsymbol{\beta}_j^{(0)}, \gamma_j^{(0)})}.$$

For $(\boldsymbol{\beta}_j, \gamma_j)$ near $(\boldsymbol{\beta}_j^{(0)}, \gamma_j^{(0)})$, we have

$$\begin{aligned}
&R_{nj}(\boldsymbol{\beta}_j, \gamma_j)\\
&\quad\approx \int w_{nj}^{(0)}(y) g_{nj}(y, \boldsymbol{\beta}_j, \gamma_j^{(0)})\mathrm{d}y\\
&\quad\approx R_{nj}(\boldsymbol{\beta}_j^{(0)}, \gamma_j^{(0)}) + (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{(0)})^T \int w_{nj}^{(0)}(y) \frac{\partial g_{nj}(y, \boldsymbol{\beta}_j, \gamma_j^{(0)})}{\partial \boldsymbol{\beta}_j}\bigg|_{\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(0)}} \mathrm{d}y\\
&\qquad + (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{(0)})^T \int w_{nj}^{(0)}(y) \frac{\partial g_{nj}^2(y, \boldsymbol{\beta}_j, \gamma_j^{(0)})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^T}\bigg|_{\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(0)}} \mathrm{d}y (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{(0)})\ .
\end{aligned}$$

Since $\int w_{nj}^{(0)}(y) \frac{\partial g_{nj}(y, \boldsymbol{\beta}_j, \gamma_j^{(0)})}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(0)}} \mathrm{d}y = 0$ and

$$\frac{\partial g_{nj}^2(y, \boldsymbol{\beta}_j, \gamma_j^{(0)})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j^T}\bigg|_{\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(0)}} = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2 g(y, z, \gamma_j^{(0)})}{dz^2}\bigg|_{z = X_i^T \boldsymbol{\beta}_j^{(0)}} X_i X_i^T,$$

we set $Y_{ij}^{(0)} = X_i^T \boldsymbol{\beta}_j^{(0)}$ and $\Gamma_{iij} = -\int w_{nj}^{(0)}(y) \frac{d^2 g(y,z,\gamma_j^{(0)})}{dz^2}|_{z=X_i^T \boldsymbol{\beta}_j^{(0)}} \mathrm{d}y$ and then solve the penalized minimization problem

$$\hat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Gamma_{iij} (Y_{ij}^{(0)} - \mathbf{X}_i^T \boldsymbol{\beta}_j)^2 + \sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}| \right\}.$$

Now $\hat{\boldsymbol{\beta}}_j$ can be obtained using the procedure given in Zou and Li (2008). In this paper, we compute $\hat{\boldsymbol{\beta}}_j$ using the pathwise coordinate optimization technique introduced in Friedman et al. (2007). After computing $\hat{\boldsymbol{\beta}}_j$, $j = 1, \ldots, J$, we solve the maximization problem $\hat{\gamma}_j = \operatorname{argmax}_{\gamma_j} \int \sqrt{\alpha_j(y) g_{nj}(y, \hat{\boldsymbol{\beta}}_j, \gamma_j) \hat{f}_n(y)} \mathrm{d}y$ to obtain $\hat{\gamma}_j$. Using the above results, we now propose the following revised algorithm for variable selection and parameter estimation.

Algorithm:

Let $\boldsymbol{\theta}^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_J^{(0)}, \boldsymbol{\beta}_1^{(0)}, \ldots, \boldsymbol{\beta}_J^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_J^{(0)})$ be an initial estimator of $\boldsymbol{\theta}$.

Step 1. For $m = 1, 2, \ldots,$ compute

$$\alpha_j^{(m-1)}(y) = \frac{\alpha_j^{(m-1)} g_{nj}(y, \boldsymbol{\beta}_j^{(m-1)}, \gamma_j^{(m-1)})}{\sum_{l=1}^{J} \alpha_l^{(m-1)} g_{nl}(y, \boldsymbol{\beta}_l^{(m-1)}, \gamma_l^{(m-1)})}.$$

Step 2. Solve the following optimization problem to obtain $\boldsymbol{\beta}_j^{(m)}$, $\gamma_j^{(m)}$, and $\alpha_j^{(m)}$, $j = 1, \ldots, J$:

$$\boldsymbol{\beta}_j^{(m)} = \operatorname{argmin}_{\boldsymbol{\beta}_j} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Gamma_{iij}^{(m)} (Y_{ij}^{(m)} - \mathbf{X}_i^T \boldsymbol{\beta}_j)^2 + \sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(m-1)}|)|\beta_{jk}| \right\} \quad (16)$$

and

$$\gamma_j^{(m)} = \operatorname{argmax}_{\gamma_j} \int \sqrt{\alpha_j^{(m-1)}(y) g_{nj}(y, \boldsymbol{\beta}_j^{(m)}, \gamma_j) \hat{f}_n(y)} \mathrm{d}y,$$
$$\alpha_j^{(m)} = \frac{[H_{nj}^{(m)}(\boldsymbol{\beta}_j^{(m)}, \gamma_j^{(m)})]^2}{\sum_{l=1}^{J} [H_{nl}^{(m)}(\boldsymbol{\beta}_l^{(m)}, \gamma_l^{(m)})]^2},$$

where $Y_{ij}^{(m)} = \mathbf{X}_i^T \boldsymbol{\beta}_j^{(m-1)}$, $\Gamma_{iij}^{(m)} = -\int w_{nj}^{(m-1)}(y) \frac{d^2 g(y,z,\gamma_j^{(m-1)})}{dz^2}|_{z=X_i^T \boldsymbol{\beta}_j^{(m-1)}} \mathrm{d}y$, and

$$w_{nj}^{(m-1)}(y) = \sqrt{\alpha_j^{(m-1)}(y) \hat{f}_n(y) / g_{nj}(y, \boldsymbol{\beta}_j^{(m-1)}, \gamma_j^{(m-1)})}.$$

Step 3. If $\|\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}\| < \varepsilon$ for a small prespecified threshold value ε, stop; otherwise, set $m = m + 1$ and return to Step 1.

To implement our algorithm, we need to select a bandwidth h_n and tuning parameters λ_{nj}. The bandwidth h_n can be selected using a general bandwidth-choice method for kernel density estimation such as the cross-validation method or the plug-in bandwidth-selection method. If we use the plug-in bandwidth-selection method, then the ideal optimal bandwidth is $h_{opt} = b_0 n^{-1/5}$ for some constant b_0, which satisfies condition (C5) in "Appendix". Since $1/5 \in (1/4, 1/2)$, according to Theorem 3 the estimators may be biased. Hence, the ideal bandwidth is not the optimal choice for such problems. For simulations in Sect. 5, we choose $h_n = n^{-2/5}$, which gives good empirical results. The tuning parameters λ_{nj} in (16) can be chosen using methods such as cross-validation, GCV, AIC or BIC.

One can use the MHD estimator (see circa (6)) or the least-absolute-deviation estimator of $\boldsymbol{\theta}_0$ as the initial estimator in the above algorithm. In our simulations, these two initial estimators gave comparable results.

