Ann Inst Stat Math (2018) 70:131–154
https://doi.org/10.1007/s10463-016-0584-7

# Semiparametric mixtures of nonparametric regressions

**Sijia Xiang**[1] · **Weixin Yao**[2]

Received: 3 March 2016 / Revised: 23 September 2016 / Published online: 5 November 2016


**Abstract** In this article, we propose and study a new class of semiparametric mixture of regression models, where the mixing proportions and variances are constants, but the component regression functions are smooth functions of a covariate. A one-step backfitting estimate and two EM-type algorithms have been proposed to achieve the optimal convergence rate for both the global parameters and the nonparametric regression functions. We derive the asymptotic property of the proposed estimates and show that both the proposed EM-type algorithms preserve the asymptotic ascent property. A generalized likelihood ratio test is proposed for semiparametric inferences. We prove that the test follows an asymptotic $\chi^2$-distribution under the null hypothesis, which is independent of the nuisance parameters. A simulation study and two real data examples have been conducted to demonstrate the finite sample performance of the proposed model.

**Keywords** EM algorithm · Kernel regression · Mixture of regression models · Semiparametric mixture models

**Electronic supplementary material** The online version of this article (doi:10.1007/s10463-016-0584-7) contains supplementary material, which is available to authorized users.

✉ Sijia Xiang
sjxiang@zufe.edu.cn

[1] School of Mathematics and Statistics, Zhejiang University of Finance and Economics, Hangzhou 310018, Zhejiang, People's Republic of China

[2] Department of Statistics, University of California, Riverside, CA 92887, USA

## 1 Introduction

Finite mixture of regression models, also known as switching regression models in econometrics, have been widely applied in various fields, see, for example, in econometrics (Wedel and DeSarbo 1993; Frühwirth-Schnatter 2001), and in epidemiology (Green and Richardson 2002). Since Goldfeld and Quandt (1973) first introduced the mixture regression model, many efforts have been made to extend the traditional parametric mixture of linear regression models. For example, Young and Hunter (2010), and Huang and Yao (2012) studied models which allow the mixing proportions to depend on the covariates nonparametrically; Huang et al. (2013) proposed a fully nonparametric mixture of regression models by assuming the mixing proportions, the regression functions, and the variance functions to be nonparametric functions of a covariate; Cao and Yao (2012) suggested a semiparametric mixture of binomial regression models for binary data.

In this article, we propose a new semiparametric mixture of regression models, where the mixing proportions and variances are constants, but the component regression functions are nonparametric functions of a covariate. Compared to traditional finite mixture of linear regression models, the newly proposed model relaxes the parametric assumption on the regression functions, and allows the regression function in each component to be an unknown but smooth function of covariates. Compared to the fully nonparametric mixture of regression models proposed by Huang et al. (2013), our new model improves the efficiency of the estimates of the mean functions by assuming the mixing proportions and variances to be constants, which are also presumed by the traditional mixture of linear regressions. The new model is more challenging to estimate due to the existence of both global parameters and local parameters. The comparison of our paper to Huang et al. (2013) is similar to the comparison between semiparametric regression and fully nonparametric regression. Although the parametric parts of our model have stronger assumption than the nonparametric parts of Huang et al. (2013), they can provide more homogeneous model and more efficient estimate. Therefore, the proposed semiparametric model can combine the good properties of both parametric models and nonparametric models.

Our new model is motivated by a US house price index data, which is also used by Huang et al. (2013). The data set contains the monthly change of S&P/Case-Shiller House Price Index (HPI) and monthly growth rate of United States Gross Domestic Product (GDP) from January 1990 to December 2002, see Fig. 3a for a scatter plot. Based on the plot, it can be seen that there are two homogeneous groups and the relationship between HPI and GDP are different in different groups. In addition, it is clear that the relationship in each group is not linear. Therefore, the traditional mixture of linear regression models can not be applied. In Fig. 3b, we added the two fitted component regression curves based on our new model, and it is clear that the new model successfully recovered the two-component regression curves. In addition, the observations were classified into two groups corresponding to two different macroeconomic cycles, which possibly explains that the impact of GDP growth rate on HPI change may be different in different macroeconomic cycles.

We will show the identifiability of the proposed model under some regularity conditions. To estimate the unknown smoothing functions, we propose both a regression

spline based estimator and a local likelihood estimator using the kernel regression technique. To achieve the optimal convergence rate for both the global parameters and the nonparametric functions, we propose a one-step backfitting estimation procedure. The asymptotic properties of the one-step backfitting estimate are investigated. In addition, we propose two EM-type algorithms to compute the proposed estimates and prove their asymptotic ascent properties. A generalized likelihood ratio test is proposed for testing whether the mixing proportions and variances are indeed constants. We investigate the asymptotic behavior of the test and prove that its limiting null distribution follows a $\chi^2$-distribution independent of the nuisance parameters. A simulation study and two real data applications are used to demonstrate the effectiveness of the new model.

The rest of the paper is organized as follows. In Sect. 2, we introduce the new semiparametric mixture of regression models and the estimation procedure. In particular, we propose a regression spline estimate and a one-step backfitting estimate. A generalized likelihood ratio test is also introduced for some semiparametric inferences. In Sect. 3, we use a Monte Carlo study and two real data examples to demonstrate the finite sample performance of the proposed model and estimates. We conclude the paper with a brief discussion in Sect. 4 and defer the proofs to the Appendix.

## 2 Estimation procedure and asymptotic properties

### 2.1 The semiparametric mixture of regression models

Assume $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ are a random sample from the population $(X, Y)$. Let $Z$ be a latent variable with $P(Z = j) = \pi_j$ for $j = 1, \ldots, k$. Suppose $E(Y|X = x, Z = j) = m_j(x)$ and conditioning on $Z = j$ and $X = x$, $Y$ follows a normal distribution with mean $m_j(x)$ and variance $\sigma_j^2$. Then, without observing $Z$, the conditional distribution of $Y$ given $X = x$ can be written as

$$Y|_{X=x} \sim \sum_{j=1}^{k} \pi_j \phi(Y|m_j(x), \sigma_j^2), \tag{1}$$

where $\phi(y|\mu, \sigma^2)$ is the normal density with mean $\mu$ and variance $\sigma^2$. In this paper, we only considered the case when $X$ is univariate. The estimation methodology and theoretical results discussed can be readily extended to multivariate $X$, but due to the "curse of dimensionality", the extension is less applicable and thus omitted here. Throughout the paper, we assume that $k$ is fixed, and therefore, refer to (1) as a finite semiparametric mixture of regression models, since $m_j(x)$ is a nonparametric function of $x$, while $\pi_j$ and $\sigma_j$ are global parameters. If $m_j(x)$ is indeed linear in $x$, model (1) boils down to a regular finite mixture of linear regression models. When $k = 1$, then model (1) is a nonparametric regression model. Therefore, model (1) is a natural extension of the finite mixture of linear regression models and the nonparametric regression model.

Huang et al. (2013) studied a nonparametric mixture of regression models (NMR),

$$Y|_{X=x} \sim \sum_{j=1}^{k} \pi_j(x)\phi(Y|m_j(x), \sigma_j^2(x)), \tag{2}$$

where $\pi_j(\cdot)$, $m_j(\cdot)$, and $\sigma_j^2(\cdot)$ are unknown but smooth functions. Compared to model (2), model (1) improves the efficiency of the estimates of $\pi_j$, $\sigma_j$ and $m_j(x)$ by assuming the mixing proportions and variances to be constants, which are also presumed by the traditional mixture of linear regressions. We will demonstrate such improvement in Sect. 3. However, the new model (1) is more challenging to estimate than model (2) due to the existence of both global parameters and local parameters. In fact, we will demonstrate later that the model estimate of (2) is an intermediate result of the proposed one-step backfitting estimate. In this article, we will also develop a generalized likelihood ratio test to compare the proposed model with model (2) and illustrate its use in Sect. 3.

Identifiability is a critical issue in many mixture models. Some well known results of identifiability of finite mixture models include: mixture of univariate normals is identifiable (Titterington et al. 1985), and finite mixture of linear regression models is identifiable provided that covariates have a certain level of variability (Hennig 2000). Based on Theorem 1 in Huang et al. (2013) and Theorem 3.2 in Wang et al. (2014), we can get the following result on the identifiability of model (1).

**Proposition 1** *Assume that*

(1) $m_j(x)$ *are differentiable functions,* $j = 1, \ldots, k$.
(2) *One of the following conditions holds:*
   (a) *For any* $i \neq j$, $\sigma_i \neq \sigma_j$;
   (b) *If there exists* $i \neq j$ *such that* $\sigma_i = \sigma_j$, *then* $\|m_i(x) - m_j(x)\| + \|m_i'(x) - m_j'(x)\| \neq 0$ *for any* $x$.
(3) *The domain* $\mathcal{X}$ *of* $x$ *is an interval in* $\mathbb{R}$.

*Then, model* (1) *is identifiable.*

### 2.2 Estimation procedure and asymptotic properties

#### *2.2.1 Regression spline based estimator*

We first introduce a regression spline based estimator, which uses the regression spline (Hastie et al 2003; de Boor 2001) to transfer the semiparametric mixture model to a parametric mixture model. A cubic spline approximation for $m_j(x)$ can be expressed as

$$m_j(x) \approx \sum_{q=1}^{Q+4} \beta_{jq} B_q(x), \quad j = 1, \ldots, k, \tag{3}$$

where $B_1(x)$, ..., $B_{Q+4}(x)$ is a cubic spline basis and $Q$ is the number of internal knots. Many spline bases can be used here, such as a truncated power spline basis or a B-spline basis. In this paper, we mainly focus on the B-spline basis.

