Journal of Multivariate Analysis 114 (2013) 288–302

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Kernel smoothers and bootstrapping for semiparametric mixed effects models[☆]

Wenceslao González Manteiga[a], María José Lombardía[b], María Dolores Martínez Miranda[c,*], Stefan Sperlich[d]

[a] Departamento de Estadística e I.O., Universidad de Santiago de Compostela, E - 15782 Santiago de Compostela, Spain
[b] Departamento de Matemáticas, Universidade da Coruña, E - 15071 A Coruña, Spain
[c] Departamento de Estadística e I.O., Universidad de Granada, E - 18071 Granada, Spain
[d] Département des sciences économiques & Institut de Recherche en Statistique, Université de Genève, CH - 1211 Genève 4, Switzerland

ARTICLE INFO

Article history:
Received 14 December 2011
Available online 9 August 2012

AMS 2010 subject classifications:
62G08
62H15
62P12

Keywords:
Mixed effects models
Non- and semiparametric models
Bootstrap inference
Bandwidth choice
Small area statistics

ABSTRACT

While today linear mixed effects models are frequently used tools in different fields of statistics, in particular for studying data with clusters, longitudinal or multi-level structure, the nonparametric formulation of mixed effects models is still quite recent. In this paper we discuss and compare different nonparametric estimation methods. In this context we introduce a computationally inexpensive bootstrap method, which is used to estimate local mean squared errors, to construct confidence intervals and to find locally optimal smoothing parameters. The theoretical considerations are accompanied by the provision of algorithms and simulation studies of the finite sample behavior of the methods. We show that our confidence intervals have good coverage probabilities, and that our bandwidth selection method succeeds to minimize the mean squared error for the nonparametric function locally.

1. Introduction

Over the past twenty years linear mixed effects models and their extensions have become quite popular statistical models for analyzing data with an a priori specified correlation structure. The accounting for the so-called "within-subject correlation" allows to deal with longitudinal and clustered data which naturally arise for example in biomedical studies. In fact, statistical inference with linear mixed effects models has been widely studied in the context of longitudinal data and biometrical data with repeated measurement; see for example [27] or [5]. At the same time, mixed effects models are also particularly suitable for small-areas estimation [15] and data matching or poverty mapping [6]. In the last five to ten years there has been a notable interest in extending parametric models to more flexible nonparametric formulations; see for example [30]. Along with the different non- and semiparametric formulations (see [33,17,18,29,2,24]), there is growing demand for feasible methods to do inference based on them. Likelihood based approaches were proposed for example by Lin and Zhang [20], and Lombardía and Sperlich [21,26].

☆ The authors gratefully acknowledge the financial support of the Spanish "Ministerio de Ciencia e Innovación" MTM2008-03010, ECO2010-15455 and AAII DE2009-0030; "Grupos de referencia competitiva" (2007/132) of the "Consellería de Educación e Ordenación Universitaria", the Belgian network IAP-Network P6/03, and the DAAD acciones integradas 50119348. Also the authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

* Corresponding author.
E-mail address: mmiranda@ugr.es (M.D. Martínez Miranda).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.08.005

The main problem when using smoothing methods is to choose an appropriate smoothing (or say, regularization) parameter. Gu and Ma [10,11] proposed generalized cross validation in the context of spline smoothing. Later Xu and Zhu [32] proposed cross validation for kernel based nonparametric mixed effects models. In this article we introduce the use of a computationally inexpensive bootstrap method for non- and semiparametric mixed effects models to facilitate both inference (here illustrated by the construction of confidence intervals) and bandwidth selection.

We start from the one-way model with a fully nonparametric formulation, i.e.

$$y_{ij} = m(\mathbf{x}_{ij}) + v_i(\mathbf{z}_{ij}) + \varepsilon_{ij}, \quad j = 1, \dots, n_i, i = 1, \dots, q, \qquad \sum_{i=1}^{q} n_i = n, \tag{1.1}$$

with y_{ij} being the observed responses, and $\mathbf{x}_{ij}$ $(k \times 1)$, $\mathbf{z}_{ij}$ $(r \times 1)$ observable covariates, where often $\mathbf{z}_{ij}$ consists of a constant and some elements of $\mathbf{x}_{ij}$. Here, $m(\cdot)$ represents the fixed effect or population function, and $v_i(\cdot)$ are the random-effects functions. The errors, ε_{ij}, are supposed to be independent with mean 0 and conditional variances $\sigma^2(\mathbf{x}_{ij})$. The $v_i(\mathbf{z}_{ij})$ can be considered as realizations of a zero-mean smooth process with a covariance function $\gamma(\mathbf{z}_{ij_1}, \mathbf{z}_{ij_2}) = E[v_i(\mathbf{z}_{ij_1})v_i(\mathbf{z}_{ij_2})]$. As it is common in these models, $v_i(\mathbf{z}_{ij})$ and ε_{ij} are assumed to be independent, conditionally to $\mathbf{x}_{ij}$. Model (1.1) comprises several often used models in longitudinal studies or for clustered data, including nested-error and random regression coefficient models, among others. More specifically we have the following.