5 Monte Carlo studies

In this section, we study the finite-sample performance and robustness properties of the proposed penalized MHD estimator defined by (9) using simulation studies. We compare them with those of the likelihood-based method proposed by Khalili and Chen (2007). We examine the finite-sample properties under the following two-component normal mixture regression model:

$$f_{\boldsymbol{\theta}}(y|\mathbf{X}) = \alpha_1 g(y, \mathbf{X}^T\boldsymbol{\beta}_1, \gamma_1) + (1-\alpha_1) g(y, X^T\boldsymbol{\beta}_2, \gamma_2), \tag{17}$$

where $g(y, \mu, \gamma) = \frac{1}{\sqrt{2\pi}\gamma}\exp(-(y-\mu)^2/2\gamma^2)$, $\mathbf{X} = (X_1, \ldots, X_5)^T$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j5})^T$ for $j = 1, 2$. We considered two distributions for the vector $\mathbf{X}$: $X_1, \ldots, X_5$ mutually independent: $X_j \sim U[0, 2]$ distribution and $\mathbf{X} \sim N(\mathbf{0}, \Pi)$ for $j = 1, \ldots, 5$, with $\Pi = (\rho_{jk})_{5\times 5}$ and $\rho_{jk} = \text{cor}(X_j, X_k) = (0.5)^{|j-k|}$. We also considered two cases for the regression parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$: (i) $\boldsymbol{\beta}_1 = (1.2, 0, 0, 0.8, 0)^T$ and $\boldsymbol{\beta}_2 = (2.5, 1, 0, 0, 2)^T$ (this is a case of components being well separated), and (ii) $\boldsymbol{\beta}_1 = (1.6, 0, 0.8, 0, 0)^T$ and $\boldsymbol{\beta}_2 = (1.5, 0, 1, 0, 0)^T$ (this is a case of components being not well separated). The mixing probabilities were $\alpha_1 = 0.25, 0.5, 0.75$ and $\gamma_1 = 1$, $\gamma_2 = 1.2$. In all the simulated designs, we used the SCAD penalty function with $a = 3.7$. We set the sample size n to be 200 and based the simulated results on 500 replications. We also examined the MCP penalty function with $a = 2$, but the results were similar to those for the SCAD penalty, so they were omitted here to save space.

For each simulated data set, we compared the performance of two variable selection methods: the proposed MHD method and the likelihood-based method of Khalili and Chen (2007). The SCAD penalty was used in both methods. We computed the penalized MHD estimator using the algorithm given in Sect. 4 with the MHD estimator (defined circa (6)) and the least-absolute-deviation estimator as initial estimators and the Epanechnikov kernel function, $K(t) = .75(1-t^2)$ for $|t| \leq 1$, as the kernel function in (6). The bandwidth was chosen as $h_n = n^{-2/5}$; this selection satisfies the

Table 1 Comparison of MHD and ML for model (17) with X from a uniform distribution

α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
0.25	MHDM	0.928	0.414	0.074	0.036	0.165	1.829	0.761	0.327	0.153
	MHDL	0.904	0.436	0.076	0.032	0.167	1.720	0.764	0.328	0.149
	ML	0.330	0.312	0.025	0.012	0.059	0.965	0.439	0.214	0.117
0.5	MHDM	0.102	0.100	0.148	0.060	0.105	0.594	0.986	0.148	0.161
	MHDL	0.106	0.088	0.164	0.089	0.107	0.580	0.968	0.154	0.150
	ML	0.070	0.066	0.058	0.022	0.038	0.377	0.546	0.116	0.124
0.75	MHDM	0.002	0.082	0.348	0.164	0.092	0.336	1.687	0.124	0.208
	MHDL	0.006	0.076	0.324	0.168	0.086	0.350	1.545	0.112	0.215
	ML	0.028	0.030	0.392	0.084	0.049	0.290	1.108	0.149	0.282

bandwidth assumptions in the theorems of Sect. 2. To select the tuning parameters λ_{nj} for $j = 1, 2$ in (16), we generated an independent validation set of sample size $n = 200$, following Fan et al. (2014). The validation error of the estimator $\hat{\boldsymbol{\beta}}_j$ is defined by $\sum_{i \in \text{validation}} \Gamma_{ij}^{(m)} (\tilde{Y}_{ij}^{(m)} - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j)^2$ with $\tilde{Y}_{ij}^{(m)} = \mathbf{X}_i^T \boldsymbol{\beta}_j^{(m-1)}$ and $\hat{f}_n(y)$ in $\Gamma_{ij}^{(m)}$ replaced by $\tilde{f}_n(y) = \frac{1}{nh_n} \sum_{i \in \text{validation}}^{n} K(\frac{y - Y_i}{h_n})$. We chose the penalization parameters λ_{nj} by minimizing the validation error. We computed the penalized maximum likelihood estimator (MLE) using the method proposed in Khalili and Chen (2007).

For each model, we measured the estimation accuracy by the average l_1-losses: $|\hat{\alpha}_1 - \alpha_1|$, $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1$, and $|\hat{\gamma}_j - \gamma_j|$ for $j = 1, 2$ over 500 replications. These are labeled as $l_1(\alpha_1)$, $l_1(\boldsymbol{\beta}_j)$ and $l_1(\gamma_j)$, $j = 1, 2$, in all tables. We also evaluated the selection accuracy by the average counts of false positive (FP) and false negative (FN) over the 500 replications; that is, the number of noise covariates included in the model and the number of signal covariates not included. These are labeled as FP1, FP2, FN1 and FN2 in all tables. Table 1 displays the simulation results for the normal model (17) with X having the uniform $[0, 2]$ distribution, $\boldsymbol{\beta}_1 = (1.2, 0, 0, 0.8, 0)^T$ and $\boldsymbol{\beta}_2 = (2.5, 1, 0, 0, 2)^T$. In Table 1, MHDM denotes the MHD variable selection method with the MHD estimator as the initial estimator, MHDL denotes the MHD variable selection method with the least-absolute-deviation estimator as the initial estimator, and ML denotes the maximum likelihood method (Khalili and Chen 2007).

Table 1 shows that, when the two components of the mixture model are well separated, the maximum likelihood method outperforms the minimum Hellinger distance method. Table 1 also shows that MHDL and MHDM give similar results. Thus, in other simulations we employed the least-absolute-deviation estimator as the initial estimator, as it was much easier to compute.

Table 2 presents the simulation results for $\mathbf{X}$ having the multivariate normal distribution $N(\mathbf{0}, \Pi)$, with $\boldsymbol{\beta}_1 = (1.2, 0, 0, 0.8, 0)^T$ and $\boldsymbol{\beta}_2 = (2.5, 1, 0, 0, 2)^T$. By comparing the values in Tables 1 and 2, we note that the two procedures give poor estimation and variable selection results. This is in part because the means of the two components of the mixture model are all zero, so the two components are not well separated.

Table 2 Comparison of MHD and ML for model (17) with X from a normal distribution

α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
0.25	MHD	1.800	0.976	0.036	0.158	0.257	4.450	0.995	1.218	0.905
	ML	1.760	0.862	0.240	0.400	0.298	3.907	1.484	0.451	0.420
0.5	MHD	1.152	0.770	0.312	0.802	0.126	2.859	2.519	1.103	0.727
	ML	1.424	0.702	0.456	0.940	0.168	3.253	2.718	0.394	0.364
0.75	MHD	0.420	0.378	0.674	1.582	0.238	1.222	4.131	0.712	0.391
	ML	0.922	0.380	0.804	1.290	0.294	2.262	3.683	0.340	0.409

Table 3 Comparison of MHD and ML for model (17) with X from a uniform distribution, $\boldsymbol{\beta}_1 = (1.6, 0, 0.8, 0, 0)^T$, and $\boldsymbol{\beta}_2 = (1.5, 0, 1, 0, 0)^T$

α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
0.25	MHD	0.110	0.262	0.100	0.070	0.203	1.060	0.614	0.136	0.259
	ML	0.206	0.434	0.192	0.098	0.256	0.888	0.654	0.271	0.257
0.5	MHD	0.074	0.206	0.082	0.074	0.134	0.973	0.567	0.136	0.257
	ML	0.254	0.384	0.176	0.102	0.278	0.846	0.678	0.283	0.275
0.75	MHD	0.092	0.244	0.070	0.132	0.291	0.971	0.614	0.119	0.297
	ML	0.234	0.526	0.152	0.104	0.398	0.929	0.635	0.234	0.299

From Table 2, we also observe that FP2 and $l_1(\alpha_1)$ for the MHD are less than those for the ML. Moreover, FP1, FN2, $l_1(\boldsymbol{\beta}_1)$ and $l_1(\boldsymbol{\beta}_2)$ for the MHD are less than those for the ML in most cases. On the other hand, $l_1(\gamma_1)$ and $l_1(\gamma_2)$ for the MHD are larger than those for the ML, and FN1 for the MHD is larger than that for the ML in most cases. When the two components of the mixture model are not well separated, a proportion of the data from one component are used for the variable selection and estimation of the regression coefficients of the other component. Furthermore, Table 2 shows that ML is sensitive to outlying distributions, whereas the proposed MHD method is more robust.

