Bernoulli **23**(2), 2017, 825–862
DOI: 10.3150/15-BEJ760

Semiparametric topographical mixture models with symmetric errors

C. BUTUCEA[1], R. NGUEYEP TZOUMPE[2] and P. VANDEKERKHOVE[1,3]

[1]*Université Paris-Est, LAMA (UMR 8050), UPEMLV, F-77454, Marne-la-Vallée, France. E-mail:* *Cristina.Butucea@u-pem.fr
[2]*IBM Watson Research Center, 1101 Kitchawan Road, Yorktown Heights NY 10598, USA. E-mail: ngueyep@us.ibm.com*
[3]*UMI Georgia Tech – CNRS 2958, School of aerospace, Georgia Institute of Technology, 270 Ferst Drive Atlanta GA 30332-0150, USA. E-mail:* **Pierre.Vandekerkhove@u-pem.fr*

Motivated by the analysis of a Positron Emission Tomography (PET) imaging data considered in Bowen *et al*. [*Radiother. Oncol.* **105** (2012) 41–48], we introduce a semiparametric topographical mixture model able to capture the characteristics of dichotomous shifted response-type experiments. We propose a pointwise estimation procedure of the proportion and location functions involved in our model. Our estimation procedure is only based on the symmetry of the local noise and does not require any finite moments on the errors (e.g., Cauchy-type errors). We establish under mild conditions minimax properties and asymptotic normality of our estimators. Moreover, Monte Carlo simulations are conducted to examine their finite sample performance. Finally, a statistical analysis of the PET imaging data in Bowen *et al*. is illustrated for the proposed method.

Keywords: asymptotic normality; consistency; contrast estimators; finite mixture of regressions; Fourier transform; identifiability; inverse problem; mixture model; semiparametric; symmetric errors

1. Introduction

The model we propose to investigate in this paper is a semiparametric topographical mixture model able to capture the characteristics of dichotomous shifted response-type experiments such as the tumor data in [5], Figure 4. Let suppose that we visit at random the space $\mathbb{R}^d$ ($d \geq 1$) by sampling a sequence of i.i.d. random variables $\mathbf{X}_i$, $i = 1, \ldots, n$, having common probability distribution function (p.d.f.) $\ell : \mathbb{R}^d \to \mathbb{R}_+$. For each $\mathbf{X}_i$ we observe an output response Y_i whose distribution is a mixture model with probability parameters depending on the design $\mathbf{X}_i$. For simplicity, let us consider first a mixture of two nonlinear regression model:

$$Y_i = W(\mathbf{X}_i)\big(a(\mathbf{X}_i) + \tilde{\varepsilon}_{1,i}\big) + \big(1 - W(\mathbf{X}_i)\big)\big(b(\mathbf{X}_i) + \tilde{\varepsilon}_{2,i}\big), \tag{1.1}$$

where locations are $a, b : \mathbb{R}^d \to \mathbb{R}$, the errors $\{\tilde{\varepsilon}_{1,i}, \tilde{\varepsilon}_{2,i}\}_{i=1,\ldots,n}$ are supposed to be i.i.d. with zero-symmetric common p.d.f. f. The mixture in model (1.1) occurs according to the random variable $W(\mathbf{x})$ at point $\mathbf{x}$, with probability $\pi : \mathbb{R}^d \to (0, 1)$,

$$W(\mathbf{x}) = \begin{cases} 1, & \text{with probability } \pi(\mathbf{x}), \\ 0, & \text{with probability } 1 - \pi(\mathbf{x}). \end{cases}$$

1350-7265 © 2017 ISI/BS

Moreover we assume that, conditionally on the $\mathbf{X}_i$'s, the $\{\tilde{\varepsilon}_{1,i}, \tilde{\varepsilon}_{2,i}\}_i$'s and the $W(\mathbf{X}_i)$'s are independent. Such a model is related to the class of Finite Mixtures of Regression (FMR), see [14] for a good overview. Briefly, statistical inference for the class of parametric FMR model was first considered by [28] who proposed a moment generating function based estimation method. An EM estimating approach was proposed by [11] in the two-component case. Variations of the latter approach were also considered in [24] and [35]. Hawkins *et al*. [16] studied the estimation problem of the number of components in the parametric FMR model using approaches derived from the likelihood equation. In [22], the authors investigated a Bayesian approach to estimate the regression coefficients and also proposed an extension of the model in which the number of components is unknown. Zhu and Zhang [38] established the asymptotic theory for maximum likelihood estimators in parametric FMR models. More recently, Städler, Bühlmann and van de Geer [29] proposed an ℓ_1-penalized method based on a Lasso-type estimator for a high-dimensional FMR model with $d \geq n$. As an alternative to parametric approaches to the estimation of a FMR model, some authors suggested the use of more flexible semiparametric approaches. These approaches can actually be classified into two groups: semiparametric FMR (SFMR) of type I and type II. We say a mixture model is of type I when the mixture probability and location parameters are Euclidean, but the mixing distribution is nonparametric, whereas a model is of type II when, the other way around, the mixture probability and location are nonparametric but the mixing density is known or belongs to a parametric family.

The study of SFMR of type I comes from the seminal work of [15] in which d-variate semiparametric mixture models of random vectors with independent components were considered. These authors proved in particular that, for $d \geq 3$, we can identify a two-component mixture model without parametrizing the distributions of the component random vectors. To the best of our knowledge, Leung and Qin [26] were the first in estimating a FMR model semiparametrically in that sense. In the two-component case, they studied the case where the components are related by Anderson's [1] exponential tilt model. Hunter and Young [21] studied the identifiability of an m-component type I SFMR model and numerically investigated a Expectation–Maximization (EM) type algorithm for estimating its parameters. Vandekerkhove [36] proposed an M-estimation method for a two-component semiparametric mixture of linear regressions with symmetric errors (type I) in which one component is known. Bordes, Kojadinovic and Vandekerkhove [3] revisited the same model by establishing new moment-based identifiability results from which they derived explicit $\sqrt{n}$-convergent estimators.

The study of type II SFMR models started with [19] who considered a semiparametric linear FMR model with Gaussian noise in which the mixing proportions are possibly covariates-dependent. They established also the asymptotic normality of their local maximum likelihood estimator and investigated a modified EM-type algorithm. Huang, Li and Wang [18] generalized the latter work to nonlinear FMR with possibly covariates-dependent noises. Toshiya [33] considered a Gaussian FMR model where the joint distribution of the response and the covariate (possibly functional) is itself modeled as a mixture. More recently [27] considered a penalized maximum likelihood approach for Gaussian FMR models with logistic weights.

To improve the flexibility of our FMR model (1.1) and address the study of models involving design-dependent noises, such as the radiotherapy application from [5] displayed below in Figure 1, we will consider a slightly more general model:

$$Y_i = W(\mathbf{X}_i)\big(a(\mathbf{X}_i) + \varepsilon_{1,i}(\mathbf{X}_i)\big) + \big(1 - W(\mathbf{X}_i)\big)\big(b(\mathbf{X}_i) + \varepsilon_{2,i}(\mathbf{X}_i)\big), \tag{1.2}$$

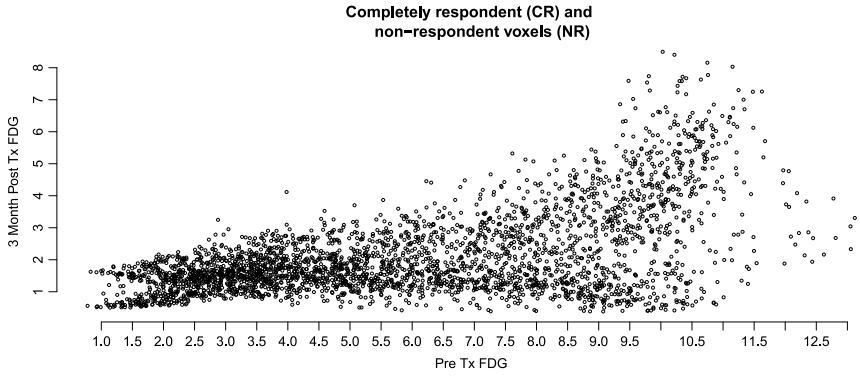


Figure 1. Display of the original PET-radiotherapy data from [5].

such that, given $\{\mathbf{X} = \mathbf{x}\}$, the common p.d.f. of the $\varepsilon_{j,i}(\mathbf{x})$, $j = 1, 2$, denoted $f_{\mathbf{x}}$, is zero-symmetric. We will say that the above model is of type III, that is, it combines type I and type II properties. Indeed, no parametric assumption is made about the mixing distribution of the errors nor about the mixing proportion and the location parameters, which are possibly design dependent. Our model is still said *semiparametric* because, given $\{\mathbf{X} = \mathbf{x}\}$, the vector $\theta(\mathbf{x}) = (\pi(\mathbf{x},)a(\mathbf{x}), b(\mathbf{x}))$ will be viewed as an Euclidean parameter to be estimated.

Examples of design-point noise dependency:

(i) (Topographical scaling). The most natural transformation is probably when considering a topographical scaling of the errors, with $\sigma : \mathbb{R}^d \to \mathbb{R}_+^*$, such that $\varepsilon_{j,i}(\mathbf{X}_i) = \sigma(\mathbf{X}_i)\tilde{\varepsilon}_{j,i}$, $j = 1, 2$, where the $\tilde{\varepsilon}_{j,i}$'s are similar to those involved in (1.1). The conditional p.d.f of the errors $\varepsilon_{j,i}$ given $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$f_{\mathbf{x}}(y) = \frac{1}{\sigma(\mathbf{x})} f\left(\frac{y}{\sigma(\mathbf{x})}\right), \qquad y \in \mathbb{R}. \tag{1.3}$$

Indeed, if f is zero-symmetric then the errors' distribution inherits trivially the same symmetry property.

(ii) (Zero-symmetric varying mixture). Another useful example could be the varying mixing proportion mixture model of r zero-symmetric distributions. For $k = 1, \ldots, r$, we consider proportion functions $\lambda_k : \mathbb{R}^d \to (0, 1)$ with $\sum_{k=1}^r \lambda_k(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^d$. The conditional p.d.f of the errors $\varepsilon_{j,i}$ given $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$f_{\mathbf{x}}(y) = \sum_{k=1}^{r} \lambda_k(\mathbf{x}) f_k(y), \qquad y \in \mathbb{R},$$

where the f_k functions are zero-symmetric p.d.f.'s.

(iii) (Antithetic location model). Consider a location function $\mu : \mathbb{R}^d \to \mathbb{R}$ and f any arbitrary p.d.f. The conditional p.d.f of the errors $\varepsilon_{j,i}$ given $\{\mathbf{X} = \mathbf{x}\}$ is defined by

$$f_{\mathbf{x}}(y) = \tfrac{1}{2} f\big(y - \mu(\mathbf{x})\big) + \tfrac{1}{2} f\big(-y + \mu(\mathbf{x})\big), \qquad y \in \mathbb{R},$$

and also results into a zero-symmetric p.d.f.

Note that any combination of the above situations could be considered in model (1.2) free from specifying any parametric family (provided the resulting zero-symmetry hold). This last remark reveals, in our opinion, the main strength of our model in the sense that it could prove to be a very flexible exploratory tool for the analysis of shifted response-type experiments. Our paper is organized as follows. Section 2 is devoted to identifiability results and a detailed description of our estimation method, while Section 3 is concerned with its asymptotic properties. The finite-sample performance of the proposed estimation method is studied for various scenarios through Monte Carlo experiments in Section 4. In Section 5, we propose to analyze the Positron Emission Tomography (PET) imaging data considered in [5]. Finally, Section 6 is devoted to auxiliary results and main proofs.

2. Estimation method

Let us define the joint density of a couple $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, designed from model (1.2):

$$g(y, \mathbf{x}) = \big[\pi(\mathbf{x}) f_{\mathbf{x}}\big(y - a(\mathbf{x})\big) + \big(1 - \pi(\mathbf{x})\big) f_{\mathbf{x}}\big(y - b(\mathbf{x})\big)\big]\ell(\mathbf{x}), \qquad (y, \mathbf{x}) \in \mathbb{R}^{d+1}, \quad (2.1)$$

while the conditional density of Y given $\{\mathbf{X} = \mathbf{x}\}$ (denoted for simplicity $Y/\mathbf{X} = \mathbf{x}$) is

$$g_{\mathbf{x}}(y) = g(y, \mathbf{x})/\ell(\mathbf{x}) = \pi(\mathbf{x}) f_{\mathbf{x}}\big(y - a(\mathbf{x})\big) + \big(1 - \pi(\mathbf{x})\big) f_{\mathbf{x}}\big(y - b(\mathbf{x})\big). \qquad (2.2)$$

We are interested in estimating the parameter $\theta_0 = \theta(\mathbf{x}_0) = (\pi(\mathbf{x}_0), a(\mathbf{x}_0), b(\mathbf{x}_0))$ at some fixed point $\mathbf{x}_0$ belonging to the interior of the support of ℓ ($\ell(\mathbf{x}_0) > 0$), denoted $\mathrm{supp}(\ell)$. For simplicity and identifiability matters, we will suppose that θ_0 belongs to the interior of the parametric space $\Xi = [p, P] \times \Delta$, where $0 < p \le P < 1$ and Δ denotes a compact set of $\mathbb{R}^2 \setminus \{(a, a) : a \in \mathbb{R}\}$.

At fixed $\mathbf{x}_0$, we prove, following [4], that identifiability holds up to label switching. Indeed, in [7] authors restricted the set of parameters to $[p, P] \times \Delta$, where $0 < p \le P < 1/2$. Another way to avoid label switching is to assume $0 < p \le P < 1$ and $a < b$. In order to have global identifiability of our model, we assume that at some fixed point $\mathbf{x}$ we have $a(\mathbf{x}) < b(\mathbf{x})$ and that functions a and b are differentiable and transversal (i.e., at each crossing point $\mathbf{x}$ where $a(\mathbf{x}) = b(\mathbf{x})$ gradients are different). The rest of this section is dedicated to identifiability of the model and the estimation procedure.

2.1. Mixture of regression models as an inverse problem

We see in formula (2.2), that the conditional density of Y given $\{\mathbf{X} = \mathbf{x}\}$ can be viewed as a mixture of the errors distribution $f_{\mathbf{x}}$ given $\{\mathbf{X} = \mathbf{x}\}$ with locations $(a(\mathbf{x}), b(\mathbf{x}))$ and mixing proportion $\pi(\mathbf{x})$. Mixture of populations with different locations is a well known inverse problem. Our inversion procedure is done in Fourier domain.

For any function g in $\mathbb{L}_1 \cap \mathbb{L}_2$, let us define its Fourier transform by

$$g^*(u) = \int \exp(iuy) g(y)\, dy \qquad \text{for all } u \in \mathbb{R}.$$

Here, the estimation method is based on the Fourier transform of the conditional density $g_{\mathbf{x}}(y)$ of $Y/\mathbf{X} = \mathbf{x}$. If the p.d.f. $f_{\mathbf{x}}$ belongs to $\mathbb{L}_1 \cap \mathbb{L}_2$ then so does $g_{\mathbf{x}}$. Denote its Fourier transform by $g^*_{\mathbf{x}}(u)$ for all $u \in \mathbb{R}$. In our model, we observe that

$$g^*_{\mathbf{x}}(u) = \big(\pi(\mathbf{x}) e^{iua(\mathbf{x})} + \big(1 - \pi(\mathbf{x})\big) e^{iub(\mathbf{x})}\big) f^*_{\mathbf{x}}(u), \qquad u \in \mathbb{R}.$$

Let us denote, for all $t = (\pi, a, b)$ in Ξ and u in $\mathbb{R}$,

$$M(t,u) := \pi e^{iua} + (1-\pi) e^{iub}. \tag{2.3}$$

Note that $|M(t,u)| \le 1$ for all $(t,u) \in \Xi \times \mathbb{R}$. Then, we have

$$g^*_{\mathbf{x}}(u) = M\big(\theta(\mathbf{x}), u\big) f^*_{\mathbf{x}}(u).$$

We introduce for convenience $\omega := \{\omega(1), \omega(2)\}$ a permutation of set $\{1,2\}$, that is, $\omega \in \{id, s\}$ where $s(1) = 2$ and $s(2) = 1$. For $t = (\pi, a, b)$, we denote $[t]_\omega := t\mathbb{I}_{\omega = id} + (1 - \pi, b, a)\mathbb{I}_{\omega = s}$ the parameter affected by a permutation ω of the labels (label 1 corresponding to location a and label 2 corresponding to location b). Let us fix $\mathbf{x}_0 \in \operatorname{supp}(\ell)$ such that $\theta(\mathbf{x}_0)$ belongs to the interior of Ξ, denoted $\mathring{\Xi}$. Noticing that the p.d.f. $f_{\mathbf{x}_0}$ is zero-symmetric we therefore have that $f^*_{\mathbf{x}_0}(u) \in \mathbb{R}$, for all $u \in \mathbb{R}$. If t belongs to Ξ, we prove in the next theorem the *picking* property

$$\Im\big(g^*_{\mathbf{x}_0}(u) \bar{M}(t,u)\big) = 0 \qquad \text{for all } u \in \mathbb{R}, \quad \text{if and only if} \quad \exists \omega \in \{id, s\} : t = \big[\theta(\mathbf{x}_0)\big]_\omega,$$

where $\Im : \mathbb{C} \to \mathbb{R}$ denotes the imaginary part of a complex number and $\bar{M}$ the complex conjugate of M. This result, which is a consequence of the linear independence of the $\{\sin(\alpha_1 u), \ldots, \sin(\alpha_p u)\}$ family proved in [4] using standard Vandermonde type determinant properties, allows us to build a *contrast* function for the parameter $t \in \Xi$:

$$S(t) := S_{\mathbf{x}_0}(t) := \int \Im\big(g^*_{\mathbf{x}_0}(u) \bar{M}(t,u)\big)^2 \ell^2(\mathbf{x}_0) w(u)\, du. \tag{2.4}$$

The function $w : \mathbb{R}^d \to \mathbb{R}_+$ is a bounded p.d.f. which helps in computing the integral via Monte-Carlo method and solves integrability issues. We stress the fact that using ℓ^2 instead of ℓ comes from the fact that the contrast estimates a quadratic functional, rather than an expected value.