Based on the approximation (3), model (1) becomes

$$Y|_{X=x} \sim \sum_{j=1}^{k} \pi_j \phi \left( Y | \sum_{q=1}^{Q+4} \beta_{jq} B_q(x), \sigma_j^2 \right).$$

The log likelihood of the collected data $\{(X_i, Y_i), i = 1, .., n\}$ is

$$\ell(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j \phi(Y_i | \sum_{q=1}^{Q+4} \beta_{jq} B_q(X_i), \sigma_j^2) \right\},$$

where $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{k-1}\}^T$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}^T$, $\boldsymbol{\beta}_j = \left(\beta_{j1}, \ldots, \beta_{j,Q+4}\right)^T$, and $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \ldots, \sigma_k^2\}^T$. The parameters $(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ can be estimated by the traditional EM algorithm for mixtures of linear regression models.

The estimation method based on the regression spline approximation is easy to implement, and therefore, will be used as an initial value for our other estimation procedures.

### *2.2.2 One-step backfitting estimation procedure*

In this section, we propose a one-step backfitting estimation procedure to achieve the optimal convergence rates for both the global parameters and the nonparametric component regression functions.

Let $\ell^*(\boldsymbol{\pi}, \boldsymbol{m}(\cdot), \boldsymbol{\sigma}^2)$ be the log-likelihood of the collected data $\{(X_i, Y_i), i = 1, .., n\}$. That is,

$$\ell^*(\boldsymbol{\pi}, \boldsymbol{m}(\cdot), \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j \phi(Y_i | m_j(X_i), \sigma_j^2) \right\}, \tag{4}$$

where $\boldsymbol{\pi} = \{\pi_1, ..., \pi_{k-1}\}^T$, $\boldsymbol{m}(\cdot) = \{m_1(\cdot), ..., m_k(\cdot)\}^T$, and $\boldsymbol{\sigma}^2 = \{\sigma_1^2, ..., \sigma_k^2\}^T$. Since $\boldsymbol{m}(\cdot)$ consists of nonparametric functions, (4) is not ready for maximization. Next, we propose a one-step backfitting procedure. First, we estimate $\boldsymbol{\pi}$, $\boldsymbol{m}$ and $\boldsymbol{\sigma}^2$ locally by maximizing the following local log-likelihood function:

$$\ell_1(\boldsymbol{\pi}(x), \boldsymbol{m}(x), \boldsymbol{\sigma}^2(x)) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j \phi(Y_i | m_j, \sigma_j^2) \right\} K_h(X_i - x), \tag{5}$$

where $K_h(t) = h^{-1} K(t/h)$, $K(\cdot)$ is a kernel density function, and $h$ is a tuning parameter.

Let $\tilde{\boldsymbol{\pi}}(x)$, $\tilde{\boldsymbol{m}}(x)$, and $\tilde{\boldsymbol{\sigma}}^2(x)$ be the maximizer of (5), which are in fact the model estimates of (2) proposed by Huang et al. (2013). Note that, in (5), the global parameters

$\boldsymbol{\pi}$ and $\boldsymbol{\sigma}^2$ are estimated locally. To improve the efficiency, we propose to update the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}^2$ by maximizing the following log-likelihood function:

$$\ell_2(\boldsymbol{\pi}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \pi_j \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^2) \right\}, \tag{6}$$

which, compared to (4), replaces $m_j(\cdot)$ by $\tilde{m}_j(\cdot)$.

Denote by $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$ the solution of maximizing (6). We can then further improve the estimate of $\boldsymbol{m}(\cdot)$ by maximizing the following local log-likelihood function:

$$\ell_3(\boldsymbol{m}(x)) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \hat{\pi}_j \phi(Y_i | m_j, \hat{\sigma}_j^2) \right\} K_h(X_i - x). \tag{7}$$

which, compared to (5), replaces $\pi_j$ and $\sigma_j^2$ by $\hat{\pi}_j$ and $\hat{\sigma}_j^2$, respectively.

Let $\hat{\boldsymbol{m}}(x)$ be the solution of (7), and we refer to $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{m}}(x)$, and $\hat{\boldsymbol{\sigma}}^2$ as the one-step backfitting estimates. In Sect. 2.2.4, we show that the one-step backfitting estimates achieve the optimal convergence rate for both the global parameters, and the nonparametric mean functions. In (7), since $\hat{\pi}_j$ and $\hat{\sigma}_j^2$ have root $n$ convergence rate, unlike $\tilde{\boldsymbol{m}}(x)$, $\hat{\boldsymbol{m}}(x)$ does not need to adjust the uncertainty of estimating $\pi_j$ and $\sigma_j^2$. Therefore, $\hat{\boldsymbol{m}}(x)$ can have better estimation accuracy than $\tilde{\boldsymbol{m}}(x)$ proposed by Huang et al. (2013).

### *2.2.3 Computing algorithms*

In this section, we propose a local EM-type algorithm (LEM) and a global EM-type algorithm (GEM) to perform the one-step backfitting.

**Local EM-type algorithm (LEM)**
In practice, we usually want to evaluate unknown functions at a set of grid points, which in this case, requires us to maximize local log-likelihood functions at a set of grid points. If we simply employ an EM algorithm separately for different grid points, the labels in the found estimators may change at different grid points, and we may not be able to get smoothed estimated curves (Huang and Yao 2012). Next, we propose a modified EM-type algorithm, which estimates the nonparametric functions simultaneously at a set of grid points. Let $\{u_t, t = 1, \ldots, N\}$ be a set of grid points where some unknown functions are evaluated, and $N$ be the number of grid points.

**Step 1: modified EM-type algorithm to maximize $\ell_1$ in (5)**
In Step 1, we use the modified EM-type algorithm of Huang et al. (2013) to maximize $\ell_1$ and obtain the estimates $\tilde{\boldsymbol{\pi}}(\cdot)$, $\tilde{\boldsymbol{m}}(\cdot)$, and $\tilde{\boldsymbol{\sigma}}^2(\cdot)$. Specifically, at the $(l+1)$th iteration,
**E-step** Calculate the expectations of component labels based on estimates from the $l$th iteration:

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)}(X_i)\phi(Y_i | m_j^{(l)}(X_i), \sigma_j^{2(l)}(X_i))}{\sum_{j=1}^{k} \pi_j^{(l)}(X_i)\phi(Y_i | m_j^{(l)}(X_i), \sigma_j^{2(l)}(X_i))}, \quad i = 1, \ldots, n, \; j = 1, \ldots, k.$$

**M-step** Update the estimates

$$\pi_j^{(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}, \tag{8}$$

$$m_j^{(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}, \tag{9}$$

$$\sigma_j^{2(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - m_j^{(l+1)}(x))^2 K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}, \tag{10}$$

for $x \in \{u_t, t = 1, \ldots, N\}$. We then update $\pi_j^{(l+1)}(X_i)$, $m_j^{(l+1)}(X_i)$, and $\sigma_j^{2(l+1)}(X_i)$, $i = 1, \ldots, n$, by linear interpolating $\pi_j^{(l+1)}(u_t)$, $m_j^{(l+1)}(u_t)$, and $\sigma_j^{2(l+1)}(u_t)$, $t = 1, \ldots, N$, respectively.

Note that in the M-step, the nonparametric functions are estimated simultaneously at a set of grid points, and therefore, the classification probabilities in the the E-step can be estimated globally to avoid the label switching problem (Celeux et al. 2000; Stephens 2000; Yao 2012, 2015; Yao and Lindsay 2009).

**Step 2: EM algorithm to maximize $\ell_2$ in (6)**
In Step 2, given $\tilde{m}_j(x)$ from Step 1, a regular EM algorithm can be used to maximize $\ell_2$ and update the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}^2$ as $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$. At the $(l+1)$th iteration,
**E-step** Calculate the expectations of component labels based on the estimates from the $l$th iteration:

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)} \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^{2(l)})}{\sum_{j=1}^k \pi_j^{(l)} \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^{2(l)})}, i = 1, \ldots, n, j = 1, \ldots, k.$$

**M-step** Update the estimates

$$\pi_j^{(l+1)} = \frac{\sum_{i=1}^n p_{ij}^{(l+1)}}{n},$$

$$\sigma_j^{2(l+1)} = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - \tilde{m}_j(X_i))^2}{\sum_{i=1}^n p_{ij}^{(l+1)}}.$$

The ascent property of the above algorithm follows from the theory of the ordinary EM algorithm.

**Step 3: Modified EM-type algorithm to maximize $\ell_3$ in (7)**
In Step 3, given $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$ from Step 2, we would then maximize $\ell_3$ to find the estimates $\hat{\boldsymbol{m}}(x)$. At the $(l+1)$th iteration,

**E-step** Calculate the expectations of component labels based on estimates from the $l$th iteration:

$$p_{ij}^{(l+1)} = \frac{\hat{\pi}_j \phi(Y_i | m_j^{(l)}(X_i), \hat{\sigma}_j^2)}{\sum_{j=1}^k \hat{\pi}_j \phi(Y_i | m_j^{(l)}(X_i), \hat{\sigma}_j^2)}, \quad i = 1, \ldots, n, \, j = 1, \ldots, k. \tag{11}$$

**M-step** Update the estimate

$$m_j^{(l+1)}(x) = \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)},$$

for $x \in \{u_t, t = 1, ..., N\}$. Similar to Step 1, we update the estimates at a set of grid points first, and then update $m_j^{(l+1)}(X_i)$, $i = 1, \ldots, n$, by linear interpolating $m_j^{(l+1)}(u_t)$, $t = 1, \ldots, N$.

