mathematics

MDPI

Article

A Novel EM-Type Algorithm to Estimate Semi-Parametric Mixtures of Partially Linear Models

Sphiwe B. Skhosana *, Salomon M. Millard and Frans H. J. Kanfer

Department of Statistics, University of Pretoria, Pretoria 0002, South Africa
* Correspondence: spiwe.skhosana@up.ac.za

Abstract: Semi- and non-parametric mixture of normal regression models are a flexible class of mixture of regression models. These models assume that the component mixing proportions, regression functions and/or variances are non-parametric functions of the covariates. Among this class of models, the semi-parametric mixture of partially linear models (SPMPLMs) combine the desirable interpretability of a parametric model and the flexibility of a non-parametric model. However, local-likelihood estimation of the non-parametric term poses a computational challenge. Traditional EM optimisation of the local-likelihood functions is not appropriate due to the label-switching problem. Separately applying the EM algorithm on each local-likelihood function will likely result in non-smooth function estimates. This is because the local responsibilities calculated at the E-step of each local EM are not guaranteed to be aligned. To prevent this, the EM algorithm must be modified so that the same (global) responsibilities are used at each local M-step. In this paper, we propose a one-step backfitting EM-type algorithm to estimate the SPMPLMs and effectively address the label-switching problem. The proposed algorithm estimates the non-parametric term using each set of local responsibilities in turn and then incorporates a smoothing step to obtain the smoothest estimate. In addition, to reduce the computational burden imposed by the use of the partial-residuals estimator of the parametric term, we propose a plug-in estimator. The performance and practical usefulness of the proposed methods was tested using a simulated dataset and two real datasets, respectively. Our finite sample analysis revealed that the proposed methods are effective at solving the label-switching problem and producing reasonable and interpretable results in a reasonable amount of time.

Keywords: mixture of regressions; EM algorithm; partial linear models; semi-parametric; local likelihood; label-switching

MSC: 62-08

Citation: Skhosana, S.B.; Millard, S.M.; Kanfer, F.H.J. A Novel EM-Type Algorithm to Estimate Semi-Parametric Mixtures of Partially Linear Models. *Mathematics* **2023**, *11*, 1087. https://doi.org/10.3390/math11051087

Academic Editors: Davide Valenti and Christophe Chesneau

Received: 24 December 2022
Revised: 15 February 2023
Accepted: 19 February 2023
Published: 22 February 2023

1. Introduction

The classical normal linear regression model (CNLRM) is useful for modelling data from a homogeneous population. However, heterogeneous populations, comprised of various sub-populations, are common in fields such as economics and environmental sciences, among others. To model data from such populations, Quandt [1] and subsequently Goldfeld and Quandt [2] introduced mixtures of NLRMs (MNLRMs). Estimation and inference of these models were studied by Hurn et al. [3] and Frühwirth-Schnatter [4] using Bayesian methods and DeSarbo and Cron [5] and De Veaux [6] using maximum likelihood via the expectation maximisation (EM) algorithm [7]. A systematic study of these models can be found in [4].

The MNLRMs work well when the component regression functions (CRFs) can be assumed to be linear parametric functions of the covariates. However, the latter assumption seldom holds for all covariates. The effect of one or more of the covariate(s) may be characterised by a non-parametric function. For this reason, Wu and Liu [8] proposed

the semi-parametric mixture of partial linear models (SPMPLMs), where each CRF is a linear combination of a parametric function of some of the covariates and a non-parametric function of one of the covariates. The model assumes that each component has a Gaussian distribution, and hence symmetric. The SPMPLMs combine the advantages of a parametric and a non-parametric function. In addition to providing a flexible approach for detecting unobserved regression relationships, this model alleviates the curse of dimensionality (COD) to a certain extent. Moreover, the fitted non-parametric functions can be used to suggest a suitable functional form for the covariate. However, estimation of this model poses a computational challenge. A local-likelihood approach to estimating the non-parametric functions is prone to experience a form of label switching. Maximising each local-likelihood function separately using the EM algorithm does not guarantee that the labels will be the same at each local point. This is because, at the E-step of the EM algorithm, each local point generates a unique set of responsibilities. The latter will be referred to as local responsibilities. The resulting estimated non-parametric functions are non-smooth and not useful in practice.

This problem was first mentioned by Huang and Yao [9] in the context of estimating the MNLRMs where the mixing proportions are non-parametric functions of a covariate. This phenomena is akin to the label-switching problem that occurs when estimating Bayesian mixtures using MCMC procedures [10].

To solve the label-switching problem, the EM algorithm must be modified. This is achieved by making use of the same responsibilities to maximise each local-likelihood function at the M-step of the EM algorithm [11]. This implies that a global set of responsibilities must be obtained. Following this idea, Huang et al. [11] proposed the effective EM algorithm for estimating the non-parametric mixture of normal regressions (NPMNRs), where the mixing proportions, variances and regression functions are non-parametric functions of a covariate. Xiang and Yao [12] proposed a local EM-type and global EM-type algorithm for estimating the semi-parametric mixture of non-parametric regressions (SPMNPRs), where only the CRFs are non-parametric. To estimate the SPMPLMs, Wu and Liu [8] proposed the profile likelihood EM (PL-EM) algorithm. Huang et al. [13] proposed a local-likelihood estimation procedure via a modified EM algorithm to estimate the mixture of varying coefficient models. To estimate the non-parametric and parametric terms of the mixture of single index models (MSIMs) and the mixture of regression models with varying single index models (MRSIP), Xiang and Yao [14] proposed a one-step backfitting estimation procedure via a modified EM algorithm. Zhang and Zheng [15] and Zhang and Pan [16] proposed a spline-backfitted kernel (SBK) estimation procedure via a modified EM algorithm to estimate the semi-parametric mixture of additive regressions (SPMARs) and the semi-parametric mixture of partially linear additive regressions (SPMPLARs). The above algorithms are local-likelihood based EM-type algorithms. More recently, Xue and Yao [17] proposed a neural network-based EM-type algorithm to estimate the MNLRMs with non-parametrically covariate-varying mixing proportions. Xue [18] also proposed a neural network EM-type algorithm to estimate a new form of a mixture of experts (MoE) model [19] with non-parametric CRFs.

In each of the above-mentioned local-likelihood based fitting algorithms, the global responsibilities are calculated without taking into account the information from the local responsibilities. To estimate the NPMNRs, Skhosana et al. [20] proposed a novel EM-type algorithm that takes into account the local information. A simulation study was used to show that the performance of the algorithm is at least as good as that of the effective EM algorithm.

Motivated by the practical importance of the SPMPLMs, as mentioned above, the present paper is concerned with estimating this model. Our first research question was: how can we obtain smooth estimates of the non-parametric functions in the presence of label switching? Our response to this question follows the same ideas as in [20]. Speckman [21] showed that a less biased and more efficient estimator of the parametric term of the CRF can be obtained via partial residuals. However, this requires the computation of an $n \times n$ smoother matrix,

where n is the sample size. This can be computationally costly for a large n, especially when applied within an iterative algorithm. Our second research question was: in an effort to reduce the computational burden imposed by the partial-residuals estimator, is there an estimator that can perform at least as well as the former estimator? In answer to the above questions, the current study makes the following contributions to estimation and computation for the SPMPLMs. Firstly, we propose a one-step backfitting EM-type algorithm to address the label-switching problem. The proposed algorithm estimates the non-parametric term as follows:

1. At the E-step, we calculate the local responsibilities over each point in the set of grid points;
2. At the M-step, we simultaneously maximise all the local-likelihood functions using each set of local responsibilities in turn. In other words, for each set of local responsibilities, we will have an estimate of the non-parametric term.
3. Among the estimates obtained in step 1, we choose the smoothest estimate as the final estimate of the non-parametric term.

We repeat the above three steps until convergence. Steps 1 and 2 are the regular E- and M-steps of the EM algorithm. Step 3 can be seen as a smoothing step. Indeed, at each iteration, the algorithm can be viewed as removing the roughness (or wiggliness) from the non-parametric term due to the effect of label switching. Using the resulting estimate of the non-parametric term, the algorithm continues by estimating the parametric term. Finally, we use the latter to improve the estimate of the non-parametric term.

Secondly, we propose a regression spline-based estimation procedure for the SPMPLMs. This approach does not use local-likelihood estimation, and thus it is free from the label-switching problem. Simulation results show that this approach performs at least as well as the profile likelihood approach in estimating the parametric term. However, the former approach performs better than the latter approach in estimating the non-parametric term. To demonstrate the practical use of the proposed methods, we consider an analysis of two climate datasets.

Lastly, to reduce the computational burden imposed by the partial-residuals estimator of the parametric part of the CRF, we propose a plug-in estimator.

The rest of the paper is organised as follows. In Section 2, we give a brief definition of the SPMPLMs. In Section 3, we discuss likelihood estimation for the SPMPLMs. We present two approaches, regression spline-based estimation and profile likelihood-based estimation. For the latter, we highlight the label-switching problem and then present the proposed solution. Section 4 presents a data-driven method used to select the appropriate smoothing parameter. Sections 5 and 6 present the results of our simulation studies and an test on two climate datasets, respectively. Section 7 concludes and gives directions for future research.

2. Model Definition

Let $(\mathbf{X}, T, Y)$ be a set of random variables from a population consisting of K subpopulations. We assume throughout the paper that both Y and T are univariate and $\mathbf{X}$ is $p-$dimensional. Let $\mathcal{K}$ be a latent variable. For $k = 1, 2, \ldots, K$, $\mathcal{K}$ has a discrete distribution $P(\mathcal{K} = k|\mathbf{X} = \mathbf{x}) = \pi_k$, where $0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Conditional on $\mathcal{K}$, $\mathbf{X} = \mathbf{x}$ and $T = t$, Y follows a Gaussian distribution with CRF $m_k(\mathbf{x}, t) = \mathbf{x}\boldsymbol{\beta}_k + g_k(t)$ and variance σ_k^2. The function $g_k(t)$ is assumed to be a smooth unknown function of the covariate t and $\boldsymbol{\beta}_k$ is a $p-$dimensional vector of regression coefficients. Conditional on $\mathbf{X} = \mathbf{x}$ and $T = t$, Y follows a mixture of Gaussians

$$p(Y|\mathbf{X} = \mathbf{x}, T = t) = \sum_{k=1}^{K} \pi_k \mathcal{N}_k\{\mathbf{x}\boldsymbol{\beta}_k + g_k(t), \sigma_k^2\}. \quad (1)$$

Note that, in order to ensure identifiability, $\boldsymbol{\beta}_k$ does not include the intercept term. In model (1), the CRF, $m_k(\mathbf{x}, t)$, is semi-parametric, being a linear combination of a parametric term ($\mathbf{x}\boldsymbol{\beta}_k$) and a non-parametric ($g_k(t)$) term. In effect, the p covariates $\mathbf{x}$ are assumed linearly related to y, whereas the relationship between t and y is assumed to be characterised by a non-parametric function. The component variances and mixing proportions are assumed to be constant and thus parametric. For $K = 1$, model (1) is a partial linear regression model. For $g_k(t)$ constant, model (1) is a mixture of linear regressions model. Thus, model (1) is a natural extension of the partial linear model and the mixture of linear regressions model.