Remark. The idea of using Fourier transform in order to solve the inverse mixture problem was introduced in [7] for density models. In the regression models, we deal with the conditional density of $Y/\mathbf{X} = \mathbf{x}_0$ and consider that it could possibly exist $\mathbf{x}_0 \in \operatorname{supp}(\ell)$ such that $\pi(\mathbf{x}_0) = 1/2$ and then $M(\theta(\mathbf{x}_0), u)$ can be 0. This has a major incidence on the definition of the function $S(t)$ where $\bar{M}(t,u)$ appears at the numerator (contrarily to Butucea and Vandekerkhove, [7] where $M(t,u)$ appeared at the denominator). Moreover, smoothing of the information that data bring at a fixed design point $\mathbf{x}_0$ changes dramatically the behavior of the estimators as we shall see later on.

2.2. Local and global identifiability

We prove in the following theorem that our model is identifiable (up to a permutation of the labels) and that $S(t)$ defines a contrast on the parametric space Ξ.

Theorem 1 (Identifiability and contrast property). *Consider model (1.2) provided with $f_{\mathbf{x}}(\cdot) \in \mathbb{L}_2$ for all $\mathbf{x} \in \mathbb{R}^d$. For a fixed point $\mathbf{x}_0$ in the interior of the support of ℓ, we assume that $f_{\mathbf{x}_0}(\cdot)$ is zero-symmetric and that $\theta_0 = \theta(\mathbf{x}_0)$ is an interior point of Ξ. Then we have the following properties*:

(i) *The Euclidean parameter $\theta_0 = (\pi(\mathbf{x}_0), a(\mathbf{x}_0), b(\mathbf{x}_0))$ is identifiable up to a permutation of the labels when the function $f_{\mathbf{x}_0}(\cdot)$ is uniquely identified.*

(ii) *The function S in (2.4) is a contrast function, that is, for all $t \in \Xi$, $S(t) \geq 0$ and $S(t) = 0$ if and only if there exists $\omega \in \{id, s\}$ such that $t = [\theta_0]_\omega$.*

Proof. (i) The local (for fixed $\mathbf{x}_0$) identifiability of model (2.2) over Ξ and the set $\mathcal{F}$ of zero-symmetric densities, that is, using notations involved in (2.3), for all $(t, t') \in \Xi^2$ and $(f, f') \in \mathcal{F}^2$,

$$M(t,u)f^*(u) = M\big(t',u\big)f'^*(u) \quad \Rightarrow \quad \exists \omega \in \{id, s\} : t' = t_\omega \text{ and } f = f',$$

is deduced from the proof of Theorem 2.1 in [4]. The main difference here is that we allow π to lie in $(0, 1)$ whereas in [4] the proportion mixing parameter was constrained to belong to $[0, 1/2)$. This constraint was also an implicit lexicographical ordering to avoid multiple label-permuted mixture representation. When revisiting step by step the proof of the latter theorem, it appears that the condition $\pi \neq 1/2$ is essentially used to avoid spurious model representation when the mixing proportion is allowed to be equal to zero (see discussion of Case 1, top of page 1223, and the counter-example, page 1206, in [4]). When $\pi > 0$, the discussion of equation (37) in [4] leads to two obvious solutions $(\pi, a, b) = (\pi', a', b')$ and $(\pi, a, b) = (1 - \pi', b', a')$. To prove that possibly spurious solutions are non-admissible, it suffices to adapt the re-parametrization in (38) of [4] to the cases $(a - a', b - b') \neq (0, 0)$ and $(a - b', b - a') \neq (0, 0)$, which basically leads to consider (by symmetry) the following conditions: for $\beta_1 = \pi\pi'$, $\beta_2 = \pi(1-\pi')$, $\beta_3 = \pi'(1-\pi)$, $\beta_4 = (1-\pi)(1-\pi')$:

- $\beta_3 + \beta_4 = 0 \Leftrightarrow \pi = 1$,
- $\beta_2 + \beta_3 = 0$ and $\beta_4 = 0 \Leftrightarrow \pi = 1$ or $\pi' = 1$,
- $\beta_3 = 0$ and $\beta_4 - \beta_2 = 0 \Leftrightarrow \pi' = 0$ and $\pi = 1/2$, or $\pi = 1$ and $\pi' = 1$,
- $\beta_2 = 0$ and $\beta_4 - \beta_3 = 0 \Leftrightarrow \pi = 0$ and $\pi' = 1/2$, or $\pi' = 1$ and $\pi = 1$.

Note that the above solutions are all non-admissible when $(\pi, \pi') \in (0, 1)^2$. From this remark, we deduce that the Euclidean part of model (2.2) is also identifiable, up to a permutation of the labels, over *our* parametric space Ξ (including $\pi = 1/2$). To identify now the local noise distribution, we proceed similarly to Step 3 in [4]. Because for $\omega \in \{id, s\}$

$$M(t,u)f^*(u) = M(t_\omega,u)f'^*(u) = M(t,u)f'^*(u), \qquad u \in \mathbb{R},$$

we have to consider the two following cases:

- $\pi \neq 1/2$. Since $|M(t,u)| \geq |1-2\pi| > 0$, we deduce $f^*_{\mathbf{x}} = f'^*_{\mathbf{x}}$ and $f_{\mathbf{x}} = f'_{\mathbf{x}}$.
- $\pi = 1/2$. Here it is to be observed that, for t fixed in Ξ, $M(t,u) = 0$ occurs to be null on a countable set of $\mathbb{R}$. Indeed,

$$M(t,u) = 0 \quad \Leftrightarrow \quad au = bu + \pi + 2k\pi, \qquad k \in \mathbb{Z} \quad \Leftrightarrow \quad u \in \left\{ \frac{\pi + 2k\pi}{a-b}, k \in \mathbb{Z} \right\}.$$

  Nevertheless, this behavior does not affect the identifiability of the noise distribution since we can conclude that the real functions f^* and f'^* coincide over $\mathbb{R}$ except on a countable set of isolated points which is equivalent, by a continuity argument, to the equality over the whole real line.

This concludes the proof of (i).

(ii) The proof is similar to the proof of Proposition 1 in [7], replacing $f^*(\cdot)$ and $g^*(\cdot)$ by $f^*_{\mathbf{x}_0}(\cdot)$ and $g^*_{\mathbf{x}_0}(\cdot)$, respectively, and noticing that $\ell(\mathbf{x}_0)$ is bounded away from zero. □

Label switching and global identifiability. The label switching phenomenon relies on the fact that the writing of the likelihood of a mixture model is invariant when permuting the label of its components. For example, when considering a k-component mixture model, there exists up to $k!$ mixture representations of the same distribution. To avoid these multiple representations (which obviously affects the estimation methods and their interpretation), there exists many different approaches: (i) in the parametric case, Teicher [32] suggest, for example, to create a lexical ordering on the parametric space, (ii) in the Bayesian case, some MCMC-based relabelling algorithms are proposed, see [8,30] or [37], (iii) in the two-component semiparametric case, the mixture proportion affected to the first component is constrained to be less than $1/2$, see [4]. In our case, since we plan to estimate the conditional model (2.2) over a grid of design-points, it would be precisely great to non-restrict the proportion mixture function $\pi(\cdot)$ to be upper-bounded by $1/2$ and also to be able to deal with intersecting curve functions $a(\cdot)$ and $b(\cdot)$. To better understand these situations and propose some practical implementations, we propose now to state, using arguments similar to [18], the global identifiability of our model (2.1) when $d = 1$. For this purpose, let us introduce the concept of *transversality*.

Definition 1. *Let* $\mathbf{x} \in \mathbb{R}$, *and let* $a(\mathbf{x})$ *and* $b(\mathbf{x})$ *two continuously differentiable real curve-functions. We say that* $a(\mathbf{x})$ *and* $b(\mathbf{x})$ *are transversal if* $(a(\mathbf{x}) - b(\mathbf{x}))^2 + \|\dot{a}(\mathbf{x}) - \dot{b}(\mathbf{x})\|^2 \neq 0$, *for any* $x \in \mathbb{R}$, *where* $\|\cdot\|$ *denotes the Euclidean norm.*

The transversality condition imposed on two real curve-functions $a(\mathbf{x})$ and $b(\mathbf{x})$ implies that if $a(\mathbf{x}) = b(\mathbf{x})$, then $\dot{a}(\mathbf{x}) \neq \dot{b}(\mathbf{x})$.

Proposition 1. *Let us suppose that* $\mathrm{supp}(\ell)$ *is an interval of* $\mathbb{R}$ *and that* $\pi(\mathbf{x}) \in (0,1)$, *respectively* $a(\mathbf{x})$ *and* $b(\mathbf{x})$, *is a continuous function, respectively are both differentiable real-functions. If* $a(\mathbf{x}_0) < b(\mathbf{x}_0)$ *at some fixed point* $\mathbf{x}_0$ *in the interior of the* $\mathrm{supp}(\ell)$ *and if* $a(\mathbf{x})$ *and* $b(\mathbf{x})$ *are transversal then our model* (2.1) *is globally identifiable over* $\mathrm{supp}(\ell)$.

Proof. Let us consider the subset of $\mathbb{R}$

$$\mathcal{E} = \big\{\mathbf{x}_k : a(\mathbf{x}_k) = b(\mathbf{x}_k)\big\},$$

where the parameter curves intersect. Since parameter curves are transversal, any point in $\mathcal{E}$ is an isolated point. This implies that the set $\mathcal{E} \subset \mathbb{R}$ has no finite accumulation (limit) point and contains at most countably many points. Therefore, without loss of generality, we assume that: $\mathbf{x}_k < \mathbf{x}_{k+1}$ and $(\mathbf{x}_k, \mathbf{x}_{k+1}) \cap \mathcal{E} = \varnothing$, $k \in \mathbb{Z}$. Assume that the conditional model (2.2) admits another representation, that is, there exist functions $(\pi', a', b', f'_{\mathbf{x}})$ such that

$$g_{\mathbf{x}}(y) = \pi'(\mathbf{x}) f'_{\mathbf{x}}\big(y - a'(\mathbf{x})\big) + \big(1 - \pi'(\mathbf{x})\big) f_{\mathbf{x}}\big(y - b'(\mathbf{x})\big).$$

We proved in (i) of Theorem 1, that for any $\mathbf{x} \notin \mathcal{E}$, model (2.2) is identifiable, it follows that there exists a permutation $\omega_{\mathbf{x}} := \{\omega_{\mathbf{x}}(1), \omega_{\mathbf{x}}(2)\}$ of set $\{1, 2\}$, that is, $\omega_{\mathbf{x}} \in \{id, s\}$ where $s(1) = 2$ and $s(2) = 1$, depending on $\mathbf{x}$ such that:

$$\begin{cases} \pi'(\mathbf{x}) = \pi(\mathbf{x}), a'(\mathbf{x}) = a(\mathbf{x}), b'(\mathbf{x}) = b(\mathbf{x}), & \text{if } \omega_{\mathbf{x}} = id, \\ \pi'(\mathbf{x}) = 1 - \pi(\mathbf{x}), a'(\mathbf{x}) = b(\mathbf{x}), b'(\mathbf{x}) = a(\mathbf{x}), & \text{if } \omega_{\mathbf{x}} = s. \end{cases}$$

Since the parameter curves $(a(\mathbf{x}), b(\mathbf{x}))$ are continuous and do not intersect on any interval $(\mathbf{x}_k, \mathbf{x}_{k+1})$ the permutation $\omega(\mathbf{x})$ must be constant on the latter interval. In addition, for any $\mathbf{x}_k \in \mathcal{E}$, consider a small interval $(\mathbf{x}_k - \epsilon, \mathbf{x}_k + \epsilon)$ such that $(\mathbf{x}_k - \epsilon, \mathbf{x}_k + \epsilon) \in (\mathbf{x}_{k-1}, \mathbf{x}_{k+1})$. Now, since the parameter curves are transversal, they have different derivatives at $\mathbf{x}_k$, hence the permutation must be constant on the neighborhood $(\mathbf{x}_k - \epsilon, \mathbf{x}_k + \epsilon)$. Indeed, without lack of generality, suppose that $\omega_{\mathbf{x}} = id$ for $\mathbf{x} \in (\mathbf{x}_k, \mathbf{x}_k + \epsilon)$ and $\omega_{\mathbf{x}} = s$ for $\mathbf{x} \in (\mathbf{x}_k - \epsilon, \mathbf{x}_k)$, then the functions a' and b' are non-differentiable anymore since, for example:

$$\big(\dot{a}'\big)^+(\mathbf{x}_k) = \dot{a}(\mathbf{x}_k) \neq \dot{b}(\mathbf{x}_k) = \big(\dot{a}'\big)^-(\mathbf{x}_k), \tag{2.5}$$

where $(\dot{a}')^+(\mathbf{x}_k)$ and $(\dot{a}')^-(\mathbf{x}_k)$ denote, respectively, the right- and left-hand side derivative of $a'(\cdot)$ at point $\mathbf{x}_k$. Therefore, there exists a permutation ω independent of $\mathbf{x} \in \operatorname{supp}(\ell)$ such that

$$\begin{cases} \pi'(\mathbf{x}) = \pi(\mathbf{x}), a'(\mathbf{x}) = a(\mathbf{x}), b'(\mathbf{x}) = b(\mathbf{x}), & \text{if } \omega = id, \\ \pi'(\mathbf{x}) = 1 - \pi(\mathbf{x}), a'(\mathbf{x}) = b(\mathbf{x}), b'(\mathbf{x}) = a(\mathbf{x}), & \text{if } \omega = s, \end{cases}$$

which concludes the proof of the global identifiability. □

Rules under the thumb. The proof of the above proposition inspires us two practical approaches to handle the label switching problem and lack of identifiability at curve intersection points.

- Label switching. Let us consider, without loss of generality, two nearest neighbors $(\mathbf{x}_1, \mathbf{x}_2)$ over a grid of testing points. Suppose that $a(\mathbf{x}_1)$ and $b(\mathbf{x}_1)$ are identified well separated and (λ, α, β) is a minimizer of $S_{\mathbf{x}_2}(\cdot)$, that is, $S_{\mathbf{x}_2}(\lambda, \alpha, \beta) = 0$. Since no big jump is expected by moving from $\mathbf{x}_1$ to $\mathbf{x}_2$, a way to decide which solution is more likely acceptable between $t_1 = (t_{1,i})_{1 \le i \le 3} = (\lambda, \alpha, \beta)$ and $t_2 = (t_{2,i})_{1 \le i \le 3} = (1 - \lambda, \beta, \alpha)$ could be to select the t with index $r \in \{1, 2\}$ satisfying

$$r = \arg\min_{i \in \{1,2\}} \big\{\big|t_{i,2} - a(\mathbf{x}_1)\big| + \big|t_{i,3} - b(\mathbf{x}_1)\big|\big\}. \tag{2.6}$$

This approach allows actually to get a sort of prior ordering very similar to the lexicographical ordering proposed by Teicher [32].

- Crossing point. Let us consider, without loss of generality, three points $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ for which we guess, based on local (estimated) adjacent patterns of functions $a(\cdot)$ and $b(\cdot)$, that $a(\mathbf{x}_1) < b(\mathbf{x}_1)$ and $a(\mathbf{x}_3) > b(\mathbf{x}_3)$. If $\mathbf{x}_1$ and $\mathbf{x}_3$ are close enough, we can suspect that $\mathbf{x}_2$ is in the neighborhood of a crossing point, that is, $a(\mathbf{x}_2) \simeq b(\mathbf{x}_2)$, and decide to estimate $\theta(\mathbf{x}_2)$ by using an estimate-based linear interpolation:

$$\tilde{\theta}(\mathbf{x}_2) := \frac{\hat{\theta}(\mathbf{x}_3) - \hat{\theta}(\mathbf{x}_1)}{\mathbf{x}_3 - \mathbf{x}_1}(\mathbf{x}_2 - \mathbf{x}_1) + \hat{\theta}(\mathbf{x}_1), \tag{2.7}$$

where the general estimator $\hat{\theta}(\cdot)$ is to be defined in (2.10).