**Global EM-type algorithm (GEM)**
To improve the estimation efficiency, one might further iterate Step 1 to Step 3 until convergence. Next, we propose a global EM-type algorithm (GEM) to approximate such iteration, but with much less computation. At the $(l + 1)$th iteration,
**E-step** Calculate the expectations of component labels based on estimates from the $l$th iteration:

$$p_{ij}^{(l+1)} = \frac{\pi_j^{(l)} \phi\left(Y_i | m_j^{(l)}(X_i), \sigma_j^{2(l)}\right)}{\sum_{j=1}^k \pi_j^{(l)} \phi\left(Y_i | m_j^{(l)}(X_i), \sigma_j^{2(l)}\right)}, i = 1, \ldots, n, \, j = 1, \ldots, k.$$

**M-step** Simultaneously update the estimates

$$\begin{aligned}
\pi_j^{(l+1)} &= \frac{\sum_{i=1}^n p_{ij}^{(l+1)}}{n}, \\
m_j^{(l+1)}(x) &= \frac{\sum_{i=1}^n p_{ij}^{(l+1)} Y_i K_h(X_i - x)}{\sum_{i=1}^n p_{ij}^{(l+1)} K_h(X_i - x)}, \\
\sigma_j^{2(l+1)} &= \frac{\sum_{i=1}^n p_{ij}^{(l+1)} (Y_i - m_j^{(l+1)}(X_i))^2}{\sum_{i=1}^n p_{ij}^{(l+1)}},
\end{aligned}$$

for $x \in \{u_t, j = 1, \ldots, N\}$. We then update $m_j^{(l+1)}(X_i)$, $i = 1, ..., n$ by linear interpolating $m_j^{(l+1)}(u_t)$, $t = 1, \ldots, N$.

*2.2.4 Asymptotic properties*

Next, we investigate the asymptotic properties of the proposed one-step backfitting estimates and the asymptotic ascent properties of the two proposed EM-type algorithms.

Let $\boldsymbol{\theta} = (\boldsymbol{m}^T, \boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\pi}^T, (\boldsymbol{\sigma}^2)^T)^T$, then $\boldsymbol{\theta} = (\boldsymbol{m}^T, \boldsymbol{\beta}^T)^T$. Define

$$\ell(\boldsymbol{\theta}, y) = \log \sum_{j=1}^{k} \pi_j \phi(y|m_j, \sigma_j^2), \tag{12}$$

and let

$$I_\theta(x) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}|X = x\right], \quad I_\beta(x) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|X = x\right],$$
$$I_m(x) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{m} \partial \boldsymbol{m}^T}|X = x\right],$$
$$I_{\beta m}(x) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T}|X = x\right], \Lambda(u|x) = E\left[\frac{\partial \ell(\boldsymbol{\theta}(x), y)}{\partial \boldsymbol{m}}|X = u\right].$$

Define

$$\kappa_l = \int t^l K(t) dt, \quad \nu_l = \int t^l K^2(t) dt.$$

Under further conditions defined in the Appendix, the consistency and asymptotic normality of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}^2$ are established in the next theorem.

**Theorem 1** *Suppose that conditions (C1) and (C3)|(C10) in the* Appendix *are satisfied, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(0, B^{-1} \Sigma B^{-1}),$$

*where* $B = E\{I_\beta(X)\}$, $\Sigma = Var\{\partial \ell(\boldsymbol{\theta}(X), Y)/\partial \boldsymbol{\beta} - \varpi(X, Y)\}$, $\varpi(x, y) = I_{\beta m} \varphi(x, y)$, *and* $\varphi(x, y)$ *is a* $k \times 1$ *vector consisting of the first* $k$ *elements of* $I_\theta^{-1}(x) \partial \ell(\boldsymbol{\theta}(x), y)/\partial \boldsymbol{\theta}$.

Based on the above theorem, we can see that the proposed one-step backfitting estimator of the global parameters have achieved the optimal square root $n$ convergence rate.

The next theorem gives the asymptotic property of $\hat{\boldsymbol{m}}(\cdot)$.

**Theorem 2** *Suppose that conditions (C2)|(C10) in the* Appendix *are satisfied, then*

$$\sqrt{nh}(\hat{\boldsymbol{m}}(x) - \boldsymbol{m}(x) - \Delta_m(x) + o_p(h^2)) \xrightarrow{D} N(0, f^{-1}(x) I_m^{-1}(x) \nu_0),$$

*where* $f(\cdot)$ *is the density of* $X$, $\Delta_m(x)$ *is a* $k \times 1$ *vector consisting of the first* $k$ *elements of* $\Delta(x)$ *with*

$$\Delta(x) = I_m^{-1}(x) \left\{\frac{1}{2} \Lambda''(x|x) + f^{-1}(x) f'(x) \Lambda'(x|x)\right\} \kappa_2 h^2.$$

Based on the above theorem, we can see that $\hat{\boldsymbol{m}}(x)$ has the same asymptotic properties as if $\boldsymbol{\beta}$ were known, since $\hat{\boldsymbol{\beta}}$ has faster convergence rate than $\hat{\boldsymbol{m}}(x)$.

The asymptotic ascent properties of the proposed EM-type algorithms are provided in the following theorem.

**Theorem 3** (i) *For the modified EM-type algorithm* (Step 1) *to maximize* $\ell_1$, *given condition* (C2),

$$\liminf_{n\to\infty} n^{-1}\left[\ell_1(\boldsymbol{\theta}^{(l+1)}(x)) - \ell_1(\boldsymbol{\theta}^{(l)}(x))\right] \geq 0$$

*in probability, for any given point* $x \in \mathcal{X}$, *where* $\ell_1(\cdot)$ *is defined in* (5).

(ii) *For the modified EM-type algorithm* (Step 3) *to maximize* $\ell_3$, *given condition* (C2),

$$\liminf_{n\to\infty} n^{-1}\left[\ell_3(\boldsymbol{m}^{(l+1)}(x)) - \ell_3(\boldsymbol{m}^{(l)}(x))\right] \geq 0$$

*in probability, for any given point* $x \in \mathcal{X}$, *where* $\ell_3(\cdot)$ *is defined in* (7).

(iii) *For the GEM algorithm, we have*

$$\liminf_{n\to\infty} n^{-1}\left[\ell^*(\boldsymbol{m}^{(l+1)}(\cdot), \boldsymbol{\pi}^{(l+1)}, \boldsymbol{\sigma}^{2(l+1)}) - \ell^*(\boldsymbol{m}^{(l)}(\cdot), \boldsymbol{\pi}^{(l)}, \boldsymbol{\sigma}^{2(l)})\right] \geq 0$$

*in probability, for any given point* $x \in \mathcal{X}$, *where* $\ell^*(\cdot)$ *is defined in* (4).

### 2.3 Hypothesis testing

Huang et al. (2013) proposed a nonparametric mixture of regression models where mixing proportions, means, and variances are all unknown but smooth functions of a covariate. Compared to Huang et al. (2013), our model can be more efficient by assuming the mixing proportions and variances to be constants. Then, a natural question to ask is whether or not the mixing proportions and variances indeed depend on the covariate. This amounts to testing the following hypothesis:

$$\begin{aligned} H_0 :& \pi_j(x) \equiv \pi_j,\ j = 1, \ldots, k-1; \\ & \sigma_j^2(x) \equiv \sigma_j^2,\ j = 1, \ldots, k; \\ & \pi_j \text{ and } \sigma_j^2 \text{ are unknown in } (0,1) \text{ and } \mathbb{R}^+. \\ H_1 :& \pi_j(x) \text{ or } \sigma_j^2(x) \text{ is not constant for some } j. \end{aligned}$$

Next, we propose to use the idea of the generalized likelihood ratio test (Fan et al. 2001) to compare model (1) with model (2).

Let $\ell_n(H_0)$ and $\ell_n(H_1)$ be the log-likelihood functions computed under the null and alternative hypothesis, respectively. Then, we can construct a likelihood ratio test statistic

$$T = \ell_n(H_1) - \ell_n(H_0). \tag{13}$$

Note that this likelihood ratio statistic is different from the parametric likelihood ratio statistics, since the null and alternative are both semiparametric models, and the number of parameters under $H_0$ or $H_1$ are undefined. The following theorem establishes the Wilks types of results for (13), that is, the asymptotic null distribution is independent of the nuisance parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$, and the nuisance nonparametric mean functions $\boldsymbol{m}(x)$.

**Theorem 4** *Suppose that conditions* (C9)–(C13) *in the* Appendix *hold and that* $nh^4 \to 0$ *and* $nh^2 \log(1/h) \to \infty$, *then*

$$r_K T \overset{a}{\sim} \chi^2_\delta,$$

*where* $r_K = [K(0) - 0.5\int K^2(t)dt]/\int[K(t) - 0.5K * K(t)]^2 dt$, $\delta = r_K(2k-1)|\mathscr{X}|[K(0) - 0.5\int K^2(t)dt]/h$, $|\mathscr{X}|$ *denotes the length of the support of* $X$, *and* $K * K$ *is the 2nd convolution of* $K(\cdot)$.

Theorem 4 unveils a new Wilks type of phenomenon, and provides a simple and useful method for semiparametric inferences. We will demonstrate its application in Sect. 3.

## 3 Examples

### 3.1 Simulation study

In this section, we use a simulation study to investigate the finite sample performance of the proposed regression spline estimate (Spline), the one-step backfitting estimate using local EM-type algorithm (LEM), and the global EM-type algorithm (GEM), and compare them with the traditional mixture of linear regressionss estimate (MLR), and the nonparametric mixture of regression models (NMR, Huang et al. 2013). For the regression spline, we use $Q = 5$, where $Q$ is the number of internal knots. For LEM, GEM and NMR, we use both the true value and the regression spline estimate as initial values, denoted by $(T)$ and $(S)$, respectively.

We conduct a simulation study for a two-component semiparametric mixture of regression models:

$$\begin{aligned}
&\pi_1 = 0.5 \text{ or } \pi_1 = 0.7,\\
&m_1(x) = 4 - \sin(2\pi x) \text{ and } m_2(x) = 1.5 + \cos(3\pi x),\\
&\sigma_1^2 = 0.09 \text{ and } \sigma_2^2 = 0.16.
\end{aligned}$$

The covariate $X$ is generated from the one-dimensional uniform distribution in [0, 1], and the Gaussian kernel is used in the simulation. The sample sizes $n = 200$ and $n = 400$ are conducted over 500 repetitions.