3. Model Estimation

In order to estimate model (1), we must estimate both the parametric ($\boldsymbol{\beta}_k, \pi_k, \sigma_k^2$) and the non-parametric ($g_k(t)$) terms of the model. A likelihood approach to estimation is adopted for this purpose. Let $\boldsymbol{\vartheta}_1 = (\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2, \mathbf{g})$ be a vector of the model parameters and non-parametric functions, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$ and $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_K)$ with $\mathbf{g}_k = (g_k(t_1), g_k(t_2), \ldots, g_k(t_n))$. For a random sample $\{(\mathbf{x}_i, t_i, y_i) : i = 1, 2, \ldots, n\}$ generated from model (1), the log-likelihood function is given by

$$\ell(\boldsymbol{\vartheta}_1) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}_k\{\mathbf{x}_i\boldsymbol{\beta}_k + g_k(t_i), \sigma_k^2\}\right]. \tag{2}$$

Due to the presence of the non-parametric function in (2), it is not straightforward to optimise the log likelihood function to obtain the maximum likelihood estimates. In the following, we discuss two likelihood-based approaches to estimate both parametric and non-parametric terms of model (1).

3.1. Regression Spline-Based Estimation

We begin by discussing a regression spline-based estimator. A regression spline is a non-parametric estimator that works by parametrising a function using a set of piecewise polynomial functions joined at a set of points, or knots, in the domain of the function (see Wu and Zhang [22] and James et al. [23] for more details). We make use of a cubic spline. The non-parametric function $g_k(t)$ can be written as

$$g_k(t) = \sum_{j=1}^{J+4} \eta_{jk} B_j(t) \tag{3}$$

where J is the number of internal knots, the $B_j(t)$s are the basis functions and the η_{jk}s are the coefficients for the k^{th} component. We make use of the B-spline basis functions, mainly because of their numerical stability (see Fan and Gijbels [24]).

After substituting (3) into (2), the log likelihood function can be written as

$$\ell(\boldsymbol{\vartheta}_2) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}_k\{\mathbf{x}_i\boldsymbol{\beta}_k + \sum_{j=1}^{J+4} \eta_{jk} B_j(t_i), \sigma_k^2\}\right], \tag{4}$$

where $\boldsymbol{\vartheta}_2 = (\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2, \boldsymbol{\eta})$ is a vector of all the model's parameters, where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K)$ with $\boldsymbol{\eta}_k = (\eta_{1k}, \eta_{2k}, \ldots, \eta_{(J+4)k})$. Note that the regression function is now completely a linear combination of parametric terms, where $\boldsymbol{\beta}_k$ and η_{jk} are both regression coefficients. To optimise the log likelihood function (4), we make use of the EM algorithm. Towards that end, define the latent variable as

$$z_{ik} = \begin{cases} 1 & \text{if } (\mathbf{x}_i, t_i, y_i) \text{ is in the } k^{th} \text{ component} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Thus, the complete data are given by $\{(\mathbf{x}_i, t_i, y_i, \mathbf{z}_i) : i = 1, 2, \ldots, n\}$, where $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})$.

The corresponding complete-data log likelihood is given by

$$\ell_c(\boldsymbol{\vartheta}_2) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\log\pi_k + \log\mathcal{N}_k\{\mathbf{x}_i\boldsymbol{\beta}_k + \sum_{j=1}^{J+4}\eta_{jk}B_j(t_i), \sigma_k^2\}\right]. \tag{6}$$

Estimation Algorithm

Since the z_{ik}s are latent variables, we maximise the conditional expected value of (6). Towards that end, at the t^{th} iteration of the EM algorithm, in the E-step we estimate the z_{ik}s using the conditional expected values given by

$$\gamma_{ik}^{(t)} = \frac{\hat{\pi}_k^{(t-1)}\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k^{(t-1)} + \sum_{j=1}^{J+4}\eta_{jk}^{(t-1)}B_j(t_i), \sigma_k^{2(t-1)}\}}{\sum_{j=1}^{K}\pi_j^{(t-1)}\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k^{(t-1)} + \sum_{j=1}^{J+4}\eta_{jk}^{(t-1)}B_j(t_i), \sigma_j^{2(t-1)}\}}. \tag{7}$$

In the M-step, we update the $\boldsymbol{\vartheta}_2$ by maximising the following conditional expected log likelihood:

$$Q(\boldsymbol{\vartheta}_2^{(t)}|\boldsymbol{\vartheta}_2^{(t-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}^{(t)}[\log\pi_k + \log\mathcal{N}\{y_i|\mathbf{x}_i\boldsymbol{\beta}_k + \sum_{j=1}^{J+4}\eta_{jk}B_j(t_i), \sigma_k^2\}].$$

Maximising $Q(\cdot|\cdot)$ with respect to π_k, $\boldsymbol{\beta}_k$, $\boldsymbol{\eta_k}$ and σ_k^2, respectively, yields

$$\begin{aligned}
\pi_k^{(t)} &= \frac{\sum_{i=1}\gamma_{ik}^{(t)}}{n} && (8)\\
\boldsymbol{\beta}_k^{(t)} &= (\mathbf{X}^\intercal W_k^{(t)}\mathbf{X})^{-1}\mathbf{X}^\intercal W_k^{(t)}(y - \mathbf{B}\boldsymbol{\eta}_k^{(t-1)}) && (9)\\
\boldsymbol{\eta}_k^{(t)} &= (\mathbf{B}^\intercal W_k^{(t)}\mathbf{B})^{-1}\mathbf{B}^\intercal W_k^{(t)}(y - \mathbf{X}\boldsymbol{\beta}_k^{(t)}) && (10)\\
\sigma_k^{2(t)} &= \frac{\sum_{i=1}^{n}\gamma_{ik}^{(t)}(y_i - \mathbf{x}_i\boldsymbol{\beta}_k^{(t)} - \mathbf{b}_i\boldsymbol{\eta}_k^{(t)})}{\sum_{i=1}^{n}\gamma_{ik}^{(t)}} && (11)
\end{aligned}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\intercal$, $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n)^\intercal$, $\mathbf{b}_i = (B_1(t_i), B_2(t_i), \ldots, B_{J+4}(t_i))$ and $W_k^{(t)} = diag\{\gamma_{1k}^{(t)}, \gamma_{2k}^{(t)}, \ldots, \gamma_{nk}^{(t)}\}$.

Upon convergence of the above EM algorithm, the estimate of $\boldsymbol{\vartheta}$ is given by the vector $\hat{\boldsymbol{\vartheta}}_2 = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{\sigma}^2, \hat{\boldsymbol{\eta}})$, whence the estimator of the non-parametric function, $g_k(t)$, can be obtained as

$$\hat{\mathbf{g}}_k = \mathbf{B}\hat{\boldsymbol{\eta}}_k \text{ for } k = 1, 2, \ldots, K. \tag{12}$$

Consequently, $\hat{\boldsymbol{\vartheta}}_1 = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{\sigma}^2, \hat{\mathbf{g}})$.

3.2. Profile Likelihood Estimation

We now discuss a profile likelihood approach to the estimation of model (1). When estimating a semi-parametric model, the profile likelihood approach proceeds in two-steps, estimating each term in turn. First, the non-parametric function ($g_k(t)$) is profiled-out as a nuisance parameter. That is, for a fixed value of the parametric term, an estimate of the function $g_k(t)$, referred to as a least favourable curve (denoted by $g_{\beta_k}(t)$), is obtained. The profile likelihood is then derived as

$$\begin{aligned}
p\ell(\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2) &= \ell(\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2, \mathbf{g}_\beta)\\
&= \sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}\pi_k\mathcal{N}_k\{\mathbf{x}_i\boldsymbol{\beta}_k + g_{\beta_k}(t_i), \sigma_k^2\}\right] && (13)
\end{aligned}$$

where $\mathbf{g}_{\boldsymbol{\beta}} = (\mathbf{g}_{\boldsymbol{\beta}_1}, \mathbf{g}_{\boldsymbol{\beta}_2}, \ldots, \mathbf{g}_{\boldsymbol{\beta}_K})$ with $\mathbf{g}_{\boldsymbol{\beta}_k} = (g_{\boldsymbol{\beta}_k}(t_1), g_{\boldsymbol{\beta}_k}(t_2), \ldots, g_{\boldsymbol{\beta}_k}(t_n))$. Note that the log likelihood function in Equation (13) is the same as (2), where $\mathbf{g}_{\boldsymbol{\beta}}$ is substituted for $\mathbf{g}$. To estimate the parametric term, the profile likelihood function (13) is maximised to obtain the maximum likelihood estimates (see Severini and Wong [25] for more details).

We make use of the local-likelihood kernel approach [26] to estimate the non-parametric function $\mathbf{g}_{\boldsymbol{\beta}_k}$. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be a set of N grid points in the domain of the covariate t. For a random sample $\{(\mathbf{x}_i, t_i, y_i) : i = 1, 2, \ldots, n\}$ from model (1), the log of the local-likelihood function for the non-parametric function $g_k(t)$ is defined as

$$\ell(\mathbf{g}(u_j)) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k + g_k(u), \sigma_k^2\}\right] K_h(t_i - u) \tag{14}$$

where $\mathbf{g}(u_j) = (g_1(u_j), g_2(u_j), \ldots, g_K(u_j))$. By maximising (14), we obtain the local-likelihood estimate of $g_k(u_j)$, denoted $g_{\boldsymbol{\beta}_k}(u_j)$, for each $u_j \in \mathcal{U}$. We make use of the local mean estimator (see Fan and Gijbels [24] for more details about this estimator). By linearly interpolating $g_{\boldsymbol{\beta}_k}(u_j)$ for $j = 1, 2, \ldots, N$, we can obtain $g_{\boldsymbol{\beta}_k}(t_i)$ for $i = 1, 2, \ldots, n$. By substituting the latter into (13), we obtain the profile likelihood function, whence we can derive the maximum likelihood estimators of the parameters $\boldsymbol{\pi}$, $\boldsymbol{\beta}$ and σ^2. For a given random sample, the estimates of the latter parameters, together with the non-parametric estimate $g_{\boldsymbol{\beta}}$, are referred to as profile likelihood estimates. In the following section, we discuss how the above profile likelihood estimation procedure is implemented.

3.2.1. Local-Likelihood EM (LL-EM) Algorithm

In this section, we present a naive implementation of the above profile likelihood estimation procedure. The EM algorithm is applied separately to maximise each of the N local-likelihood functions (14) followed by another separate application of the algorithm to the maximisation of the profile likelihood function (13). As we demonstrate below, this approach is subject to the label-switching problem, as mentioned in [11] and comprehensively discussed in [20]. Let $\{(\mathbf{x}_i, t_i, y_i, \mathbf{z}_i) : i = 1, 2, \ldots, n\}$ be the complete data where $\mathbf{z}_i$ is the latent indicator variable, as defined in (5). The log of the complete local-likelihood function is given by

$$\ell_c(\mathbf{g}(u)) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\log\pi_k + \log\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k + g_{\boldsymbol{\beta}_k}(u), \sigma_k^2\}\right] K_h(t_i - u). \tag{15}$$

For each grid point u, we maximise (15) using the EM algorithm. In the E-step of the t^{th} iteration, we estimate the latent variable z_{ik} using the responsibilities

$$\gamma_{ik}^{(t)}(u) = \frac{\hat{\pi}_k^{(t-1)}\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k^{(t-1)} + g_{\boldsymbol{\beta}_k}^{(t-1)}(u), \hat{\sigma}_k^{2(t-1)}\}}{\sum_{\ell=1}^{K}\hat{\pi}_\ell^{(t-1)}\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_\ell^{(t-1)} + g_{\boldsymbol{\beta}_\ell}^{(t-1)}(u), \hat{\sigma}_\ell^{2(t-1)}\}}. \tag{16}$$

In the M-step, we update $\mathbf{g}_{\boldsymbol{\beta}_k}(u)$, for $u \in \mathcal{U}$, by maximising the conditional expectation of (15) given by

$$Q[\mathbf{g}^{(t)}(u)|\mathbf{g}^{(t-1)}(u)] = \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}^{(t)}(u)[\log\pi_k^{(t-1)} + \log\mathcal{N}\{y_i|\mathbf{x}_i\boldsymbol{\beta}_k^{(t-1)} + g_k(u), \sigma_k^{2(t-1)}\}]K_h(t_i - u). \tag{17}$$

Maximising $Q(\cdot|\cdot)$ with respect to $g_k(u_j)$, for each $u_j \in \mathcal{U}$, yields the following estimator:

$$g_{\boldsymbol{\beta}_k}^{(t)}(u_j) = \frac{\sum_{i=1}^{n}\gamma_{ik}^{(t)}(u_j)K_h(t_i - u_j)(y_i - \mathbf{x}_i\boldsymbol{\beta}_k)}{\sum_{i=1}^{n}\gamma_{ik}^{(t)}(u)K_h(t_i - u_j)}. \tag{18}$$

To obtain $g_{\beta_k}^{(t)}(t_i) : i = 1, 2, \ldots, n$, we interpolate over $g_{\beta_k}^{(t)}(u_j) : j = 1, 2, \ldots, N$.