Remark. For mixture models with higher number of components, that is,

$$Y_i = \sum_{j=1}^{J} W_j(\mathbf{X}_i)\big(\gamma_j(\mathbf{X}_i) + \varepsilon_{j,i}(\mathbf{X}_i)\big), \qquad i = 1, \dots, n,$$

where $(W_1(\mathbf{x}), \dots, W_J(\mathbf{x}))$ are distributed according to a J-components ($J > 2$) multinomial distribution with parameters $(\pi_1(\mathbf{x}), \dots, \pi_J(\mathbf{x}))$, and noises $(\varepsilon_{j,i})$, $j = 1, \dots, J$, i.i.d. according to $f_{\mathbf{x}}$, we assume that there exists a compact set $\Psi \subset]0,1[^{J-1} \times \mathbb{R}^J$ of parameters $(\pi_1(\mathbf{x}), \dots, \pi_{J-1}(\mathbf{x}), \gamma_1(\mathbf{x}), \dots, \gamma_J(\mathbf{x}))$ where the model is *identifiable*. Note that the 3-components mixture model has been studied closely in [4] and [20] where sufficient identifiability conditions were given. The case where $d > 3$ is more involved for full description and it is still an open question. Identifiability of a location mixture of probability densities was proven in [2] when the mixing density is a Pólya frequency function. In this setup, if the conditional density of the errors is a symmetric Pólya frequency function, the estimation procedure described hereafter can be adapted over the parameter space Ψ with analogous results.

2.3. Estimation procedure

In order to build an estimator of the contrast $S(t)$ defined in (2.4), a local smoothing has to be performed in order to extract the information that the random design $\mathbf{X}_1, \dots, \mathbf{X}_n$ brings to the knowledge of the conditional law of $Y/\mathbf{X} = \mathbf{x}_0$. We use a kernel smoothing approach, but local polynomials or wavelet methods could also be employed. This smoothing is a major difference with respect to the density model considered in [7] and all the rates will depend on the smoothing parameter applied to the kernel function.

Estimation of $\theta(\mathbf{x}_0)$. We choose a kernel function $K : \mathbb{R}^d \to \mathbb{R}$ belonging to $\mathbb{L}_1$ and to $\mathbb{L}_4$ and some bandwidth parameter $h > 0$ to be described later on. For $\mathbf{x}_0 \in \operatorname{supp}(\ell)$ fixed, we denote

$$\begin{aligned} Z_k(t,u,h) &:= \big(e^{iuY_k}\bar{M}(t,u) - e^{-iuY_k}\bar{M}(t,-u)\big)K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &= \big(e^{iuY_k}M(t,-u) - e^{-iuY_k}M(t,u)\big)K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &= 2 \cdot \Im\big(e^{iuY_k}M(t,-u)\big)K_h(\mathbf{X}_k - \mathbf{x}_0), \end{aligned} \tag{2.8}$$

where $K_h(\mathbf{x}) := h^{-d}K(\mathbf{x}/h)$. Indeed, $\bar{M}(t,u) = M(t,-u)$ for all t and u. The empirical contrast of $S(t)$ is defined by

$$S_n(t) = -\frac{1}{4n(n-1)} \sum_{j\neq k, j,k=1}^{n} \int Z_k(t,u,h) Z_j(t,u,h) w(u)\, du, \tag{2.9}$$

where $w : \mathbb{R} \to \mathbb{R}_+^*$ is a bounded p.d.f., having a finite moment of order 4, that is, $\int u^4 w(u)\, du < \infty$. From this empirical contrast, we then define the estimator

$$\hat{\theta}_n = \arg\inf_{t\in\Theta} S_n(t), \tag{2.10}$$

of $\theta_0 = \theta(\mathbf{x}_0)$ where the parametric space Θ is now constrained, for unicity of solution, according to a prior knowledge provided by the rule (2.6). For simplicity, we will suppose that at the point of interest $\mathbf{x}_0$ we have $a(\mathbf{x}_0) < b(\mathbf{x}_0)$, which translates into:

$$\Theta = [p, P] \times \Delta_{\text{ord}}, \tag{2.11}$$

where $0 < p \leq P < 1$ and Δ_{ord} denotes a compact set of $\{(a,b) \in \mathbb{R}^2 : a < b\}$. We shall study successively the properties of $S_n(t)$ as an estimator of $S(t)$ and deduce consistency and asymptotic normality of $\hat{\theta}_n$ as an estimator of θ_0.

Estimation methodology for $f_{\mathbf{x}_0}$. For the estimation of the local noise density $f_{\mathbf{x}_0}$ we suggest to consider the natural smoothed version of the plug-in density estimate given in [36], Section 2.2, under the assumption that $\pi(\mathbf{x}_0) \neq 1/2$.

Let us denote by $\varphi(\mathbf{x}, y) = \ell(\mathbf{x}) f_{\mathbf{x}}(y)$. We plug $\hat{\theta}_n$ in the natural smoothed nonparametric kernel estimator of $\varphi(\mathbf{x}, y)$ deduced from (2.3), whenever the unknown parameter θ_0 is required. For $\mathbf{x}_0$ fixed, we consider the Fourier transform of $\varphi(\mathbf{x}_0, y)$: $\varphi^*_{\mathbf{x}_0}(u) = \ell(\mathbf{x}_0) f^*_{\mathbf{x}_0}(u) = \ell(\mathbf{x}_0) g^*_{\mathbf{x}_0}(u)/M(\theta_0, u)$. Provided that $\hat{\pi}_n \neq 1/2$, which insures $|M(\hat{\theta}_n, u)| \geq |1 - 2\hat{\pi}_n| \neq 0$, we estimate by

$$\varphi^*_{\mathbf{x}_0,n}(u) = \frac{1}{n} \sum_{k=1}^{n} \frac{Q^*(h_{1,n}u) e^{iuY_k}}{M(\hat{\theta}_n, u)} K_{h_{2,n}}(\mathbf{X}_k - \mathbf{x}_0),$$

where Q is a univariate kernel ($\int Q = 1$ and $Q \in \mathbb{L}_2$) and $(h_{1,n}, h_{2,n})$ are bandwidth parameters properly chosen. Note that $G^*_n(u) := Q^*(h_{1,n}u)/M(\hat{\theta}_n, u)$ is in $\mathbb{L}_1$ and $\mathbb{L}_2$ and has an inverse Fourier transform which we denote by $G_n(u/h_{1,n})/h_{1,n}$. Therefore, the estimator of $\varphi(\mathbf{x}_0, y)$ is

$$\varphi_n(\mathbf{x}_0, y) = \frac{1}{nh_{1,n}} \sum_{k=1}^{n} G_n\left(\frac{y - Y_k}{h_{1,n}}\right) K_{h_{2,n}}(\mathbf{X}_k - \mathbf{x}_0).$$

Finally the estimator of $f_{\mathbf{x}_0}$ is obtained by considering

$$\hat{f}_{\mathbf{x}_0}(y) = \frac{f_n(y|\mathbf{x}_0) \mathbb{I}_{f_n(y|\mathbf{x}_0)\geq 0}}{\int_{\mathbb{R}} f_n(y|\mathbf{x}_0) \mathbb{I}_{f_n(y|\mathbf{x}_0)\geq 0}\, dy}, \qquad \text{where } f_n(y|\mathbf{x}_0) = \frac{\varphi_n(\mathbf{x}_0, y)}{\ell_n(\mathbf{x}_0)}, \tag{2.12}$$

where $\ell_n(\mathbf{x}_0) = \frac{1}{n}\sum_{k=1}^{n} K_{h_{2,n}}(\mathbf{X}_k - \mathbf{x}_0)$. The asymptotic properties of this local density estimator are not established yet but we strongly guess that the bandwidth conditions required to prove its convergence and classical convergence rate are similar to those found in the conditional density estimation literature, see [6] or [9].

3. Performance of the method

We give upper bounds for the mean squared error of $S_n(t)$. We are interested in consistency and asymptotic normality of $\hat{\theta}_n$ and this requires some small amount of smoothness $\alpha > 1$ for the functions π, a and b and for the p.d.f. of the errors. From now on, $\|v\|$ denotes the Euclidean norm of vector v.

We say that a function F is Hölder α-smooth if it belongs to the set of functions $L(\alpha, M)$ with $\alpha = k + \beta > 0$ ($k \in \mathbb{N}$ and $\beta \in (0, 1]$) and $M > 0$, such that F has k bounded derivatives and, for all multi-index $j = (j_1, \ldots, j_d) \in \mathbb{N}^d$ with $|j| := j_1 + \cdots + j_d = k$, we have

$$\left|F^{(j)}(\mathbf{x}) - F^{(j)}(\mathbf{y})\right| \leq M\|\mathbf{x} - \mathbf{y}\|^{\beta}, \qquad (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d.$$

A1. We assume that the functions π, a, b, ℓ belong to $L(\alpha, M)$ with $\alpha, M > 0$.

Remark. We may actually suppose that the functions appearing in our model have different smoothness parameters, but the rate will be governed by the smallest smoothness parameter.

An important consequence of this assumption is that the density ℓ is uniformly bounded by some constant depending only on α and M, that is, $\sup_{\ell \in L(\alpha,M)} \|\ell\|_{\infty} < \infty$.

A2. Assume that $f_{\mathbf{x}}(\cdot) \in \mathbb{L}_1 \cap \mathbb{L}_2$ for all $\mathbf{x} \in \mathbb{R}^d$. In addition, we require that there exists a w-integrable function φ such that

$$\left|f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)\right| \leq \varphi(u)\left\|\mathbf{x} - \mathbf{x}'\right\|^{\alpha}, \qquad \left(\mathbf{x}, \mathbf{x}'\right) \in \mathbb{R}^d \times \mathbb{R}^d, u \in \mathbb{R}.$$

Remark. Note that for the scaling model (1.3), if f is the $\mathcal{N}(0, 1)$ p.d.f. and $\sigma(\cdot)$ is bounded and Hölder α-smooth, we have:

$$\left|f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)\right| \leq \frac{u^2}{2}\left|\sigma^2(\mathbf{x}) - \sigma^2\left(\mathbf{x}'\right)\right| \leq C\frac{u^2}{2}\left\|\mathbf{x} - \mathbf{x}'\right\|^{\alpha}.$$

A3. We assume that the kernel K is such that $\int |K| < \infty$, $\int K^4 < \infty$ and that it satisfies also the moment condition

$$\int \|\mathbf{x}\|^{\alpha}\left|K(\mathbf{x})\right| d\mathbf{x} < \infty.$$

A4. The weight function w is a p.d.f. such that

$$\int \left(u^4 + \varphi(u)\right) w(u)\, du < \infty.$$

The following results will hold true under the additional assumption on the kernel (see **A3**): $\int \mathbf{x}^j K(\mathbf{x})\,d\mathbf{x} = 0$, for all j such that $|j| \le k$.

Proposition 2. *For each $t \in \Theta$ and $\mathbf{x}_0 \in \mathrm{supp}(\ell)$ fixed, suppose $\theta_0 \in \mathring{\Theta}$ and that assumptions* **A1–A4** *hold. Then, the empirical contrast function $S_n(\cdot)$ defined in (2.9) satisfies*

$$E\big[\big(S_n(t) - S(t)\big)^2\big] \le C_1 h^{2\alpha} + C_2 \frac{1}{nh^d},$$

if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, where constants C_1, C_2 depend on Θ, K, w, α and M but are free from n, h, t and $\mathbf{x}_0$.

Theorem 2 (Consistency). *Let suppose that assumptions of Proposition 2 hold. The estimator $\hat{\theta}_n$ defined in (2.9)–(2.10) converges in probability to $\theta(\mathbf{x}_0) = \theta_0$ if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.*

The following theorem establishes the asymptotic normality of the estimator $\hat{\theta}_n$ of θ_0. Recall that $\theta_0 = \theta(\mathbf{x}_0)$ belongs to Θ and that there exists $L_* > 0$ such that $\ell(\mathbf{x}_0) \ge L_*$. We see that the local smoothing with bandwidth $h > 0$ deteriorates the rate of convergence to $\sqrt{nh^d}$ instead of $\sqrt{n}$ for the density model. In the asymptotic variance, we will use the following notation:

$$\dot{J}(\theta_0, u) := \Im\big(-\dot{M}(\theta_0, u)\bar{M}(\theta_0, u)\big) f^*_{\mathbf{x}_0}(u)\ell(\mathbf{x}_0), \tag{3.1}$$

and

$$V(\theta_0, u_1, u_2) := 4 \cdot \int \Im\big(e^{iu_1 y}\bar{M}(\theta_0, u_1)\big) \cdot \Im\big(e^{iu_2 y}\bar{M}(\theta_0, u_2)\big) g_{\mathbf{x}_0}(y)\,dy, \tag{3.2}$$

where the function $M(\cdot,\cdot)$ is defined in (2.3). Note that $\dot{J}(\theta_0, \cdot)$ is uniformly bounded by some constant and that V is well defined for all $(u_1, u_2) \in \mathbb{R} \times \mathbb{R}$ and also uniformly bounded by some constant.

Theorem 3 (Asymptotic normality). *Suppose that assumptions of Proposition 2 hold. The estimator $\hat{\theta}_n$ of θ_0 defined by (2.9)–(2.10), with $h \to 0$ such that $nh^d \to \infty$ and such that $h^{2\alpha+d} = o(n^{-1})$, as $n \to \infty$, is asymptotically normally distributed*:

$$\sqrt{nh^d}(\hat{\theta}_n - \theta_0) \to N(0, \mathcal{S}) \qquad \textit{in distribution},$$

where $\mathcal{S} = \frac{1}{4}\mathcal{I}^{-1}\Sigma\mathcal{I}$, with

$$\mathcal{I} = -\frac{1}{2}\int \dot{J}(\theta_0, u)\dot{J}(\theta_0, u)^\top w(u)\,du,$$

and

$$\Sigma := \iint \dot{J}(\theta_0, u_1)\dot{J}^\top(\theta_0, u_2)V(\theta_0, u_1, u_2)w(u_1)w(u_2)\,du_1\,du_2,$$

for $\dot{J}$ defined in (3.1) and V in (3.2).

The above results show that our estimator of θ_0 behaves like any nonparametric pointwise estimator. This is indeed the case and we provide in the next theorem the best achievable convergence rates uniformly over the large set of functions involved in our model, see assumptions **A1**–**A2**.

Theorem 4 (Minimax rates). *Suppose* **A1**–**A4** *and consider* $\mathbf{x}_0 \in \operatorname{supp}(\ell)$ *fixed such that* $\ell(\mathbf{x}_0) \geq L_* > 0$ *for all* $\ell \in L(\alpha, M)$ *and* $\theta_0 = \theta(\mathbf{x}_0) \in \mathring{\Theta} \setminus \{1/2\}$. *The estimator* $\hat{\theta}_n$ *of* θ_0 *defined by* (2.9)–(2.10), *with* $h \asymp n^{-1/(2\alpha+d)}$, *as* $n \to \infty$, *is such that*

$$\sup E\big[\|\hat{\theta}_n - \theta_0\|^2\big] \leq C n^{-2\alpha/(2\alpha+d)},$$

where the supremum is taken over all the functions π, a, b, ℓ *and* f^* *checking assumptions* **A1**–**A2**. *Moreover,*

$$\inf_{T_n} \sup E\big[\|T_n - \theta_0\|^2\big] \geq c n^{-2\alpha/(2\alpha+d)},$$

where $C, c > 0$ *depend only on* α, M, Θ, K *and* w, *and the infimum is taken over the set of all the estimators* T_n (*measurable function of the observations* $(X_1, \ldots, X_n)$) *of* θ_0.

Proof hints. Throughout the proofs of the previous results, we learn that the estimator $\hat{\theta}_n$ of θ_0, behaves asymptotically as $\dot{S}_n(\theta_0)$ which is a U-statistic with a dominant term whose bias is of order $h^{2\alpha}$ and whose variance is smaller than $C_2(nh^d)^{-1}$. The bias-variance compromise will produce an optimal choice of the bandwidth h of order $n^{-1/(2\alpha+d)}$ and a rate $n^{-2\alpha/(2\alpha+d)}$. It is the optimal rate for estimating a Hölder α-smooth regression function at a fixed point and the optimality results in the previous theorem are a consequence of the general nonparametric problem, see [23,31,34].

4. Practical behaviour

4.1. Algorithm

We describe below the initialization scheme and the optimization method used to determine the estimates of the locations $a(\mathbf{x}_k)$, $b(\mathbf{x}_k)$ and the weight functions $\pi(\mathbf{x}_k)$ for a fixed sequence of testing points $\{\mathbf{x}_k, k = 1, \ldots, K\}$. To simply differentiate these testing points from the design data points, we will allocate specifically the index k for the numbering of the testing points and the index i for the numbering of the dataset points, that is, $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$.

Initialization:

1. For each design data point $\mathbf{x}_i$, $i = 1, \ldots, n$, fit a kernel regression smoothing $\bar{m}(\mathbf{x}_i)$ with local bandwidth $\bar{h}_{\mathbf{x}_i}$. The `R` package `lokerns`, see [17], can be used.
2. Classify each data point $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ according to: if $y_i > \bar{m}(\mathbf{x}_i)$ classify $(\mathbf{x}_i, y_i)$ in group 1 associated with location $a(\cdot)$, otherwise classify it in group 2 associated with $b(\cdot)$.

3. For each $\mathbf{x}_k$, $k = 1, \ldots, K$, obtain initial value $\bar{a}(\mathbf{x}_k)$, respectively $\bar{b}(\mathbf{x}_k)$, by fitting a kernel regression smoothing based on the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, previously classified in group 1 with local bandwidth $\bar{h}_{1,\mathbf{x}_k}$, respectively in group 2 with local bandwidth $\bar{h}_{2,\mathbf{x}_k}$.
4. Compute the local bandwidth $h_{\mathbf{x}_k} = \min(\bar{h}_{1,\mathbf{x}_k}, \bar{h}_{2,\mathbf{x}_k})$.
5. Fix an arbitrary single value $\bar{\pi}$ for all the $\pi(\mathbf{x}_k)$'s.