The performance of the estimates of the mean functions $\boldsymbol{m}(x)$ is measured by the square root of the average squared errors (RASE),

$$\text{RASE}_m^2 = N^{-1}\sum_{j=1}^{2}\sum_{t=1}^{N}\left[\hat{m}_j(u_t) - m_j(u_t)\right]^2,$$

where $\{u_t, t = 1, \ldots, N\}$ are a set of grid points at which the unknown functions are evaluated. In our simulation, we set $N = 100$. To compare between model (1) and the nonparametric mixture of regression models proposed by Huang et al. (2013), we also report the RASE of $\pi$ and $\sigma^2$, denoted by $\text{RASE}_\pi$ and $\text{RASE}_{\sigma^2}$, respectively.

Bandwidth plays an important role in the estimation of $\boldsymbol{m}(\cdot)$. There are ways to calculate the theoretical optimal bandwidth, but in practice, data driven methods, such as cross-validation (CV), are popularly used. Please see Zhang and Yang (2015) and the reference therein for the application and properties of cross-validation. Let $\mathscr{D}$ be the full data set, and divide $\mathscr{D}$ into a training set $\mathscr{R}_l$ and a test set $\mathscr{T}_l$. That is, $\mathscr{R}_l \cup \mathscr{T}_l = \mathscr{D}$ for $l = 1, ..., L$. We use the training set $\mathscr{R}_l$ to obtain the estimates $\{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{m}}(\cdot), \hat{\boldsymbol{\sigma}}^2\}$, then consider a likelihood version CV, which is defined by

$$\mathrm{CV}(h) = \sum_{l=1}^{L} \sum_{t \in \mathscr{T}_l} \log \left\{ \sum_{j=1}^{k} \hat{\pi}_j \phi(y_t | \hat{m}_j(x_t), \hat{\sigma}_j^2) \right\}.$$

In the simulation, we set $L = 10$ and randomly partition the data. We repeat the procedure 30 times, and take the average of the selected bandwidths as the optimal bandwidth, denoted by $\hat{h}$. In the simulation, we consider three different bandwidths, $\hat{h} \times n^{-2/15}$, $\hat{h}$, and $1.5\hat{h}$, which correspond to under-smoothing (US), appropriate smoothing (AS), and over-smoothing (OS), respectively.

Tables 1 and 2 report the average of $\mathrm{RASE}_\pi$, $\mathrm{RASE}_m$, and $\mathrm{RASE}_{\sigma^2}$, for $\pi_1 = 0.5$ and $\pi_1 = 0.7$, respectively. All the values are multiplied by 100. From Tables 1 and 2, we can see that LEM, GEM, and the regression spline estimates give better results than the mixture of linear regressions estimate. Compared to NMR, model (1) improves the efficiency of the estimation of mixing proportions and variances, and provides slightly better estimates for the mean functions. In addition, both LEM and GEM provide better results for the mean functions than the regression spline estimate when the sample size is small. We further notice that LEM(S) and GEM(S) provide similar results to LEM(T) and GEM(T). Therefore, the spline estimate provides good initial values for other estimates.

From Tables 1 and 2, LEM and GEM have similar performance in terms of model fitting. However, in terms of computation time, GEM has an absolute advantage over LEM. For example, on a personal laptop with an i7-3610QM CPU and 8GB of RAM, the average calculation time (in s) for each repetition when $n = 200$ is 0.072 and 0.017 for LEM and GEM, respectively, and 0.105 and 0.028 when $n = 400$.

Next, we test the accuracy of the standard error estimation and the confidence interval construction for $\pi_1$, $\sigma_1$ and $\sigma_2$ via a conditional bootstrap procedure. Given the covariate $X = x$, the response $Y^*$ can be generated from the estimated distribution $\sum_{j=1}^{k} \hat{\pi}_j \phi(Y | \hat{m}_j(x), \hat{\sigma}_j^2)$. For the simplicity of presentation, we only report the results for GEM(T). We apply the proposed estimation procedure to each of the 200 bootstrap samples, and further obtain the confidence intervals.

Table 3 reports the results from the bootstrap procedure. SD contains the standard deviation of 500 replicates, and can be considered as true standard errors. SE and STD contain the mean and standard deviation of the 500 estimated standard errors based on the conditional bootstrap procedure. In addition, the coverage probability of the 95% confidence intervals based on the estimated standard errors are also reported. From Table 3 we can see that the bootstrap procedure estimates the true standard error quite well, since all the differences between the true value and the estimates are less than two standard errors of the estimates. The coverage probabilities are satisfactory for $\pi_1$, but a bit low for $\sigma_1$ and $\sigma_2$, especially for over-smoothing bandwidth.

**Table 1** The average of $\text{RASE}_{\pi}$, $\text{RASE}_{\sigma^2}$ & $\text{RASE}_{m}$ when $\pi_1 = 0.5$ (true values times 100)

| $n$ | h | | LEM(T) | GEM(T) | LEM(S) | GEM(S) | Spline | MLR | NMR(T) | NMR(S) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{RASE}_{\pi}$ | 2.82 | 2.84 | 2.83 | 2.84 | 2.85 | 4.40 | 13.38 | 13.37 |
| | US | $\text{RASE}_{\sigma^2}$ | 4.34 | 4.39 | 4.35 | 4.40 | 2.72 | 65.62 | 9.88 | 9.94 |
| | | $\text{RASE}_{m}$ | 20.81 | 20.84 | 20.98 | 21.02 | 39.98 | 87.32 | 20.48 | 21.35 |
| | | $\text{RASE}_{\pi}$ | 2.83 | 2.81 | 2.84 | 2.81 | 2.83 | 4.37 | 9.55 | 9.53 |
| 200 | AS | $\text{RASE}_{\sigma^2}$ | 2.69 | 2.73 | 2.70 | 2.73 | 2.78 | 63.29 | 12.62 | 12.66 |
| | | $\text{RASE}_{m}$ | 17.72 | 17.67 | 17.73 | 17.67 | 45.60 | 87.13 | 18.77 | 19.52 |
| | | $\text{RASE}_{\pi}$ | 2.79 | 2.69 | 2.78 | 2.69 | 2.76 | 4.57 | 8.39 | 8.38 |
| | OS | $\text{RASE}_{\sigma^2}$ | 2.73 | 2.42 | 2.73 | 2.42 | 2.74 | 64.52 | 20.77 | 20.81 |
| | | $\text{RASE}_{m}$ | 23.12 | 22.99 | 23.14 | 22.99 | 32.33 | 87.48 | 25.30 | 25.39 |
| | | $\text{RASE}_{\pi}$ | 2.02 | 2.00 | 2.03 | 2.00 | 1.98 | 3.39 | 10.56 | 10.54 |
| | US | $\text{RASE}_{\sigma^2}$ | 2.88 | 2.91 | 2.89 | 2.91 | 1.80 | 66.15 | 7.99 | 7.98 |
| | | $\text{RASE}_{m}$ | 15.76 | 15.78 | 15.77 | 15.78 | 15.10 | 85.88 | 15.82 | 15.85 |
| | | $\text{RASE}_{\pi}$ | 2.03 | 2.02 | 2.04 | 2.02 | 2.03 | 3.41 | 7.35 | 7.35 |
| 400 | AS | $\text{RASE}_{\sigma^2}$ | 1.87 | 1.88 | 1.87 | 1.88 | 1.77 | 65.54 | 9.87 | 9.89 |
| | | $\text{RASE}_{m}$ | 13.20 | 13.19 | 13.20 | 13.19 | 17.65 | 85.77 | 14.11 | 14.15 |
| | | $\text{RASE}_{\pi}$ | 2.19 | 2.15 | 2.20 | 2.15 | 2.14 | 3.38 | 6.54 | 6.54 |
| | OS | $\text{RASE}_{\sigma^2}$ | 1.92 | 1.76 | 1.92 | 1.76 | 1.85 | 65.46 | 16.21 | 16.22 |
| | | $\text{RASE}_{m}$ | 16.86 | 16.78 | 16.86 | 16.78 | 15.85 | 85.73 | 18.56 | 18.56 |

We also apply the bootstrap procedure to investigate the point-wise coverage probability of the mean functions, at a set of evenly distributed grid points. Table 4 shows the results of the 95% confidence interval for the two-component mean functions. From the table, we can see that the mean function of the first component tends to have higher coverage probability than the second component, especially for over-smoothing bandwidth. In addition, the coverage probability is generally lower than the nominal level for over-smoothing bandwidth.

Next, we assess the performance of the testing procedure proposed in Sect. 2.3. Under the null hypothesis, the mixing proportion $\pi_1$ and variances $\sigma_1^2$ and $\sigma_2^2$ are constants. We compute the distribution of $T$ with $n = 200$ and $n = 400$ via 500 repetitions, and compare it with the $\chi^2$-approximation. The histogram of the null distribution is shown in Fig. 1, where the solid line corresponds to a density of the $\chi^2$-distribution with degrees of freedom $\delta$ defined in Theorem 4. Figure 2 shows the Q-Q plot for the two cases. From Figs. 1 and 2, the finite sample null distribution is quite close to a $\chi^2$-distribution with degrees of freedom $\delta$, especially for the case of $n = 400$.