Next, we considered another case in which the regression coefficients are not well separated: $\boldsymbol{\beta}_1 = (1.6, 0, 0.8, 0, 0)^T$ and $\boldsymbol{\beta}_2 = (1.5, 0, 1, 0, 0)^T$. Tables 3 and 4 display the simulation results with X having the uniform[0, 2] distribution and the multivariate normal distribution $N(\mathbf{0}, \Pi)$, respectively. It is clear from Tables 3 and 4 that MHD outperforms ML in both cases. The estimates in Table 4 in fact are much better than those in Table 2. This may be interpreted as follows: when the regression coefficients of the two components of a mixture model are close, computing the estimators of the regression coefficients of one component using the data from the other component introduces a relatively small error.

To investigate the effect of data contamination on the two estimators, we added a third component, w, to model (17) with a contamination rate α, i.e.,

$$f_{\alpha\boldsymbol{\theta}}(y|\mathbf{X}) = (1-\alpha)[\alpha_1 g(y, \mathbf{X}^T\boldsymbol{\beta}_1, \gamma_1)+(1-\alpha_1)g(y, \mathbf{X}^T\boldsymbol{\beta}_2, \gamma_2)]+\alpha w(y|\mathbf{X}). \quad (18)$$

Table 4 Comparison of MHD and ML for model (17) with X from a normal distribution, $\boldsymbol{\beta}_1 = (1.6, 0, 0.8, 0, 0)^T$, and $\boldsymbol{\beta}_2 = (1.5, 0, 1, 0, 0)^T$

α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
0.25	MHD	0.038	0.008	0.010	0.020	0.256	0.363	0.329	0.422	0.226
	ML	0.538	0.104	0.626	0.090	0.345	0.532	0.572	0.370	0.413
0.5	MHD	0.032	0.010	0.010	0.050	0.137	0.332	0.379	0.397	0.225
	ML	0.524	0.060	0.690	0.088	0.304	0.456	0.601	0.347	0.441
0.75	MHD	0.018	0.012	0.002	0.170	0.252	0.289	0.472	0.384	0.202
	ML	0.336	0.198	0.460	0.238	0.331	0.550	0.765	0.322	0.440

Table 5 Comparison of MHD and ML for model (18) with $\mu = 0$ and $\sigma = 8$

n	α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
100	0.25	MHD	0.888	0.726	0.266	0.126	0.214	2.181	1.269	1.532	0.817
		ML	0.960	0.954	0.334	0.152	0.187	2.801	1.275	2.510	0.587
	0.5	MHD	0.268	0.390	0.364	0.226	0.166	1.140	1.767	1.260	0.961
		ML	0.604	0.636	0.682	0.262	0.209	1.818	2.023	1.643	0.985
	0.75	MHD	0.040	0.266	0.608	0.450	0.122	0.662	3.040	0.871	1.436
		ML	0.260	0.404	1.178	0.464	0.138	1.105	3.968	1.121	1.423
200	0.25	MHD	0.798	0.542	0.100	0.038	0.189	1.903	0.815	1.484	0.714
		ML	0.712	0.874	0.110	0.090	0.165	2.333	0.704	2.369	0.499
	0.5	MHD	0.096	0.150	0.200	0.084	0.146	0.659	1.108	1.016	0.924
		ML	0.318	0.398	0.304	0.138	0.184	1.277	1.136	1.532	0.740
	0.75	MHD	0.003	0.086	0.422	0.234	0.111	0.353	2.024	0.745	1.326
		ML	0.054	0.252	0.802	0.452	0.134	0.746	2.778	1.059	1.356

In the simulation, we set the contamination rate to be $\alpha = 0.05$ and w is taken as $w(y|x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(y-\mu)^2/(2\sigma^2))$. In each simulation, we computed the MHD and ML estimators based on samples of sizes 100 and 200 from model (17), but we replaced $100\alpha\%$ of the observations by random samples from the $N(\mu, \sigma^2)$ distribution, which can be interpreted as an outlier distribution. For X having the uniform[0, 2] distribution, with $\boldsymbol{\beta}_1 = (1.2, 0, 0, 0.8, 0)^T$ and $\boldsymbol{\beta}_2 = (2.5, 1, 0, 0, 2)^T$ being well separated, Tables 5 and 6 give the simulation results under model (18) with $\mu = 0$, $\sigma = 8$ and $\mu = -5$, $\sigma = 0.5$, respectively, which correspond to adding outliers on both ends and on the left end, respectively. It is clear from the FP, FN, and l_1-loss values in Tables 5 and 6 that the MHD has a better performance than the ML for these cases. Simulation results for adding outliers on the right end also gave a similar conclusion, but these results are omitted here to save space.

By comparing Table 1 with Tables 5 and 6, we note that a small proportion of contamination in the data can greatly affect the efficiency of the variable selection and parameter estimation of the ML method, whereas it has little influence on the proposed MHD method.

Table 6 Comparison of MHD and ML for model (18) with $\mu = -5$ and $\sigma = 0.5$

n	α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
100	0.25	MHD	0.962	1.300	0.198	0.074	0.204	3.958	1.061	2.825	1.154
		ML	1.138	1.424	0.280	0.176	0.172	6.298	1.221	1.537	0.681
	0.5	MHD	0.522	0.306	0.440	0.222	0.187	1.197	1.762	1.513	0.653
		ML	0.612	0.872	0.792	0.244	0.241	2.670	1.792	1.644	0.697
	0.75	MHD	0.448	0.842	0.638	0.818	0.324	2.088	3.276	1.992	0.879
		ML	0.820	0.972	0.914	0.840	0.329	3.709	3.436	1.216	0.964
200	0.25	MHD	0.768	1.446	0.028	0.072	0.218	3.049	0.780	2.248	1.039
		ML	1.094	1.066	0.094	0.140	0.182	4.522	0.912	1.376	0.589
	0.5	MHD	0.304	0.206	0.162	0.140	0.156	0.762	1.109	1.417	0.639
		ML	0.476	0.636	0.246	0.234	0.225	2.039	1.145	1.567	0.539
	0.75	MHD	0.276	0.758	0.624	0.780	0.333	1.427	2.801	1.618	0.723
		ML	0.652	0.874	0.890	0.810	0.389	3.532	2.835	1.224	0.854

In (18), let $X = (X_1, \ldots, X_{35})^T$, where $X_1, \ldots, X_{35}$ are mutually independent and $X_j \sim U[0, 2]$ for $j = 1, \ldots, 35$, $\boldsymbol{\beta}_1 = (1.2, 0, 0, 0.8, 0, \beta_6^{(1)}, \ldots, \beta_{35}^{(1)})^T$ and $\boldsymbol{\beta}_2 = (2.5, 1, 0, 0, 2, \beta_6^{(2)}, \ldots, \beta_{35}^{(2)})^T$ with $\beta_6^{(1)} = \cdots = \beta_{35}^{(1)} = 0$ and $\beta_6^{(2)} = \ldots = \beta_{35}^{(2)} = 0$. Table 7 displays the simulation results with $\mu = 0$ and $\sigma = 8$ for $n = 200$ and $n = 300$. We see from Table 7 that the MHD performs better than the ML, overall.

In summary, our simulation study shows that the ML method outperforms the MHD method when the components of the mixture model are well separated, whereas the MHD method performs better than the ML method when the components are not well separated. Furthermore, the MHD method is more robust than the ML method to data contamination.

6 Real-data examples

In this section, we analyze a real data set using the proposed methodology and compare the results with those of the penalized likelihood-based method (Khalili and Chen 2007). We analyzed the plasma beta-carotene level data set from a cross-sectional study (http://lib.stat.cmu.edu/datasets/PlasmaRetinol). It has observations for 315 patients, 273 females and 42 males.