Let $\mathbf{g}_{\beta_k} = (g_{\beta_k}(t_1), g_{\beta_k}(t_2), \ldots, g_{\beta_k}(t_n))$ be the estimate at convergence of the above EM algorithm. Given $\mathbf{g}_{\beta_k}$, the complete-data profile log-likelihood function corresponding to (13) is given by

$$p\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\log\pi_k + \log\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k + g_{\beta_k}(t_i), \sigma_k^2\}\right] \tag{19}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$. In the E-step, we estimate the latent variable by the responsibilities

$$r_{ik}^{(t)} = \frac{\hat{\pi}_k^{(t-1)}\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k^{(t-1)} + g_{\beta_k}(t_i), \hat{\sigma}_k^{2(t-1)}\}}{\sum_{\ell=1}^{K}\hat{\pi}_\ell^{(t-1)}\mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_\ell^{(t-1)} + g_{\beta_\ell}(t_i), \hat{\sigma}_\ell^{2(t-1)}\}} \tag{20}$$

from where we derive the expected value of (19) as

$$Q^p(\boldsymbol{\theta}_u^{(t)}|\boldsymbol{\theta}_u^{(t-1)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t)}[\log\pi_k + \log\mathcal{N}\{y_i|\mathbf{x}_i\boldsymbol{\beta}_k + g_{\beta_k}(t_i), \sigma_k^2\}]. \tag{21}$$

In the M-step, we maximise $Q^p(\cdot|\cdot)$ to update $\hat{\boldsymbol{\theta}}$. The EM update equations for π_k, $\boldsymbol{\beta}_k$ and σ_k^2 are, respectively, given as

$$\pi_k^{(t)} = \frac{\sum_{i=1}^{n} r_{ik}^{(t)}}{n} \tag{22}$$

$$\boldsymbol{\beta}_k^{(t)} = (\tilde{\mathbf{X}}_k^\intercal W_k^{(t)}\tilde{\mathbf{X}}_k)^{-1}\tilde{\mathbf{X}}_k^\intercal W_k^{(t)}\tilde{\mathbf{y}}_k \tag{23}$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} r_{ik}^{(t)}(y_i - \mathbf{x}_i\boldsymbol{\beta}_k - g_{\beta_k}(t_i))^2}{\sum_{i=1}^{n} r_{ik}^{(t)}} \tag{24}$$

where $\tilde{\mathbf{X}}_k = (\mathbf{I} - \mathbf{S}_k^\intercal)\mathbf{X}$, $\tilde{\mathbf{y}}_k = (\mathbf{I} - \mathbf{S}_k^\intercal)\mathbf{y}$, $W_k^{(t)} = diag\{r_{1k}^{(t)}, r_{2k}^{(t)}, \ldots, r_{nk}^{(t)}\}$ and $\mathbf{S}_k$, for $k = 1, 2, \ldots, K$, is a kernel smoother matrix with elements given by

$$(\mathbf{S}_k)_{ij} = \frac{r_{ik}^{(t)}K_h(t_i - u_j)}{\sum_{i=1}^{n} r_{ik}^{(t)}K_h(t_i - u_j)}. \tag{25}$$

Note that each complete-data local likelihood (15) has its own set of responsibilities $\gamma_{ik}(u)$, for each $u \in \mathcal{U}$. This is a possible source of label switching, as the component labels based on the $\gamma_{ik}(u)$s at each $u \in \mathcal{U}$ are not guaranteed to be aligned. In our simulations using this naive approach, we can observe non-smooth estimates of the non-parametric function. Thus, the approach is quite sensitive to label switching. The solution to this problem is to make use of the same responsibilities to maximise each local-likelihood function. The latter responsibilities can be used to maximise (17) at each $u \in \mathcal{U}$. The objective is to obtain this unique set of responsibilities (referred to as the global responsibilities). Using this idea, Wu and Liu [8] proposed a modified EM (PL-EM) algorithm to address this problem. Their algorithm simultaneously maximises (17) and (21). In the following, we propose a novel approach in which the local responsibilities at each grid point $\gamma_{ik}(u)$s are used to calculate the global responsibilities. The latter are chosen from the former by imposing a smoothness constraint on the estimated non-parametric function. Our approach here follows the same idea used in [20].

3.2.2. One-Step Backfitting EM (OB-PL-EM) Algorithm

We now propose a one-step backfitting profile likelihood EM (OB-PL-EM) algorithm to implement the profile likelihood estimation procedure outlined above while addressing

the label-switching problem. The algorithm proceeds in three steps. In the first step, we use Algorithm 2 in [20] to obtain the least favourable curve, $\mathbf{g}_{\beta_k}$, for $k = 1, 2, \ldots, K$.

Let $\hat{\mathbf{g}}_{\boldsymbol{\beta}}(v)$ and $\hat{\boldsymbol{\gamma}}(v)$ be the resulting estimates of $\mathbf{g}_{\boldsymbol{\beta}}$ and the corresponding set of local responsibilities, obtained at the local point $v \in \mathcal{U}$, respectively. The latter will be used as the global responsibilities in what follows.

In the second step, we use the EM algorithm to obtain the estimate of $\boldsymbol{\theta}$. Use the estimated global responsibilities $\hat{\boldsymbol{\gamma}}(v)$ as the initial responsibilities at the E-step. Calculate the smoother matrix $\mathbf{S}_k$, for $k = 1, 2, \ldots, K$ as in (25) using $\hat{\boldsymbol{\gamma}}(v)$, for each $u_j : j = 1, 2, \ldots, N$. Given $\hat{\mathbf{g}}_{\boldsymbol{\beta}}(v)$ and the kernel smoother matrix $\mathbf{S}_k : k = 1, 2, \ldots, K$, at the M-step, maximise (19) to estimate the elements of the parameter vector $\boldsymbol{\theta}$. The above E- and M-steps are repeated until convergence. We denote $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ as the estimate of $\boldsymbol{\theta}$ obtained from above.

Note that the estimate $\hat{\mathbf{g}}_{\boldsymbol{\beta}}(v)$ is based on a set of local responsibilities $\hat{\boldsymbol{\gamma}}(v)$. At the third step of the algorithm, we improve the efficiency of this estimate by maximising the following local log-likelihood function:

$$\ell(\mathbf{g}(u)) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \hat{\pi}_k \mathcal{N}\{\mathbf{x}_i \hat{\boldsymbol{\beta}}_k + g_k(u), \hat{\sigma}_k^2\} \right] K_h(t_i - u). \tag{26}$$

where $\boldsymbol{\pi}$, $\boldsymbol{\beta}$ and σ^2 are replaced with $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ in (14), respectively. To maximise (26), we alternate the following E- and M-steps until convergence. Using $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{g}}_{\boldsymbol{\beta}}(v)$ to initialise the algorithm, in the E-step, calculate the responsibilities as

$$\gamma_{ik}^{(t)} = \frac{\hat{\pi}_k \mathcal{N}\{\mathbf{x}_i \hat{\boldsymbol{\beta}}_k + g_{\beta_k}^{(t-1)}(t_i), \hat{\sigma}_k^2\}}{\sum_{\ell=1}^{K} \hat{\pi}_\ell \mathcal{N}\{\mathbf{x}_i \hat{\boldsymbol{\beta}}_\ell + g_{\beta_\ell}^{(t-1)}(t_i), \hat{\sigma}_\ell^2\}}. \tag{27}$$

In the M-step, update the $\mathbf{g}_{\beta_k} : k = 1, 2, \ldots, K$ by maximising $Q[\mathbf{g}^{(t)}(u) | \mathbf{g}^{(t-1)}(u)]$ using the responsibilities in (27) for all $u_j : j = 1, 2, \ldots, N$. The E- and M-steps are repeated until convergence.

Denote $\hat{\mathbf{g}}_{\boldsymbol{\beta}}$ as the resulting estimate. Let $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\theta}}, \hat{\mathbf{g}}_{\boldsymbol{\beta}})$. We refer to $\hat{\boldsymbol{\vartheta}}$ as the one-step backfitting profile likelihood estimator.

A summary of the above one-step backfitting PL-EM algorithm is given in Algorithm 1.

Remark 1. *Note that each stage of Algorithm 1 consists of performing a regular EM algorithm. It follows that the desired ascent property of the algorithm is achieved at each stage.*

Remark 2. *It is interesting to note at this point that if we incorporate the proposed regression spline estimation in the first stage of Algorithm 1, the proposed estimation procedure is similar to the SBK method of Zhang and Pan [16]. Thus, the results on the asymptotic behaviour of the resulting estimators $\hat{\boldsymbol{\vartheta}}$ can be assumed to follow from those presented in [16]. This is also supported by the finite sample performance of $\hat{\boldsymbol{\vartheta}}$ from our simulations.*

Remark 3. *The second stage of the algorithm becomes computationally intensive as the sample size n increases, since the number of grid points N must be equal to n. This is because the estimator $\hat{\boldsymbol{\beta}}_k$ requires the smoother matrix $\mathbf{S}_k$ to be an $n \times n$ matrix. For large sample sizes ($n > 10{,}000$ true for big datasets), it may be impossible to run the algorithm in a reasonable time. This is also true for the PL-EM algorithm proposed by Wu and Liu [8].*

In the following section, we propose an alternative estimation procedure that does not require the computation of the smoother matrix.

Algorithm 1 One-step backfitting PL-EM (OB-PL-EM) algorithm for fitting model (1)

Step 1: Estimating the non-parametric function: $g_k(t) : k = 1, 2, \ldots, K$
Repeat Steps 1–3 of Algorithm 2 in [20] until convergence to obtain $\hat{\mathbf{g}}_{\beta}(v)$ and $\hat{\gamma}(v)$.

Step 2: Estimating the parameters: π, β and σ^2

Initialisation: Given $\hat{\mathbf{g}}_{\beta}(v)$, let $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ and $\sigma^{2(0)}$ be the initial states of the parameters.
E-step: Calculate the responsibilities using (20).
M-step: Update the parameters $\boldsymbol{\pi}$, $\boldsymbol{\beta}$ and σ^2 using (22)–(24).
Iteration: Repeat the E- and M-step until convergence.

Step 3: Re-estimate the non-parametric function: $g_k(t) : k = 1, 2, \ldots, K$

Initialisation: Fix the values of the parameters $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$ and let $\mathbf{g}_{\boldsymbol{\beta}}(v)$ be the initial state of the non-parametric function.
E-Step: Calculate the responsibilities γ_{ik} using (27).
M-Step: Update $\mathbf{g}_{\boldsymbol{\beta}_k} : k = 1, 2, \ldots, K$ using (18) by making use of γ_{ik} for all $u \in \mathcal{U}$.
Iteration: Repeat the E- and M-step until convergence.