For *longitudinal data*, may it be balanced or unbalanced panels or just repeated measurements, one considers a study involving q subjects. In this case, y_{ij} is taken from subject i $(i = 1, \dots, q)$, either at time point t_{ij} being an element of the explanatory part $\mathbf{x}_{ij}$ $(j = 1, \dots, n_i)$, or – as is typical panel studies – simply at time $t = j$. Observations from different subjects are assumed to be independent, while observations from the same subject are naturally correlated. This intra-subject dependence is often modeled by $v_i(\mathbf{z}_{ij}) = v_i(t_{ij})$, and these random effects can be interpreted as unobservable subject effects, sometimes also called "real effects".

For *clustered data*, a multi-level structure or small area model, one considers observations from q clusters (areas, groups, ...), where y_{ij} is taken from ith cluster with covariate $\mathbf{x}_{ij}$ $(j = 1, \dots, n_i, i = 1, \dots, q)$. Now, observations from different clusters are assumed to be independent, whereas those from the same cluster are supposed to be correlated to various degrees. Then, this intra-cluster correlation is modeled by a random term $v_i(\mathbf{z}_{ij})$, which is often called "latent effects".

Another special and quite popular case is the so-called *nested-error regression model*, where $m(\cdot)$ is linear (or a polynomial), and v_i is simply an indicator function for ith $(i = 1, \dots, q)$ cluster. Such model is considered for example in the context of small area statistics where the population is divided into q small areas, with n_i being the number of sampled units (individuals) in the ith area. Also in econometric panel studies this is, maybe, the most often applied model. In this case q units have been observed over n_i periods.

As we will concentrate on nonparametric forms of $m(\cdot)$ and $v_i(\cdot)$ based on local polynomials (see Section 2), it is useful to mention also the more general *random regression coefficient model* including a random slope, i.e.

$$y_{ij} = \beta x_{ij} + b_i x_{ij} + \varepsilon_{ij},$$

where β and b_i are the fixed-effects and the random-effects coefficients, respectively. Compared to them, the flexibility of model (1.1) offers also new perspectives for various problems faced in applied statistics. Apart from the rather flexible modeling to avoid specification problems of the functional form in the mean function [23] or in the variances [9], it can further be used for data mining and specification testing; see [26] and references therein. The contribution of our article to the literature can be summarized as follows.

First, we will review the existing approaches to estimate model (1.1) by local polynomial estimation. There, the main question is how to choose the weights, i.e. how to combine correlation structure and kernel weights in the estimating equations. We investigate mainly three different strategies for the estimation and explore both the statistical properties and some practical issues.

Second, we introduce a computationally inexpensive bootstrap procedure in order to do inference. So far, in this kind of models bootstrap based inference has focused mainly on testing distributional assumptions on the random effects, see [3] for a review, or on likelihood based tests for functional misspecification, see [26]. Here we consider the application of the bootstrap methods to construct confidence intervals.

Third, we propose to solve the bandwidth choice problem for the estimation of the fixed-effects function by bootstrap estimation of the mean squared error. In non- and semiparametric mixed effects models, the choice of smoothing parameters becomes a more critical but also challenging task than in common non- and semiparametric estimation problems. Due to the more complex data structure, neither intuition nor eye balling will help here. The standard nonparametric methods ignoring the correlation structure are not suitable because they cannot pick up the extra variability.

This is, probably, the first article considering local optimal smoothing for mixed effects models. The resampling scheme follows the spirit of [8,22,21]. Evidently, the estimates of the mean squared errors can equally well be used to construct confidence and prediction intervals. We will mainly deal with the estimation and inference of the population function $m(\cdot)$, but also discuss the mixed effects or individual functions, say $\eta_i(\mathbf{x}, \mathbf{z}) = m(\mathbf{x}) + v_i(\mathbf{z})$, to analyze particular population parameters such as $\Theta_i = \dot{m}(\mathbf{x}_i) + \dot{v}_i(\mathbf{z}_i)$ with $\dot{m}(\mathbf{x}_i) = \sum_{j=1}^{n_i} m(\mathbf{x}_{ij})/n_i$ and $\dot{v}_i(\mathbf{z}_i) = \sum_{j=1}^{n_i} v_i(\mathbf{z}_{ij})/n_i$, which arise mainly in small area statistics. All together these methods provide us with powerful tools for data analysis in mixed effects models for estimation, inference, and local bandwidth selection.