Estimation:

1. Generate one w-distributed i.i.d. sample (U_r), $r = 1, \ldots, N$ dedicated to the pointwize Monte Carlo estimation of $S_n(t)$ defined by:

$$S_n^{\mathrm{MC}}(t) = -\frac{1}{4n(n-1)N} \sum_{j \neq k, j,k=1}^{n} \sum_{r=1}^{N} Z_k(t, U_r, h) Z_j(t, U_r, h).$$

   In the Sections 4.2 and 5, we will consider $N = n$ and w the p.d.f. corresponding to the mixture $0.1 \cdot \mathcal{N}(0, 1) + 0.9 \cdot \mathcal{U}_{[-2,2]}$.
2. Compute the minimizer $\hat{\theta}(\mathbf{x}_k) = (\hat{\pi}(\mathbf{x}_k), \hat{a}(\mathbf{x}_k), \hat{b}(\mathbf{x}_k))$ of $S_n^{\mathrm{MC}}(\cdot)$ evaluated at each point $\mathbf{x}_0 = \mathbf{x}_k$, by using the starting values $(\bar{\pi}, \bar{a}(\mathbf{x}_k), \bar{b}(\mathbf{x}_k))$ and the local bandwidth $h_{\mathbf{x}_k}$.

In our simulations, the above minimization will be deliberately done over a non-constrained space, that is, generically $\theta(\cdot) \in [0.05, 0.95] \times [A, B]^2$, with $A < B$. Our goal is to analyze experimentally if a performant initialization procedure is able to prevent from spurious phenomenons like the label switching or component merging occurring when $\pi(\mathbf{x}_0)$ is close to 0.5. This kind of information is actually very relevant to interpret correctly some cross-over effects as the one we will observe in Figure 10(a). Note that other initialization methods can be figured out. We can for instance use, similarly to [18], a mixture of polynomial regressions with constant proportions and variances to pick initial values $\bar{a}(x)$ and $\bar{b}(x)$, or the `R` package `flexmix`, see [13], that implements a general framework for finite mixture of regression models based on EM-type algorithms (we selected this latter approach for the analysis of radiotherapy application in Section 5).

4.2. Simulations

In this section, we propose to measure the performances of our estimator $\hat{\theta}_n(\cdot)$ over a testing sequence $\{\mathbf{x}_k = k/K\}$, $k = 1, \ldots, K = 20$. Given that in the simulation setting the true function $\theta(\cdot)$ is known, we can compute, similarly to [18], the Root Average Squared Errors (RASE) of our estimator. To this end we generate $M = 100$ datasets $(\mathbf{X}_i^{[z]}, Y_i^{[z]})_{1 \leq i \leq n}$, $z = 1, \ldots, M$ of sizes $n = 400, 800, 1200$, for each of the scenario described below and, for each scalar parameter $s = a, b, \pi$, denote by $\mathrm{RASE}_s^{[z]}$ the RASE performance associated to the zth dataset, defined by $\mathrm{RASE}_s^{[z]} = (1/K \sum_{k=1}^{K} R_s^{[z]}(k))^{1/2}$, where $R_s^{[z]}(k) = (\hat{s}^{[z]}(\mathbf{x}_k) - s(\mathbf{x}_k))^2$, and the empirical RASE by

$$\mathrm{RASE}_s = \frac{1}{M} \sum_{z=1}^{M} \mathrm{RASE}_s^{[z]}. \tag{4.1}$$

Let us also define the empirical squared deviation at point $\mathbf{x}_k$ by $\nu_k = \frac{1}{M}\sum_{z=1}^{M} R_s^{[z]}(k)$, and empirical variance of the squared deviation at $\mathbf{x}_k$ by $\sigma_s^2(k) = \frac{1}{M-1}\sum_{z=1}^{M}(R_s^{[z]}(k) - \nu_k)^2$. From these quantities we deduce the averaged variance of the squared deviations defined by

$$\sigma_s^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_s^2(k). \tag{4.2}$$

In all the simulation setups, we use the same mixing proportion function $\pi(\cdot)$:

$$\pi(\mathbf{x}) = \frac{\sin(3\pi\mathbf{x}) - 1}{15} + 0.4, \qquad \mathbf{x} \in [0, 1].$$

Gaussian setup (G). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Gaussian topographical scaling model corresponding to (1.3), that is, f is the $\mathcal{N}(0, 1)$ p.d.f. when the location and scaling functions are

$$a(\mathbf{x}) = 4 - 2\sin(2\pi\mathbf{x}), \qquad b(\mathbf{x}) = 1.5\cos(3\pi\mathbf{x}) - 3, \qquad \sigma(\mathbf{x}) = 0.9\exp(\mathbf{x}), \qquad \mathbf{x} \in [0, 1].$$

Student setup (T). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Student distribution with continuous degrees of freedom function denoted $df(\mathbf{x})$. The locations and degrees of freedom functions are

$$a(\mathbf{x}) = 3 - 2\sin(2\pi\mathbf{x}), \qquad b(\mathbf{x}) = 1.5\cos(3\pi\mathbf{x}) - 2, \qquad df(\mathbf{x}) = -5\mathbf{x} + 8, \qquad \mathbf{x} \in [0, 1].$$

Laplace setup (L). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Laplace distribution with scaling function $\nu(\mathbf{x})$. The locations and scaling functions are

$$a(\mathbf{x}) = 5 - 3\sin(2\pi\mathbf{x}), \qquad b(\mathbf{x}) = 2\cos(3\pi\mathbf{x}) - 4, \qquad \nu(\mathbf{x}) = \mathbf{x} + 1, \qquad \mathbf{x} \in [0, 1].$$

The selected bandwidths, whose mean and standard deviation are reported in Table 1, are obtained at the initialization step and are extracted from the function `lokerns` of the `R`-library `lokern`. This function calculates an estimator of the regression function with an automatically chosen local plugin bandwidth function. The automatically chosen bandwidths are calculated by finding the bandwidths that minimize the asymptotically optimal mean squared error. To estimate the variance component in the mean squared error this method estimates a functional of a smooth variance function for our heteroscedastic errors.

Table 1. Mean and standard deviation of the `lokerns`-selected Bandwidth

Sample size	Gauss	Student	Laplace
$n = 400$	0.0915 (0.0185)	0.0812 (0.0147)	0.0877 (0.0220)
$n = 800$	0.0860 (0.0099)	0.0780 (0.0091)	0.0823 (0.0151)
$n = 1200$	0.0813 (0.0072)	0.0743 (0.0061)	0.0791 (0.0122)

Table 2. Mean and Standard Deviation of RASEs for data with Gaussian Errors

Sample size	Method	RASE_π	RASE_a	RASE_b
$n = 400$	NMRG	0.011 (0.015)	0.579 (1.064)	0.228 (0.374)
	NMR-SE	0.007 (0.011)	1.031 (2.061)	0.293 (0.581)
$n = 800$	NMRG	0.010 (0.013)	0.505 (0.986)	0.219 (0.401)
	NMR-SE	0.003 (0.005)	0.492 (0.998)	0.150 (0.269)
$n = 1200$	NMRG	0.009 (0.012)	0.474 (0.892)	0.215 (0.401)
	NMR-SE	0.002 (0.003)	0.311 (0.572)	0.123 (0.264)

Comments on Tables 2–4. We report for the simulation setups (**G**), (**T**) and (**L**) the quantities RASE_s defined in (4.1), and between parenthesis σ_s^2 defined in (4.2), for $s = \pi, a, b$. In these tables, we label our method as NMR-SE (Nonparametric Mixture of Regression with Symmetric Errors). To illustrate the contribution of our method, we compare our results with the RASE obtained by using the local EM-type algorithm proposed by Huang, Li and Wang [18] for Nonparametric Mixture of Regression models with Gaussian noises (method labeled for simplicity NMRG). When the errors of the simulated model are Gaussian, the NMRG estimation should outperform our method, since the NMRG method assumes correctly that the errors are normally distributed, while our method does not make any parametric assumption on the distribution of the errors. When the sample size $n = 400$, the NMRG is more precise than our method, since the RASE_s's and σ_s^2's are both smaller for the NMRG. When we increase the sample size of the simulated datasets to $n = 800, 1200$, our method becomes more competitive and yields RASE_s's and σ_s^2's that are lower than those obtained by NMRG. This surprising behavior is probably due to the fact that in model (1.2) we impose the equality in law of the noises up to a shift parameter, when in the NMRG approach possibly different variances are fitted to each kind of noise, increasing by the way drastically the degrees of freedom of the model to be addressed.

In Tables 3 and 4, we observe that our method has globally smaller RASE_s's and σ_s^2's. This result is not surprising, given that in the estimation methodology of Huang, Li and Wang [18], the distribution of the noise are then completely misspecified under the simulation setups (**T**) and (**L**). Note however, that when the sample size is small $n = 400$, the NMRG displays better results, which can be explained by the fact that when we generate small size datasets, the points

Table 3. Mean and Standard Deviation of RASEs for data with Student Errors

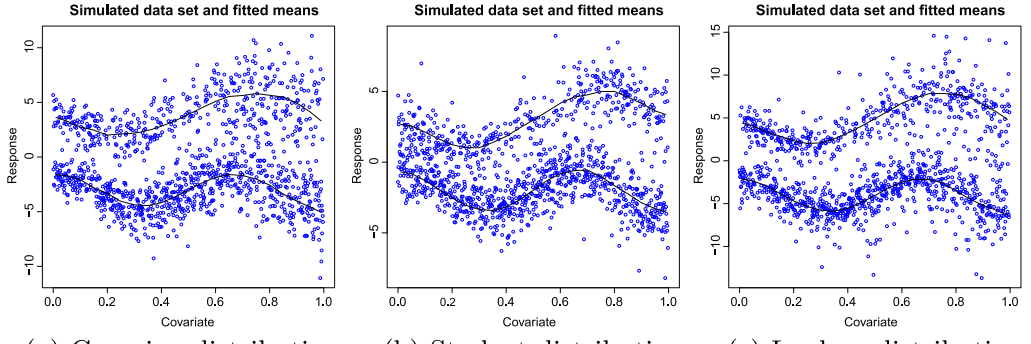
Sample size	Method	RASE_π	RASE_a	RASE_b
$n = 400$	NMRG	0.013 (0.018)	0.330 (0.557)	0.135 (0.196)
	NMR-SE	0.010 (0.016)	0.454 (0.932)	0.217 (0.473)
$n = 800$	NMRG	0.011 (0.014)	0.276 (0.530)	0.101 (0.156)
	NMR-SE	0.004 (0.007)	0.192 (0.374)	0.175 (0.561)
$n = 1200$	NMRG	0.010 (0.014)	0.216 (0.433)	0.111 (0.165)
	NMR-SE	0.003 (0.005)	0.127 (0.255)	0.053 (0.096)

Table 4. Mean and Standard Deviation of RASEs for data with Laplacian Errors

Sample size	Method	RASE_π	RASE_a	RASE_b
$n = 400$	NMRG	0.011 (0.014)	0.815 (1.527)	0.323 (0.493)
	NMR-SE	0.007 (0.001)	1.242 (2.420)	0.376 (0.714)
$n = 800$	NMRG	0.010 (0.013)	0.659 (0.192)	0.283 (0.428)
	NMR-SE	0.003 (0.005)	0.489 (0.870)	0.191 (0.398)
$n = 1200$	NMRG	0.009 (0.012)	0.592 (1.072)	0.236 (0.346)
	NMR-SE	0.002 (0.003)	0.308 (0.566)	0.127 (0.2548)

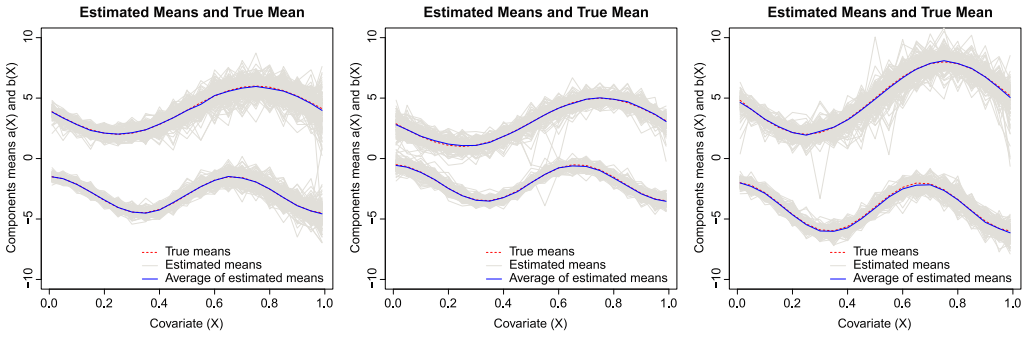
that are supposed to be in the tails of the non-normal distributions are less likely to appear in the dataset. So in that case it can be reasonable to assume that the Gaussian distribution approximates the errors distribution well.

Comments on Figures 2–6. To illustrate the sensitivity of our method and compare it graphically to the NMRG approach, we plot in Figure 1 different samples coming from the setups (**G**), (**T**), and (**L**) for $n = 1200$, and in blue lines the corresponding true location functions $a(\cdot)$ and $b(\cdot)$. In Figure 2, respectively Figure 3, we plot in grey the $M = 200$ segment-line interpolation curves obtained by connecting the points $(\mathbf{x}_k, \hat{s}^{[z]}(\mathbf{x}_k))$, $k = 1, \ldots, K$ where $s(\cdot) = a(\cdot)$, $b(\cdot)$ for the NMRG method, respectively our NMR-SE method. In Figures 4 and 5 we do the same for $s(\cdot) = \pi(\cdot)$. In Figures 2–5, the dashed red lines represent the mean curves obtained by connecting the points $(\mathbf{x}_k, \bar{s}(\mathbf{x}_k))$, $k = 1, \ldots, K$ with $\bar{s}(\mathbf{x}_k) = 1/M \sum_{z=1}^{M} \hat{s}^{[z]}(\mathbf{x}_k)$ and $s(\cdot) = a(\cdot)$, $b(\cdot)$ and $\pi(\cdot)$. Let us observe first that the good behavior of the NMR-SE method is confirmed by the small variability of the curves in Figures 3 and 5 compared to those in Figures 2 and 4 corresponding to the NMRG method. Second, it is important to notice that sometimes, since we did not constrained our method to have $\pi \in [p, P]$ with $0 < p < P < 1/2$, we run into some spurious estimation due to label switching or component merging phenomenon.



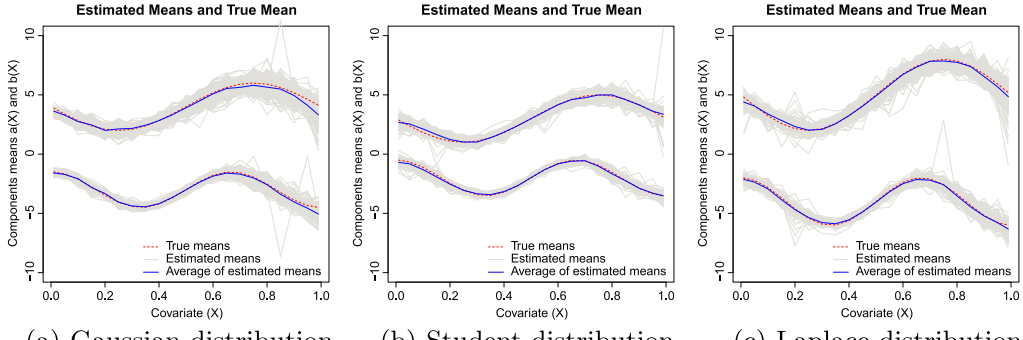
(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure 2. Examples of simulated datasets with different distribution errors.



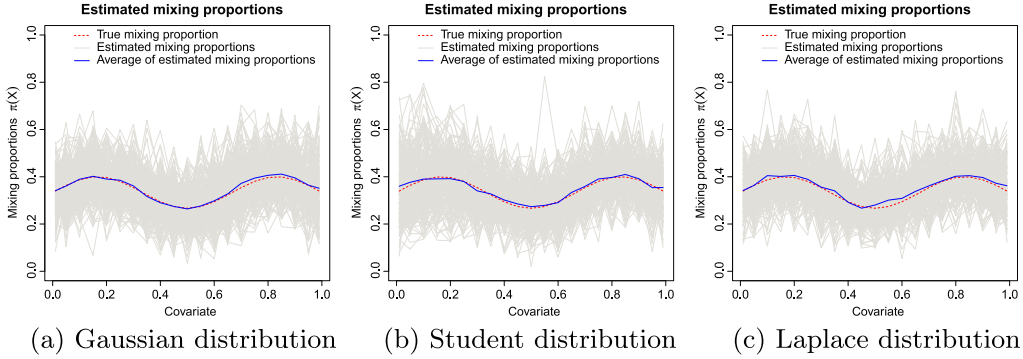
(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure 3. Mean Curves estimated with NMRG.