3.2.3. Profile Likelihood Plug-in EM (PL-p-EM) Algorithm

To reduce the computational burden imposed by the OB-PL-EM estimation procedure, we propose an alternative estimation procedure. To estimate $\boldsymbol{\beta}_k$, we make use of a plug-in estimator which does not require the computation of the $n \times n$ smoother matrix.

Let $\tilde{\mathbf{g}}_{\boldsymbol{\beta}_k}(t) : k = 1, 2, \ldots, K$ be an estimate of $g_k(t) : k = 1, 2, \ldots, K$ obtained as in the OB-PL-EM or PL-EM estimation procedure. To estimate $\boldsymbol{\theta}$, we maximise the following profile log likelihood function:

$$p\ell^1(\boldsymbol{\theta}) = \sum_{i=1}^{n} log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}\{\mathbf{x}_i\boldsymbol{\beta}_k + \tilde{g}_{\boldsymbol{\beta}_k}(t_i), \sigma_k^2\}\right]. \tag{28}$$

Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ be the estimator obtained from maximising (28).

The fitting algorithm to perform the above procedure proceeds in two steps. In the first step, in the t^{th} iteration, we obtain $\tilde{\mathbf{g}}_{\boldsymbol{\beta}_k}^{(t)} : k = 1, 2, \ldots, K$ as in the OB-PL-EM algorithm or PL-EM algorithm. Let $\tilde{\gamma}_{ik} : i = 1, 2, \ldots, n; k = 1, 2, \ldots, K$ be the resulting global responsibilities. In the second step, we update $\tilde{\boldsymbol{\theta}}^{(t-1)}$ by maximising the expected complete-data version of (28) using the global responsibilities. The resulting update equations for $\tilde{\pi}_k^{(t)}$ and $\tilde{\sigma^{2(t)}}_k$, for $k = 1, 2, \ldots, K$, are the same as (22) and (24), respectively. The resulting update equation for $\tilde{\boldsymbol{\beta}}_k^{(t)}$, for $k = 1, 2, \ldots, K$, is given by

$$\tilde{\boldsymbol{\beta}}_k^{(t)} \quad = \quad (\mathbf{X}^\intercal W_k^{(t)} \mathbf{X})^{-1} \mathbf{X}^\intercal W_k^{(t)} (\mathbf{y} - \tilde{\mathbf{g}}_{\boldsymbol{\beta}_k}^{(t)}). \tag{29}$$

From a comparison of (23) and (29), it is not hard to see why we refer to $\tilde{\boldsymbol{\beta}}_k$ as a plug-in estimator. The asymptotic behaviour of $\tilde{\boldsymbol{\beta}}_k$ has not yet been studied. However, we are encouraged by its finite sample performance presented in our simulations.

Let $\tilde{\boldsymbol{\vartheta}} = (\tilde{\mathbf{g}}_{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$ be the resulting estimator of $\boldsymbol{\vartheta}$. We refer to $\tilde{\boldsymbol{\vartheta}}$ as the profile likelihood plug-in estimator. The above estimation procedure is summarised in Algorithm 2.

Algorithm 2 Profile likelihood plug-in EM-type (PL-p-EM) algorithm for fitting model (1)

Initialisation: Let $\tilde{\boldsymbol{\theta}}^{(0)}$ and $\tilde{\mathbf{g}}_{\boldsymbol{\beta}}^{(0)}$ be the initial parameter vector and non-parametric function. These can be obtained using the proposed regression-spline based estimator, for instance.
Step 1: At the t^{th} iteration, obtain $\tilde{\mathbf{g}}_{\boldsymbol{\beta}_k}^{(t)}$ and the resulting global responsibilities $\tilde{\gamma}_{ik}^{(t)}$, for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$. This can be done as in the OB-PL-EM or PL-EM algorithm for $N < n$.
Step 2: Using the global responsibilities, update $\tilde{\boldsymbol{\beta}}^{(t-1)}$, $\tilde{\pi}^{(t-1)}$ and $\tilde{\sigma}^{2(t-1)}$ using (29), (22) and (24), respectively.
Iteration: Repeat Steps 1 and 2 until convergence.

4. Choosing the Bandwidth

To estimate the non-parametric function $g_k(\cdot)$, we need to choose an appropriate value for the smoothing parameter, h. In practice, this is usually data-dependent based on the cross-validation (CV) or generalised CV (GCV). For estimating model (1), Wu and Liu [8] proposed a multi-fold CV approach to choose h. For the $K = 1$ case, Speckman [21] proposed a GCV approach to choose h and provided theoretical evidence to support its application. As for its simplicity, we also propose a GCV method to choose h for estimating model (1). GCV provides a data-based estimate of h in order to minimise the following unobservable mean squared error (MSE):

$$\text{MSE}(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}[m_k(x_i, t_i) - \hat{m}_k(x_i, t_i)]^2 \tag{30}$$

where $m_k(\cdot,\cdot)$ and $\hat{m}_k(\cdot,\cdot)$ are the regression function and its estimator for the k^{th} component. In matrix notation, $\hat{m}_k(\cdot,\cdot)$ can be expressed as

$$\begin{aligned}\hat{\mathbf{M}}_k &= \mathbf{X}\hat{\boldsymbol{\beta}}_k + \hat{\mathbf{g}}_{\boldsymbol{\beta}_k} \\ &= \mathbf{X}\hat{\boldsymbol{\beta}}_k + \mathbf{S}_k(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_k) \\ &= \mathbf{A}_k\mathbf{y}\end{aligned} \tag{31}$$

where $\mathbf{A}_k = \mathbf{S}_k + \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\intercal R_k\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\mathbf{R}_k(\mathbf{I} - \mathbf{S}_k)$ is a linear smoother matrix (see Buja et al. [27] for more details on linear smoothers). We here define the GCV function as

$$\text{GCV}(h) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(1 - \text{df}/n)^2} \tag{32}$$

where

$$\text{df} = \sum_{k=1}^{K}\text{trace}(\mathbf{A}_k) \tag{33}$$

$$\hat{y}_i = \sum_{k=1}^{K}\gamma_{ik}\{\mathbf{x}_i\boldsymbol{\beta}_k + g_{\boldsymbol{\beta}_k}(t_i)\} \tag{34}$$

denote the degrees of freedom and the i^{th} fitted value, respectively. In analogy with parametric regression, df represents the effective number of parameters used to estimate the regression function. The GCV criteria selects the bandwidth that minimises GCV(h).

5. Simulations

We performed an extensive simulation study to demonstrate the finite sample performance of the methods proposed in this paper. Throughout our simulations, we considered a $K = 2$ component mixture environment and a univariate $\mathbf{x}$ (that is, $p = 1$), denoted x. We compare the performance of the proposed procedure (OB-PL-EM) with that of the PL-EM procedure proposed by Wu and Liu [8]. To initialise the algorithms, we made use of the regression spline-based estimator (R-spline-EM) with $Q = 3$ internal knots chosen to be the 1st, 2nd and 3rd quartiles of the covariate t. In order to improve the stability of the model estimate and alleviate the issue of the dependence on the initial solution, we made use of the following initialisation strategy: Fit a mixture of regression splines for a 100 times from random starts and choose as the initial solution the model with the smallest BIC. We also show the results obtained using the plug-in (PL-p-EM) procedure. The algorithms were implemented using the R programming language (version 3.6.1 released 2019-07-05 [28]). The *bs* function from the splines R package was used within the R-spline-EM function to compute the basis functions. Throughout our simulations, the two covariates, x and t, were generated from a uniform distribution on the interval $(0, 1)$. We generated 500 samples of sizes $n = 200$, 400, 800 and 1000. We made use of the Epanechnikov kernel function and

$N = 100$ grid points. The set of grid points $\mathcal{U}$ was chosen uniformly from the domain of the covariate t.

5.1. Performance Assessment

To assess the performance of the estimator $\hat{g}_k(\cdot)$ or $\tilde{g}_k(\cdot)$, we make use of the root of the average squared errors (RASE):

$$\text{RASE}^2_{\mathbf{g}_\beta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}[g_k(t_i) - \hat{g}_k(t_i)]^2. \tag{35}$$

For the parametric estimators, we made use of the MSE:

$$\text{MSE}_\theta = (\hat{\theta} - \theta)^2 \tag{36}$$

and Bias

$$\text{Bias}_\theta = \hat{\theta} - \theta. \tag{37}$$

5.2. Simulation Study

The first aim of this simulation study was to illustrate the performance of the proposed one-step backfitting profile likelihood estimators and the profile likelihood plug-in estimators. The data used in this example were generated from model (1) using the $K = 2$ component setting given in Table 1.

Table 1. The $K = 2$ component SPMPLMs.

k	1	2
$m_k(x,t)$	$\beta_1 x + g_1(t)$	$\beta_2 x + g_2(t)$
$g_k(t)$	$\exp(2t-1)+1$	$\sin(2\pi t)$
σ_k^2	0.09	0.04
π_k	0.65	0.35
β_k	5	-1

To show the effectiveness of the proposed estimation procedure, we considered the following bandwidths: $\frac{2}{3}h_{GCV}$, h_{GCV} and $\frac{3}{2}h_{GCV}$ corresponding to under-smoothing (US), appropriate smoothing (AS) and over-smoothing (OS), respectively, where h_{GCV} denotes the bandwidth selected by the GCV method. Tables 2–5 reports the averages and standard deviations of the performance measures. The results show that the proposed estimation procedures have generally good performance. The proposed profile likelihood-based estimation procedures provided similar results to the PL-EM estimation procedure under all three bandwidths. The regression spline-based estimation procedure performed similarly to the profile likelihood procedure for estimating the parametric part. However, the R-spline-EM procedure generally has better performance when estimating the non-parametric part. This shows that it is a good choice for initialising the profile likelihood-based procedures.

Figure 1 presents the non-parametric function estimates based on the profile likelihood-based estimators for a typical sample of size $n = 400$. Included in Figure 1 are also the 95% point-wise bootstrap confidence intervals (CI). Based on the fitted model $\hat{\pi}_1\mathcal{N}\{y|x\hat{\beta}_1 + \hat{g}_1(t), \hat{\sigma_1^2}\} + \hat{\pi}_2\mathcal{N}\{y|x\hat{\beta}_2 + \hat{g}_2(t), \hat{\sigma_2^2}\}$, for each (x_i, t_i), generate y_i^* for all $i = 1, 2, \ldots, n$. Let $\{(y_i^*, x_i, t_i) : i = 1, 2, \ldots, n\}$ be the bootstrap sample obtained in the above manner. We repeated this process 1000 times. The component function estimates are virtually the same. Noticeably, the 95% point-wise bootstrap CIs based on the parametric estimator $\hat{\boldsymbol{\beta}}_k$ do not contain the points at the boundaries. This might indicate that the non-parametric estimator based on this parametric estimator has boundary bias. This in turn gives support for the usefulness of the plug-in estimator $\tilde{\boldsymbol{\beta}}_k$.