Observational studies have suggested that a low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with an increased risk of developing certain types of cancer. We are interested in the relationship between the plasma beta-carotene level (betaplasma) and 11 covariates given as age (x_1), gender (x_2), smoking status (x_3), quetelet (x_4), vitamin use (x_5), number of calories consumed per day (x_6), grams of fat consumed per day (x_7), grams of fiber consumed per day (x_8), number of alcoholic drinks consumed per week (x_9), cholesterol consumed (x_{10}), and dietary beta-carotene consumed (x_{11}). Since the observational values of some predictors such as fiber and alcohol are small, other predictors such as

Table 7 Comparison of MHD and ML for model (18) with $X = (X_1, \ldots, X_{35})$ from a uniform distribution

n	α_1		FP1	FN1	FP2	FN2	$l_1(\alpha_1)$	$l_1(\boldsymbol{\beta}_1)$	$l_1(\boldsymbol{\beta}_2)$	$l_1(\gamma_1)$	$l_1(\gamma_2)$
200	0.25	MHD	1.752	1.252	2.204	0.308	0.186	2.952	3.274	1.808	0.427
		ML	1.688	1.388	2.874	0.338	0.159	3.572	2.002	2.468	0.335
	0.5	MHD	1.104	0.739	3.936	0.682	0.178	1.786	5.488	1.153	0.942
		ML	2.086	0.748	4.728	0.750	0.191	2.460	7.043	1.568	0.718
	0.75	MHD	0.474	0.224	7.934	1.386	0.123	0.819	9.106	0.709	1.658
		ML	1.032	0.330	10.546	1.378	0.118	1.220	27.694	1.097	1.592
300	0.25	MHD	1.591	1.173	0.202	0.288	0.167	2.632	1.470	1.921	0.777
		ML	1.642	1.232	0.336	0.322	0.136	3.503	1.067	2.364	0.252
	0.5	MHD	1.024	0.357	1.426	0.466	0.141	1.266	3.227	1.124	0.926
		ML	1.582	0.576	2.156	0.558	0.181	1.959	2.592	1.529	0.656
	0.75	MHD	0.362	0.070	5.054	1.246	0.118	0.556	6.620	0.627	1.564
		ML	0.634	0.196	5.782	1.342	0.116	0.847	18.638	1.077	1.529

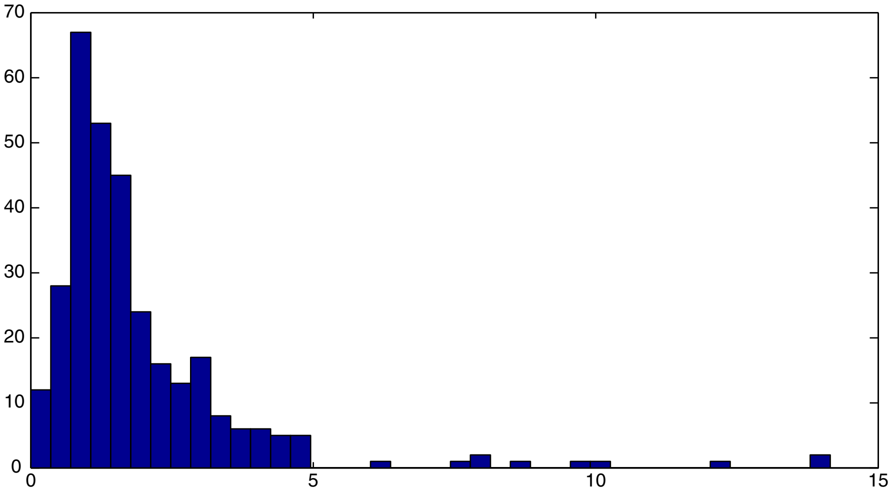

Fig. 1 Histogram of the response Y

Table 8 Parameter estimates for plasma beta-carotene level data set

		x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
MHD	C1	1.232	0	1.879	0	0	0	0	0	3.340	0	0	0
	C2	3.868	0	8.798	5.258	0	0	0	0	3.293	0	0	0
ML	C1	1.382	0	1.718	0	0	0	0	0	4.576	0	0	0
	C2	4.806	0	4.111	0	0	0	0	0	16.050	71.375	0	0

calories and dietary beta-carotene are large. We set Y = betaplasma/100, and all the predictors are first scaled to have a mean of zero and unit variance.

Figure 1 shows a histogram of Y. It indicates that there are some unusual points in the response and suggests a mixture of two normal linear models:

$$Y \sim \alpha N(\mathbf{x}^T \boldsymbol{\beta}_1, \sigma_1^2) + (1-\alpha) N(\mathbf{x}^T \boldsymbol{\beta}_2, \sigma_2^2), \tag{19}$$

with $\mathbf{x}$ being a 12×1 vector containing all the 11 potential covariates plus an intercept. Table 8 presents the parameter estimates for the mixture model (19) calculated using the proposed penalized MHD method and the penalized ML method, with the SCAD penalty for both methods. For the MHD method, we have $\hat{\alpha} = 0.8555$, $\hat{\sigma}_1 = 0.4985$, and $\hat{\sigma}_2 = 2.6623$; and for the ML method, we obtain $\hat{\alpha} = 0.8310$, $\hat{\sigma}_1 = 0.5434$, and $\hat{\sigma}_2 = 2.2519$. By comparing Table 8 with Table 6 of Wang et al. (2013), we find that all three methods, i.e., MHD, ML, and ESL-LASSO, select fiber. Table 8 shows that MHD and ML both select gender. As indicated by the coefficients of the second component of the model in Table 8, the ML method is more sensitive than the MHD method to the outliers.

To evaluate the prediction performance of the selected models and the three methods, we applied a combination of the bootstrap and a cross-validation method to the data set. For each bootstrap sample, we randomly divided the data into five partitions.

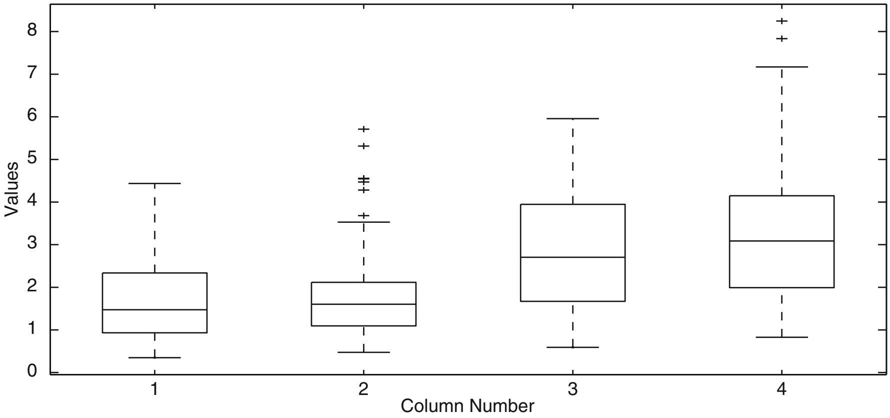

Fig. 2 Boxplots for MSPE. *1* MHD method based on two components; *2* ML method based on two components; *3* MHD method based on one component; *4* ML method based on one component

We used four folds of the data to estimate the model and the remaining fold for the testing data set. We classified each element of the testing data set into one of the two components based on their estimated posterior probabilities (i.e., to component 1 (2) if the estimated posterior probability for component 1 (2) is greater than 0.5). We calculated the mean squared prediction error (MSPE) for the testing data set. The MSPEs for the MHD and ML methods over the 200 replications are reported as boxplots in Fig. 2. For comparison, Fig. 2 also displays the boxplots of the MSPEs for the MHD and ML over the 200 replications based on a one-component linear model. This figure shows that a finite mixture of two regressions fits the data better than the ordinary linear model does. It also shows that the MHD has better predicting power than the ML for this data set, which has some unusual data points.

7 Concluding remarks

We have proposed a robust variable selection procedure for FMR models using a minimum-distance technique. We have investigated the asymptotic properties of the proposed estimator. We have established some global robustness properties of our estimator by finding its finite-sample breakdown point. Our breakdown-point results appear to be new; to the best of our knowledge, there is no existing work on breakdown-point analysis for FMR models.