The rest of the paper is organized as follows. In Section 2 we introduce the nonparametric mixed-effects model with the particular case of a semiparametric model where the random-effects are just linear. In that context, we study the different proposals for marginal nonparametric estimation and compare them with respect to efficiency aspects, implementation and finite sample performance. In Section 3 we introduce a fast bootstrap-based method for inference. This will be used to construct confidence intervals, and to get local optimal bandwidths. All this is accompanied by studies of its finite sample behavior. Technical proofs, the estimation of the variance–covariance matrices, and the joint estimation of fixed effects together with the prediction of random effects are deferred to the Appendix.

2. Marginal estimation in nonparametric mixed-effects models

2.1. A local linear mixed-effects model

Assuming that the functions m and v_i have $(p+1)$th continuous derivatives (for simplicity set $p=1$), for any fixed $\mathbf{x}$ in a generic domain $\mathcal{X}$, m can be approximated by a linear function within a neighborhood of $\mathbf{x}$ by a Taylor expansion, i.e.

$$m(\mathbf{x}_{ij}) \approx m(\mathbf{x}) + (\mathbf{x}_{ij} - \mathbf{x})^t \bigtriangledown m(\mathbf{x}), \tag{2.1}$$

where $\bigtriangledown m$ denotes the gradient of the function m.

Similarly, for the random function v_i one has

$$v_i(\mathbf{z}_{ij}) \approx v_i(\mathbf{z}) + (\mathbf{z}_{ij} - \mathbf{z})^t \bigtriangledown v_i(\mathbf{z}),$$

for any fixed $\mathbf{z}$ in a generic domain $\mathcal{Z}$.

Setting $\mathbf{X}_{ij} = (1, (\mathbf{x}_{ij} - \mathbf{x})^t)$, $\boldsymbol{\beta} = (m(\mathbf{x}), \bigtriangledown m(\mathbf{x})^t)^t$, and $\mathbf{Z}_{ij} = (1, (\mathbf{z}_{ij} - \mathbf{z})^t)$, $\mathbf{b}_i = (v_i(\mathbf{z}), \bigtriangledown v_i(\mathbf{z})^t)^t$, model (1.1) can be approximated by the local linear mixed effects model (LME):

$$y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i + \varepsilon_{ij} \quad j = 1, \dots, n_i, i = 1, \dots, q. \tag{2.2}$$

A most crucial assumption is that the random effects, $\mathbf{b}_i$, are independent from $\mathbf{X}_{ij}$ $(j = 1, \dots, n_i; i = 1, \dots, q)$ with mean zero and $E[\mathbf{b}_i\mathbf{b}_i^t] = \mathbf{B}_i$. The errors, ε_{ij}, are both mutually independent and independent from $\mathbf{b}_i$, have conditional mean zero, $E[\varepsilon_{ij}|\mathbf{x}_{ij}] = 0$, and finite conditional variances $\sigma^2(\mathbf{x}_{ij})$. By stacking the observations in the q groups, the model can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \dots, q, \tag{2.3}$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^t$, $\mathbf{X}_i$ is a $(n_i \times 2)$ matrix with rows $\mathbf{X}_{ij} = (1, (\mathbf{x}_{ij} - \mathbf{x})^t)$, $\mathbf{Z}_i$ a matrix with rows $\mathbf{Z}_{ij} = (1, (\mathbf{z}_{ij} - \mathbf{z})^t)$ $(j = 1, \dots, n_i, i = 1, \dots, q)$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})^t$ an error vector with conditional covariance matrix $\boldsymbol{\Sigma}_i = \mathrm{diag}(\sigma^2(\mathbf{x}_{i1}), \dots, \sigma^2(\mathbf{x}_{in_i}))$. The model (2.3) can be written more compactly, i.e.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{2.4}$$

with $\mathbf{y} = (\mathbf{y}_1^t, \dots, \mathbf{y}_q^t)^t$, matrix $\mathbf{X}$ $(n \times (k+1))$ with block rows $\mathbf{X}_i$, and matrix $\mathbf{Z}$ $(n \times q(r+1))$ with diagonal blocks $\mathbf{Z}_i$. The coefficient of the random effects is given by $\mathbf{b} = (\mathbf{b}_1^t, \dots, \mathbf{b}_q^t)^t$, and the global vector of errors by $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_q)^t$.

2.2. Three ways to estimate the fixed-effects function

For the estimation of such a local linear mixed-effects model let us start by considering the fixed-effects $\boldsymbol{\beta}$ and the marginal model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2.5}$$

with $\mathbf{u} = \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$, involving a within-subjects correlation expressed by $E[\mathbf{u}\mathbf{u}^t|\mathbf{X}, \mathbf{Z}] = \mathbf{V} = \mathbf{Z}\mathbf{B}\mathbf{Z}^t + \boldsymbol{\Sigma}$. Here $\mathbf{B}$ is a symmetric matrix of dimension $n(r+1) \times n(r+1)$ defined by diagonal blocks $\mathbf{B}_i$.