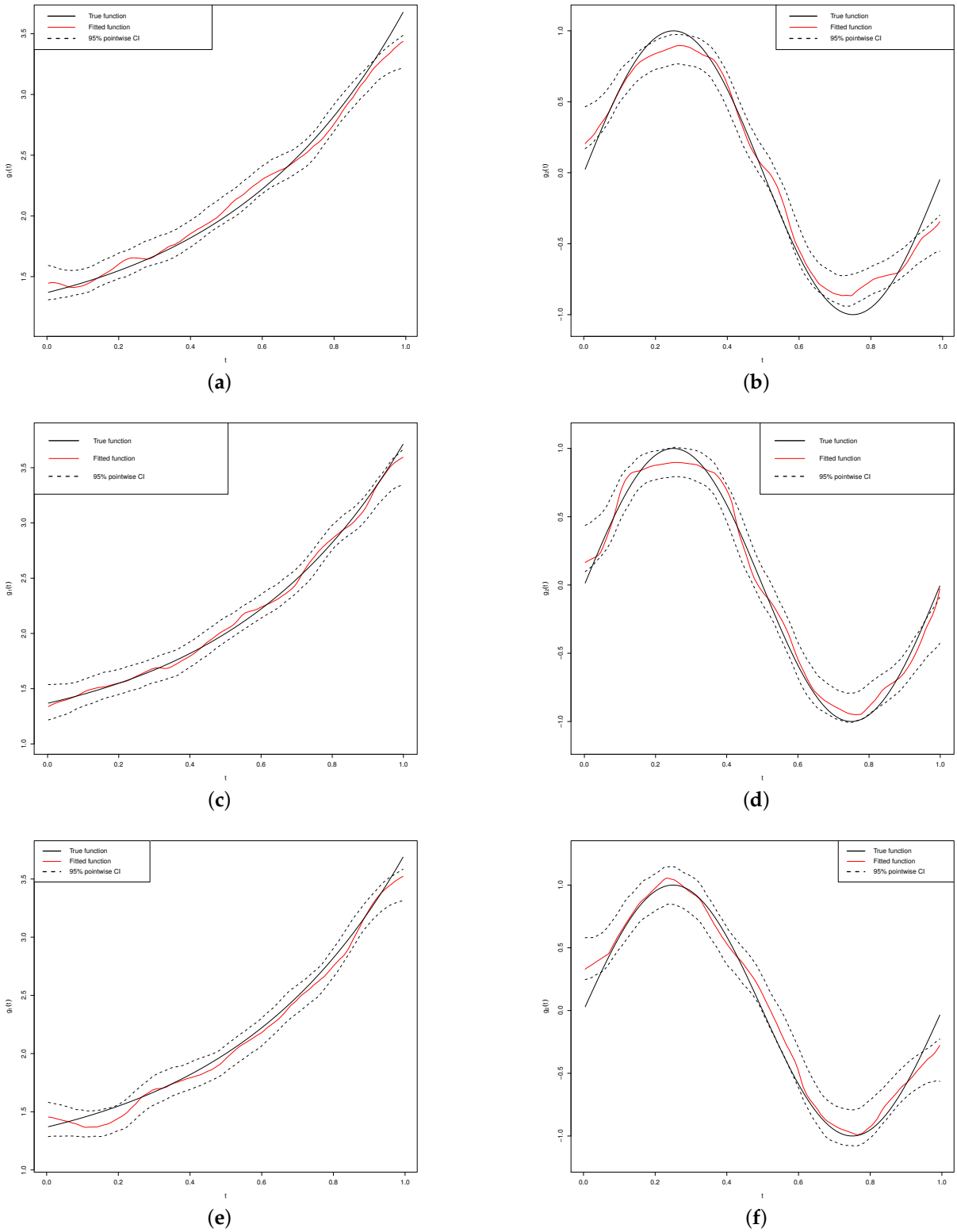


Figure 1. Fitted non-parametric component functions (red solid curve) for a typical sample of size $n = 400$ based on the (**a**,**b**) OB-PL-EM algorithm, (**c**,**d**) PL-p-EM algorithm and (**e**,**f**) PL-EM algorithm. The black solid curve gives the true component function. The dotted lines give the 95% bootstrap confidence intervals.

Table 2. Averages (and standard deviations) of the performance measures for $n = 200$.

	h	$\mathbf{MSE}_{\beta_1}$	$\mathbf{MSE}_{\beta_2}$	$\mathbf{MSE}_{\pi_1}$	$\mathbf{MSE}_{\sigma_1^2}$	$\mathbf{MSE}_{\sigma_2^2}$	$\mathbf{Bias}_{\beta_1}$	$\mathbf{Bias}_{\beta_2}$	$\mathbf{Bias}_{\pi_1}$	$\mathbf{Bias}_{\sigma_1^2}$	$\mathbf{Bias}_{\sigma_2^2}$	$\mathbf{RASE}_{\mathbf{g}_\beta}$
OB-PL-EM	US	0.0087 (0.0123)	0.0090 (0.0117)	0.0010 (0.0015)	0.0002 (0.0015)	0.0001 (0.0001)	−0.0071 (0.0929)	−0.0106 (0.0946)	−0.0015 (0.0319)	−0.0085 (0.0109)	−0.0024 (0.0078)	0.1403 (0.0244)
	AS	0.0099 (0.0135)	0.0080 (0.0112)	0.0011 (0.0015)	0.0001 (0.0002)	0.0001 (0.0002)	−0.0082 (0.0990)	−0.0142 (0.0881)	−0.0007 (0.0333)	−0.0046 (0.0105)	0.0037 (0.0078)	0.1405 (0.0247)
	OS	0.0091 (0.0129)	0.0096 (0.0149)	0.0010 (0.0015)	0.0001 (0.0002)	0.0003 (0.0004)	−0.0074 (0.0953)	−0.0055 (0.0978)	−0.0008 (0.0324)	−0.0012 (0.0117)	0.0148 (0.0097)	0.1747 (0.0264)
PL-p-EM	US	0.0086 (0.0120)	0.0093 (0.0129)	0.0010 (0.0015)	0.0002 (0.0002)	0.0001 (0.0001)	−0.0010 (0.0930)	−0.0099 (0.0962)	−0.0014 (0.0319)	−0.0088 (0.0109)	−0.0042 (0.0064)	0.1407 (0.0246)
	AS	0.0100 (0.0136)	0.0095 (0.0137)	0.0011 (0.0015)	0.0001 (0.0002)	0.0000 (0.0001)	0.0042 (0.0999)	−0.0120 (0.0968)	−0.0005 (0.0333)	−0.0048 (0.0105)	0.0010 (0.0065)	0.1425 (0.0266)
	OS	0.0096 (0.0140)	0.0134 (0.0221)	0.0010 (0.0015)	0.0001 (0.0002)	0.0002 (0.0003)	0.0185 (0.0964)	−0.0056 (0.1159)	−0.0006 (0.0324)	−0.0012 (0.0117)	0.0120 (0.0088)	0.1799 (0.0300)
PL-EM	US	0.0086 (0.0122)	0.0089 (0.0117)	0.0010 (0.0015)	0.0002 (0.0002)	0.0001 (0.0001)	−0.0071 (0.0928)	−0.0089 (0.0941)	−0.0014 (0.0319)	−0.0088 (0.0108)	−0.0043 (0.0063)	0.1402 (0.0243)
	AS	0.0098 (0.0134)	0.0078 (0.0110)	0.0011 (0.0015)	0.0001 (0.0002)	0.0000 (0.0001)	−0.0078 (0.0990)	−0.0119 (0.0878)	−0.0006 (0.0333)	−0.0048 (0.0105)	0.0009 (0.0065)	0.1404 (0.0247)
	OS	0.0091 (0.0130)	0.0095 (0.0146)	0.0010 (0.0015)	0.0001 (0.0002)	0.0002 (0.0003)	−0.0069 (0.0954)	−0.0026 (0.0973)	−0.0006 (0.0324)	−0.0013 (0.0117)	0.0118 (0.0087)	0.1746 (0.0262)
R-Spline-EM		0.0091 (0.0128)	0.0082 (0.0118)	0.0010 (0.0015)	0.0002 (0.0002)	0.0001 (0.0001)	−0.0081 (0.0951)	−0.0094 (0.0899)	−0.0011 (0.0325)	−0.0057 (0.0110)	−0.0033 (0.0067)	0.1191 (0.0385)

Table 3. Averages (and standard deviations) of the performance measures for $n = 400$.

	h	$\mathbf{MSE}_{\beta_1}$	$\mathbf{MSE}_{\beta_2}$	$\mathbf{MSE}_{\pi_1}$	$\mathbf{MSE}_{\sigma_1^2}$	$\mathbf{MSE}_{\sigma_2^2}$	$\mathbf{Bias}_{\beta_1}$	$\mathbf{Bias}_{\beta_2}$	$\mathbf{Bias}_{\pi_1}$	$\mathbf{Bias}_{\sigma_1^2}$	$\mathbf{Bias}_{\sigma_2^2}$	$\mathbf{RASE}_{\mathbf{g}_\beta}$
OB-PL-EM	US	0.0052 (0.0080)	0.0043 (0.0060)	0.0005 (0.0008)	0.0001 (0.0001)	0.0001 (0.0002)	−0.0103 (0.0718)	−0.0150 (0.0639)	−0.0022 (0.0233)	−0.0052 (0.0077)	0.0044 (0.0085)	0.1034 (0.0176)
	AS	0.0047 (0.0065)	0.0037 (0.0059)	0.0006 (0.0008)	0.0001 (0.0001)	<0.0001 (0.0001)	−0.0044 (0.0688)	−0.0045 (0.0609)	<0.0001 (0.0236)	−0.0039 (0.0080)	0.0013 (0.0051)	0.1050 (0.0412)
	OS	0.0046 (0.0069)	0.0046 (0.0063)	0.0006 (0.0008)	0.0001 (0.0001)	0.0002 (0.0002)	−0.0081 (0.0674)	−0.0123 (0.0671)	−0.0026 (0.0238)	−0.0004 (0.0077)	0.0106 (0.0066)	0.1282 (0.0184)
PL-p-EM	US	0.0050 (0.0077)	0.0042 (0.0060)	0.0005 (0.0008)	0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0053 (0.0702)	−0.0096 (0.0642)	−0.0017 (0.0234)	−0.0054 (0.0076)	−0.0023 (0.0047)	0.1031 (0.0172)
	AS	0.0048 (0.0063)	0.0039 (0.0060)	0.0006 (0.0008)	0.0001 (0.0001)	<0.0001 (<0.0001)	0.0031 (0.0689)	−0.0052 (0.0625)	−0.0001 (0.0236)	−0.0040 (0.0080)	0.0009 (0.0049)	0.1056 (0.0412)
	OS	0.0048 (0.0071)	0.0062 (0.0083)	0.0006 (0.0008)	0.0001 (0.0001)	0.0001 (0.0001)	0.0108 (0.0684)	−0.0112 (0.0778)	−0.0024 (0.0238)	−0.0006 (0.0077)	0.0070 (0.0052)	0.1308 (0.0200)
PL-EM	US	0.0052 (0.0080)	0.0040 (0.0054)	0.0005 (0.0008)	0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0088 (0.0717)	−0.0092 (0.0628)	−0.0017 (0.0234)	−0.0054 (0.0076)	−0.0023 (0.0047)	0.1030 (0.0172)
	AS	0.0048 (0.0065)	0.0038 (0.0061)	0.0006 (0.0008)	0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0051 (0.0689)	−0.0055 (0.0613)	−0.0001 (0.0236)	−0.0040 (0.0080)	0.0009 (0.0049)	0.1051 (0.0413)
	OS	0.0046 (0.0069)	0.0046 (0.0062)	0.0006 (0.0008)	0.0001 (0.0001)	0.0001 (0.0001)	−0.0076 (0.0674)	−0.0100 (0.0671)	−0.0025 (0.0237)	−0.0006 (0.0077)	0.0070 (0.0052)	0.1282 (0.0184)
R-Spline-EM		0.0047 (0.0069)	0.0042 (0.0082)	0.0006 (0.0008)	0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0073 (0.0683)	−0.0106 (0.0635)	−0.0016 (0.0235)	−0.0031 (0.0078)	−0.0010 (0.0048)	0.0817 (0.0204)