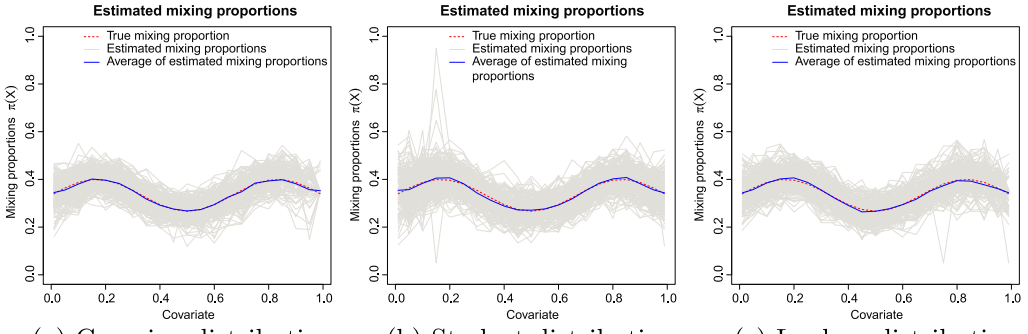
(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure 4. Mean Curves estimated with NMR-SE.



(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure 5. Mixing proportions estimated with NMRG.



(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure 6. Mixing proportions curves estimated with NMR-SE.

Around the value $\pi(\mathbf{x}) = 1/2$. We propose in this paragraph to investigate the practical behavior of our methodology when $\pi(\mathbf{x})$ lies possibly in a neighborhood of $1/2$. In such a case it is known that our model is not identifiable locally. Indeed, there exist various representations of the same observed conditional density, *that is*:

$$g_{\mathbf{x}}(y) = \tfrac{1}{2} f_{\mathbf{x}}\big(y - a(\mathbf{x})\big) + \tfrac{1}{2} f_{\mathbf{x}}\big(y - b(\mathbf{x})\big) = 1h_{\mathbf{x}}\big(y - \mu(\mathbf{x})\big) + \underbrace{0h_{\mathbf{x}}\big(y + \mu(\mathbf{x})\big)}_{\substack{\text{degenerated 2nd component} \\ \text{in model (2.2)}}}, \tag{4.3}$$

where $\mu(\mathbf{x}) := [a(\mathbf{x}) + b(\mathbf{x})]/2$ and

$$h_{\mathbf{x}}(y) := g_{\mathbf{x}}\big(y + \mu(\mathbf{x})\big) = \tfrac{1}{2} f_{\mathbf{x}}\big(y - \big[a(\mathbf{x}) - b(\mathbf{x})\big]/2\big) + \tfrac{1}{2} f_{\mathbf{x}}\big(y + \big[a(\mathbf{x}) - b(\mathbf{x})\big]/2\big).$$

Practically, for moderate values of n, it is reasonable to think that for $\pi(\mathbf{x})$ close to 0.5 our method could detect a $(1-\epsilon)h_{\mathbf{x}}(y-\mu(\mathbf{x})) + \epsilon h_{\mathbf{x}}(y+\mu(\mathbf{x}))$-type model, where ϵ denotes a generic small quantity, close to the degenerated representation in (4.3) instead of the true model (2.2).

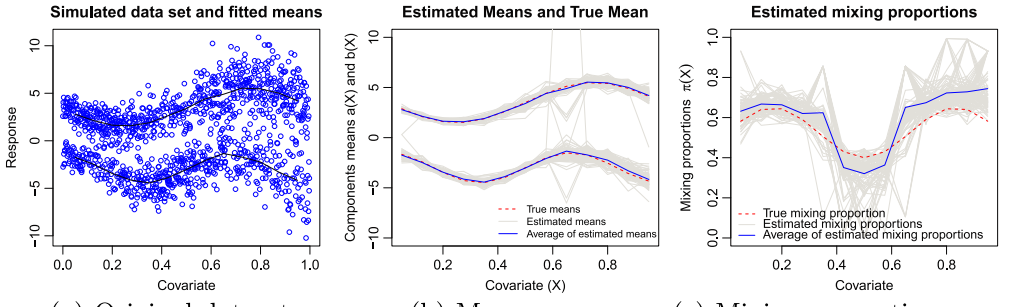
To illustrate the behavior of our method in circumstances "close" to the above situation, we consider the following setup.

Balanced Gaussian setup (BG). The errors $\varepsilon_{j,i}(\mathbf{x})$'s are distributed according to a Gaussian topographical scaling model corresponding to (1.3), that is, f is the $\mathcal{N}(0,1)$ p.d.f. when the location and scaling functions are

$$a(\mathbf{x}) = 5 - 2\sin(2\pi\mathbf{x}), \qquad b(\mathbf{x}) = 1.5\cos(3\pi\mathbf{x}), \qquad \sigma(\mathbf{x}) = 0.9\exp(\mathbf{x}), \qquad \mathbf{x} \in [0,1].$$

and the mixing proportion function is

$$\pi(x) = \frac{\sin(3\pi\mathbf{x}) - 1}{8} + 0.65. \tag{4.4}$$



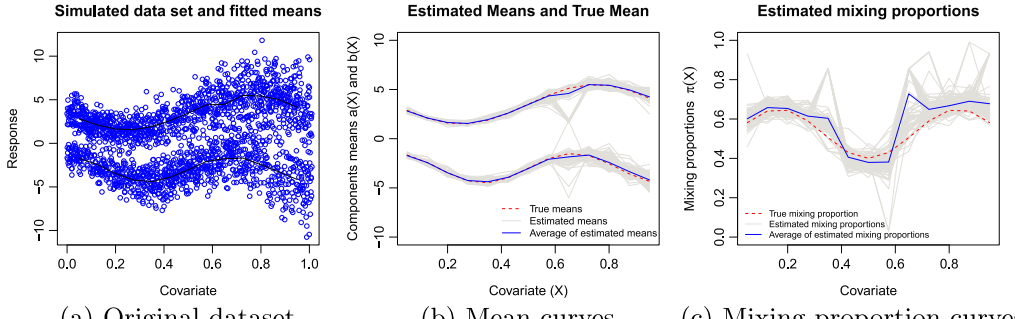
(a) Original dataset (b) Mean curves (c) Mixing proportion curves

Figure 7. Simulated situations with some mixing proportions close to 0.5 and $n = 1200$.

Our methodology is applied on the (**BG**) setup with $n = 1200$, corresponding to Figure 7, and $n = 2000$, corresponding to Figure 8. As it is shown in Figure 7(c), the above parametrization allows to get mixing proportion close to $1/2$ for $\mathbf{x}$ values basically lying in the interval $[0.3, 0.7]$.

Comments on Figures 7 and 8. For simplicity, let us point out first that on the interval $[0, 0.3)$ the variance of the noises is basically rather small and the mean curves near to each other when the variances observed on $(0.7, 1]$ are large and the mean curves distant from each other. When we examine the performances of our method on these intervals it appears that the results are reasonably good on $[0, 0.3)$ for both $n = 1200$ and 2000 while they turn to be much more unstable, specially when considering the mixing proportion curves, on $(0.7, 1]$ for $n = 1200$ with a significant improvement for $n = 2000$. This behavior is probably due to the poor quality of the empirical kernel estimate $S_n(t)$ in (2.9) for $\mathbf{x}$ design values in the range corresponding to bumpy and highly dispersed conditional densities $g_{\mathbf{x}}$, which is precisely the case for $\mathbf{x} \in (0.7, 1]$ compared to $\mathbf{x} \in [0, 0.3)$.

Now, for $\mathbf{x} \in [0.3, 0.7]$ corresponding to $\pi(\mathbf{x})$ values qualitatively close to 0.5, we observe the consequences of the lack of identifiability described in (4.3) since the estimate of $\hat{\pi}(\mathbf{x})$ are



(a) Original dataset (b) Mean curves (c) Mixing proportion curves

Figure 8. Simulated situations with some mixing proportions close to 0.5 and $n = 2000$.

strongly attracted by 0 or 1 when we also observe that $\hat{a}(\mathbf{x})$ or $\hat{b}(\mathbf{x})$ are attracted by the spurious value $\mu(\mathbf{x}) = (a(\mathbf{x}) + b(\mathbf{x}))/2$. Nevertheless, it is important to notice that this drawback is significantly reduced when the sample size n increases from 1200 to 2000, which is in agreement with our asymptotic results proved under the identification constraint $\pi \in [p, P]$ with $0 < p < P < 1$ expressed in Theorems 1–3.

5. Application in radiotherapy

In this section, we implement the proposed methodology to a dataset obtained from applying Positron Emission Radiotherapy (PET) to a canine patient with locally advanced Sinonasal Neoplasia. These data were provided by [5], Figure 4, who used them to quantify the associations between pre-radiotherapy and post-radiotherapy PET-parameters via spatially resolved mixture of linear regressions. Intensity Modulated Radiotherapy is an advanced radiotherapy method that uses computer controlled device to deliver radiation of varying intensities to tumor or smaller areas within the tumor. There is evidence showing that the tumor is not homogeneous in its response to the radiation, and that some regions are more resistant than others. Functional imaging techniques (such as Positron Emission Tomography) can be used to identify the radiotherapy resistant regions within the tumor. For instance, an uptake in PET imaging of follow-up 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG) is empirically linked to a local recurrence of the disease. Bowen *et al*. [5], use this approach to construct a prescription function that maps the image intensity values into a local radiation dose that will maximize the probability of a desired clinical outcome. In their manuscript, they validate the use of molecular imaging based prescription function against clinical outcome by establishing an association between imaging biomarkers (PET imaging pre-radiotherapy) and regional imaging response to known dosage of therapy (PET imaging post-radiotherapy). The regional imaging response captures the change in imaging signal over an individual image volume element (called a voxel). In our model of interest (1.2), the pre-radiotherapy PET imaging intensities correspond to the input $\mathbf{X}_i$'s, and the post-radiotherapy PET imaging levels are the outputs Y_i's. For many patients, the empirical link between post-treatment PET of FDG (regional imaging response) and pre-treatment PET of FDG (imaging biomarker at baseline) is well captured by a mixture regression model with two components. For a set of voxels with similar pre-treatment PET intensities, the nature of the response to the radiotherapy leads to two groups of voxels. The first group corresponds to voxels that respond well to the radiotherapy, and the second group contains the non-responding voxels. In our model of interest (1.2), the non-responding voxel group corresponds to the case where $W(\mathbf{X}_i) = 1$. The location parameters of each group appears to change as the pre-radiotherapy imaging intensity $\mathbf{X}_i$ varies. These changes in location are captured in our model by the location functions $a(\cdot)$ or $b(\cdot)$, where $a(\cdot)$, respectively $b(\cdot)$, is the component mean function for the completely responding (CR), respectively non-responding (NR), voxel. Additionally, the proportion of voxels $\pi(\mathbf{X}_i)$ that respond well to treatment depends on the pre-treatment level of the PET, so the mixture model should also account for a mixing proportion that depends on the input $\mathbf{X}_i$. For a given input $\mathbf{x}$, we assume that the intensity level of the completely responding and the non-responding voxel have approximately the same p.d.f. $f_{\mathbf{x}}$ up to a shift parameter, with the topographical scaling structure (1.3) presented in the Introduction. The variance of the distribution also changes with the level of the covariate (pre-treatment PET FDG). In many cases the variance increases as the intensity

of a voxel's PET pre-radiotherapy increases, this is simply due to the fact the responding voxels will have a low post-treatment PET intensity, while the non-responding voxels will not. The aforementioned topographical scaling property, will allow to model this behavior. To obtain initial values for the location curves $a(\cdot)$ and $b(\cdot)$, we first use the `R` package `flexmix`, see [13], which allows us to fit defined parametric functions to the mixture. For the mixing proportion function we set a fixed constant value $\bar{\pi}(\mathbf{x}) = 0.4$. The bandwidths are computed according to the methodology described in Section 4.1, except that the groups are now determined as an output of the `flexmix` package. The behavior of the local bandwidths selected by the `flexmix` package is displayed in Figure 9.

We propose to apply the NMRG and NMR-SE to this dataset. In Figure 10(a), we show the PET image response to radiotherapy at 3 months, measured by FDG PET uptake, versus the pre-treatment FDG PET uptake. We also display component means obtained by fitting the NMRG and the NMR-SE. For both methods, we observe that the location functions $b(x)$ corresponding to the completely responding voxels, show little variation across the range of values of pre-treatment FDG PET. NMRG and NMR-SE yield fitted means $b(x)$ that are pretty similar to each other.

The fitted location functions $a(x)$ are associated with the non-responding voxels. For both methods, the estimated component means $a(x)$ increase with the pre-treatment FDG PET uptake. A significant difference between NMR-SE and NMRG lies in the fact that the estimated location function $a(x)$ of NMR-SE is slightly greater than the estimated location function obtained with NMRG. This implies that more voxels will be attributed to the non-responding group when we use NMRG instead of NMR-SE. This is confirmed by the Figure 10(b), where we display the mixing proportions $\pi(x)$ for each method. As expected, we see that the NMRG yields mixing proportions of non-responding voxels that are larger than the mixing proportions obtained by using our method. The NMRG mixing proportions lies between (40% and 70 %), while the NMR-SE mixing proportions is between (18% and 60%). The NMR-SE mixing proportion of non-responding voxels is less than 40% for this patient when pre-treatment FDG PET uptake is between 2.75 SUV and 6.875 SUV. We can conclude based on the results from our method that the current radiation dose could be appropriate for patients that exhibit pre-treatment FDG PET uptake close to the range aforementioned. On the other hand, NMRG does not present a wide range of pre-treatment FDG uptake where the non-responding mixing proportion is less than 50%. We see in addition in Figure 11 that the conditional distributions, obtained from formula (2.12) with $h_{1,n} = h_{2,n} = 0.2$, are about zero-symmetric with reasonably small trimming effect due to $\mathbb{I}_{f_n(y|\mathbf{x}_0)\geq 0}$ in (2.12) (tiny wave effect on both sides of the main mode). This is a good model validation tool since we are actually able to recover, after local Fourier inversion, the basic symmetry assumption technically made on the distributions of the errors; see for quality comparison other existing (nonconditional) semiparametric inversion density estimates performed on real datasets: Figures 1–2(a) in [4], Figure 3 in [7], Figure 5 in [36], or Figures 2–3 in [3].

6. Auxiliary results and main proofs

Let us denote by $\|\cdot\|$ the Euclidean norm of a vector and by $\|\cdot\|_2$ the Frobenius norm of any squared matrix. Recall the definition of Z_k in (2.8) and let $J(t,u,h) := E[Z_1(t,u,h)]$. Let $\dot{Z}_k$ and $\dot{J}$ denote respectively, the gradient of Z_k and J with respect to their first argument t.

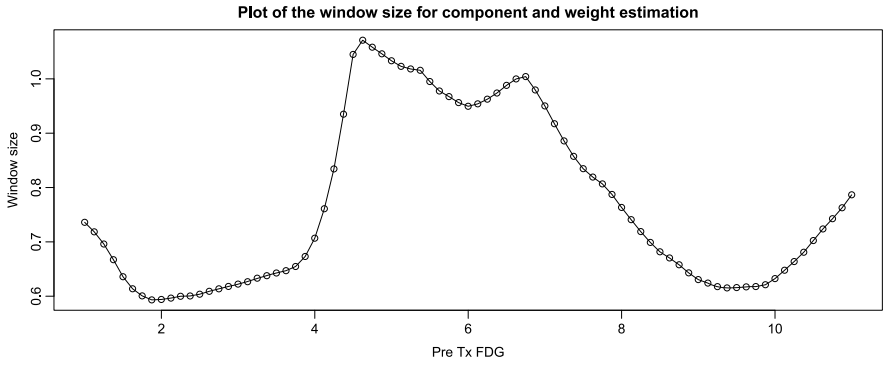
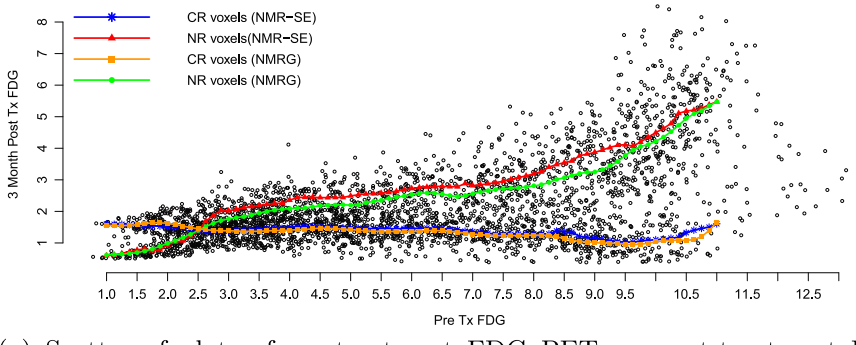
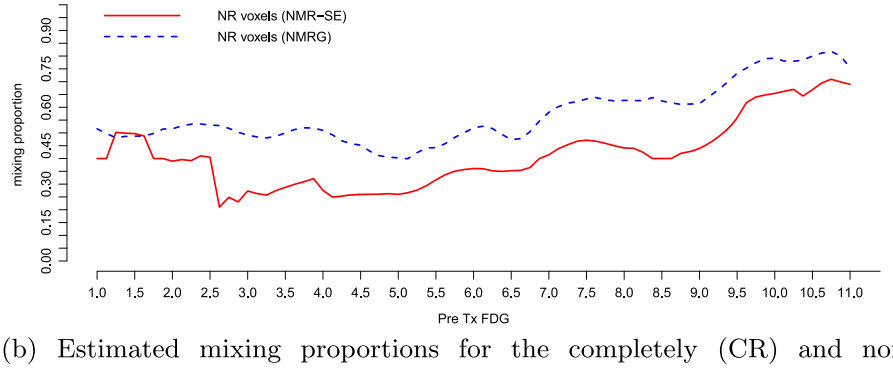


Figure 9. Behavior of the local bandwidths selected by the `flexmix` package in the PET application.