Lin and Carroll [17] propose a formal extension of the parametric generalized estimating equations approach (GEE) introducing kernel weights for clustered data. They consider the standard GEE based on quasi-likelihood for inference of the fixed-effect parameter $\boldsymbol{\beta}$, given by

$$0 = \sum_{i=1}^{q} \mathbf{X}_i^t (\mathbf{V}_i^c)^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{2.6}$$

with $\mathbf{V}_i^c$ not necessarily being the true intra-subject correlation matrix $\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^t + \boldsymbol{\Sigma}_i$ but rather a working matrix. Indeed, it was defined by $\mathbf{V}_i^c = \mathbf{S}_i^{1/2}\mathbf{R}_i(\delta)\mathbf{S}_i^{1/2}$ with $\mathbf{S}_i^{1/2}$ being a diagonal matrix depending on the diagonal elements of $\mathbf{V}_i$, and $\mathbf{R}_i(\delta)$ an invertible "working-correlation" matrix, possibly depending on an additional parameter δ. The solution can be obtained by iteratively reweighted least squares, and they showed that the resulting estimator is asymptotically Gaussian under mild regularity conditions.

If the parametric (linear) model is only assumed locally, it is necessary to introduce kernel-weights to take into account only observations in a small neighborhood. Lin and Carroll [17] suggest two different ways to introduce the weights, namely

$$0 = \sum_{i=1}^{q} \mathbf{X}_i^t (\mathbf{V}_i^c)^{-1} \mathbf{W}_{ih} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}) \tag{2.7}$$

$$\text{and} \quad 0 = \sum_{i=1}^{q} \mathbf{X}_i^t \mathbf{W}_{ih}^{1/2} (\mathbf{V}_i^c)^{-1} \mathbf{W}_{ih}^{1/2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \tag{2.8}$$

where $\mathbf{W}_{ih} = \operatorname{diag}(K_h(\mathbf{x}_{i1} - \mathbf{x}), \ldots, K_h(\mathbf{x}_{in_i} - \mathbf{x}))$, with $K_h(\cdot) = h^{-1}K(\cdot/h)$, K is a (multivariate) kernel function and $h > 0$ is a bandwidth parameter. Note that the two approximations are different if $\mathbf{V}_i^c$ is not diagonal. The authors recommend to ignore the within-subject correlation completely as the estimators from (2.7) or (2.8) reach asymptotically the minimum variance with $\mathbf{R}_i(\delta) = \mathbf{I}$. Wang [29] explains the reason why asymptotically this is still optimal and therefore different from the parametric GEE in this respect. Asymptotically (i.e. when $h \to 0$) there is effectively only one single observation (i, j) for each $i = 1, \ldots, q$ that contributes to estimate $m(\mathbf{x})$. This unique observation is weighted by $K_h(\mathbf{x}_{ij} - \mathbf{x})v_i^{jj}$, with v_i^{jj} being the (j, j)-element of $\mathbf{V}^{-1}$. Consequently, for asymptotic efficiency improvements, alternative weighting methods are necessary which make properly use of the correlation within-subjects. In the following we will present three different ways of making use of the present correlation.

The idea of generalized (or weighted) least squares regression (say GLS) is to transform (2.5) such that one gets uncorrelated observations. Vilar and Francisco [28] adopted this methodology to a local polynomial estimator for an AR(1) time series structure. Following Vilar and Francisco's strategy but adapted to the current underlying model, given that $\mathbf{V}$ is a symmetric and positive defined matrix, its inverse has a squared root $\mathbf{V}^{-1/2}$ which satisfies $\mathbf{V}^{-1} = \mathbf{V}^{-1/2}(\mathbf{V}^{-1/2})^t = (\mathbf{V}^{-1/2})^t\mathbf{V}^{-1/2}$. Then one considers

$$\widetilde{\mathbf{y}} = \mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\mathbf{u} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{u}}, \tag{2.9}$$

where it is satisfied that $E[\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^t|\widetilde{\mathbf{X}}] = \mathbf{V}^{-1/2}\mathbf{V}(\mathbf{V}^{-1/2})^t = \mathbf{I}$, i.e. one has uncorrelated errors. Afterward, the kernel weights are introduced in order to obtain weighted least squares regression of

$$\widetilde{\widetilde{\mathbf{y}}} = \mathbf{W}_h^{1/2}\widetilde{\mathbf{y}} = \mathbf{W}_h^{1/2}\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{W}_h^{1/2}\mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}_h^{1/2}\mathbf{V}^{-1/2}\mathbf{u} = \widetilde{\widetilde{\mathbf{X}}}\boldsymbol{\beta} + \widetilde{\widetilde{\mathbf{u}}}$$

with $\mathbf{W}_h = \operatorname{diag}(\mathbf{W}_{ih})$. The weighted least squares regression leads to a marginal local linear estimator for the fixed effects, namely

$$\widehat{\boldsymbol{\beta}}_{M1} = \left(\sum_{i=1}^{q} \mathbf{X}_i^t \mathbf{V}_i^{-1/2} \mathbf{W}_{ih} \mathbf{V}_i^{-1/2} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{q} \mathbf{X}_i^t \mathbf{V}_i^{-1/2} \mathbf{W}_{ih} \mathbf{V}_i^{-1/2}\, \mathbf{y}_i.$$