Table 4. Averages (and standard deviations) of the performance measures for $n = 800$.

	h	$\mathbf{MSE}_{\beta_1}$	$\mathbf{MSE}_{\beta_2}$	$\mathbf{MSE}_{\pi_1}$	$\mathbf{MSE}_{\sigma_1^2}$	$\mathbf{MSE}_{\sigma_2^2}$	$\mathbf{Bias}_{\beta_1}$	$\mathbf{Bias}_{\beta_2}$	$\mathbf{Bias}_{\pi_1}$	$\mathbf{Bias}_{\sigma_1^2}$	$\mathbf{Bias}_{\sigma_2^2}$	$\mathbf{RASE}_{\mathbf{g}_\beta}$
OB-PL-EM	US	0.0021 (0.0030)	0.0026 (0.0036)	0.0003 (0.0004)	<0.0001 (<0.0001)	0.0001 (0.0002)	−0.0099 (0.0450)	−0.0197 (0.0474)	−0.0026 (0.0159)	−0.0040 (0.0057)	0.0077 (0.0078)	0.0789 (0.0116)
	AS	0.0022 (0.0030)	0.0020 (0.0029)	0.0003 (0.0004)	<0.0001 (<0.0001)	0.0001 (0.0001)	−0.0116 (0.0450)	−0.0136 (0.0421)	−0.0021 (0.0168)	−0.0022 (0.0054)	0.0060 (0.0065)	0.0756 (0.0120)
	OS	0.0022 (0.0030)	0.0022 (0.0030)	0.0003 (0.0004)	<0.0001 (<0.0001)	0.0001 (0.0001)	−0.0054 (0.0465)	−0.0138 (0.0450)	−0.0015 (0.0172)	−0.0012 (0.0056)	0.0092 (0.0054)	0.0906 (0.0122)
PL-p-EM	US	0.0021 (0.0029)	0.0023 (0.0030)	0.0003 (0.0004)	<0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0051 (0.0452)	−0.0111 (0.0464)	−0.0017 (0.0160)	−0.0041 (0.0057)	−0.0016 (0.0034)	0.0781 (0.0112)
	AS	0.0021 (0.0028)	0.0019 (0.0028)	0.0003 (0.0004)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0047 (0.0451)	−0.0084 (0.0432)	−0.0017 (0.0169)	−0.0023 (0.0054)	0.0002 (0.0033)	0.0754 (0.0118)
	OS	0.0023 (0.0031)	0.0024 (0.0033)	0.0003 (0.0004)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	0.0083 (0.0468)	−0.0085 (0.0486)	−0.0009 (0.0172)	−0.0013 (0.0056)	0.0032 (0.0035)	0.0912 (0.0122)
PL-EM	US	0.0021 (0.0030)	0.0022 (0.0029)	0.0003 (0.0004)	<0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0073 (0.0451)	−0.0104 (0.0458)	−0.0018 (0.0160)	−0.0041 (0.0057)	−0.0016 (0.0034)	0.0780 (0.0111)
	AS	0.0021 (0.0029)	0.0018 (0.0027)	0.0003 (0.0004)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0101 (0.0448)	−0.0086 (0.0419)	−0.0017 (0.0169)	−0.0023 (0.0054)	0.0002 (0.0033)	0.0753 (0.0118)
	OS	0.0022 (0.0030)	0.0021 (0.0028)	0.0003 (0.0004)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0037 (0.0463)	−0.0085 (0.0446)	−0.0010 (0.0172)	−0.0013 (0.0056)	0.0032 (0.0035)	0.0903 (0.0120)
R-Spline-EM		0.0021 (0.0030)	0.0021 (0.0031)	0.0003 (0.0004)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0067 (0.0457)	−0.0109 (0.0444)	−0.0016 (0.0167)	−0.0019 (0.0056)	−0.0002 (0.0035)	0.0577 (0.0141)

Table 5. Averages (and standard deviations) of the performance measures for $n = 1000$.

	h	$\mathbf{MSE}_{\beta_1}$	$\mathbf{MSE}_{\beta_2}$	$\mathbf{MSE}_{\pi_1}$	$\mathbf{MSE}_{\sigma_1^2}$	$\mathbf{MSE}_{\sigma_2^2}$	$\mathbf{Bias}_{\beta_1}$	$\mathbf{Bias}_{\beta_2}$	$\mathbf{Bias}_{\pi_1}$	$\mathbf{Bias}_{\sigma_1^2}$	$\mathbf{Bias}_{\sigma_2^2}$	$\mathbf{RASE}_{\mathbf{g}_\beta}$
OB-PL-EM	US	0.0020 (0.0027)	0.0020 (0.0029)	0.0002 (0.0004)	<0.0001 (0.0001)	<0.0001 (0.0002)	−0.0105 (0.0433)	−0.0198 (0.0399)	−0.0009 (0.0153)	−0.0041 (0.0052)	0.0081 (0.0084)	0.0784 (0.0095)
	AS	0.0017 (0.0026)	0.0021 (0.0030)	0.0002 (0.0003)	<0.0001 (<0.0001)	0.0001 (0.0002)	−0.0087 (0.0406)	−0.0215 (0.0399)	−0.0017 (0.0148)	−0.0031 (0.0047)	0.0082 (0.0073)	0.0699 (0.0105)
	OS	0.0017 (0.0023)	0.0018 (0.0025)	0.0003 (0.0003)	<0.0001 (<0.0001)	0.0001 (0.0001)	−0.0057 (0.0406)	−0.0161 (0.0397)	−0.0012 (0.0161)	−0.0021 (0.0049)	0.0082 (0.0058)	0.0705 (0.0106)
PL-p-EM	US	0.0019 (0.0026)	0.0015 (0.0022)	0.0002 (0.0004)	<0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0066 (0.0435)	−0.0097 (0.0381)	<0.0001 (0.0153)	−0.0042 (0.0051)	−0.0021 (0.0029)	0.0774 (0.0090)
	AS	0.0017 (0.0025)	0.0016 (0.0023)	0.0002 (0.0003)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0034 (0.0407)	−0.0126 (0.0386)	−0.0009 (0.0148)	−0.0031 (0.0047)	−0.0008 (0.0030)	0.0690 (0.0098)
	OS	0.0017 (0.0022)	0.0017 (0.0023)	0.0003 (0.0003)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	0.0029 (0.0409)	−0.0093 (0.0400)	−0.0006 (0.0161)	−0.0021 (0.0049)	0.0008 (0.0031)	0.0703 (0.0103)
PL-EM	US	0.0019 (0.0026)	0.0015 (0.0022)	0.0002 (0.0004)	<0.0001 (0.0001)	<0.0001 (<0.0001)	−0.0077 (0.0433)	−0.0096 (0.0380)	<0.0001 (0.0153)	−0.0042 (0.0051)	−0.0021 (0.0029)	0.0775 (0.0090)
	AS	0.0017 (0.0025)	0.0016 (0.0022)	0.0002 (0.0003)	<0.0001 (0.0003)	<0.0001 (<0.0001)	−0.0061 (0.0405)	−0.0126 (0.0380)	−0.0009 (0.0148)	−0.0031 (0.0047)	−0.0008 (0.0030)	0.0689 (0.0098)
	OS	0.0017 (0.0022)	0.0016 (0.0022)	0.0003 (0.0003)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0036 (0.0406)	−0.0094 (0.0392)	−0.0006 (0.0161)	−0.0021 (0.0049)	0.0008 (0.0031)	0.0700 (0.0102)
R-Spline-EM		0.0018 (0.0030)	0.0019 (0.0040)	0.0002 (0.0003)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	−0.0050 (0.0421)	−0.0129 (0.0410)	−0.0006 (0.0154)	−0.0017 (0.0049)	0.0001 (0.0031)	0.0516 (0.0135)

In terms of computational time, the average run times (in seconds) of the PL-p-EM and PL-EM algorithms were 1.67 and 7.65, 6.14 and 31.49 and 7.7 and 54.65 for $n = 400$, 800 and 1000, respectively.

The second aim of this simulation study was to demonstrate the effectiveness of the proposed GCV method for smoothing parameter selection. We used the same sampling settings as in Table 1. We generated 500 samples, and for each sample we obtained h_{GCV} over a reasonable range of bandwidths. We then randomly split the 500 selected bandwidths into 10 groups, each consisting of 50. Table 6 reports the average and standard deviation for each of the 10 groups for sample sizes $n = 200, 400$ and 800. We can see that the method is consistent, as the variation between the selected bandwidth from one group to another is small. This in turn shows the effectiveness of the method. The last column gives the average and standard deviation of the 10 average bandwidths for each sample size.

Table 6. Averages (and standard deviations) of the bandwidths.

n	Group										Ave ($\bar{h}$)
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
200	0.104 (0.015)	0.096 (0.014)	0.101 (0.014)	0.098 (0.016)	0.099 (0.017)	0.101 (0.015)	0.099 (0.016)	0.102 (0.015)	0.099 (0.018)	0.097 (0.013)	**0.100 (0.002)**
400	0.080 (0.012)	0.082 (0.012)	0.082 (0.013)	0.080 (0.012)	0.078 (0.011)	0.080 (0.013)	0.081 (0.012)	0.082 (0.012)	0.077 (0.011)	0.081 (0.011)	**0.080 (0.002)**
800	0.068 (0.010)	0.063 (0.011)	0.068 (0.009)	0.064 (0.011)	0.069 (0.011)	0.063 (0.011)	0.068 (0.010)	0.064 (0.009)	0.067 (0.011)	0.064 (0.010)	**0.066 (0.002)**

6. Application

In this section, we demonstrate the practical usefulness of the methods proposed in this paper using two real data examples. The data were obtained from the Our World In Data database. To access these dataset, see the data availability statement section. The data comprises CO_2 emissions per capita (CO_2), oil consumption per capita (EnergyUse), the number of people living in urban areas (Urbanisation), real GDP per capita (GDP-per-capita) and the percentage share of primary energy attributable to renewable energy sources (RenewEnergyShare) for the period 1990 to 2019 for 7–8 OECD countries.

6.1. Climate Data 1

For our first application, we considered the impacts of EnergyUse and Urbanisation on CO_2. After pre-processing the data, we produced scatter plots of the data; see Figure 2. A two-regime (component) structure is clearly evident in the figure. Moreover, there appears to be a non-linear relationship between CO_2 and Urbanisation within each regime.

In the first regime, including countries such as Australia, per capita CO_2 emissions increase sharply up to a point and then they gradually decrease.