(a) Scatter of plots of pre-treatment FDG PET vs. post-treatment FDG PET and estimated location functions for the completely respondent and non-respondent voxel subpopulations



(b) Estimated mixing proportions for the completely (CR) and non-respondent (NR) voxel subpopulation

Figure 10. Location and mixing proportion function estimation by using NMR-SE and NMRG methods.

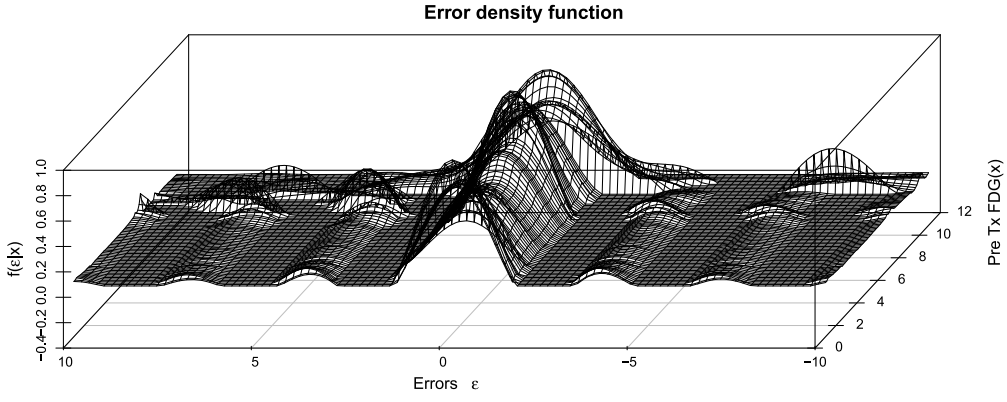


Figure 11. Density Estimates of the errors for the different levels of PET Tx FDG values.

Lemma 1. *Under assumption* **A1** *we have*:

(i) *For all* $(u,h) \in \mathbb{R} \times \mathbb{R}_+^*$ *and any* $k = 1, \dots, n$,

$$\sup_{t\in\Theta} \big| Z_k(t,u,h) \big| \le \frac{2\|K\|_\infty}{h^d}, \qquad \sup_{t\in\Theta} \big| J(t,u,h) \big| \le 2\|\ell\|_\infty \int |K|.$$

(ii) *For all* $(u,h) \in \mathbb{R} \times \mathbb{R}_+^*$ *and any* $k = 1, \dots, n$,

$$\sup_{t\in\Theta} \big\| \dot{Z}_k(t,u,h) \big\| \le 4\big(1+|u|\big)\frac{\|K\|_\infty}{h^d}, \qquad \sup_{t\in\Theta} \big\| \dot{J}(t,u,h) \big\| \le 4\big(1+|u|\big)\|\ell\|_\infty \int |K|.$$

(iii) *For all* $(u,h) \in \mathbb{R} \times \mathbb{R}_+^*$ *and any* $k = 1, \dots, n$,

$$\sup_{t\in\Theta} \big\| \ddot{Z}_k(t,u,h) \big\|_2 \le C\big(1+|u|+u^2\big)\frac{\|K\|_\infty}{h^d},$$

$$\sup_{t\in\Theta} \big\| \ddot{J}_k(t,u,h) \big\|_2 \le C\big(1+|u|+u^2\big)\|\ell\|_\infty \cdot \int |K|,$$

for some constant $C > 0$.

Proof of Lemma 1. (i) It is easy to see, from $|M(t,u)| \le 1$, that

$$\big| Z_k(t,u,h) \big| \le 2\big| K_h(\mathbf{X}_k - \mathbf{x}_0) \big| \le 2\frac{\|K\|_\infty}{h^d},$$

and that

$$\big| J(t,u,h) \big| \le 2\left| \int \Im\big(g_{\mathbf{x}}^*(u)\bar{M}(t,u)\big) K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})\,d\mathbf{x} \right| \le 2\|\ell\|_\infty \cdot \int |K|.$$

(ii) We note that

$$\dot{Z}_k(t,u,h) = \left\{ e^{iuY_k} \begin{pmatrix} e^{-iu\alpha} - e^{-iu\beta} \\ -iu\pi e^{-iu\alpha} \\ -iu(1-\pi)e^{-iu\beta} \end{pmatrix} - e^{-iuY_k} \begin{pmatrix} e^{iu\alpha} - e^{iu\beta} \\ iu\pi e^{iu\alpha} \\ iu(1-\pi)e^{iu\beta} \end{pmatrix} \right\} K_h(\mathbf{X}_k - \mathbf{x}_0),$$

and that

$$\begin{aligned} E\big[\dot{Z}_k(t,u,h)\big] &= \dot{J}_k(t,u,h) \\ &= \int \left\{ g_{\mathbf{x}^*}(u) \begin{pmatrix} e^{-iu\alpha} - e^{-iu\beta} \\ -iu\pi e^{-iu\alpha} \\ -iu(1-\pi)e^{-iu\beta} \end{pmatrix} - g_{\mathbf{x}^*}(-u) \begin{pmatrix} e^{iu\alpha} - e^{iu\beta} \\ iu\pi e^{iu\alpha} \\ iu(1-\pi)e^{iu\beta} \end{pmatrix} \right\} \\ &\quad \times K_h(\mathbf{x} - \mathbf{x}_0)\ell(\mathbf{x})\,d\mathbf{x}. \end{aligned}$$

We thus have

$$\begin{aligned} \big\|\dot{Z}_k(t,u,h)\big\| &= \big\|e^{iuY_k}\dot{M}(t,-u) - e^{-iuY_k}\dot{M}(t,u)\big\| K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &\le \big(2\big(2^2 + P^2u^2 + (1-p)^2u^2\big)\big)^{1/2} K_h(\mathbf{X}_k - \mathbf{x}_0) \\ &\le 4\big(1+|u|\big)\frac{\|K\|_\infty}{h^d}, \end{aligned}$$

and

$$\begin{aligned} \big\|\dot{J}_k(t,u,h)\big\| &= \int \big\|g^*_{\mathbf{x}}(u)\dot{M}(t,-u) - g^*_{\mathbf{x}}(-u)\dot{M}(t,u)\big\| \big|K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})\big|\,d\mathbf{x} \\ &\le \big(2\big(2^2 + P^2u^2 + (1-p)^2u^2\big)\big)^{1/2} \int \big|K_h(\mathbf{x}-\mathbf{x}_0)\ell(\mathbf{x})\big|\,d\mathbf{x} \\ &\le 4\big(1+|u|\big)\|\ell\|_\infty . \int |K|. \end{aligned}$$

(iii) Formula of $\ddot{M}(t,u)$ being tedious, we shortly write that

$$\ddot{Z}_k(t,u,h) = \big\{e^{iuY_k}\ddot{M}(t,-u) - e^{-iuY_k}\ddot{M}(t,u)\big\}K_h(\mathbf{X}_k - \mathbf{x}_0),$$

and deduce our bound from the above expression using arguments similar to (i) and (ii). □

Lemma 2. (i) *For all* $(t,t') \in \Theta^2$, *there exists a constant* $C_1 > 0$ *such that*

$$\big|S_n(t) - S_n\big(t'\big)\big| \le C_1\big\|t - t'\big\| \sum_{j\ne k,\,j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)}.$$

(ii) *For all* $(t,t') \in \Theta^2$, *there exists a constant* $C_2 > 0$ *such that*

$$\big\|\ddot{S}_n(t) - \ddot{S}_n\big(t'\big)\big\|_2 \le C_2\big\|t - t'\big\| \sum_{j\ne k,\,j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)}.$$

(iii) *There exists some constants $C_1, C_2 > 0$ depending on Θ, α, M, K such that*

$$E\left[\left(\sum_{j\neq k,\, j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} - \ell^2(\mathbf{x}_0)\right)^2\right] \leq C_1 h^{2\alpha} + \frac{C_2}{nh^d},$$

as $h \to 0$ and $nh^d \to \infty$.

Proof. (i) By a first order Taylor expansion we have

$$S_n(t) - S_n\big(t'\big) = -\frac{1}{2n(n-1)} \int \big(t - t'\big)^\top \sum_{j\neq k,\, j,k=1}^{n} \dot{Z}_k(t_u, u, h) Z_j(t_u, u, h) w(u)\, du,$$

where for all $u \in \mathbb{R}$, t_u lies in the line segment with extremities t and t'. Therefore, according to calculations made in the proofs of Lemma 1(i) and (ii), we obtain

$$\big|S_n(t) - S_n\big(t'\big)\big| \leq \big\|t - t'\big\| \int_{\mathbb{R}} 4\big(1 + |u|\big) w(u)\, du \left| \sum_{j\neq k,\, j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} \right|,$$

which ends the proof of (i) by using assumption **A4**.

(ii) Let recall first that

$$\ddot{S}_n(t) = \frac{-1}{2n(n-1)} \sum_{k\neq j} \int \big[\ddot{Z}_k(t, u, h) Z_j(t, u, h) + \dot{Z}_k(t, u, h) \dot{Z}_j(t, u)^\top\big] w(u)\, du.$$

We shall bound from above as follows

$$\begin{aligned}
\big\|\ddot{S}_n(t, u) - \ddot{S}_n\big(t', u\big)\big\|_2 \leq \frac{1}{2n(n-1)} \sum_{k\neq j} \Bigg\{ & \left\| \int \big(\ddot{Z}_k(t, u, h) - \ddot{Z}_k\big(t', u, h\big)\big) Z_j(t, u) w(u)\, du \right\|_2 \\
& + \left\| \int \ddot{Z}_k\big(t', u, h\big)\big(Z_j(t, u, h) - Z_j\big(t', u, h\big)\big) w(u)\, du \right\|_2 \\
& + \left\| \int \dot{Z}_k(t, u, h)\big(\dot{Z}_j(t, u, h) - \dot{Z}_j\big(t', u, h\big)\big)^\top w(u)\, du \right\|_2 \\
& + \left\| \int \big(\dot{Z}_k(t, u, h) - \dot{Z}_k\big(t', u, h\big)\big) \dot{Z}_j\big(t', u, h\big)^\top w(u)\, du \right\|_2 \Bigg\}.
\end{aligned}$$

For each term in the previous sum, we use Taylor expansion and upper-bounds similar to those developed in the proof of Lemma 1, and get

$$\begin{aligned}
&\big\|\ddot{S}_n(t, u) - \ddot{S}_n\big(t', u\big)\big\|_2 \\
&\quad \leq \big\|t - t'\big\| C \int \big(1 + |u| + u^2 + |u|^3\big) w(u)\, du \left| \sum_{j\neq k,\, j,k=1}^{n} \frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} \right|,
\end{aligned}$$

for some constant $C > 0$, which finishes the proof by using assumption **A4**.

(iii) The proof is a consequence of Proposition 2 hereafter. □

Proof of Proposition 2. We shall bound from above the mean square error by the usual decomposition into squared bias plus variance.

Note that

$$E\big[S_n(t)\big] = -\frac{1}{4}\int \big(E\big[Z_1(t,u,h)\big]\big)^2 w(u)\,du$$

as $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ are independent. Moreover,

$$\begin{aligned} E\big[Z_1(t,u,h)\big] &= \int\!\!\int \big(e^{iuy}M(t,-u) - e^{-iuy}M(t,u)\big)K_h(\mathbf{x}-\mathbf{x}_0)g(y,\mathbf{x})\,dy\,d\mathbf{x} \\ &= \int \left(\int \big(e^{iuy}M(t,-u) - e^{-iuy}M(t,u)\big)g_{\mathbf{x}}(y)\,dy\right)\ell(\mathbf{x})K_h(\mathbf{x}-\mathbf{x}_0)\,d\mathbf{x} \\ &= \int \big(g^*_{\mathbf{x}}(u)M(t,-u) - g^*_{\mathbf{x}}(-u)M(t,u)\big)\ell(\mathbf{x})K_h(\mathbf{x}-\mathbf{x}_0)\,d\mathbf{x}. \end{aligned}$$

Let us denote by $L(\mathbf{x},t,u) := g^*_{\mathbf{x}}(u)M(t,-u) - g^*_{\mathbf{x}}(-u)M(t,u)$, which is further equal to

$$L(\mathbf{x},t,u) = 2i\cdot\Im\big(g^*_{\mathbf{x}}(u)M(t,-u)\big) = 2i\cdot\Im\big(M\big(\theta(\mathbf{x}),u\big)M(t,-u)\big)f^*_{\mathbf{x}}(u).$$

We can write $E[Z_1(t,u,h)] = [(L(\cdot,t,u)\ell)\star K_h](\mathbf{x}_0)$, where $\star$ denotes the convolution product. The bias of $S_n(t)$ is bounded from above as follows:

$$\begin{aligned} \big|E\big[S_n(t)\big] - S(t)\big| &= \frac{1}{4}\left|\int \big(\big[\big(L(\cdot,t,u)\ell\big)\star K_h\big]^2(\mathbf{x}_0) - L^2(\mathbf{x}_0,t,u)\ell^2(\mathbf{x}_0)\big)w(u)\,du\right| \\ &\le \frac{1}{4}\int \big|\big[\big(L(\cdot,t,u)\ell\big)\star K_h\big](\mathbf{x}_0) - L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)\big| \\ &\quad\times \big|\big[\big(L(\cdot,t,u)\ell\big)\star K_h\big](\mathbf{x}_0) + L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)\big|w(u)\,du. \end{aligned}$$

Now

$$\big|L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)\big| \le 2\|\ell\|_\infty \le 2C,$$

as $\|\ell\|_\infty$ is further bounded by a constant $C = C(\alpha, M)$ depending only on α, $M > 0$, uniformly over $\ell \in L(\alpha, M)$ (see remark following condition **A1**). We also have

$$\begin{aligned} E\big[Z_1(t,u,h)\big] = \big|\big[\big(L(\cdot,t,u)\ell\big)\star K_h\big](\mathbf{x}_0)\big| &\le \int \big|L(\mathbf{x},t,u)\big|l(\mathbf{x})|K|_h(\mathbf{x}-\mathbf{x}_0)\,d\mathbf{x} \\ &\le 2C\int |K|. \end{aligned} \tag{6.1}$$

Moreover, for all $u \in \mathbb{R}$,

$$\begin{aligned}
&\big|\big[\big(L(\cdot,t,u)\ell\big) \star K_h\big](\mathbf{x}_0) - L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)\big| \\
&\quad \le \int \big|L(\mathbf{x}+\mathbf{x}_0,t,u)\ell(\mathbf{x}+\mathbf{x}_0) - L(\mathbf{x}_0,t,u)\ell(\mathbf{x}_0)\big| \cdot |K|_h(\mathbf{x})\,d\mathbf{x} \\
&\quad \le c\big(|u|+\varphi(u)\big) \int \|\mathbf{x}\|^\alpha \cdot |K|_h(\mathbf{x})\,d\mathbf{x} \le c \cdot h^\alpha\big(|u|+\varphi(u)\big) \int \|\mathbf{x}\|^\alpha \cdot |K|(\mathbf{x})\,d\mathbf{x},
\end{aligned}$$

under our assumptions **A1**–**A4**. Indeed, that implies that $L(\cdot,t,u)\ell(\cdot)$ is Hölder α-smooth for all $(t,u) \in \Theta \times \mathbb{R}$, with some constant $c > 0$, see Lemma 3. Therefore, we get

$$\big|E\big[S_n(t)\big] - S(t)\big| \le 2C\bigg(1 + \int |K|\bigg)c\bigg(\int \|\mathbf{x}\|^\alpha \cdot |K|(\mathbf{x})\,d\mathbf{x}\bigg) \cdot \bigg(\int |u| w(u)\,du\bigg) \cdot h^\alpha.$$

Similarly to $S_n(t)$ variance decomposition, we write

$$\begin{aligned}
&S_n(t) - E\big[S_n(t)\big] \\
&\quad = \frac{-1}{4n(n-1)} \sum_{j \ne k} \bigg(\int \big(Z_j(t,u,h)Z_k(t,u,h) - E^2\big[Z_1(t,u,h)\big]\big)w(u)\,du\bigg) \\
&\quad = \frac{-1}{2n} \sum_j \int \big(Z_j(t,u,h) - E\big[Z_1(t,u,h)\big]\big)E\big[Z_1(t,u,h)\big]w(u)\,du \\
&\qquad + \frac{-1}{4n(n-1)} \\
&\qquad \times \sum_{j \ne k} \bigg(\int \big(Z_j(t,u,h) - E\big[Z_1(t,u,h)\big]\big)\big(Z_k(t,u,h) - E\big[Z_1(t,u,h)\big]\big)w(u)\,du\bigg) \\
&\quad = T_1 + T_2, \qquad \text{say.}
\end{aligned}$$

Terms in T_1 and T_2 are uncorrelated and thus $\operatorname{Var}(S_n(t)) = \operatorname{Var}(T_1) + \operatorname{Var}(T_2)$.