This gives our first marginal local linear estimator

$$\widehat{m}_{M1}(\mathbf{x}) = \mathbf{e}_1^t \widehat{\boldsymbol{\beta}}_{M1} = \sum_{i=1}^{q} \mathbf{w}_i^{M1}(\mathbf{x})\mathbf{y}_i \tag{2.10}$$

with weights

$$\mathbf{w}_i^{M1}(\mathbf{x}) = \mathbf{e}_1^t \left(\sum_{l=1}^{q} \mathbf{X}_l^t \mathbf{V}_l^{-1/2} \mathbf{W}_{lh} \mathbf{V}_l^{-1/2} \mathbf{X}_l \right)^{-1} \mathbf{X}_i^t \mathbf{V}_i^{-1/2} \mathbf{W}_{ih} \mathbf{V}_i^{-1/2}. \tag{2.11}$$

Several authors have pointed out that such a transformation does not improve the asymptotic first-order properties of the estimator; but it does improve the estimation in finite samples. Linton et al. [19] propose a two-step estimator which first calculates the "working independence" estimator of Lin and Carroll [17], then uses the results to construct a linear transformation of $\mathbf{y}$ that exhibits a diagonal covariance matrix, and finally runs a local linear regression on the transformed data. Alternatively to the GLS, Wang [29] improves on the GEE approach proposing a kernel-type estimator which makes an efficient use of the correlation structure achieving the best results when the true correlation is known. However, her estimator is difficult to calculate, as it requires an iterative procedure initialized with the estimates from Lin and Carroll [17]. We decided to stick here to computationally less demanding procedures.

In a similar spirit as Wang [29], Chen and Jin [2] suggest to use a weighting function that is based on local variances instead of the global ones. This actually comes closer to the idea of generalized least squares estimators and leads to the following marginal local linear estimator. Consider

$$\mathbf{W}_h^{1/2}\mathbf{y} = \mathbf{W}_h^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}_h^{1/2}\mathbf{u}$$

and transform it to get uncorrelated observations by calculating

$$\boldsymbol{\Omega}_h^{-1/2}\mathbf{W}_h^{1/2}\mathbf{y} = \boldsymbol{\Omega}_h^{-1/2}\mathbf{W}_h^{1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}_h^{-1/2}\mathbf{W}_h^{1/2}\mathbf{u},$$

with $\boldsymbol{\Omega}_h = E[\mathbf{W}_h^{1/2}\mathbf{u}(\mathbf{W}_h^{1/2}\mathbf{u})^t] = \mathbf{W}_h^{1/2}\mathbf{V}\mathbf{W}_h^{1/2}$. Here $\boldsymbol{\Omega}_h^{-1/2}$ is the Moore–Penrose generalized inverse of $\boldsymbol{\Omega}_h^{1/2}$. Finally, introduce kernel weights by

$$\widetilde{\mathbf{y}} = \mathbf{W}_h^{1/2}\boldsymbol{\Omega}_h^{-1/2}\mathbf{W}_h^{1/2}\mathbf{y} = \mathbf{W}_h^{1/2}\boldsymbol{\Omega}_h^{-1/2}\mathbf{W}_h^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}_h^{1/2}\boldsymbol{\Omega}_h^{-1/2}\mathbf{W}_h^{1/2}\mathbf{u} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{u}}.$$

The resulting marginal local linear estimator is

$$\widehat{m}_{M2}(\mathbf{x}) = \mathbf{e}_1^t\widehat{\boldsymbol{\beta}}_{M2} = \sum_{i=1}^q \mathbf{w}_i^{M2}(\mathbf{x})\mathbf{y}_i, \tag{2.12}$$

$$\widehat{\boldsymbol{\beta}}_{M2} = \left(\sum_{i=1}^q \mathbf{X}_i^t\mathbf{W}_{ih}^{1/2}\boldsymbol{\Omega}_{ih}^{-1/2}\mathbf{W}_{ih}\boldsymbol{\Omega}_{ih}^{-1/2}\mathbf{W}_{ih}^{1/2}\mathbf{X}_i\right)^{-1}\sum_{i=1}^q \mathbf{X}_i^t\mathbf{W}_{ih}^{1/2}\boldsymbol{\Omega}_{ih}^{-1/2}\mathbf{W}_{ih}\boldsymbol{\Omega}_{ih}^{-1/2}\mathbf{W}_{ih}^{1/2}\mathbf{y}_i$$