Variable selection is fundamentally important for knowledge discovery with high-dimensional data (Fan and Li 2006). In spite of considerable progress on variable selection in various models for high-dimensional data, there has been little work on FMR models in this context. In fact, it appears that only the excellent work of Städler et al. (2010) and Khalili and Lin (2013) addresses variable selection for high-dimensional FMR models. From our simulations, we observed that the proposed method does not perform quite well for high-dimensional cases such as $p > n$. Thus, it would be interesting and useful to develop results similar to those in the present paper for high-dimensional settings.

Compared to likelihood-based approaches, inferences for FMR models with a distance measure, especially with the Hellinger distance, are generally more involved from a computational point of view. Our algorithm reduces the computational burden somewhat. Indeed, all the computations reported in Sects. 5 and 6 were carried out on a personal computer. Although we have discussed only the continuous case in detail, similar results can be easily established for the discrete case.

It is well known now that the minimum Hellinger distance approach leads to good robust estimation results for various models. This approach, however, requires a nonparametric density estimator and generally involves with some complications such as the bandwidth-selection problem. To avoid density estimation, some alternative approaches have been proposed; see, for example, the density power divergence technique defined in Basu et al. (1988) and the decomposable pseudo-distances method proposed in Broniatowski et al. (2012), among others. It would be interesting to develop results for robust variable selection and estimation in FMR models based on these methods.

Acknowledgements We wish to thank the Chief Editor, Professor Kenji Fukumizu, an Associate Editor, and two reviewers for their helpful comments and suggestions that led to substantial improvements in this paper. Q. Tang's research was supported in part by the National Social Science Foundation of China (16BTJ019) and Jiangsu Natural Science Foundation of China (BK20151481). R.J. Karunamuni's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

In this "Appendix", we list the conditions used in the theorems and outline the proofs of main results. For convenience of notation, we write $f_{\boldsymbol{\theta}}(y)$ for $f_{\boldsymbol{\theta},\eta}(y)$ defined by (4). The following technical conditions are imposed:

(C1) $E\|\mathbf{X}\|^2 < +\infty$ and $\max_{1\le i\le n}\|X_i\| = O_p(1)$. $\int \sup_{t\in\Theta}|\frac{\partial f_t(y)}{\partial t}|\mathrm{d}y < +\infty$, where $|b| = \max_{1\le i\le k}|b_i|$ for a vector $b = (b_1,\ldots,b_k)^T$. $\ddot{S}_{\boldsymbol{\theta}}^{(l,m)}(y) \in L_2$ for $1 \le l, m \le J(p+2)-1$ and $H(\boldsymbol{\theta}_0) = -\int \ddot{S}_{\boldsymbol{\theta}_0}(y) f_{\boldsymbol{\theta}_0}^{1/2}(y)\mathrm{d}y$ is a positive definite matrix, where $\ddot{S}_{\boldsymbol{\theta}}^{(l,m)}(y)$ denotes the $(l,m)^{th}$ component of $\ddot{S}_{\boldsymbol{\theta}}(y)$, $\ddot{S}_{\boldsymbol{\theta}}(y) = \frac{\partial^2 S_{\boldsymbol{\theta}}(y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$, and $S_{\boldsymbol{\theta}}(y) = f_{\boldsymbol{\theta}}^{1/2}(y)$.

(C2) The second and third continuous partial derivatives of $g(y,z,u)$ exist w.r.t. y and z, u, respectively. For a given $\tilde{L} > 0$ and some ϵ-neighborhood of θ, $B(\theta,\epsilon)$, define $\tilde{g}(y) = \inf_{\|x\|\le\tilde{L},t\in B(\theta,\epsilon)}\min_{1\le j\le k} g(y,\mathbf{x}^T t_j,u_j)$. Suppose that $1/\tilde{g}(y)$ is bounded on any compact subset of R and that, as $L\to\infty$,

$$
\begin{array}{ll}
\int_{|y|>L}\int_{\|x\|\le\tilde{L}}|x_r|\breve{g}_z(y,\mathbf{x})\mathrm{d}\eta(\mathbf{x})\mathrm{d}y \to 0, & \int_{|y|>L}\int_{\|x\|\le\tilde{L}}\breve{g}_u(y,\mathbf{x})\mathrm{d}\eta(\mathbf{x})\mathrm{d}y \to 0,\\
\int_{|y|>L}\int_{\|x\|\le\tilde{L}} g^*(y,\mathbf{x})\mathrm{d}\eta(\mathbf{x})\mathrm{d}y \to 0, & \int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}} x_r^2(\dot{g}_z^*(y,\mathbf{x}))^2\mathrm{d}\eta(\mathbf{x})}{\tilde{g}(y)}\mathrm{d}y \to 0,\\
\int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}}(\dot{g}_u^*(y,\mathbf{x}))^2\mathrm{d}\eta(\mathbf{x})}{\tilde{g}(y)}\mathrm{d}y \to 0, & \int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}} x_r^4(\dot{g}_z^*(y,\mathbf{x}))^4\mathrm{d}\eta(\mathbf{x})}{\tilde{g}^3(y)}\mathrm{d}y \to 0,\\
\int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}}(\dot{g}_u^*(y,\mathbf{x}))^4\mathrm{d}\eta(\mathbf{x})}{\tilde{g}^3(y)}\mathrm{d}y \to 0, & \int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}} x_r^2x_q^2(\ddot{g}_{zz}^*(y,\mathbf{x}))^2\mathrm{d}\eta(\mathbf{x})}{\tilde{g}(y)}\mathrm{d}y \to 0,\\
\int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}}(\ddot{g}_{uu}^*(y,\mathbf{x}))^2\mathrm{d}\eta(\mathbf{x})}{\tilde{g}(y)}\mathrm{d}y \to 0, & \int_{|y|>L}\frac{\int_{\|x\|\le\tilde{L}} x_r^2(\ddot{g}_{zu}^*(y,\mathbf{x}))^2\mathrm{d}\eta(\mathbf{x})}{\tilde{g}(y)}\mathrm{d}y \to 0
\end{array}
$$

for $r, q = 0, \ldots, p$, where $x_0 = 1$, $\breve{g}_z(y, \mathbf{x}) = \sup_{t\in\Theta} \max_{1\le j\le k} |\dot{g}_z(y, \mathbf{x}^T t_j, u_j)|$, $\breve{g}_u(y, \mathbf{x}) = \sup_{t\in\Theta} \max_{1\le j\le k} |\dot{g}_u(y, \mathbf{x}^T t_j, u_j)|$,

$$g^*(y, \mathbf{x}) = \sup_{t\in B(\theta,\epsilon)} \max_{1\le j\le k} g(y, x^T t_j, u_j), \quad \dot{g}^*_z(y, x) = \sup_{t\in B(\theta,\epsilon)} \max_{1\le j\le k} |\dot{g}_z(y, x^T t_j, u_j)|,$$

$\ddot{g}^*_{zz}(y, \mathbf{x}) = \sup_{t\in B(\theta,\epsilon)} \max_{1\le j\le k} |\ddot{g}_{zz}(y, \mathbf{x}^T t_j, u_j)|$, $\dot{g}^*_u(y, \mathbf{x})$, $\ddot{g}^*_{zu}(y, \mathbf{x})$, and $\ddot{g}^*_{uu}(y, \mathbf{x})$ are defined in a similar fashion, $\dot{g}_z(y, z, u) = \frac{\partial g(y,z,u)}{\partial z}$, $\dot{g}_u(y, z, u) = \frac{\partial g(y,z,u)}{\partial u}$, $\ddot{g}_{zz}(y, z, u) = \frac{\partial^2 g(y,z,u)}{\partial z^2}$, and $\ddot{g}_{zu}(y, z, u) = \frac{\partial^2 g(y,z,u)}{\partial z \partial u}$.

(C3) The kernel function $K(\cdot)$ is a bounded symmetric density with compact support $[-M, M]$.