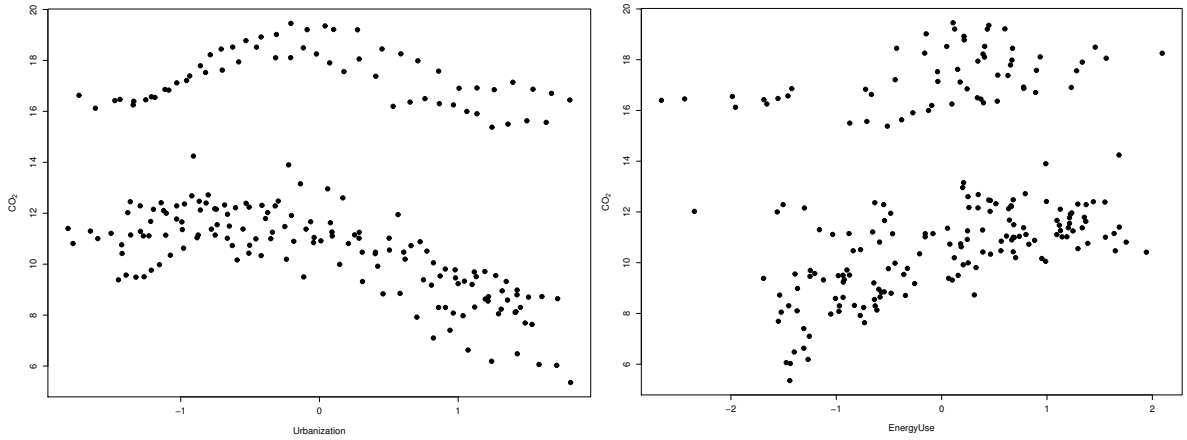


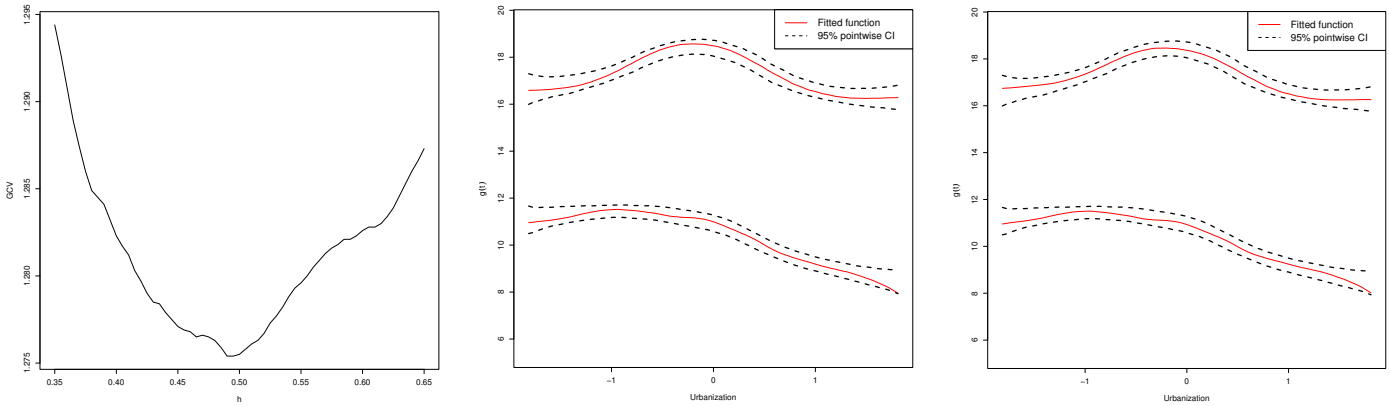
Figure 2. Scatter plots of climate data 1.

In the second regime, including countries such as Denmark, per capita CO_2 emissions increased gradually up to a point, followed by a sharp decline. This naive interpretation of the data suggests that the relationship between CO_2 and Urbanisation in the two regimes exhibits a form of the environmental Kuznets curve (EKC) (see [29]). This could be naively explained as implying that the increased number of people in the urban areas (which is where corporate and industrial activities take place) translates into an increase in skilled human capital. The latter in turn leads to an increase in productivity.

For the data in Figure 2, we propose to fit the following model:

$$\pi_1 \mathcal{N}\{y|x\beta_1 + g_1(t), \sigma_1^2\} + \pi_2 \mathcal{N}\{y|x\beta_2 + g_2(t), \sigma_2^2\} \tag{38}$$

where y, x and t are CO_2, Energyuse and Urbanisation, respectively. Figure 3a plots the GCV(h) over a range of bandwidths, and the minimum GCV(h) occurs at 0.4925. Figure 3b,c show the estimated non-parametric functions based on the OB-PL-EM procedure and the PL-p-EM procedure, respectively. It can be seen that there is virtually no difference between the estimated functions using the two procedures. This shows that the plug-in estimator (29) works as well as the estimator (9). Included in Figure 3b,c are plots of the 95% point-wise confidence intervals obtained via bootstrapping.



(**a**) (**b**) (**c**)

Figure 3. Fitted model (38): (**a**) The GCV plot over a range of bandwidths with a minimum at $h = 0.4925$. (**b**) estimated non-parametric functions (red solid lines) $\hat{g}_k(t) : k = 1, 2$ based on the OB-PL-EM procedure and (**c**) the same as in (**b**) using the PL-p-EM procedure. The dashed lines are the 95% point-wise confidence intervals.

We then checked whether the data can be explained by a simple mixture of linear regressions model. More specifically, we checked whether the non-parametric function $g_k(t)$ has a linear structure. Mathematically, we wanted to test the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: g_k(t) = \alpha_{0k} + \alpha_{1k}(t) \qquad \text{for } k = 1, 2 \\ \mathrm{H}_a &: g_k(t) \text{ is a smooth function} \end{aligned}$$

To test these hypotheses, we made use of the bootstrap specification test proposed by Wu and Liu [8]. Define the test statistic as

$$T_n = \sum_{i=1}^{n} (\hat{y}_i - \tilde{y}_i)^2 \tag{39}$$

where $\hat{y}$ and $\tilde{y}$ are the fitted values (defined as in Equation (34)) obtained from fitting the model under the null and alternative hypotheses, respectively. To ensure that our test results would not be sensitive to the choice of the bandwidth, we performed the test using the bandwidths $\frac{2}{3}h_{GCV}$, h_{GCV} and $\frac{3}{2}h_{GCV}$, where $h_{GCV} = 0.4925$. Respectively, the observed test statistics were 48.5473, 48.4349 and 48.3420 with p-values 0.01, 0.022 and 0.02. At a 5% level of significance, we can reject the null hypothesis. Therefore, model (38) provides an adequate fit for these data.

Table 7 gives the estimated parameters of the fitted model (38). Included in the table are the 95% confidence intervals obtained via the bootstrap procedure outlined above.

The first thing to note is that the estimated slope parameters are positive, as expected (see Figure 2). The second thing to note is that the 95% confidence interval for the slope parameter of the first component includes a zero. This implies that the parameter may not be significant.

To obtain further evidence in support of this conclusion, we conducted the following hypotheses test:

$$\begin{aligned} \mathrm{H}_0 &: \beta_1 = 0 \\ \mathrm{H}_a &: \beta_1 \neq 0 \end{aligned}$$

Under the null hypothesis, we fit the following reduced model:

$$\pi_1 \mathcal{N}\{y|g_1(t), \sigma_1^2\} + \pi_2 \mathcal{N}\{y|x\beta_2 + g_2(t), \sigma_2^2\} \tag{40}$$

and under the alternative hypothesis we fit model (38). The test statistic was similarly defined as in (39). The observed test statistic is 0.0015 with a p-value of 0.426. Thus, we cannot reject the null hypothesis. We can therefore conclude that the reduced model (40) provides an adequate fit for the data. The estimated value of β_2 is 0.4337, which is virtually the same as the one in Table 7. The same applies for the rest of the parameter estimates. The fitted non-parametric functions are similar to those obtained for the fitted model (38), and therefore, they are omitted.

Table 7. Parameter estimates of the fitted model (38) and the corresponding 95% bootstrap confidence intervals.

	β_1		β_2		π_1		σ_1^2		σ_2^2	
	0.126		0.433		0.286		0.288		0.827	
95% CI	−0.043	0.333	0.243	0.627	0.229	0.343	0.178	0.396	0.619	0.984

6.2. Climate Data 2

For our second application, we consider the impact of GDP-per-capita and RenewEnergyShare on CO_2. After pre-processing the data, we produced scatter plots of the data in Figure 4. From the figure, we can see that there are at least two components. Moreover, it seems that the relationship between CO_2 and GDP-per-capita is non-linear. For this data,

we propose to fit the SPMPLMs. We first obtained the number of components using the BIC as in Wu and Liu [8]. We fit the SPMPLMs for $K = 2, 3, 4$ and 5 and choose the model with the smallest BIC. The BIC scores are 939.1803, 967.2592, 1133.067 and 1220.576, respectively. We therefore fit a $K = 2$ component SPMPLMs. We use y, t and x to denote CO_2, GDP-per-capita and RenewEnergyShare, respectively. Figure 5a shows the GCV plot with the minimum GCV occurring at $h = 0.475$. Figure 5b,c gives the fitted non-parametric functions for the two components, respectively, obtained using the OB-PL-EM estimation procedure.

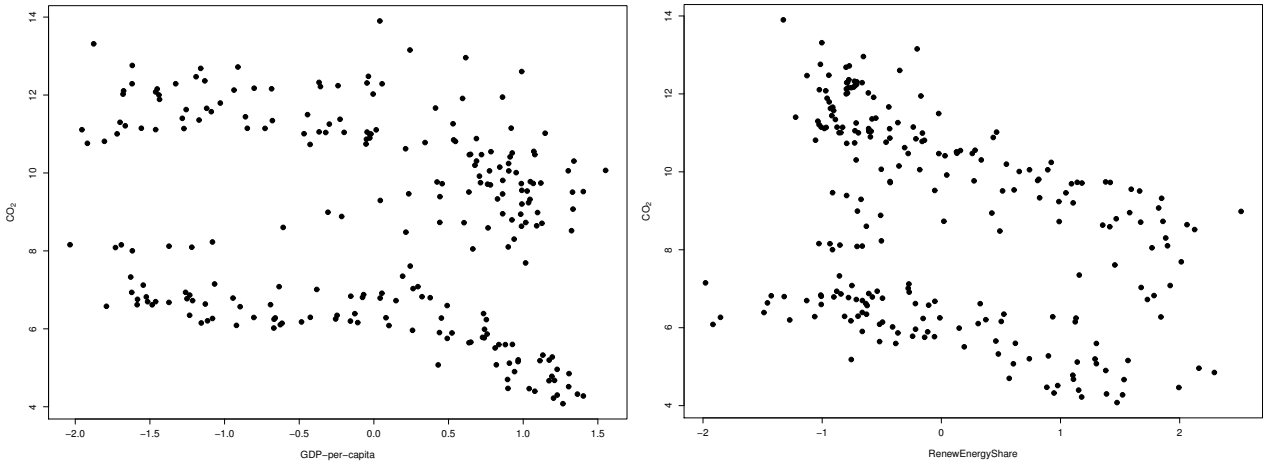
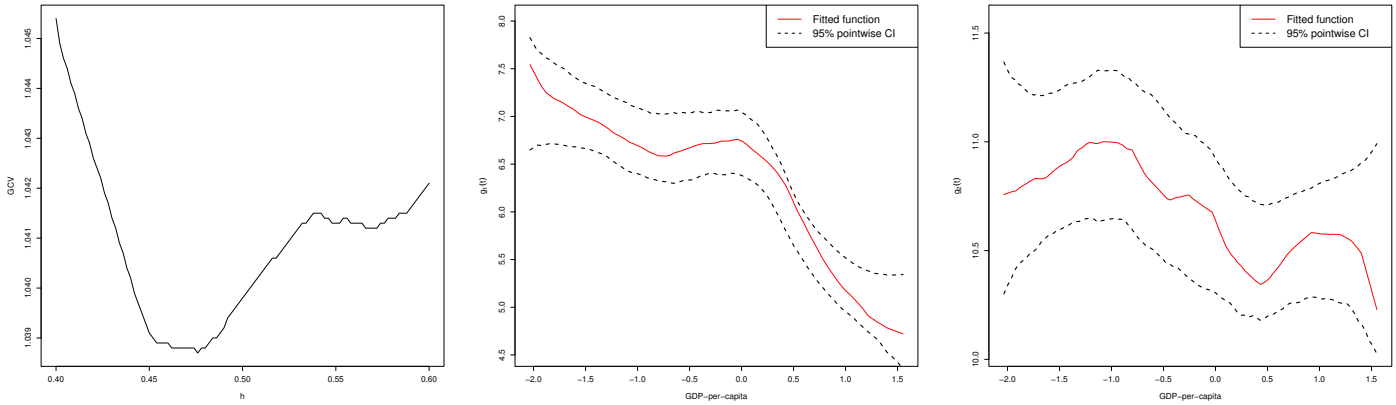


Figure 4. Scatter plots of the climate data 2.