with $\boldsymbol{\Omega}_{ih} = \mathbf{W}_{ih}^{1/2}\mathbf{V}_i\mathbf{W}_{ih}^{1/2}$ and weights $\mathbf{w}_i^{M2}(\mathbf{x}) = \mathbf{e}_1^t\left(\sum_{l=1}^q \mathbf{X}_l^t\mathbf{W}_{lh}^{1/2}\boldsymbol{\Omega}_{lh}^{-1/2}\mathbf{W}_{lh}\boldsymbol{\Omega}_{lh}^{-1/2}\mathbf{W}_{lh}^{1/2}\mathbf{X}_l\right)^{-1}\mathbf{X}_i^t\mathbf{W}_{ih}^{1/2}\boldsymbol{\Omega}_{ih}^{-1/2}\mathbf{W}_{ih}\boldsymbol{\Omega}_{ih}^{-1/2}\mathbf{W}_{ih}^{1/2}$.

Asymptotic properties can be obtained in the same way as done by Vilar and Francisco [28], but the results do not reveal the desirable intuition about the influence of the intra-subject correlation. Though it is already easier to implement and faster calculated than the versions of [29,19], the necessary use of a generalized inverse makes the theoretical study and the calculation of the estimator quite tedious.

The third approach consists in an alternative of Park and Wu [24] which is easier to implement and calculate with, and is also intuitively appealing. This gives a third marginal estimator based – along their presentation – on the likelihood

$$\mathcal{L}_M(\boldsymbol{\beta};\mathbf{y}) = -\frac{1}{2}\sum_{i=1}^q\left\{[\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}]^t\,\mathbf{W}_{ih}^{1/2}\mathbf{V}_i^{-1}\mathbf{W}_{ih}^{1/2}\,[\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}] + C_{iM}\right\}, \tag{2.13}$$

with a constant $C_{iM} = \log|\mathbf{V}_i| + 2n_i\log(2\pi)$. The estimator is given by

$$\widehat{m}_{M3}(\mathbf{x}) = \mathbf{e}_1^t\widehat{\boldsymbol{\beta}}_{M3} = \sum_{i=1}^q \mathbf{w}_i^{M3}(\mathbf{x})\mathbf{y}_i, \tag{2.14}$$

with

$$\widehat{\boldsymbol{\beta}}_{M3} = \left(\sum_{i=1}^q \mathbf{X}_i^t\mathbf{W}_{ih}^{1/2}\mathbf{V}_i^{-1}\mathbf{W}_{ih}^{1/2}\mathbf{X}_i\right)^{-1}\sum_{i=1}^q \mathbf{X}_i^t\mathbf{W}_{ih}^{1/2}\mathbf{V}_i^{-1}\mathbf{W}_{ih}^{1/2}\,\mathbf{y}_i,$$

and the weights

$$\mathbf{w}_i^{M3}(\mathbf{x}) = \mathbf{e}_1^t\left(\sum_{l=1}^q \mathbf{X}_l^t\mathbf{W}_{lh}^{1/2}\mathbf{V}_i^{-1}\mathbf{W}_{lh}^{1/2}\mathbf{X}_l\right)^{-1}\mathbf{X}_i^t\mathbf{W}_{ih}^{1/2}\mathbf{V}_i^{-1}\mathbf{W}_{ih}^{1/2}. \tag{2.15}$$

An interesting observation can be made about the profile-kernel GEEs derived from the optimization of the log-likelihood based score of (2.13). In fact, following a similar reasoning used to motivate the estimators (2.10) and (2.12), we are actually considering an estimator in the transformed model

$$\widetilde{\mathbf{y}} = \mathbf{V}^{-1/2}\mathbf{W}_h^{1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{W}_h^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\mathbf{W}_h^{1/2}\mathbf{u} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{u}}.$$

However, in this case the transformation does not lead to an uncorrelated model

$$E[\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^t|\widetilde{\mathbf{X}}] = \mathbf{V}^{-1/2}\mathbf{W}_h^{1/2}\mathbf{V}\mathbf{W}_h^{1/2}(\mathbf{V}^{-1/2})^t \neq \mathbf{W}_h\mathbf{I}.$$

Park and Wu [24] argue that regardless the asymptotic findings, in the numerical outcome this estimator should perform well, maybe not much worse than that of the much more involved estimator of Chen and Jin [2].