(C4) $\sup_{y\in\mathbb{R}} \sup_{|v|\le M} \frac{f_{\boldsymbol{\theta}_0}(y+v)}{f^{1/2}_{\boldsymbol{\theta}_0}(y)} = O(1)$, $\sup_{y\in\mathbb{R}} \sup_{|v|\le M} \frac{[f''_{\boldsymbol{\theta}_0}(y+v)]^2}{f^{7/4}_{\boldsymbol{\theta}_0}(y)} = O(1)$, and as $L \to \infty$ $\int_{|y|>L} \frac{\dot{S}^2_{\boldsymbol{\theta}_0 q}(y)}{f^{1/2}_{\boldsymbol{\theta}_0}(y)} \mathrm{d}y \to 0$ for $q = 1, \ldots, J(p+2)-1$, where $f''_{\boldsymbol{\theta}}(y) = \frac{\partial^2 f_{\boldsymbol{\theta}}(y)}{\partial y^2}$ and $\dot{S}_{\boldsymbol{\theta}q}(y)$ is the q^{th} entry of the vector $\dot{S}_{\boldsymbol{\theta}}(y)$.

(C5) The bandwidth $h_n = b_0 n^{-\gamma}$ for some $\gamma \in (1/8, 1/2)$ and constant $b_0 > 0$. $E|Y|^s < +\infty$ for $s > 6/(1-2\gamma)$. There exists some l, $1/s < l < (1-2\gamma)/6$, satisfying $\sup_{|y|\le n^l} \sup_{|v|\le M} \frac{f_{\boldsymbol{\theta}_0}(y+v)}{f_{\boldsymbol{\theta}_0}(y)} = O(1)$, and as $n \to \infty$

$$(n^{1/2}h_n)^{-1} \int_{|y|\le n^l} \frac{|\dot{g}_{jq}(y)|}{g_j(y)} \mathrm{d}y \to 0, \quad (n^{1/2}h_n)^{-1} \int_{|y|\le n^l} \frac{|\dot{g}_{j\gamma}(y)|}{g_j(y)} \mathrm{d}y \to 0,$$

where $g_j(y) = g_j(y, \beta_j, \gamma_j) = \int g(y, x^T\beta_j, \gamma_j)\mathrm{d}\eta(x)$; $\dot{g}_{jq}(y) = \frac{\partial g_j(y)}{\partial \beta_{jq}}$ for $j = 1, \ldots, J$, $q = 1, \ldots, p$; and $\dot{g}_{j\gamma}(y) = \frac{\partial g_j(y)}{\partial \gamma_j}$.

(C6) $\int \frac{\int g^4(y, \mathbf{x}^T\boldsymbol{\beta}_j, \gamma_j)\mathrm{d}\eta(x)}{g^3_j(y)} \mathrm{d}y < +\infty$; $\int \frac{\int x_r^2 \dot{g}^2_z(y, \mathbf{x}^T\boldsymbol{\beta}_j, \gamma_j)\mathrm{d}\eta(x)}{g_j(y)} \mathrm{d}y < +\infty$ for $j = 1, \ldots, J$, $r = 1, \ldots, p$; and $\int \frac{\int \dot{g}^2_u(y, \mathbf{x}^T\boldsymbol{\beta}_j, \gamma_j)\mathrm{d}\eta(x)}{g_j(y)} \mathrm{d}y < +\infty$.

Conditions (C1) and (C2) guarantee that (2.5) and (2.6) of Beran (1977) hold about an expansion of the first and second partial derivatives in some neighborhood of $\boldsymbol{\theta}_0$. Condition (C3) is also a typical assumption on kernels, including the family of symmetric beta kernel functions (Chen 1999). Conditions (C4)–(C6) are used to derive the asymptotic normality of the MHD estimators. When X is bounded and $g(y, x^T\beta_j, \gamma_j) = \exp\{-(y - x^T\beta_j)^2/(2\gamma_j^2)\}/(\sqrt{2\pi}\gamma_j)$, or X is a normal random variable and $g(y, x, \beta_{j1}, \beta_{j2}, \gamma_j) = \exp\{-[y - (\beta_{j1} + \beta_{j2}x)]^2/(2\gamma_j^2)\}/(\sqrt{2\pi}\gamma_j)$ for $j = 1, \ldots, J$, the above conditions are satisfied, see Remarks 3.4 and 3.4 of Tang and Karunamuni (2013) for details.

Lemma 1 *Under the assumptions of Theorem 3, there exists a local minimizer $\hat{\boldsymbol{\theta}}$ of (9) such that $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| = O_p(n^{-1/2})$.*

Proof Let

$$P_n(\boldsymbol{\theta}) = \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}|, \quad P_{n1}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=1}^{d_j} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}|$$

and $D_n(\boldsymbol{\theta}) = \int S_{\boldsymbol{\theta},n}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y - P_n(\boldsymbol{\theta})$. It suffices to prove that for any given $\varepsilon > 0$, there exists a constant C such that

$$P\left\{ \sup_{\|v\|=C} D_n(\boldsymbol{\theta}_0 + n^{-1/2}v) < D_n(\boldsymbol{\theta}_0) \right\} \geq 1 - \varepsilon. \tag{20}$$

Note that

$$\begin{aligned} D_n(\boldsymbol{\theta}_0 + n^{-1/2}v) - D_n(\boldsymbol{\theta}_0) \leq & \int [S_{\boldsymbol{\theta}_0+n^{-1/2}v,n}(y) - S_{\boldsymbol{\theta}_0,n}(y)]\hat{f}_n^{1/2}(y)\mathrm{d}y \\ & -[P_{n1}(\boldsymbol{\theta}_{01} + n^{-1/2}v_1) - P_{n1}(\boldsymbol{\theta}_{01})]. \end{aligned} \tag{21}$$

By a Taylor expansion,

$$\begin{aligned} & \int [S_{\boldsymbol{\theta}_0+n^{-1/2}v,n}(y) - S_{\boldsymbol{\theta}_0,n}(y)]\hat{f}_n^{1/2}(y)\mathrm{d}y \\ & = n^{-1/2}v \int \dot{S}_{\boldsymbol{\theta}_0,n}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y + \frac{1}{2n}v^T \int \ddot{S}_{\boldsymbol{\theta}^*,n}(y)\hat{f}_n^{1/2}(y)\mathrm{d}yv, \end{aligned}$$

where $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + n^{-1/2}v$. As in the proof of Lemma 3.1 of Tang and Karunamuni (2013), we have

$$\int [\ddot{S}_{\boldsymbol{\theta}^*,n}(y) - \ddot{S}_{\boldsymbol{\theta}_0,n}(y)]\hat{f}_n^{1/2}(y)\mathrm{d}y \leq \left(\int [\ddot{S}_{\boldsymbol{\theta}^*,n}(y) - \ddot{S}_{\boldsymbol{\theta}_0,n}(y)]^2\mathrm{d}y\right)^{1/2} \left(\int \hat{f}_n(y)\mathrm{d}y\right)^{1/2} = o_p(1).$$

Similar to the proof of Theorem 3.2 of Tang and Karunamuni (2013), we obtain

$$\int \ddot{S}_{\boldsymbol{\theta}_0,n}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y = -H(\boldsymbol{\theta}_0) + o_p(1),$$

where $H(\boldsymbol{\theta}_0) = -\int \ddot{S}_{\boldsymbol{\theta}_0}(y) f_{\boldsymbol{\theta}_0}^{1/2}(y)\mathrm{d}y$. Hence

$$\begin{aligned} & \int [S_{\boldsymbol{\theta}_0+n^{-1/2}v,n}(y) - S_{\boldsymbol{\theta}_0,n}(y)]\hat{f}_n^{1/2}(y)\mathrm{d}y \\ & = n^{-1/2}v \int \dot{S}_{\boldsymbol{\theta}_0,n}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y - \frac{1}{2n}v^T H(\boldsymbol{\theta}_0)v[1 + o_p(1)]. \end{aligned} \tag{22}$$

By (A.26) of Tang and Karunamuni (2013), it follows that $\int \dot{S}_{\boldsymbol{\theta}_0,n}(y)\hat{f}_n^{1/2}(y)\mathrm{d}y = O_p(n^{-1/2})$. Since $\boldsymbol{\theta}^{(0)} \to_P \boldsymbol{\theta}_0$, we have $P\{P_{n1}(\boldsymbol{\theta}_{01} + n^{-1/2}v_1) - P_{n1}(\boldsymbol{\theta}_{01}) = 0\} \to 1$ as $n \to \infty$. Hence, for sufficiently large C, (20) follows from (21) and (22) and the fact that $H(\boldsymbol{\theta}_0)$ is positive definite. The proof of Lemma 1 is complete. □