### 3.2 Real data applications

*Example 1* (The US house price index data) In this section, we illustrate the proposed methodologies with an empirical analysis of US house price index data (sample size

**Table 2** The average of $\text{RASE}_{\pi}$, $\text{RASE}_{\sigma^2}$ & $\text{RASE}_{m}$ when $\pi_1 = 0.7$ (true values times 100)

| $n$ | h | | LEM(T) | GEM(T) | LEM(S) | GEM(S) | Spline | MLR | NMR(T) | NMR(S) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{RASE}_{\pi}$ | 2.66 | 2.68 | 2.66 | 2.68 | 2.66 | 4.07 | 11.54 | 11.53 |
| | US | $\text{RASE}_{\sigma^2}$ | 5.45 | 5.58 | 5.48 | 5.58 | 3.50 | 62.56 | 11.25 | 11.33 |
| | | $\text{RASE}_{m}$ | 23.57 | 23.63 | 23.75 | 24.43 | 48.12 | 90.04 | 23.09 | 23.49 |
| | | $\text{RASE}_{\pi}$ | 2.56 | 2.54 | 2.55 | 2.54 | 2.58 | 4.21 | 8.35 | 8.36 |
| 200 | AS | $\text{RASE}_{\sigma^2}$ | 3.27 | 3.35 | 3.29 | 3.35 | 3.84 | 64.35 | 14.40 | 14.53 |
| | | $\text{RASE}_{m}$ | 20.10 | 20.09 | 20.11 | 20.09 | 47.52 | 90.16 | 21.38 | 21.41 |
| | | $\text{RASE}_{\pi}$ | 2.74 | 2.64 | 2.88 | 2.77 | 2.73 | 4.18 | 7.30 | 7.42 |
| | OS | $\text{RASE}_{\sigma^2}$ | 3.10 | 2.81 | 3.13 | 2.83 | 3.60 | 64.13 | 22.09 | 22.19 |
| | | $\text{RASE}_{m}$ | 26.18 | 25.99 | 27.02 | 26.80 | 48.17 | 90.15 | 28.82 | 29.74 |
| | | $\text{RASE}_{\pi}$ | 1.79 | 1.80 | 1.80 | 1.80 | 1.78 | 3.16 | 9.24 | 9.24 |
| | US | $\text{RASE}_{\sigma^2}$ | 3.74 | 3.81 | 3.74 | 3.81 | 2.16 | 66.98 | 9.23 | 9.24 |
| | | $\text{RASE}_{m}$ | 18.00 | 18.03 | 18.00 | 18.03 | 18.91 | 87.49 | 17.93 | 17.99 |
| | | $\text{RASE}_{\pi}$ | 1.87 | 1.86 | 1.87 | 1.86 | 1.89 | 3.20 | 6.45 | 6.45 |
| 400 | AS | $\text{RASE}_{\sigma^2}$ | 2.26 | 2.27 | 2.26 | 2.27 | 2.14 | 65.31 | 11.29 | 11.32 |
| | | $\text{RASE}_{m}$ | 14.92 | 14.90 | 14.92 | 14.90 | 19.67 | 87.57 | 16.02 | 16.00 |
| | | $\text{RASE}_{\pi}$ | 1.94 | 1.89 | 1.94 | 1.89 | 1.87 | 2.95 | 5.59 | 5.59 |
| | OS | $\text{RASE}_{\sigma^2}$ | 2.27 | 2.09 | 2.27 | 2.09 | 2.21 | 65.44 | 18.02 | 18.03 |
| | | $\text{RASE}_{m}$ | 19.48 | 19.41 | 19.48 | 19.41 | 19.79 | 87.12 | 21.59 | 21.63 |

**Table 3** Standard errors and coverage probabilities

| | h | $\pi_1$ | | | $\sigma_1$ | | | $\sigma_2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SD | SE(STD) | 95% | SD | SE(STD) | 95% | SD | SE(STD) | 95% |
| $n = 200$ | US | 0.037 | 0.036 (0.002) | 94.11 | 0.014 | 0.013 (0.003) | 88.82 | 0.024 | 0.023 (0.004) | 91.09 |
| (0.5, 0.5) | AS | 0.037 | 0.036 (0.002) | 93.40 | 0.014 | 0.013 (0.002) | 94.00 | 0.029 | 0.022 (0.004) | 91.20 |
| | OS | 0.038 | 0.035 (0.002) | 90.60 | 0.014 | 0.015 (0.002) | 96.20 | 0.022 | 0.025 (0.004) | 97.20 |
| $n = 400$ | US | 0.027 | 0.025 (0.001) | 94.40 | 0.010 | 0.009 (0.001) | 94.80 | 0.018 | 0.017 (0.003) | 96.20 |
| (0.5, 0.5) | AS | 0.026 | 0.025 (0.001) | 93.80 | 0.009 | 0.009 (0.001) | 94.00 | 0.016 | 0.016 (0.002) | 96.40 |
| | OS | 0.026 | 0.025 (0.001) | 93.20 | 0.009 | 0.010 (0.001) | 93.80 | 0.016 | 0.018 (0.002) | 94.80 |
| $n = 200$ | US | 0.031 | 0.032 (0.002) | 94.80 | 0.011 | 0.011 (0.002) | 90.20 | 0.035 | 0.029 (0.009) | 83.20 |
| (0.7, 0.3) | AS | 0.033 | 0.032 (0.002) | 94.60 | 0.011 | 0.011 (0.001) | 96.40 | 0.028 | 0.027 (0.006) | 85.60 |
| | OS | 0.033 | 0.032 (0.002) | 93.20 | 0.013 | 0.013 (0.002) | 89.60 | 0.032 | 0.033 (0.008) | 97.00 |
| $n = 400$ | US | 0.023 | 0.023 (0.001) | 94.80 | 0.008 | 0.008 (0.001) | 94.20 | 0.023 | 0.023 (0.004) | 92.20 |
| (0.7, 0.3) | AS | 0.025 | 0.023 (0.001) | 93.40 | 0.008 | 0.008 (0.001) | 95.00 | 0.021 | 0.021 (0.003) | 93.40 |
| | OS | 0.023 | 0.023 (0.001) | 94.60 | 0.009 | 0.009 (0.001) | 83.20 | 0.021 | 0.023 (0.004) | 96.20 |

$n = 141$) that are introduced in Sect. 1. GDP is a well known measure of the size of a nation's economy, as it recognizes the total goods and services produced within a nation in a given period, and HPI is known as a measure of a nation's average housing

**Table 4** Pointwise coverage probabilities

| | h | | 0.111 | 0.222 | 0.333 | 0.444 | 0.556 | 0.667 | 0.778 | 0.889 |
|---|---|---|---|---|---|---|---|---|---|---|
| | US | $m_1$ | 92.00 | 93.00 | 93.80 | 92.40 | 93.40 | 93.40 | 94.20 | 92.80 |
| $n = 200$ | | $m_2$ | 90.00 | 92.20 | 95.20 | 93.40 | 94.20 | 94.40 | 92.80 | 93.20 |
| | AS | $m_1$ | 92.20 | 91.40 | 90.80 | 87.60 | 90.00 | 91.40 | 93.00 | 90.40 |
| (0.5, 0.5) | | $m_2$ | 85.40 | 89.40 | 85.00 | 89.00 | 87.00 | 84.20 | 89.40 | 89.40 |
| | OS | $m_1$ | 92.00 | 77.00 | 80.80 | 83.00 | 87.00 | 80.80 | 79.60 | 89.80 |
| | | $m_2$ | 58.60 | 80.20 | 53.60 | 76.60 | 73.60 | 48.80 | 80.80 | 73.00 |
| | US | $m_1$ | 93.40 | 94.40 | 95.60 | 93.40 | 93.00 | 95.80 | 96.00 | 94.00 |
| $n = 400$ | | $m_2$ | 97.20 | 94.40 | 94.20 | 91.60 | 93.40 | 94.60 | 95.00 | 94.80 |
| | AS | $m_1$ | 91.40 | 93.00 | 93.60 | 91.80 | 90.60 | 92.40 | 92.00 | 91.60 |
| (0.5, 0.5) | | $m_2$ | 89.80 | 91.80 | 87.40 | 90.00 | 88.40 | 88.80 | 89.40 | 90.40 |
| | OS | $m_1$ | 88.80 | 76.60 | 81.60 | 89.00 | 86.00 | 80.80 | 79.60 | 88.80 |
| | | $m_2$ | 61.80 | 82.20 | 51.40 | 78.60 | 79.80 | 48.60 | 80.00 | 73.80 |
| | US | $m_1$ | 91.40 | 97.00 | 93.40 | 93.60 | 93.00 | 94.80 | 94.60 | 93.40 |
| $n = 200$ | | $m_2$ | 89.00 | 93.20 | 92.20 | 91.00 | 92.80 | 92.40 | 93.40 | 90.80 |
| | AS | $m_1$ | 92.40 | 88.60 | 91.40 | 90.20 | 86.40 | 89.60 | 89.60 | 89.20 |
| (0.7, 0.3) | | $m_2$ | 82.60 | 89.00 | 89.40 | 86.20 | 84.20 | 84.20 | 87.20 | 86.40 |
| | OS | $m_1$ | 91.40 | 62.20 | 67.20 | 82.80 | 82.00 | 67.00 | 62.80 | 90.00 |
| | | $m_2$ | 60.60 | 83.80 | 63.80 | 81.60 | 76.00 | 57.20 | 78.20 | 76.00 |
| | US | $m_1$ | 92.40 | 94.20 | 93.60 | 94.60 | 93.40 | 96.80 | 94.00 | 95.40 |
| $n = 400$ | | $m_2$ | 93.40 | 95.60 | 93.00 | 94.00 | 93.60 | 93.60 | 93.80 | 93.00 |
| | AS | $m_1$ | 91.80 | 90.80 | 89.40 | 91.20 | 91.80 | 92.20 | 88.80 | 92.20 |
| (0.7, 0.3) | | $m_2$ | 83.60 | 89.00 | 87.20 | 89.20 | 88.60 | 85.80 | 88.20 | 88.60 |
| | OS | $m_1$ | 90.40 | 60.80 | 67.00 | 87.20 | 86.60 | 68.20 | 62.00 | 87.20 |
| | | $m_2$ | 56.40 | 81.00 | 60.40 | 78.60 | 79.60 | 56.80 | 81.00 | 74.00 |