On the one hand,

$$\begin{aligned}
\operatorname{Var}(T_1) &= \frac{1}{4n} \operatorname{Var}\bigg(\int \big(Z_1(t,u,h) - E\big[Z_1(t,u,h)\big]\big)E\big[Z_1(t,u,h)\big]w(u)\,du\bigg) \\
&= \frac{1}{4n} E\bigg[\bigg|\int \big(Z_1(t,u,h) - E\big[Z_1(t,u,h)\big]\big)E\big[Z_1(t,u,h)\big]w(u)\,du\bigg|^2\bigg] \\
&\le \frac{1}{4n} E\bigg[\int \big|Z_1(t,u,h) - E\big[Z_1(t,u,h)\big]\big|^2 w(u)\,du\bigg] \int \big|E\big[Z_1(t,u,h)\big]\big|^2 w(u)\,du,
\end{aligned}$$

according to Cauchy–Schwarz inequality. Now we use (6.1) and obtain

$$\operatorname{Var}(T_2) \le \frac{1}{4n}\bigg(2C \int |K|\bigg)^2 \int E\big[\big|Z_1(t,u,h)\big|^2\big]w(u)\,du.$$

We have,

$$\begin{aligned} E\big[\big|Z_1(t,u,h)\big|^2\big] &= E\big[E\big[\big|2i\cdot\Im\big(e^{iuY}M(t,-u)\big)\big|^2|\mathbf{X}\big]\big(K_h(\mathbf{X}-\mathbf{x}_0)\big)^2\big] \\ &= 4E\big[\big|\Im\big(g^*_{\mathbf{X}}(u)M(t,-u)\big)\big|^2\big(K_h(\mathbf{X}-\mathbf{x}_0)\big)^2\big] \\ &\le 4\int \frac{1}{h^{2d}}K^2\bigg(\frac{\mathbf{x}-\mathbf{x}_0}{h}\bigg)\ell(\mathbf{x})\,d\mathbf{x} \\ &\le 4C\frac{\int K^2}{h^d}. \end{aligned}$$

Therefore,

$$\operatorname{Var}(T_1) \le 4C^3\frac{(\int|K|)^2\int K^2}{nh^d}, \tag{6.2}$$

for all $t\in\Theta$, $h>0$.

On the other hand,

$$\begin{aligned} \operatorname{Var}(T_2) &= \frac{1}{16n(n-1)} \\ &\quad\times E\bigg[\bigg|\int\big(Z_1(t,u,h)-E\big[Z_1(t,u,h)\big]\big)\big(Z_2(t,u,h)-E\big[Z_2(t,u,h)\big]\big)w(u)\,du\bigg|^2\bigg] \\ &\le \frac{1}{16n(n-1)} \\ &\quad\times E\bigg[\int\big|Z_1(t,u,h)-E\big[Z_1(t,u,h)\big]\big|^2\big|Z_2(t,u,h)-E\big[Z_2(t,u,h)\big]\big|^2w(u)\,du\bigg] \\ &\le \frac{1}{16n(n-1)}\int E^2\big[\big|Z_1(t,u,h)\big|^2\big]w(u)\,du \le \frac{1}{16n(n-1)}\bigg(\frac{2C\int K^2}{h^d}\bigg)^2 \\ &= \frac{C^2(\int K^2)^2}{4n(n-1)h^{2d}}, \end{aligned}$$

which is clearly a $o((nh^d)^{-1})$ and concludes the proof. □

Lemma 3 (Smoothness of $L(\mathbf{x},t,u)\ell(\mathbf{x})$). *Assume* **A1–A4**. *There exists a constant $C>0$, such that for all $(\mathbf{x},\mathbf{x}')\in\mathbb{R}^d\times\mathbb{R}^d$ and all $(t,u)\in\Theta\times\mathbb{R}$:*

$$\big|L(\mathbf{x},t,u)\ell(\mathbf{x})-L\big(\mathbf{x}',t,u\big)\ell\big(\mathbf{x}'\big)\big| \le C\big(|u|+\varphi(u)\big)\big\|\mathbf{x}-\mathbf{x}'\big\|^\alpha.$$

Proof. For $t=(\pi,a,b)\in\Theta$, and $(\mathbf{x},u)\in\mathbb{R}^d\times\mathbb{R}$ we write

$$L(\mathbf{x},t,u)\ell(\mathbf{x}) = f^*_{\mathbf{x}}(u)\ell(\mathbf{x})\mathcal{T}(\mathbf{x},t,u) \quad\text{and}\quad \mathcal{T}(\mathbf{x},t,u) := \sum_{i=1}^4\mathcal{T}_i(\mathbf{x},t,u),$$

where

$$\begin{aligned}
\mathcal{T}_1(\mathbf{x},t,u) &= \pi(\mathbf{x})\pi\,\sin\big[u\big(a(\mathbf{x})-a\big)\big],\\
\mathcal{T}_2(\mathbf{x},t,u) &= \pi(\mathbf{x})(1-\pi)\,\sin\big[u\big(a(\mathbf{x})-b\big)\big],\\
\mathcal{T}_3(\mathbf{x},t,u) &= \big(1-\pi(\mathbf{x})\big)\pi\,\sin\big[u\big(b(\mathbf{x})-a\big)\big],\\
\mathcal{T}_4(\mathbf{x},t,u) &= \big(1-\pi(\mathbf{x})\big)(1-\pi)\,\sin\big[u\big(b(\mathbf{x})-b\big)\big].
\end{aligned}$$

For all $(\mathbf{x},\mathbf{x}')\in\mathbb{R}^d\times\mathbb{R}^d$ we have

$$\begin{aligned}
&\big|L(\mathbf{x},t,u)\ell(\mathbf{x})-L\big(\mathbf{x}',t,u\big)\ell\big(\mathbf{x}'\big)\big|\\
&\quad\le 2\big|f^*_{\mathbf{x}}(u)\ell(\mathbf{x})\big|\big|\mathcal{T}(\mathbf{x},t,u)-\mathcal{T}\big(\mathbf{x}',t,u\big)\big|+2\big|\mathcal{T}\big(\mathbf{x}',t,u\big)\big|\big|f^*_{\mathbf{x}}(u)\ell(\mathbf{x})-f^*_{\mathbf{x}'}(u)\ell\big(\mathbf{x}'\big)\big|\\
&\quad\le 2\|\ell\|_\infty\big|\mathcal{T}(\mathbf{x},t,u)-\mathcal{T}\big(\mathbf{x}',t,u\big)\big|+2\big|f^*_{\mathbf{x}}(u)\ell(\mathbf{x})-f^*_{\mathbf{x}'}(u)\ell\big(\mathbf{x}'\big)\big|.
\end{aligned}$$

Let us now show the α-smooth Hölder property of $\mathcal{T}_1$, the proof for the other $\mathcal{T}_i$'s being completely similar. For all $(\mathbf{x},\mathbf{x}')\in\mathbb{R}^d\times\mathbb{R}^d$

$$\begin{aligned}
\big|\mathcal{T}_1(\mathbf{x},t,u)-\mathcal{T}_1\big(\mathbf{x}',t,u\big)\big| &\le \big|\sin\big[u\big(a(\mathbf{x})-a\big)\big]-\sin\big[u\big(a\big(\mathbf{x}'\big)-a\big)\big]\big|+\big|\pi(\mathbf{x})-\pi\big(\mathbf{x}'\big)\big|\\
&\le |u|\big|\big(a(\mathbf{x})-a\big(\mathbf{x}'\big)\big]\big|+\big|\pi(\mathbf{x})-\pi\big(\mathbf{x}'\big)\big|\\
&\le M|u|\big\|\mathbf{x}-\mathbf{x}'\big\|^\alpha+M\big\|\mathbf{x}-\mathbf{x}'\big\|^\alpha.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\big|f^*_{\mathbf{x}}(u)\ell(\mathbf{x})-f^*_{\mathbf{x}'}\ell\big(\mathbf{x}'\big)\big| &\le \big|\ell(\mathbf{x})-\ell\big(\mathbf{x}'\big)\big|+\|\ell\|_\infty\big|f^*_{\mathbf{x}}(u)-f^*_{\mathbf{x}'}(u)\big|,\\
&\le \big(M+\|\ell\|_\infty\varphi(u)\big)\big\|\mathbf{x}-\mathbf{x}'\big\|^\alpha,
\end{aligned}$$

which concludes the proof. □

Proof of Theorem 2. Our method is based on a consistency proof for mininum contrast estimators by [10], pages 94–96. Let us consider a countable dense set D in Θ, then $\inf_{t\in\Theta}S_n(t)=\inf_{t\in D}S_n(t)$, is a measurable random variable. We define in addition the random variable

$$W(n,\xi)=\sup\big\{\big|S_n(t)-S_n\big(t'\big)\big|;\,\big(t,t'\big)\in D^2,\,\big\|t-t'\big\|\le\xi\big\},$$

and recall that $S(\theta_0)=0$. Let us consider a non-empty open ball B_* centered on θ_0 such that S is bounded from below by a positive real number 2ε on $\Theta\setminus B_*$. Let us consider a sequence $(\xi_p)_{p\ge1}$ decreasing to zero, and take p such that there exists a covering of $\Theta\setminus B_*$ by a finite number κ of balls $(B_i)_{1\le i\le\kappa}$ with centers $t_i\in\Theta$, $i=1,\dots,\kappa$, and radius less than ξ_p. Then, for all $t\in B_i$, we have

$$S_n(t)\ge S_n(t_i)-\big|S_n(t)-S_n(t_i)\big|\ge S_n(t_i)-\sup_{t\in B_i}\big|S_n(t)-S_n(t_i)\big|,$$

which leads to

$$\inf_{t\in\Theta\setminus B_*} S_n(t) \geq \inf_{1\leq i\leq\kappa} S_n(t_i) - W(n,\xi_p).$$

As a consequence, we have the following events inclusions

$$\begin{aligned}\{\hat{\theta}_n \notin B_*\} &\subseteq \Big\{\inf_{t\in\Theta\setminus B_*} S_n(t) < \inf_{t\in B_*} S_n(t) < S_n(\theta_0)\Big\}\\ &\subseteq \Big\{\inf_{1\leq i\leq\kappa} S_n(t_i) - W(n,\xi_p) < S_n(\theta_0)\Big\}\\ &\subseteq \big\{W(n,\xi_p) > \varepsilon\big\} \cup \Big\{\inf_{1\leq i\leq\kappa}\big(S_n(t_i) - S_n(\theta_0)\big) \leq \varepsilon\Big\}.\end{aligned}$$

In addition, we have

$$\begin{aligned}&P\Big(\inf_{1\leq i\leq\kappa}\big(S_n(t_i) - S_n(\theta_0)\big) \leq \varepsilon\Big)\\ &\quad\leq 1 - \prod_{i=1}^{\kappa}\big(1 - \big[P\big(\big|S_n(t_i) - S(t_i)\big| \geq \varepsilon\big) + P\big(\big|S_n(\theta_0) - S(\theta_0)\big| \geq \varepsilon\big)\big]\big),\end{aligned}$$

where, according to Proposition 2, the last two terms in the right-hand side of the above inequality vanish to zero if $h^d n \to \infty$ and $h \to 0$ as $n \to \infty$. To conclude we use Lemma 2 and notice that, for all $(t,t') \in \Theta^2$, we have

$$\begin{aligned}&\big|S_n(t) - S_n\big(t'\big)\big|\\ &\quad\leq \frac{C\|t-t'\|}{n(n-1)}\left|\sum_{j\neq k,\,j,k=1}^{n} K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)\right| \qquad (6.3)\\ &\quad\leq C\big\|t-t'\big\|\ell^2(\mathbf{x}_0) + C\big\|t-t'\big\|\left|\sum_{j\neq k,\,j,k=1}^{n}\frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} - \ell^2(\mathbf{x}_0)\right|.\end{aligned}$$

We deduce from above that

$$\begin{aligned}P\big(W(n,\xi_p) > \varepsilon\big) &\leq P\left(C\xi_p\ell^2(\mathbf{x}_0) > \frac{\varepsilon}{2}\right)\\ &\quad + \left(\frac{2C\xi_p}{\varepsilon}\right)^2 E\left[\left(\sum_{j\neq k,\,j,k=1}^{n}\frac{K_h(\mathbf{X}_k - \mathbf{x}_0)K_h(\mathbf{X}_j - \mathbf{x}_0)}{n(n-1)} - \ell^2(\mathbf{x}_0)\right)^2\right],\end{aligned}$$

where the last term in the right-hand side is of order $(nh^d)^{-1} + h^{2\alpha}$ and tends to 0 by our assumption on h. Since for p sufficiently large we have $C\xi_p\ell^2(\mathbf{x}_0) < \varepsilon/2$ and thus $P(C\xi_p\ell^2(\mathbf{x}_0) > \varepsilon/2) = 0$, this concludes the proof of the consistency in probability of $\hat{\theta}_n$ when $nh^d \to \infty$ and $h \to 0$ as $n \to \infty$. □

Proof of Theorem 3. By a Taylor expansion of $\dot{S}_n$ around θ_0, we have

$$0 = \dot{S}_n(\hat{\theta}_n) = \dot{S}_n(\theta_0) + \ddot{S}_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0),$$

where $\bar{\theta}_n$ lies in the line segment with extremities $\hat{\theta}_n$ and θ_0.

Let us study the behaviour of

$$\dot{S}_n(\theta_0) = \frac{-1}{2n(n-1)} \sum_{j \neq k} \int \dot{Z}_k(\theta_0, u, h) Z_j(\theta_0, u, h) w(u)\, du,$$

where $\dot{Z}_k$ denotes the gradient of Z_k with respect to the first argument. Recall that $\theta_0 = \theta(\mathbf{x}_0) = (\pi(\mathbf{x}_0), a(\mathbf{x}_0), b(\mathbf{x}_0))$ and therefore

$$J(t,u,h) = E\big[Z_1(t,u,h)\big] = 2i \int \Im\big(M\big(\theta(\mathbf{x}), u\big) M(t,-u)\big) f^*_{\mathbf{x}}(u) \ell(\mathbf{x}) K_h(\mathbf{x} - \mathbf{x}_0)\, d\mathbf{x},$$

satisfies $J(\theta_0, u, h) \to 0$ as $h \to 0$. Indeed, the last integral may be equal to 0 if the set $\{\mathbf{x} : \theta(\mathbf{x}) = \theta(\mathbf{x}_0)\}$ has Lebesgue measure 0, or tends (by uniform continuity in $\mathbf{x}$ of the integrand) to

$$2i \Im\big(M\big(\theta(\mathbf{x}_0), u\big) M\big(\theta(\mathbf{x}_0), -u\big)\big) f^*_{\mathbf{x}_0}(u) \ell(\mathbf{x}_0) = 0.$$

Moreover,

$$\dot{Z}_k(t,u,h) = \Im\big(\dot{M}(t,-u) e^{iuY_k}\big) K_h(\mathbf{X}_k - \mathbf{x}_0).$$

Denote $\dot{J}(t,u,h) = E[\dot{Z}_k(t,u,h)]$ and observe that

$$\dot{J}(t,u,h) = \int \Im\big(\dot{M}(t,-u) M\big(\theta(\mathbf{x}), u\big) f^*_{\mathbf{x}}(u)\big) K_h(\mathbf{x} - \mathbf{x}_0) \ell(\mathbf{x})\, d\mathbf{x}.$$

Then, we decompose $\dot{S}_n(\theta_0)$ as follows

$$\begin{aligned}
&\dot{S}_n(\theta_0) \\
&\quad = \frac{-1}{2n(n-1)} \sum_{j \neq k} \int \big(\dot{Z}_k(\theta_0,u,h) - \dot{J}(\theta_0,u,h)\big)\big(Z_j(\theta_0,u,h) - E\big[Z_j(\theta_0,u,h)\big]\big) w(u)\, du \\
&\qquad - \frac{1}{2n} \sum_{j=1}^{n} \int \dot{J}(\theta_0,u,h)\big(Z_j(\theta_0,u,h) - E\big[Z_j(\theta_0,u,h)\big]\big) w(u)\, du \\
&\quad := -\frac{1}{2}\big(A_n(h) + B_n(h)\big),
\end{aligned} \tag{6.4}$$

where terms in $A_n(h)$ and $B_n(h)$ are uncorrelated. On the one hand, we use a multivariate central limit theorem for independent random variables taking values in a Hilbert space, following [25] or [12], Theorem 4, page 396. This will give us the limit behavior of the term

$$B_n(h) = \frac{1}{n} \sum_{j=1}^{n} U_j(h), \qquad U_j(h) := \int \dot{J}(\theta_0,u,h)\big(Z_j(\theta_0,u,h) - E\big[Z_j(\theta_0,u,h)\big]\big) w(u)\, du.$$

The random variables $U_j(h)$, $j = 1, \dots, n$ are independent, centered, but their common law depend on n via h. Our goal is to show that

$$nh^d \operatorname{Var}\big(B_n(h)\big) = \sum_{j=1}^{n} \operatorname{Var}\left(\sqrt{\frac{h^d}{n}} U_j(h)\right) \to \Sigma, \qquad \text{as } n \to \infty \tag{6.5}$$

and that

$$\sum_{j=1}^{n} E\left[\left\|\sqrt{\frac{h^d}{n}} U_j(h)\right\|^4\right] = \frac{h^{2d}}{n} E\big[\|U_1(h)\|^4\big] \to 0, \qquad \text{as } n \to \infty. \tag{6.6}$$

Indeed, (6.6) implies the Lindeberg's condition in [25]:

$$\sum_{j=1}^{n} E\left[\left\|\sqrt{\frac{h^d}{n}} U_j(h)\right\|^2 \cdot \mathbb{I}_{\|\sqrt{h^d/n} U_j(h)\| \geq \varepsilon}\right] \to 0, \qquad \text{as } n \to \infty, \text{ for any } \varepsilon > 0.$$

On the other hand, we prove that

$$\sqrt{nh^d} A_n(h) \to 0, \qquad \text{in probability, as } n \to \infty, \tag{6.7}$$

stating that $\sqrt{nh^d} A_n(h)$ negligible term and that, as a consequence, the limiting behavior of $\sqrt{nh^d} \dot{S}_n(\theta_0)$ is only driven by $\sqrt{nh^d} B_n(h)$. This will end the proof of the theorem.