Lemma 2 *Under the assumptions of Theorem 3, for any $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O(n^{-1/2})$ and $\boldsymbol{\theta}_2 \neq \mathbf{0}$, with probability tending to 1, we have*

$$D_n((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) < D_n((\boldsymbol{\theta}_1, \mathbf{0})).$$

Proof By a Taylor expansion, we obtain

$$S_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2),n}(y) = S_{(\boldsymbol{\theta}_1,\mathbf{0}),n}(y) + \boldsymbol{\theta}_2^T \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta}_2}\bigg|_{\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\mathbf{0})} + \frac{1}{2}\boldsymbol{\theta}_2^T \frac{\partial^2 S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}\bigg|_{\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2^*)} \boldsymbol{\theta}_2,$$

where $\boldsymbol{\theta}_2^*$ is between $\mathbf{0}$ and $\boldsymbol{\theta}_2$. As in the proof of (22), it follows that

$$\int \frac{\partial^2 S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}\bigg|_{\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2^*)} \hat{f}_n^{1/2}(y)\mathrm{d}y = \int \frac{\partial^2 S_{\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f_{\boldsymbol{\theta}_0}^{1/2}(y)\mathrm{d}y[1 + o_p(1)] = O_p(1).$$

By (A.26) of Tang and Karunamuni (2013), we have

$$\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta}_2}\bigg|_{\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\mathbf{0})} \hat{f}_n^{1/2}(y)\mathrm{d}y = O_p(n^{-1/2})\ .$$

Using the fact that $\|\boldsymbol{\theta}_2\| = O(n^{-1/2})$ and $n^{1/2}\lambda_{nj} \to +\infty$, we deduce that with probability tending to 1, it holds that

$$\begin{aligned}
&D_n((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) - D_n((\boldsymbol{\theta}_1, \mathbf{0}))\\
&\quad = \int [S_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2),n}(y) - S_{(\boldsymbol{\theta}_1,\mathbf{0}),n}(y)]\hat{f}_n^{1/2}(y)\mathrm{d}y - \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=d_j}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}|\\
&\quad = O_p(n^{-1/2}) \sum_{j=1}^{J}\sum_{k=d_j}^{p} |\beta_{jk}| - \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=d_j}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\beta_{jk}|\\
&\quad = \sum_{j=1}^{J} \lambda_{nj} \sum_{k=d_j}^{p} \left[O_p((n^{1/2}\lambda_{nj})^{-1}) - \alpha_j^{1/2}\lambda_{nj}^{-1} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)\right] |\beta_{jk}| < 0.
\end{aligned}$$

This completes the proof of Lemma 2. □

Proof of Theorem 3. By Lemmas 1 and 2, there exists a $\sqrt{n}$-consistent local maximizer $\check{\boldsymbol{\theta}} = (\check{\boldsymbol{\theta}}_1, \mathbf{0}^T)^T$ of (9). By a Taylor expansion, with probability tending 1, we have

$$\begin{aligned}
&D_n((\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2))\\
&\quad = D_n((\check{\boldsymbol{\theta}}_1, \mathbf{0})) + (\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})^T \int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=(\check{\boldsymbol{\theta}}_1,0)} \hat{f}_n^{1/2}(y)\mathrm{d}y\\
&\qquad + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})^T \int \frac{\partial^2 S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \hat{f}_n^{1/2}(y)\mathrm{d}y(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}) - \sum_{j=1}^{J} \alpha_j^{1/2} \sum_{k=d_k+1}^{p} p'_{\lambda_{nj}}\left(|\beta_{jk}^{(0)}|\right)|\hat{\beta}_{jk}|,
\end{aligned}$$

where $\boldsymbol{\theta}^*$ is between $\hat{\boldsymbol{\theta}}$ and $\check{\boldsymbol{\theta}}$. By Theorem 2, it follows that $\hat{\boldsymbol{\theta}} \rightarrow_P \boldsymbol{\theta}_0$. Using an argument similar to the one used in the proof of (22), we obtain $\int \frac{\partial^2 S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}{}_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\hat{f}_n^{1/2}(y)\mathrm{d}y = -H(\boldsymbol{\theta}_0)[1+o_p(1)]$. Noting that with probability tending to 1, $\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_1}{}_{\boldsymbol{\theta}=(\check{\boldsymbol{\theta}}_1,0)}\hat{f}_n^{1/2}(y)\mathrm{d}y = 0$, we have

$$\begin{aligned}
&D_n((\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2))\\
&\quad = D_n((\check{\boldsymbol{\theta}}_1, \mathbf{0})) + \hat{\boldsymbol{\theta}}_2^T \int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=(\check{\boldsymbol{\theta}}_1,0)}\hat{f}_n^{1/2}(y)\mathrm{d}y\\
&\qquad -\frac{1}{2}(\hat{\boldsymbol{\theta}}-\check{\boldsymbol{\theta}})^T H(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\check{\boldsymbol{\theta}})[1+o_p(1)] - \sum_{j=1}^{J}\alpha_j^{1/2}\sum_{k=d_k+1}^{p} p'_{\lambda_{nj}}(|\beta_{jk}^{(0)}|)|\hat{\beta}_{jk}|,
\end{aligned} \tag{23}$$

Using a Taylor expansion, we have

$$\begin{aligned}
\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=(\check{\boldsymbol{\theta}}_1,0)}\hat{f}_n^{1/2}(y)\mathrm{d}y = \int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}&\\
\hat{f}_n^{1/2}(y)\mathrm{d}y + H_{21}(\boldsymbol{\theta}_0)\check{\boldsymbol{\theta}}_1[1+o_p(1)],&
\end{aligned}$$

where $H_{21}(\boldsymbol{\theta}_0) = \int \frac{\partial^2 S_{\boldsymbol{\theta}}(y)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1^T}{}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f_{\boldsymbol{\theta}_0}^{1/2}(y)\mathrm{d}y$. By (A.26) of Tang and Karunamuni (2013), we have $\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\hat{f}_n^{1/2}(y)\mathrm{d}y = O_p(n^{-1/2})$. Then $\check{\boldsymbol{\theta}}_1 = O_p(n^{-1/2})$ implies that $\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_2}{}_{\boldsymbol{\theta}=(\check{\boldsymbol{\theta}}_1,0)}\hat{f}_n^{1/2}(y)\mathrm{d}y = O_p(n^{-1/2})$. If $\hat{\boldsymbol{\theta}} \neq \check{\boldsymbol{\theta}}$, then by (23), we have $D_n((\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)) < D_n((\check{\boldsymbol{\theta}}_1, \mathbf{0}))$. This is a contradiction to the fact that $\hat{\boldsymbol{\theta}}$ is a maximizer of (10). So $\hat{\boldsymbol{\theta}}_2 = 0$ and $\hat{\boldsymbol{\theta}}_1 = \check{\boldsymbol{\theta}}_1$.