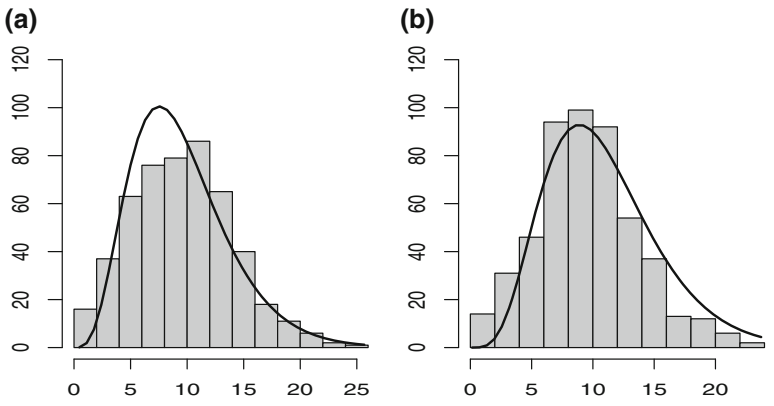


**Fig. 1** Histogram of $T_n$ and $\chi^2$-approximation of $T_n$: **a** $n = 200$, **b** $n = 400$

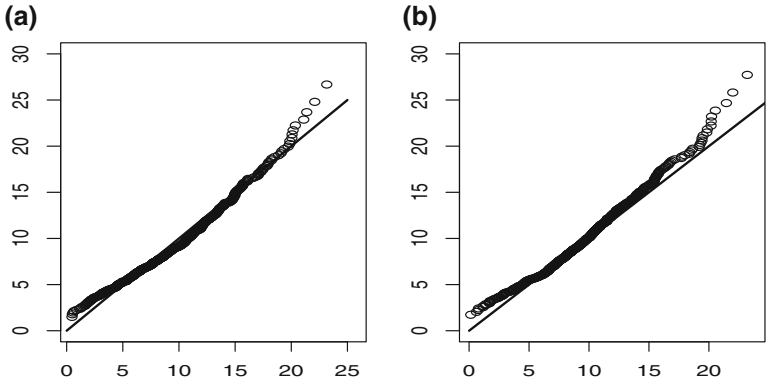


**Fig. 2** Q–Q plot: **a** $n = 200$, **b** $n = 400$

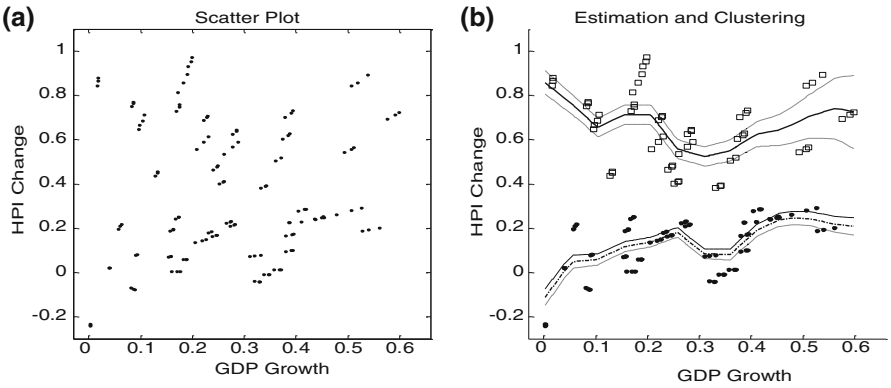


**Fig. 3** **a** Scatterplot of US house price index data; **b** estimated mean functions with 95% confidence intervals and a clustering result

price in repeat sales. It is believed that the housing price and GDP are correlated, and so it is of interest to study how GDP growth rate helps to predict HPI change.

First, a two-component mixture of nonparametric regression models is fitted to the data. For real data sets, we use Monte-Carlo cross-validation (MCCV) (Shao 1993) to select the bandwidths. In MCCV, the data are randomly partitioned into disjoint training subsets with size $n(1 - p)$ and test subsets with size $np$, where $p$ is the percentage of data used for testing. The procedure is repeated for 100 times, and we take the average as the selected bandwidth. For estimation and testing purpose, we use MCCV with $p = 10\%$, and the selected bandwidth is 0.030. Figure 3b contains the estimated mean functions and their 95% point-wise confidence intervals through the conditional bootstrap procedure, and the 95% confidence interval for $\pi_1$, $\sigma_1$ and $\sigma_2$ are (0.347, 0.518), (0.009, 0.020) and (0.004, 0.008), respectively. Figure 3b also reports the hard-clustering results, denoted by dots and squares, respectively, for the two components. The hard-clustering results are obtained by maximizing classification probabilities $\{p_{i1}, p_{i2}\}$ for all $i = 1, \ldots, n$. It can be checked that the dots in the lower cluster are mainly from Jan 1990 to Sep 1997, while the squares in the upper cluster

**Table 5** Average (standard deviation) of MSPE

| | $p = 10\%$ | $p = 20\%$ | $p = 25\%$ | $p = 33\%$ |
|---|---|---|---|---|
| US House Price Index Data | | | | |
| Model (1) | 0.086 (0.021) | 0.086 (0.014) | 0.085 (0.012) | 0.086 (0.011) |
| NMR (Huang et al. 2013) | 0.089 (0.025) | 0.090 (0.018) | 0.089 (0.015) | 0.089 (0.013) |
| NO data | | | | |
| Model (1) | 0.930 (0.285) | 1.011 (0.194) | 1.033 (0.182) | 1.037 (0.153) |
| NMR (Huang et al. 2013) | 1.330 (0.357) | 1.446 (0.246) | 1.511 (0.233) | 1.504 (0.200) |

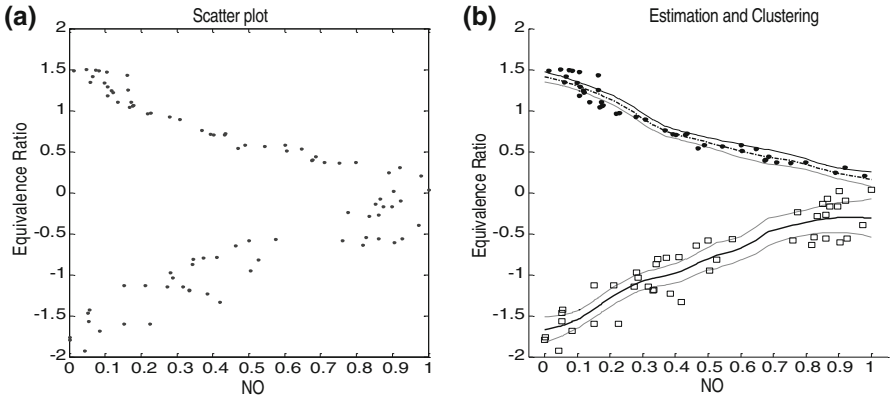


**Fig. 4** **a** Scatterplot of NO data; **b** estimated mean functions with 95% confidence intervals and a clustering result

are mainly from Oct 1997 to Dec 2002, when the economy experienced an internet boom and bust. In addition, it can be seen that in the first cycle of lower component, GDP growth has an overall positive impact on HPI change. However, in the second cycle of the upper component, GDP growth has a negative impact on HPI change, if GDP growth is smaller than 0.3; when GDP growth is larger than 0.3, it then has a similar positive impact on HPI change as the first cycle.

To examine whether the mixing proportions and variances are indeed constant, we apply the generalized likelihood ratio test developed in Sect. 2.3. The $p$-value is 0.331, and shows that model (1) is more appropriate for the data. To evaluate the prediction performance of the proposed model and compare it to the NMR model proposed by Huang et al. (2013), in Table 5, we use MCCV with repetition time 500 to report the average and standard deviation of the mean squared prediction error (MSPE) evaluated at the testing sets. It can be seen that the prediction performance of model (1) is slightly better than that of the NMR model (Huang et al. 2013).

*Example 2 (NO data)* This data set gives the equivalence ratio, a measure of the richness of the air-ethanol mix in an engine against the concentration of nitrogen oxide emissions in a study using pure ethanol as a spark-ignition engine fuel. The data set contains 99 observations and is presented in Hurvich et al. (1998). Figure 4a shows the scatter plot of the data, which clearly indicates two different nitrous oxide

concentration dependencies, with no clear linear trend. As a result, a two-component mixture of nonparametric regression models is fitted to the data.

Similar to the above example, the selected bandwidth is 0.091 based on MCCV with $p = 10\%$. The confidence intervals for parameter estimates are $(0.395, 0.608)$, $(0.005, 0.012)$, $(0.025, 0.053)$ for $\pi_1$, $\sigma_1$ and $\sigma_2$, respectively. Figure 4b contains the estimated mean functions and their 95% point-wise confidence intervals through the bootstrap procedure. The $p$-value of the generalized likelihood ratio test is 0.219, indicating that model (1) is the preferred model. Table 5 reports the average and standard deviation of MSPE evaluated at the testing sets based on MCCV. Based on Table 5, the new model has better prediction performance than the NMR model.

## 4 Discussion

Motivated by a US house index data, in this article, we proposed a new class of semiparametric mixture of regression models, where mixing proportions and variances are constants, but the component regression functions are smooth functions of a covariate. The identifiability of the proposed model is established and a one-step backfitting estimation procedure is proposed to achieve the optimal convergence rate for both the global parameters and the nonparametric regression functions. The proposed regression spline estimate is simple to calculate and can be easily extended to some other semiparametric and nonparametric mixture of regression models (Young and Hunter 2010; Huang et al. 2013; Huang and Yao 2012). But it requires more research to derive the asymptotic results for such regression spline based estimators for mixture models. A generalized likelihood ratio test has been proposed for semiparametric inferences.