(**a**) (**b**) (**c**)

Figure 5. Fitted model: (**a**) The GCV plot over a range of bandwidths with a minimum at $h = 0.475$. (**b**) Estimated non-parametric function (red solid lines) $\hat{g}_1(t)$ and (**c**) estimated non-parametric function (red solid lines) $\hat{g}_2(t)$. The dashed lines are the 95% point-wise confidence intervals.

As in the previous application, we conducted a hypothesis test to check whether the non-parametric functions $g_k(t)$ have linear structure. Using (39), the observed test statistics are 0.3457, 0.3407 and 0.3273 with p-values 0.02, 0.004 and 0.004, respectively. As before, we used a range of bandwidths to ensure that our results are not sensitive to the choice of bandwidth. The hypothesis test results show that, at a 5% level of significance, we can reject the null hypothesis that the non-parametric functions have a linear structure. It follows that the $K = 2$ component SPMPLMs is adequate for this data.

7. Discussion and Conclusions

In this section, we give a summary of the paper and a brief discussion of the results. Moreover, we also provide directions for future research.

7.1. Summary

This paper considered maximum likelihood estimation of the semi-parametric mixture of partial linear models (SPMPLMs) via the EM algorithm. The mixture component regression function (CRF) of each linear model consists of a parametric and non-parametric term. We considered both global and local estimation of the non-parametric term. For the former, we proposed a regression spline-based estimation procedure. For the latter, we first identified the label switching problem involved when separately maximising each local-likelihood function. The general solution to this problem is to simultaneously maximise each local-likelihood function using the same responsibilities obtained at the E-step of the EM algorithm. Thus, a global set of responsibilities must be obtained. The proposed one-step backfitting profile likelihood estimation procedure makes use of the local responsibilities to compute global responsibilities.

In addition, the non-parametric estimator requires a smoothing parameter. We proposed a data-driven approach using the GCV method to select this parameter. To reduce the computational burden imposed by the partial residuals estimator of the parametric term of the CRF, we proposed a plug-in estimator.

7.2. Discussion of the Results

We demonstrated the performances of the proposed methods through a simulation study. Based on the results, the proposed methods achieved accurate estimation of both the parametric and non-parametric terms of the model. Moreover, the latter performed at least as well as the competitive methods. In general, the regression spline-based estimator performs better than the profile likelihood estimators for estimating the non-parametric term. Furthermore, the non-parametric estimator based on the plug-in estimator performs better than that based on the partial residuals estimator. In terms of computational time, the plug-in procedure reduces the computational burden drastically.

To illustrate the practical use of the proposed methods, we used them to estimate the SPMPLMs for two climate datasets. For the first dataset, we considered the effects of urbanisation and energy consumption per capita on CO_2 emissions. The estimated model identified clearly a mixture structure consisting of two groups of countries. For the first group (top of Figure 3b), carbon emissions increase rapidly as more people move into the urban areas and thereafter declines slowly. For the second group (bottom of Figure 3b), carbon emissions increase slowly as urbanisation increases and then rapidly declines for further increases in urbanisation.

For the second dataset, we considered the effect of GDP per capita and the share of total energy from renewable sources on CO_2 emissions. We proposed a $K = 2$ component SPMPLMs for this data. The estimated model revealed two groups of countries. For the first group (see Figure 5b), carbon emissions decrease rapidly as per capita GDP increases, followed by an increase and thereafter a further decrease in CO_2 with further increases in per capita GDP. For the second group (see Figure 5c), carbon emissions increase up to a point and then decrease, followed by another increase and then a decrease.

For both datasets, the resulting functions exhibit the environmental Kuznets curve (EKC) hypothesis, which says that carbon emissions increase up to a point of national income and decrease beyond that point.

7.3. Future Work

For future studies, it will be of interest to adapt the proposed ideas to estimate more flexible models where the CRFs are comprised of additive non-parametric functions. The proposed algorithm computes the global estimate of the non-parametric function by selecting the smoothest locally estimated function. This procedure is discrete in that it depends on a single set of local responsibilities. Efforts are now underway to develop methods that continuously combine all the sets of local responsibilities to compute the global responsibilities. Note that the proposed algorithm can also be applied to address label switching when estimating parametric Bayesian mixtures using MCMC procedures.

However, this version of the algorithm has factorial complexity, growing with the size of the Markov chain. This gives rise to a further line of research to develop a computationally efficient procedure.

Author Contributions: Conceptualisation, S.B.S., S.M.M. and F.H.J.K.; methodology, S.B.S., S.M.M. and F.H.J.K.; software, S.B.S.; formal analysis, S.B.S.; investigation, S.B.S.; data curation, S.B.S.; writing—original draft preparation, S.B.S.; writing—review and editing, S.B.S., S.M.M. and F.H.J.K.; visualization, S.B.S.; supervision, S.M.M. and F.H.J.K.; funding acquisition, S.B.S., S.M.M. and F.H.J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the South African National Research Foundation (NRF), grant number 149091.

Data Availability Statement: The data used in the application can be obtained from a public database: https://ourworldindata.org/urbanization (accessed on 2 September 2022); https://ourworldindata.org/energy (accessed on 2 September 2022); https://ourworldindata.org/co2-emissions (accessed on 2 September 2022); https://ourworldindata.org/renewable-energy (accessed on 2 September 2022); https://ourworldindata.org/grapher/real-gdp-per-capita-pennwt (accessed on 2 September 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CI	Confidence Interval
CNLRM	Classical Normal Linear Regression Model
CO_2	Carbon Dioxide
COD	Curse Of Dimensionality
CRF	Component Regression Function
EKC	Environmental Kuznets Curve
EM	Expectation-Maximization
GCV	Generalized Cross Validation
GDP	Gross Domestic Product
SPMPLM	Semi-parametric Mixture Of Partially Linear Models
SPMNPR	Semi-parametric Mixture of Non-parametric Regressions
SPMAR	Semi-parametric Mixture of Additive Regressions
SPMPLAR	Semi-parametric Mixture of Partially Linear Additive Regressions
SBK	Spline-Backfitted Kernel
MoE	Mixture of Experts
MSIM	Mixture of Single Index Models
MNLRM	Mixtures of Normal Linear Regression
MRSIP	Mixture of Regressions Models with Varying Single Index Proportions
NPMNR	Non-parametric Mixture of Normal Regressions
OB-PL-EM	One-step Backfitting Profile Likelihood EM
PL-EM	Profile Likelihood Expectation Maximization
PL-p-EM	Profile Likelihood plug-in Expectation Maximization

References

1. Quandt, R.E. A New Approach to Estimating Switching Regressions. *J. Am. Stat. Assoc.* **1972**, *67*, 306–310. [CrossRef]
2. Goldfeld, S.M.; Quandt, R.E. A Markov model for switching regressions. *J. Econom.* **1973**, *1*, 3–15. [CrossRef]
3. Hurn, M.; Justel, A.; Robert, C.P. Estimating mixtures of regressions. *J. Comput. Graph. Stat.* **2003**, *12*, 55–79. [CrossRef]
4. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*; Springer: New York, NY, USA, 2006.
5. DeSarbo, W.S.; Cron, W.L. A maximum likelihood methodology for clusterwise linear regression. *J. Classif.* **1988**, *5*, 249–282. [CrossRef]
6. De Veaux, R.D. Mixtures of linear regressions. *Comput. Stat. Data Anal.* **1989**, *8*, 227–245. [CrossRef]
7. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–38. [CrossRef]
8. Wu, X.; Liu, T. Estimation and testing for semiparametric mixtures of partially linear models. *Commun. Stat.-Theory Methods* **2017**, *46*, 8690–8705. [CrossRef]

9. Huang, M.; Yao, W. Mixture of regression models with varying mixing proportions: A semiparametric approach. *J. Am. Stat. Assoc.* **2012**, *107*, 711–724. [CrossRef]
10. Stephens, M. Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2000**, *62*, 795–809. [CrossRef]
11. Huang, M.; Li, R.; Wang, S. Nonparametric mixture of regression models. *J. Am. Stat. Assoc.* **2013**, *108*, 929–941. [CrossRef] [PubMed]
12. Xiang, S.; Yao, W. Semiparametric mixtures of nonparametric regressions. *Ann. Inst. Stat. Math.* **2016**, *70*, 131–154. [CrossRef]
13. Huang, M.; Yao, W.; Wang, S.; Chen, Y. Statistical inference and applications of mixture of varying coefficient models. *Scand. J. Stat.* **2018**, *45*, 618–643. [CrossRef]
14. Xiang, S.; Yao, W. Semiparametric mixtures of regressions with single-index for model based clustering. *Adv. Data Anal. Classif.* **2020**, *14*, 261–292. [CrossRef]
15. Zhang, Y.; Zheng, Q. Semiparametric mixture of additive regression models. *Commun. Stat.-Theory Methods* **2018**, *47*, 681–697. [CrossRef]
16. Zhang, Y.; Pan, W. Estimation and inference for mixture of partially linear additive models. *Commun. Stat.-Theory Methods* **2022**, *51*, 2519–2533. [CrossRef]
17. Xue, J.; Yao, W. Machine Learning Embedded Semiparametric Mixtures of Regressions with Covariate-Varying Mixing Proportions. *Econom. Stat.* **2022**, *22*, 159–171. [CrossRef]
18. Xue, J. Machine Learning Embedded Nonparametric Mixture Regression Models. Ph.D. Thesis, UC Riverside, Riverside, CA, USA, 2022.
19. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef]
20. Skhosana, S.B.; Kanfer, F.H.J.; Millard, S.M. Fitting Non-Parametric Mixture of Regressions: Introducing an EM-Type Algorithm to Address the Label-Switching Problem. *Symmetry* **2022**, *14*, 1058. [CrossRef]
21. Speckman, P. Kernel smoothing in partial linear models. *J. R. Stat. Soc. Ser. B (Methodol.)* **1988**, *50*, 413–436. [CrossRef]
22. Wu, H.; Zhang, J.T. *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*; John Wiley & Sons: Hoboken, NJ, USA, 2006; Volume 515.
23. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer: New York, NY, USA, 2021. [CrossRef]
24. Fan, J.; Gijbels, I. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*; CRC Press: Boca Raton, FL, USA, 1996; Volume 66.
25. Severini, T.A.; Wong, W.H. Profile likelihood and conditionally parametric models. *Ann. Stat.* **1992**, *20*, 1768–1802. [CrossRef]
26. Tibshirani, R.; Hastie, T. Local likelihood estimation. *J. Am. Stat. Assoc.* **1987**, *82*, 559–567. [CrossRef]
27. Buja, A.; Hastie, T.; Tibshirani, R. Linear smoothers and additive models. *Ann. Stat.* **1989**, *17*, 453–510. [CrossRef]
28. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
29. Dinda, S. Environmental Kuznets curve hypothesis: A survey. *Ecol. Econ.* **2004**, *49*, 431–455. [CrossRef]