We now prove the asymptotic normality part. Consider $D_n((\boldsymbol{\theta}_1, \mathbf{0}))$ as a function of $\boldsymbol{\theta}_1$. Noting that with probability tending 1, $\hat{\boldsymbol{\theta}}_1$ is the $\sqrt{n}$-consistent maximizer of $D_n((\boldsymbol{\theta}_1, \mathbf{0}))$ and satisfies

$$\frac{\partial D_n((\boldsymbol{\theta}_1, \mathbf{0}))}{\partial\boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1} = \int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1,\mathbf{0})}\hat{f}_n^{1/2}(y)\mathrm{d}y = 0.$$

By an argument similar to the one used in the proof of (22), we obtain

$$\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_1,\mathbf{0})}\hat{f}_n^{1/2}(y)\mathrm{d}y = \int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\hat{f}_n^{1/2}(y)\mathrm{d}y - H_1(\boldsymbol{\theta}_{01})(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01})[1+o_p(1)].$$

Hence

$$H_1(\boldsymbol{\theta}_{01})(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01})[1+o_p(1)] = \int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial\boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\hat{f}_n^{1/2}(y)\mathrm{d}y\ . \tag{24}$$

Using an argument similar to the one used in the proof of Theorem 3.3 of Tang and Karunamuni (2013), we obtain

$$n^{1/2}\left(\int \frac{\partial S_{\boldsymbol{\theta},n}(y)}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \hat{f}_n^{1/2}(y)\mathrm{d}y - A_{n1}(\boldsymbol{\theta}_{01})\right) \rightarrow_d N\left(\mathbf{0}, \Sigma_1(\boldsymbol{\theta}_{01})\right).$$

Now (11) follows from the preceding expression and (24). This completes the proof of Theorem 3. □

Proof of Theorem 5 Note that

$$\begin{aligned} &2\left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 \\ &\quad \geq 2\left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n))\right\|^2 - 2\left\| f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2. \end{aligned} \tag{25}$$

Since

$$\begin{aligned} &\left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 \\ &\quad + 2\sum_{j=1}^{J} \hat{\alpha}_{j,n+n_1}^{1/2} \sum_{k=1}^{p} p'_{\lambda_{(n+n_1)j}}(|\beta_{jk,n+n_1}^{(0)}|)|\hat{\beta}_{jk,n+n_1}| \\ &\leq \left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 + 2\sum_{j=1}^{J} \hat{\alpha}_{j,n}^{1/2} \sum_{k=1}^{p} p'_{\lambda_{(n+n_1)j}}(|\beta_{jk,n+n_1}^{(0)}|)|\hat{\beta}_{jk,n}|, \end{aligned}$$

we have

$$\begin{aligned} &\left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 \\ &\quad \leq \left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 + \left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n))\right\|^2 + 2P_{n,n_1}^{*}. \end{aligned} \tag{26}$$

Clearly

$$\begin{aligned} &\left\| f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 \\ &\quad \leq \left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1})\right\|^2 + \left\| f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n))\right\|^2 \end{aligned} \tag{27}$$

and

$$\begin{aligned}
&\left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 \\
&\geq \left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 - \left\| f_{n+n_1}^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) \right. \\
&\left. - f^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) \right\|^2 - \left\| f_n^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 . \qquad (28)
\end{aligned}$$

Combining (25)–(28), we conclude that

$$4 \left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1}) \right\|^2 \geq \left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 - \varepsilon_{n,n_1}$$

If $\hat{\boldsymbol{\theta}}(h_n)$ breaks down, then $\sup_{\#\boldsymbol{V}_{n_1}=n_1} d(\hat{\boldsymbol{\theta}}(h_n), \hat{\boldsymbol{\theta}}(h_{n+n_1})) = \infty$. So by the definition of δ^*, $\left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_{n+n_1})) - f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) \right\|^2 \geq \delta^*$. Hence, $\left\| f^{1/2}(\hat{\boldsymbol{\theta}}(h_n)) - \hat{f}_{n+n_1}^{1/2}(h_{n+n_1}) \right\|^2 \geq (\delta^* - \varepsilon_{n,n_1})/4$. The rest of the proof is similar to that of Tamura and Boos (1986) and is therefore omitted to save space. This completes the proof of Theorem 5. □

References

Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, *85*, 549–559.

Beran, R. (1977). Minimum Hellinger distance estimators for parametric models. *Annals of Statistics*, *5*, 445–463.

Beran, R. (1978). An efficient and robust adaptive estimator of location. *Annals of Statistics*, *6*, 292–313.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Broniatowski, M., Toma, A., Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, *142*, 2574–2585.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics and Data Analysis*, *31*, 131–145.

Cutler, A., Cordero-Braña, O. I. (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, *91*, 1716–1723.

Devlin, S. J., Gnandesikan, R., Kettenring, J. R. (1981). Robust Estimation of Dispersion Matrices and Principal Components. *Journal of the American Statistical Association*, *76*, 354–362.

Devroye, L. P., Wagner, T. J. (1979). The L_1 convergence of kernel density estimates. *Annals of Statistics*, *7*, 1136–1139.

Donoho, D. (1982). *Breakdown properties of multivariate location estimators. Unpublished qualifying paper*. Cambridge, Massachusetts, USA: Harvard University, Department of Statistics.

Donoho, D., Huber, P. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, J. L. Hodges Jr. (Eds.), *A Festschrift for E. L. Lehmann* (pp. 157–184). Belmont, CA: Wadsworth.

Donoho, D. L., Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *Annals of Statistics*, *16*, 552–586.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *In International Congress of Mathematicians*, *3*, 595–622.

Fan, J., Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transaction Information Theory*, *57*, 5467–5484.

Fan, J., Xue, L., Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, *42*, 819–849.

Frank, I., Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, *35*, 109–135.

Friedman, J., Hastie, T., Höflinng, H., Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*, 302–332.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust statistics*: *The approach based on influence functions*. New York: Wiley.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, *17*, 273–296.

Jiang, W., Tanner, M. A. (1999). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Machine Learning*, *11*, 1183–1198.

Karlis, D., Xekalaki, E. (2001). Robust inference for finite mixtures. *Journal of Statistical Planning and Inference*, *93*, 93–115.

Karunamuni, R. J., Wu, J. (2011). One-step minimum Hellinger distance estimation. *Computational Statistics and Data Analysis*, *55*, 3148–3164.

Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *The Canadian Journal of Statistics*, *38*, 519–539.

Khalili, A., Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, *102*, 1025–1038.

Khalili, A., Lin, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics*, *69*, 436–446.

Khalili, A., Chen, J., Lin, S. (2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics*, *12*, 156–172.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, *11*(8), 1–18.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, *22*, 1081–1114.

Lu, Z., Hui, Y. V., Lee, A. H. (2003). Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics*, *59*, 1016–1026.

Lv, J., Fan, J. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, *37*, 3498–3528.

Markatou, M. (2000). Mixture models, robustness and the weighted likelihood methodology. *Biometrics*, *56*, 483–486.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics*, *4*, 51–67.

McLachlan, G. J., Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Pollard, D. (1981). Stong consistency of k-means clustering. *Annals of Statistics*, *9*, 135–140.

Shen, L. Z. (1995). On optimal B-robust influence functions in semiparametric models. *Annals of Statistics*, *23*, 968–989.

Städler, N., Bühlmann, P., van de Geer, S. (2010). l_1-penalization for mixture regression models. *Test*, *19*, 209–256.

Tamura, R., Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, *81*, 223–229.

Tang, Q., Karunamuni, R. J. (2013). Minimum distance estimation in a finite mixture regression model. *Journal of Multivariate Analysis*, *120*, 185–204.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*: *Series B*, *58*, 267–288.

Titterington, D. M., Smith, A. F. M., Markov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

Toma, A. (2008). Minimum Hellinger distance estimators for multivariate distributions from Johnson system. *Journal of Statistical Planning and Inference*, *138*, 803–816.

van der Vaart, A. (1996). Efficient maximum likelihood estimation in semiparametric models. *Annals of Statistics*, *24*, 862–878.

Wang, H., Li, G., Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics*, *25*, 347–355.

Wang, X., Jiang, Y., Huang, M., Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, *108*, 632–643.

Wu, J., Karunamuni, R. J. (2012). Efficient Hellinger distance estimates for semiparametric models. *Journal of Multivariate Analysis*, *107*, 1–23.

Wu, J., Karunamuni, R. J. (2015). Profile Hellinger distance estimation. *Statistics*, *49*(4), 711–740.

Wu, J., Karunamuni, R. J., Zhang, B. (2010). Minimum Hellinger distance estimation in a two-sample semiparametric model. *Journal of Multivariate Analysis*, *101*, 1102–1122.

Zhang, C.-H. (2010). Nearly unbiased variable selection under mini-max concave penalty. *Annals of Statistics*, *38*, 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: *Series B*, *67*, 301–320.

Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*, 1509–1533.

Zou, H., Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*, 1733–1751.