When the dimension of the predictors is high, due to the curse of dimensionality, it is unpractical to estimate the component regression functions fully nonparametrically. Therefore, it is our interest to further extend the proposed mixture of nonparametric regression models to some other nonparametric or semiparametric models, such as mixture of partial linear regression models, mixture of additive models, and mixture of varying coefficient partial linear models.

In this paper, we assume that the number of components is known. However, in some applications, it might be infeasible to assume a known number of components in advance. Therefore, more research is needed to select the number of components for the proposed semiparametric mixture model. One possible way is to use AIC or BIC to choose the number of components. However, it is not clear how to define the degree of freedom for a semiparametric mixture model. Similar to Huang et al. (2013), one might also fit a mixture of linear regression using local data and choose the number of components based on traditional AIC or BIC. In addition, as one reviewer pointed out that when the number of components, $k$, is too large, the variance of model parameter estimates may be very large and the asymptotic results might not hold for the finite-sample setting. In this case, one might use a bootstrap procedure to estimate the standard errors of parameter estimates. Furthermore, it will be also interesting to investigate whether there are any minimax properties of the proposed estimation procedure.

**Acknowledgements** The authors thank the editor, the associate editor, and reviewers for their constructive comments that have led to dramatic improvement of the earlier version of this article. Xiang's research is supported by NSF of China Grant 11601477 and Zhejiang Provincial NSF of China Grant LQ16A010002. Yao's research is supported by NSF Grant DMS-1461677 and also funded by Department of Energy, Award no. 10006272.

## Appendix

In this section, the brief proofs of Theorems 1, 2 and 4 are presented, and please refer to the supplement file for more detailed proof. The conditions required by Theorems 1, 2, 3 and 4 are listed below. They are not the weakest sufficient conditions, but could easily facilitate the proofs.

### Technical conditions

(C1) $nh^4 \to 0$ and $nh^2 \log(1/h) \to \infty$ as $n \to \infty$ and $h \to 0$.
(C2) $nh \to \infty$ as $n \to \infty$ and $h \to 0$.
(C3) The sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ are independently and identically distributed from $f(x, y)$ with finite sixth moments. The support for $x$, denoted by $\mathscr{X} \in \mathbb{R}$, is bounded and closed.
(C4) $f(x, y) > 0$ in its support and has continuous first derivative.
(C5) $|\partial^3 \ell(\boldsymbol{\theta}, X, Y)/\partial\theta_i \partial\theta_j \partial\theta_k| \le M_{ijk}(X, Y)$, where $E(M_{ijk}(X, Y))$ is bounded for all $i, j, k$ and all $X, Y$.
(C6) The unknown functions $m_j(x)$, $j = 1, \ldots, k$, have continuous second derivative.
(C7) $\sigma_j^2 > 0$ and $\pi_j > 0$ for $j = 1, \ldots, k$ and $\sum_{j=1}^k \pi_j = 1$.
(C8) $E(X^{2r}) < \infty$ for some $\epsilon < 1 - r^{-1}$, $n^{2\epsilon - 1} h \to \infty$.
(C9) $I_\theta(x)$ and $I_m(x)$ are positive definite.
(C10) The kernel function $K(\cdot)$ is symmetric, continuous with compact support.
(C11) The marginal density $f(x)$ of $X$ is Lipschitz continuous and bounded away from 0. $X$ has a bounded support $\mathscr{X}$.
(C12) $t^3 K(t)$ and $t^3 K'(t)$ are bounded and $\int t^4 K(t) dt < \infty$.
(C13) $E|q_\theta|^4 < \infty$, $E|q_m|^4 < \infty$, where $\frac{\partial \ell(\theta(X), Y)}{\partial \theta} = q_\theta$, and $\frac{\partial \ell(\theta(X), Y)}{\partial m} = q_m$.

*Proof of Theorem 1.* Define $\hat{\boldsymbol{\beta}}^* = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, where $\hat{\boldsymbol{\beta}}$ maximizes $\ell_2(\boldsymbol{\beta})$ in (6). Let

$$\ell(\tilde{\boldsymbol{m}}(X_i), \boldsymbol{\beta}, Y_i) = \log \left\{ \sum_{j=1}^k \pi_j \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^2 \right\},$$

$$\ell(\tilde{\boldsymbol{m}}(X_i), \boldsymbol{\beta} + \boldsymbol{\beta}^*/\sqrt{n}, Y_i) = \log \left\{ \sum_{j=1}^k (\pi_j + \pi_j^*/\sqrt{n}) \phi(Y_i | \tilde{m}_j(X_i), \sigma_j^2 + \sigma_j^{2*}/\sqrt{n} \right\}.$$

Since $\hat{\boldsymbol{\beta}}$ maximizes $\ell_2$, it is easy to see that $\hat{\boldsymbol{\beta}}^*$ maximizes

$$\begin{aligned}\ell_n(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \left\{ \ell(\tilde{\boldsymbol{m}}(X_i), \boldsymbol{\beta} + \boldsymbol{\beta}^*/\sqrt{n}, Y_i) - \ell(\tilde{\boldsymbol{m}}(X_i), \boldsymbol{\beta}, Y_i \right\} \\ &= A_n \boldsymbol{\beta}^* + \frac{1}{2} \boldsymbol{\beta}^{*T} B_n \boldsymbol{\beta}^* + o_p(\|\boldsymbol{\beta}^*\|^2),\end{aligned}$$

where $A_n = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\tilde{\boldsymbol{m}}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta}}$ and $B_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\tilde{\boldsymbol{m}}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$. It can be easily seen that $B_n = -B + o_p(1)$ with $B = E\{I_\beta(X)\}$, therefore, by quadratic approximation lemma,

$$\hat{\boldsymbol{\beta}}^* = B^{-1} A_n + o_p(1). \tag{14}$$

Define $R_{1n} = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T} (\tilde{\boldsymbol{m}}(X_i) - \boldsymbol{m}(X_i))$, then $A_n = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta}} + R_{1n} + O_p(\sqrt{\frac{1}{n}} \|\tilde{\boldsymbol{m}} - \boldsymbol{m}\|_\infty^2)$. Let $\varphi(X_t, Y_t)$ be a $k \times 1$ vector whose elements are the first $k$ entries of $I_\theta^{-1}(X_t) \frac{\partial \ell(\boldsymbol{\theta}(X_t), Y_t)}{\partial \boldsymbol{\theta}}$. From assumption (C1), we know that $O_p\{n^{1/2}[\gamma_n h^2 + \gamma_n^2 \log^{1/2}(1/h)]\} = o_p(1)$, where $\gamma_n = (nh)^{-1/2}$. By similar argument as the proof of Theorem 2 in Huang et al. (2013), it can be shown that $\tilde{\boldsymbol{\theta}}(X_i) - \boldsymbol{\theta}(X_i) = \frac{1}{n} f^{-1}(X_i) I_\theta^{-1}(X_i) \sum_{t=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}(X_i), Y_t)}{\partial \boldsymbol{\theta}} K_h(X_t - X_i) + O_p\{\gamma_n h^2 + \gamma_n^2 \log^{1/2}(1/h)\}$. Since $\boldsymbol{m}(X_i) - \boldsymbol{m}(X_t) = O(X_i - X_t)$,

$$\begin{aligned}R_{1n} &= n^{-3/2} \sum_{t=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T} f^{-1}(X_i) \varphi(X_t, Y_t) K_h(X_i - X_t) + O_p(n^{1/2} h^2) \\ &= R_{2n} + O_p(n^{1/2} h^2).\end{aligned}$$

It can be shown that $E[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{m}^T} f^{-1}(X_i) K_h(X_i - X_t)] = I_{\beta m}(X_t)$. Let $\varpi(X_t, Y_t) = I_{\beta m}(X_t) \varphi(X_t, Y_t)$, and $R_{n3} = -n^{-1/2} \sum_{j=1}^{n} \varpi(X_t, Y_t)$, then $R_{n2} - R_{n3} \xrightarrow{P} 0$, and therefore

$$A_n = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} \left\{ \frac{\partial \ell(\boldsymbol{m}(X_i), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta}} - \varpi(X_i, Y_i) \right\} + o_p(1),$$

given $nh^4 \to 0$. Let $\Sigma = \mathrm{Var}\{\frac{\partial \ell(\boldsymbol{\theta}(X), Y)}{\partial \boldsymbol{\beta}} - \varpi(X, Y)\}$, then $\mathrm{Var}(A_n) = \Sigma$. It can be easily seen that $E(A_n) = 0$, therefore by (14),

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(0, B^{-1} \Sigma B^{-1}).$$

□

*Proof of Theorem 2.* Define $\hat{\boldsymbol{m}}^* = \sqrt{nh}(\hat{\boldsymbol{m}}(x) - \boldsymbol{m}(x))$, where $\hat{\boldsymbol{m}}(x)$ maximizes (7). It can be shown that

$$\hat{\boldsymbol{m}}^*(x) = f(x)^{-1} I_m(x)^{-1} \hat{S}_n + o_p(1), \tag{15}$$

where

$$\hat{S}_n = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{m}(x), \hat{\boldsymbol{\beta}}, Y_i)}{\partial \boldsymbol{m}} K_h(X_i - x). \tag{16}$$