Let us prove (6.5) and (6.6). Note that $nh^d \operatorname{Var}(B_n(h)) = h^d \operatorname{Var}(U_1(h))$ and that

$$\begin{aligned} &\operatorname{Var}\big(U_1(h)\big) \\ &\quad = \iint \dot{\jmath}(\theta_0, u_1, h) \dot{\jmath}^{\top}(\theta_0, u_2, h) \operatorname{Cov}\big(Z_1(\theta_0, u_1, h), Z_1(\theta_0, u_2, h)\big) w(u_1) w(u_2)\, du_1\, du_2. \end{aligned}$$

Similarly to Proposition 2, by uniform continuity in $\mathbf{x}$ of the integrand in $\dot{\jmath}$, we get

$$\lim_{h \to 0} \dot{\jmath}(\theta_0, u, h) := \dot{\jmath}(\theta_0, u).$$

See that $\|\dot{\jmath}(\theta_0, u)\| \leq 2(1 + |u|)\|\ell\|_{\infty}$ and that the latter upper bound is integrable with respect to the measure $w(u)\, du$ by assumption on w. It remains to study:

$$\begin{aligned} &\operatorname{Cov}\big(Z_1(\theta_0, u_1, h), Z_1(\theta_0, u_2, h)\big) \\ &\quad = E\big[Z_1(\theta_0, u_1, h) Z_1(\theta_0, u_2, h)\big] - E\big[Z_1(\theta_0, u_1, h)\big] E\big[Z_1(\theta_0, u_2, h)\big]. \end{aligned}$$

From (6.1), we deduce that

$$h^d \big|E\big[Z_1(\theta_0, u_1, h)\big] E\big[Z_1(\theta_0, u_2, h)\big]\big| \leq h^d \left(2C \int |K|\right)^2 \to 0,$$

when $h \to 0$ as $n \to \infty$. We also have

$$\begin{aligned}
&h^d E\big[Z_1(\theta_0,u_1,h)Z_1(\theta_0,u_2,h)\big] \\
&\quad = 4\cdot \iint \Im\big(e^{iu_1 y}M(\theta_0,-u_1)\big)\Im\big(e^{iu_2 y}M(\theta_0,-u_2)\big)\frac{1}{h^d}K^2\left(\frac{\mathbf{x}-\mathbf{x}_0}{h}\right)g(y,\mathbf{x})\,dy\,d\mathbf{x} \\
&\quad = 4\cdot \int \Im\big(e^{iu_1 y}M(\theta_0,-u_1)\big)\cdot\Im\big(e^{iu_2 y}M(\theta_0,-u_2)\big)g(y,\mathbf{x}_0)\,dy\left(\int K^2\right)\big(1+o(1)\big) \\
&\quad = 4\cdot \int \Im\big(e^{iu_1 y}M(\theta_0,-u_1)\big)\cdot\Im\big(e^{iu_2 y}M(\theta_0,-u_2)\big)g_{\mathbf{x}_0}(y)\,dy\cdot \ell(\mathbf{x}_0)\left(\int K^2\right)\big(1+o(1)\big),
\end{aligned}$$

as $h \to 0$. See also that we can write

$$\begin{aligned}
V(\theta_0,u_1,u_2) := &\int \big(e^{iu_1 y}M(\theta_0,-u_1)-e^{-iu_1 y}M(\theta_0,u_1)\big) \\
&\times \big(e^{iu_2 y}M(\theta_0,-u_2)-e^{-iu_2 y}M(\theta_0,u_2)\big)g_{\mathbf{x}_0}(y)\,dy \\
= {}& M(\theta_0,u_1+u_2)M(\theta_0,-u_1)M(\theta_0,-u_2)f^*_{\mathbf{x}_0}(u_1+u_2) \\
&- M(\theta_0,u_1-u_2)M(\theta_0,-u_1)M(\theta_0,u_2)f^*_{\mathbf{x}_0}(u_1-u_2) \\
&- M(\theta_0,-u_1+u_2)M(\theta_0,u_1)M(\theta_0,-u_2)f^*_{\mathbf{x}_0}(-u_1+u_2) \\
&+ M(\theta_0,-u_1-u_2)M(\theta_0,u_1)M(\theta_0,u_2)f^*_{\mathbf{x}_0}(-u_1-u_2)
\end{aligned}$$

and this is a bounded function with respect to u_1 and u_2. Therefore

$$h^d \operatorname{Var}\big(U_1(h)\big) \to \iint \dot{\jmath}(\theta_0,u_1)\dot{\jmath}^{\top}(\theta_0,u_2)V(\theta_0,u_1,u_2)w(u_1)w(u_2)\,du_1\,du_2 =: \Sigma,$$

as $h \to 0$. This proves (6.5).

Now, denote by $v^{(k)}$ the kth coordinate of a vector v and use Jensen inequality to see that

$$\begin{aligned}
E\big[\|U_1(h)\|^4\big] &\le 3\big(E\big[\big(U_1^{(1)}(h)\big)^4\big]+E\big[\big(U_1^{(2)}(h)\big)^4\big]+E\big[\big(U_1^{(3)}(h)\big)^4\big]\big) \\
&\le 3\sum_{k=1}^{3} E\left[\left(\int \dot{\jmath}^{(k)}(\theta_0,u,h)\big(Z_1(\theta_0,u,h)-E\big[Z_1(\theta_0,u,h)\big]\big)w(u)\,du\right)^4\right] \\
&\le 3\sum_{k=1}^{3}\int \big|\dot{\jmath}^{(k)}(\theta_0,u,h)\big|^4 E\big[\big|Z_1(\theta_0,u,h)\big|^4\big]w(u)\,du.
\end{aligned}$$

We have $|\dot{\jmath}^{(k)}(\theta_0,u,h)| \le 4(1+|u|)(\int |K|)\|\ell\|_\infty$ by Lemma 1 and

$$\begin{aligned}
E\big[\big|Z_1(\theta_0,u,h)\big|^4\big] &= \iint 4\big|\Im\big(e^{iuy}M(\theta_0,-u)\big)\big|^4\frac{1}{h^{4d}}K^4\left(\frac{\mathbf{x}-\mathbf{x}_0}{h}\right)g(y,\mathbf{x})\,dy\,d\mathbf{x} \\
&\le \frac{4}{h^{3d}}\int \frac{1}{h^d}K^4\left(\frac{\mathbf{x}-\mathbf{x}_0}{h}\right)\ell(\mathbf{x})\,d\mathbf{x} \le \frac{O(1)}{h^{3d}}\left(\int K^4\right)\|\ell\|_\infty,
\end{aligned}$$

as $h \to 0$. Therefore,

$$\frac{h^{2d}}{n} E\big[\big\|U_1(h)\big\|^4\big] \leq \frac{O(1)}{nh^d} \int |K| \cdot \int K^4 \cdot \int \big(1+|u|\big)^4 w(u)\,du = o(1),$$

as $n \to \infty$ and $h \to 0$ such that $nh^d \to \infty$. This proves (6.6).

To prove (6.7), we notice that $A_n(h)$ defined in (6.4) can be treated similarly to T_1 in (6.2). By this remark, we easily prove that $\mathrm{Var}(A_n) = o((nh^d)^{-1})$ which insure the wanted result.

Let us prove that

$$\ddot{S}_n(\theta_n) \longrightarrow \mathcal{I}(\theta_0), \qquad \text{in probability, as } n \to \infty,$$

where $\mathcal{I} = \mathcal{I}(\theta_0) = -\frac{1}{2}\int \dot{J}(\theta_0,u)\dot{J}^\top(\theta_0,u)w(u)\,du$, and $\dot{J}(\theta_0,u)$ is defined in (3.1). We start by writing the triangular inequality

$$\big\|\ddot{S}_n(\theta_n) - \mathcal{I}\big\| \leq \big\|\ddot{S}_n(\theta_n) - \ddot{S}_n(\theta_0)\big\| + \big\|\ddot{S}_n(\theta_0) - E\big(\ddot{S}_n(\theta_0)\big)\big\| + \big\|E\big(\ddot{S}_n(\theta_0)\big) - \mathcal{I}\big\|.$$

Then using upper bounds similar to (6.3) slighly adapted to $\ddot{S}_n$ instead of S_n and the convergence in probability of $\hat{\theta}_n$ towards θ_0 established in Theorem 2, we have that $\|\ddot{S}_n(\theta_n) - \ddot{S}_n(\theta_0)\| \to 0$ in probability as $n \to \infty$. By writing

$$E\big(\ddot{S}_n(\theta_0)\big) = -\frac{1}{2}\int \big(\ddot{J}(\theta_0,u,h)J(\theta_0,u,h) + \dot{J}(\theta_0,u,h)\dot{J}(\theta_0,u,h)^\top\big)w(u)\,du$$

and noticing, according to Bochner's lemma, that $J(\theta_0,u,h) \to 0$ and $\dot{J}(\theta_0,u,h) \to \dot{J}(\theta_0,u)$ as $h \to 0$, we have, according to the Lebesgue's theorem, that $E[\ddot{S}_n(\theta_0)]$ tends to $\mathcal{I}$ as $h \to 0$. Finally, we decompose $-2n(n-1)(\ddot{S}_n(\theta_0) - E[\ddot{S}_n(\theta_0)]) = \sum_{l=1}^3 (D_{1,l} + D_{2,l})$ where

$$D_{1,1} = \sum_{k\neq j} \int \big(\ddot{Z}_k(\theta_0,u,h) - \ddot{J}(\theta_0,u,h)\big)\big(Z_j(\theta,u,h) - J(\theta_0,u,h)\big)w(u)\,du,$$

$$D_{1,2} = (n-1)\sum_{k} \int \big(\ddot{Z}_k(\theta_0,u,h) - \ddot{J}(\theta_0,u,h)\big)J(\theta_0,u,h)w(u)\,du,$$

$$D_{1,3} = (n-1)\sum_{j} \int \ddot{J}(\theta_0,u,h)\big(Z_j(\theta,u,h) - J(\theta_0,u,h)\big)w(u)\,du,$$

and

$$D_{2,1} = \sum_{k\neq j} \int \big(\dot{Z}_k(\theta_0,u,h) - \dot{J}(\theta_0,u,h)\big)\big(\dot{Z}_j(\theta,u,h) - \dot{J}(\theta_0,u,h)\big)^\top w(u)\,du,$$

$$D_{2,2} = (n-1)\sum_{k} \int \big(\dot{Z}_k(\theta_0,u,h) - \dot{J}(\theta_0,u,h)\big)J(\theta_0,u,h)^\top w(u)\,du,$$

$$D_{2,3} = (n-1)\sum_{j} \int \dot{J}(\theta_0,u,h)\big(Z_j(\theta,u,h) - J(\theta_0,u,h)\big)^\top w(u)\,du.$$

Noticing that terms $D_{i,3}$, $i = 1, 2$, respectively $D_{i,j,}$, $i = 1, 2$ and $j = 2, 3$, can be treated as T_1 respectively T_2 in the proof of Proposition 2, we obtain

$$\operatorname{Var}\big(\ddot{S}_n(\theta_0)\big) = O\left(\frac{1}{nh^d}\right),$$

which concludes the proof. □

Acknowledgments

The authors thank warmly Dr.'s Bowen and Chappell for providing the Positron Emission Tomography dataset presented in [5], Figure 4, as well as Dr. Wang for sharing the EM-type algorithm code developed in [18].

The authors want to acknowledge the anonymous referees whose insightful comments helped them to improve the results of this paper.

References

[1] Anderson, J.A. (1979). Multivariate logistic compounds. *Biometrika* **66** 17–26. MR0529143

[2] Balabdaoui, F. and Butucea, C. (2014). On location mixtures with Pólya frequency components. *Statist. Probab. Lett.* **95** 144–149. MR3262962

[3] Bordes, L., Kojadinovic, I. and Vandekerkhove, P. (2013). Semiparametric estimation of a two-component mixture of linear regressions in which one component is known. *Electron. J. Stat.* **7** 2603–2644. MR3121625

[4] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. MR2278356

[5] Bowen, R.S., Chappell, R.J., Bentzen, S.M., Deveau, M.A., Forrest, L.J. and Jeraj, R. (2012). Spatially resolved regression analysis of pre-treatment FDG, FLT and cu-ATSM PET from post-treatment FDG PET: An exploratory study. *Radiother. Oncol.* **105** 41–48.

[6] Brunel, E., Comte, F. and Lacour, C. (2010). Minimax estimation of the conditional cumulative distribution function. *Sankhya A* **72** 293–330. MR2746114

[7] Butucea, C. and Vandekerkhove, P. (2014). Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.* **41** 227–239. MR3181141

[8] Celeux, G., Hurn, M. and Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450

[9] Cohen, S. and Le Pennec, E. (2012). Conditional density estimation by penalized likelihood model selection and applications. Preprint. Available at arXiv:1103.2021.

[10] Dacunha-Castelle, D. and Duflo, M. (1983). *Probabilités et Statistiques. Tome* 2. Paris: Masson. MR0732786

[11] De Veaux, R.D. (1989). Mixtures of linear regressions. *Comput. Statist. Data Anal.* **8** 227–245. MR1028403

[12] Gikhman, I.I. and Skorokhod, A.V. (2004). *The Theory of Stochastic Processes. I.* Berlin: Springer. MR2058259

[13] Gruen, B., Leisch, F. and Sarkar, D. (2013). flexmix: Flexible Mixture Modeling. URL http://CRAN.R-project.org/package=flexmix. R package version 2.3-11.

[14] Grün, B. and Leisch, F. (2006). Fitting finite mixtures of linear regression models with varying and fixed effects in R. In *Proceedings in Computational Statistics* (A. Rizzi and M. Vichi, eds.) 853–860. Amsterdam: Elsevier.
[15] Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224. MR1962504
[16] Hawkins, D.S., Allen, D.M. and Stromberg, A.J. (2001). Determining the number of components in mixtures of linear models. *Comput. Statist. Data Anal.* **38** 15–48. MR1869478
[17] Herrmann, E. (2013). lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth, 2013. URL http://CRAN.R-project.org/package=lokern. R package version 1.1-4.
[18] Huang, M., Li, R. and Wang, S. (2013). Nonparametric mixture of regression models. *J. Amer. Statist. Assoc.* **108** 929–941. MR3174674
[19] Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: A semiparametric approach. *J. Amer. Statist. Assoc.* **107** 711–724. MR2980079
[20] Hunter, D.R., Wang, S. and Hettmansperger, T.P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275
[21] Hunter, D.R. and Young, D.S. (2012). Semiparametric mixtures of regressions. *J. Nonparametr. Stat.* **24** 19–38. MR2885823
[22] Hurn, M., Justel, A. and Robert, C.P. (2003). Estimating mixtures of regressions. *J. Comput. Graph. Statist.* **12** 55–79. MR1977206
[23] Ibragimov, I.A. and Has'minskiĭ, R.Z. (1981). *Statistical Estimation: Asymptotic Theory. Applications of Mathematics* **16**. New York–Berlin: Springer. MR0620321
[24] Jones, P.N. and McLachlan, G.J. (1992). Fitting finite mixture models in a regression context. *Australian J. Statist.* **34** 233–240.
[25] Kandelaki, N.P. and Sozanov, V.V. (1964). On a central limit theorem for random elements with values in Hilbert space. *Theory Probab. Appl.* **71** 38–46.
[26] Leung, D.H.-Y. and Qin, J. (2006). Semi-parametric inference in a bivariate (multivariate) mixture model. *Statist. Sinica* **16** 153–163. MR2256084
[27] Montuelle, L. and Le Pennec, E. (2014). Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electron. J. Stat.* **8** 1661–1695. MR3263134
[28] Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.* **73** 730–752. MR0521324
[29] Städler, N., Bühlmann, P. and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722
[30] Stephens, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293
[31] Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204
[32] Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34** 1265–1269. MR0155376
[33] Toshiya, H. (2013). Mixture regression for observational data, with application to functional regression models. Preprint. Available at arXiv:1307.0170.
[34] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359
[35] Turner, T.R. (2000). Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *J. Roy. Statist. Soc. Ser. C* **49** 371–384. MR1824547
[36] Vandekerkhove, P. (2013). Estimation of a semiparametric mixture of regressions model. *J. Nonparametr. Stat.* **25** 181–208. MR3039977
[37] Yao, W. and Lindsay, B.G. (2009). Bayesian mixture labeling by highest posterior density. *J. Amer. Statist. Assoc.* **104** 758–767. MR2751453

[38] Zhu, H.-T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 3–16. MR2035755

Received December 2014 and revised August 2015