In order to check this statement, but also for illustrative purposes, we evaluated the finite sample performance of the three here introduced estimators. We performed a small simulation study (details not shown for brevity), where we found that $\widehat{\boldsymbol{\beta}}_{M3}$ and $\widehat{\boldsymbol{\beta}}_{M2}$ behave quite similar indeed, but outperform $\widehat{\boldsymbol{\beta}}_{M1}$ by far. Not surprisingly, implementation efforts and computational expenses are much lower for $\widehat{\boldsymbol{\beta}}_{M3}$ than for $\widehat{\boldsymbol{\beta}}_{M2}$ but higher than for $\widehat{\boldsymbol{\beta}}_{M1}$. Fig. 1, left side, shows a typical outcome of the three estimators using quartic kernels for a model with $k = 1, x \sim U[-3, 3]$ and $m(x) = \sin(x)$ with random effects $v_i(\mathbf{z}_{ij}) = \gamma_i \sim N(0, 0.3)$ and error $\varepsilon_i \sim N(0, 0.1)$, where we set $q = 30$ and $n_i = 4$ for all i. The bandwidth used in the example was based on a standard plug-in rule for local linear smoothers with independent data, tending therefore to undersmooth. The critical points, i.e. those where one expects highest and lowest bias and/or variance when estimating a nonparametric function by a local linear kernel smoother, are the minimum, maximum, turning points, etc. In the right panel of Fig. 1 we can see that, not surprisingly, ignoring the dependence structure may lead to serious disturbances due to the random effects.

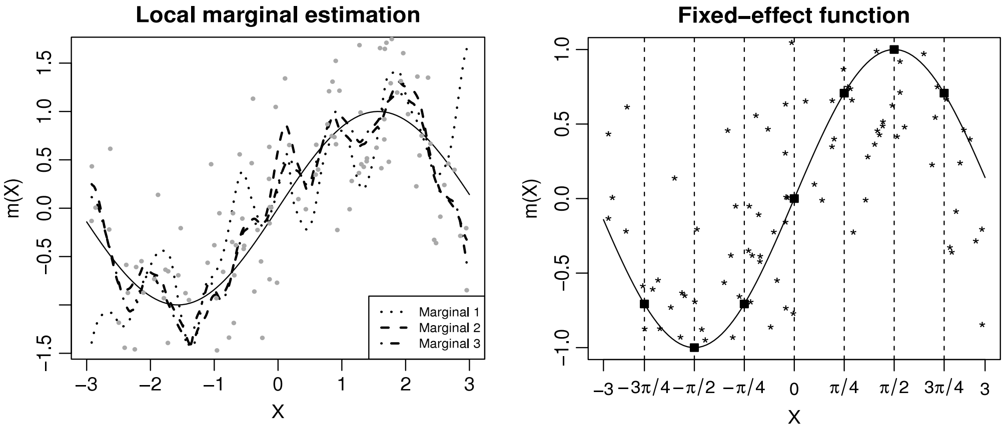

Fig. 1. Left: the three marginal local linear estimates with bandwidth $h = 0.345$ and quartic kernels. Right: simulated sample with fixed effects function and its "critical points".

2.3. A semiparametric model: linear random effects

Nonparametric functionals with nonparametric random effects are, especially for practitioners, a quite demanding challenge in either aspect: asymptotic studies, including further inferences, implementation and interpretation. It is therefore helpful to consider the particular though still quite flexible case of semiparametric models with linear random effects, i.e.

$$y_{ij} = m(\mathbf{x}_{ij}) + \mathbf{b}_i\mathbf{z}_{ij} + \varepsilon_{ij} \quad \text{with } j = 1, \ldots, n_i,\ i = 1, \ldots, q, \qquad \sum_{i=1}^{q} n_i = n, \tag{2.16}$$

where $\mathbf{b}_i$ indicate the random effects. This semiparametric model was also proposed in [30] but without any theoretical study or empirical illustration. To estimate (2.16) one can use each of the marginal estimators presented above.

As far as prediction is concerned, and in particular for small area statistics, it is of essential interest to predict the random effects. This can also be done in a purely nonparametric model; see Appendix A.3. But it is rather cumbersome, what can be seen from two simple considerations: first, for interpretation, simulation and resampling try to imagine the stochastic process that is generating $v_i(\mathbf{z}_{ij})$. Second, while most papers in the literature do not even consider the analysis of predictors for a nonparametric random function v_i, others only discuss its difficulty. In different papers we found different predictors even if they started from the same smoothed likelihood. In contrast, for a semiparametric mixed effects model like (2.16), things simplify a lot and become more interesting for the practical use.