Notice that

$$\begin{aligned} \hat{S}_n &= \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m}} K_h(X_i - x) + \sqrt{\frac{h}{n}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \times \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m} \partial \boldsymbol{\beta}^T} K_h(X_i - x) + o_p(1) \\ &\equiv S_n + D_n + o_p(1). \end{aligned}$$

where $S_n = \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}(x), Y_i)}{\partial \boldsymbol{\theta}} K_h(X_i - x)$. Since $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$ and $\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m} \partial \boldsymbol{\beta}^T} K_h(X_i - x) = -f(x) I_{\beta m}^T(x) + o_p(1)$, then $D_n = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sqrt{h} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{m} \partial \boldsymbol{\beta}^T} K_h(X_i - x) = -\sqrt{h} f(x) I_{\beta m}^T(x) + o_p(1)$. Thus, from (15), $\hat{\boldsymbol{m}}^*(x) = f(x)^{-1} I_m(x)^{-1} S_n + o_p(1)$. Let $\Lambda(u|x) = E[\frac{\partial \ell(\boldsymbol{m}(x), \boldsymbol{\beta}, Y)}{\partial \boldsymbol{m}} | X = u]$, it can be shown that

$$E(S_n) = \sqrt{nh} \left[ \frac{1}{2} f(x) \Lambda''(x|x) + f'(x) \Lambda'(x|x) \right] \kappa_2 h^2, \operatorname{Var}(S_n) = f(x) I_m(x) \nu_0, \tag{17}$$

where $\nu_0 = \int K^2(t) \mathrm{d}t$. To complete the proof, let $\Delta(x) = I_m^{-1}(x) \left[ \frac{1}{2} \Lambda''(x|x) + f^{-1}(x) f'(x) \Lambda'(x|x) \right] \kappa_2 h^2$, and $\Delta_m(x)$ be a $k \times 1$ vector whose elements are the first $k$ entries of $\Delta(x)$, then

$$\sqrt{nh}(\hat{\boldsymbol{m}}(x) - \boldsymbol{m}(x) - \Delta_m(x) + o_p(h^2)) \xrightarrow{D} N(0, f^{-1}(x) I_m^{-1}(x) \nu_0).$$

□

*Proof of Theorem 4.* Since $\hat{\boldsymbol{\beta}}$ has faster convergence rate than $\hat{\boldsymbol{m}}(\cdot)$, $\hat{\boldsymbol{m}}(\cdot)$ has the same asymptotic properties as if $\boldsymbol{\beta}$ were known. Therefore, in the following proof, we study the property of $\hat{\boldsymbol{m}}(\cdot)$ assuming $\boldsymbol{\beta}$ to be known.

Define $\frac{\partial \ell(\theta(X_i), Y_i)}{\partial \theta} = q_{\theta i}$, $\frac{\partial^2 \ell(\theta(X_i), Y_i)}{\partial \theta \partial \theta^T} = q_{\theta\theta i}$ and similarly, define $q_{\mathrm{mi}}$, $q_{\mathrm{mmi}}$ and so on. Let $\tilde{\boldsymbol{\theta}}$ be the estimator under $H_1$ (Huang et al. 2013), and $\hat{\boldsymbol{m}}$ be the estimator under $H_0$ (model (2.1). From previous proof, we have

$$\tilde{\boldsymbol{\theta}}(X_i) - \boldsymbol{\theta}(X_i) = \frac{1}{n} f^{-1}(X_i) I_\theta^{-1}(X_i) \sum_{t=1}^{n} q_{\theta t} K_h(X_t - X_i)(1 + o_p(1)), \tag{18}$$

$$\hat{\boldsymbol{m}}(X_i) - \boldsymbol{m}(X_i) = \frac{1}{n} f^{-1}(X_i) I_m^{-1}(X_i) \sum_{t=1}^{n} q_{mt} K_h(X_t - X_i)(1 + o_p(1)). \tag{19}$$

By (18) and (19), we can obtain that

$$\begin{aligned}
\sum_{i=1}^{n} \ell(\tilde{\boldsymbol{\theta}}(X_i), Y_i) - \sum_{i=1}^{n} \ell(\boldsymbol{\theta}(X_i), Y_i) &= \Bigg\{ \frac{1}{n} \sum_{i,l} q_{\theta i}^T f^{-1}(X_l) I_\theta^{-1}(X_l) q_{\theta l} K_h(X_i - X_l) \\
&\quad + \frac{1}{2n^2} \sum_{i,j,l} q_{\theta i}^T f^{-2}(X_l) I_\theta^{-1}(X_l) q_{\theta\theta l} I_\theta^{-1}(X_l) q_{\theta j} K_h(X_i - X_l) K_h(X_j - X_l) \Bigg\} (1 + o_p(1)), \\
\sum_{i=1}^{n} \ell(\hat{\boldsymbol{m}}(X_i), Y_i) - \sum_{i=1}^{n} \ell(\boldsymbol{m}(X_i), Y_i) &= \Bigg\{ \frac{1}{n} \sum_{i,l} q_{mi}^T f^{-1}(X_l) I_m^{-1}(X_l) q_{ml} K_h(X_i - X_l) \\
&\quad + \frac{1}{2n^2} \sum_{i,j,l} q_{\mathrm{mi}}^T f^{-2}(X_l) I_m^{-1}(X_l) q_{\mathrm{mml}} I_m^{-1}(X_l) q_{\mathrm{mj}} K_h(X_i - X_l) K_h(X_j - X_l) \Bigg\} (1 + o_p(1)),
\end{aligned}$$

and so,

$$\begin{aligned}
T &= \frac{1}{n} \sum_{i,l} \left[ q_{\theta i}^T I_\theta^{-1}(X_l) q_{\theta l} - q_{\mathrm{mi}}^T I_m^{-1}(X_l) q_{ml} \right] f^{-1}(X_l) K_h(X_i - X_l) \\
&\quad + \frac{1}{2n^2} \sum_{i,j,l} \Big[ q_{\theta i}^T I_\theta^{-1}(X_l) q_{\theta\theta l} \\
&\quad \times I_\theta^{-1}(X_l) q_{\theta j} - q_{\mathrm{mi}}^T I_m^{-1}(X_l) q_{\mathrm{mml}} I_m^{-1}(X_l) q_{\mathrm{mj}} \Big] f^{-2}(X_l) K_h(X_i - X_l) K_h(X_j - X_l) \\
&\equiv \Lambda_n + \frac{1}{2} \Gamma_n.
\end{aligned}$$

By similar argument as Fan et al. (2001), it can be shown that under conditions (C9)–(C12), as $h \to 0, nh^{3/2} \to \infty$,

$$\begin{aligned}
\Lambda_n &= \frac{2k-1}{h} K(0) E f(X)^{-1} \\
&\quad + \frac{1}{n} \sum_{l \neq i} \left[ q_{\theta i}^T I_\theta^{-1}(X_l) q_{\theta l} - q_{\mathrm{mi}}^T I_m^{-1}(X_l) q_{\mathrm{ml}} \right] f^{-1}(X_l) K_h(X_i - X_l) + o_p(h^{-1/2}), \\
\Gamma_n &= -\frac{(2k-1)}{h} E f(X)^{-1} \int K^2(t) dt \\
&\quad - \frac{2}{n} \sum_{i<j} [q_{\theta i}^T I_\theta^{-1}(X_i) q_{\theta j} - q_{\mathrm{mi}}^T I_m^{-1}(X_i) q_{\mathrm{mj}}] f^{-1}(X_i) \\
&\quad \times K_h * K_h(X_i - X_j) + o_p(h^{-1/2}).
\end{aligned}$$

Therefore, $T = \mu_n + W_n/2\sqrt{h} + o_p(h^{-1/2})$, where $\mu_n = \frac{(2k-1)|\mathscr{X}|}{h}[K(0) - 0.5\int K^2(t)dt]$,

$$W_n = \frac{\sqrt{h}}{n}\sum_{i\neq j}\{q_{\theta i}^T I_\theta^{-1}(X_j)[2K_h(X_i - X_j) - K_h * K_h(X_i - X_j)]f^{-1}(X_j)q_{\theta j} - q_{\mathrm{mi}}^T I_m^{-1}(X_j)[2K_h(X_i - X_j) - K_h * K_h(X_i - X_j)]f^{-1}(X_j)q_{\mathrm{mj}}\}.$$

It can be shown that $\mathrm{Var}(W_n) \to \zeta$, where $\zeta = 2(2k-1)Ef^{-1}(X)\int[2K(t) - K * K(t)]^2dt$. Apply Proposition 3.2 in de Jong (1987), we obtain that

$$W_n \xrightarrow{D} N(0, \zeta),$$

and completes the proof. □

## References


Cao, J., Yao, W. (2012). Semiparametric mixture of binomial regression with a degenerate component. *Statistica Sinica*, *22*, 27–46.

Celeux, G., Hurn, M., Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*, 957–970.

de Boor, C. (2001). *A practical guide to splines*. New York: Springer.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields*, 75, 261–277.

Fan, J., Zhang, C., Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annuals of Statistics*, *29*, 153–193.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of American and Statisical Association*, *96*, 194–209.

Goldfeld, S. M., Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, *1*, 3–6.

Green, P. J., Richardson, S. (2002). A Markov model for switching regression. *Journal of American and Statistical Association*, *97*, 1055–1070.

Hastie, T., Tibshirani, R., Friedman, J. (2003). *The elements of statistical learning, data mining, inference and prediction*. New York: Springer.

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

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

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

Hurvich, C. M., Simonoff, J. S., Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, *60*, 271–294.

Shao, J. (1993). Linear models selection by cross-validation. *Journal of the American Statistical Association*, *88*, 486–494.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Society, Series B*, *62*, 795–809.

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

Wang, S., Yao, W., Huang, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statistics and Probability Letters*, *93*, 41–45.

Wedel, M., DeSarbo, W. S. (1993). A latent class binomial logit methodology for the analysis of paired comparison data. *Decision Sciences*, *24*, 1157–1170.
Yao, W. (2012). Model based labeling for mixture models. *Statistics and Computing*, *22*, 337–347.
Yao, W. (2015). Label switching and its simple solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation*, *85*, 1000–1012.
Yao, W., Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association*, *104*, 758–767.
Young, D. S., Hunter, D. R. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*, *54*, 2253–2266.
Zhang, Y., Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*, 95–112.