Let us denote by $\widehat{m}_M(\mathbf{x})$ any of the above marginal nonparametric fixed-effects estimators. Then the random-effects component can be predicted as follows:

$$\widehat{\mathbf{b}}^{sm} = (\widehat{\mathbf{b}}_1^{sm}, \ldots, \widehat{\mathbf{b}}_q^{sm})^t = \mathbf{B}\widetilde{\mathbf{Z}}^t\mathbf{V}^{-1}(\mathbf{y} - \widehat{\mathbf{m}}_M), \tag{2.17}$$

where $\widetilde{\mathbf{Z}}$ is the $(n \times r)$-matrix with rows $(\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i})$, $i = 1, \ldots, q$, and $\widehat{\mathbf{m}}_M$ the vector of all $\widehat{m}_M(\mathbf{x}_{ij})$. Under the assumption that $(\mathbf{b}, \mathbf{y})$ follows a multivariate normal distribution, this is the Best Linear Predictor (BLP) for $\mathbf{b}$. When the marginal estimator is derived from generalized linear least squares, the resultant predictor is the BLUP. The individual mean curves can be estimated by $\widehat{\eta}_i(\mathbf{x}, \mathbf{z}) = \widehat{m}_M(\mathbf{x}) + \widehat{\mathbf{b}}_i^{sm}\mathbf{z}$, and thus the average by $\widehat{\Theta}_i = \sum_{j=1}^{n_i}\left(\widehat{m}_M(\mathbf{x}_{ij}) + \widehat{\mathbf{b}}_i^{sm}\mathbf{z}_{ij}\right)/n_i$. Estimators and predictors depend on the variance matrices $\mathbf{B}$ and $\boldsymbol{\Sigma}$. When these matrices are unknown, consistent estimators are plugged in; see Appendix A.2.

3. Mean squared error estimation with bootstrap

The usual efficiency criterion for any estimator of a function $m(\mathbf{x})$ is the prediction error, usually measured by the conditional Mean Squared Error (MSE) at each estimation point $\mathbf{x}$

$$\mathrm{MSE}(\widehat{m}_h(\mathbf{x})) = E\left[(\widehat{m}_h(\mathbf{x}) - m(\mathbf{x}))^2\right] = (\mathrm{Bias}(\widehat{m}_h(\mathbf{x})))^2 + \mathrm{Var}(\widehat{m}_h(\mathbf{x}))$$

or globally by its integrated version, i.e. $\mathrm{MISE}(\widehat{m}_h) = \int \mathrm{MSE}(\widehat{m}_h(\mathbf{x}))d\mathbf{x}$.

In nonparametric mixed-effects model one is interested in quantifying the error of the nonparametric estimator $m(\mathbf{x})$ and the two types of random functions: the individual curves η_i, and the mean parameters $\widehat{\Theta}_i$, $i = 1, \ldots, q$. For the latter two parameters one tries to control the Mean Squared Prediction Error (MSPE). In the current literature on mixed-effects models, resampling methods are becoming more and more popular to estimate the MSE and MSPE; see for example [12].

The common methods are variants of the parametric bootstrap or the moment-matching bootstrap, also known as wild bootstrap. Lombardía and Sperlich [21] used a combination of parametric and wild bootstrap for testing problems. It certainly can also be used for constructing local confidence intervals or global bands. The main disadvantage of resampling-based inference is the time consuming procedure. We will see now how this can be avoided.

The bootstrap algorithm is formulated for given variance matrices. For practical implementation the variance matrices can be estimated considering (restricted) maximum likelihood (cf. [4]) or moment methods (cf. [32]); see again Appendix A.2 for more details. Note that as long as the variance functions can be estimated at a rate faster than that for $m(\cdot)$, the asymptotic results are not affected, i.e. do not change. Consider now the following resampling scheme.

Step 1. Take a pilot estimate $\widehat{m}_{h_0}(\mathbf{x})$ with a pilot bandwidth h_0.
Step 2. Generate $v_i^*(\mathbf{z}_{ij})$ from the assumed mean zero random process with known covariance structure $\gamma(\mathbf{z}_{ij_1}, \mathbf{z}_{ij_2})$ conditioned on the original covariates $\mathbf{z}_{ij}, i = 1, \ldots, q, j = 1, \ldots, n_i$.
Step 3. Generate $\varepsilon_{ij}^* = \sigma(\mathbf{x}_{ij})W_{ij}, j = 1, \ldots, n_i;\; i = 1, \ldots, q$, from W_{ij} being independent and identically distributed (i.i.d.) random variables with $E[W_{ij}] = 0$ and $E[W_{ij}^2] = 1$, independent of $v_i^*(\mathbf{z}_{ij})$.
Step 4. Construct the bootstrap model $y_{ij}^* = \widehat{m}_{h_0}(\mathbf{x}_{ij}) + v_i^*(\mathbf{z}_{ij}) + \varepsilon_{ij}^*$, from the bootstrap sample, $\{(y_{ij}^*, \mathbf{x}_{ij}, \mathbf{z}_{ij}); j = 1, \ldots, n_i; i = 1, \ldots, q\}$, and calculate the bootstrap estimator $\widehat{m}_h^*(\mathbf{x})$.

A bootstrap estimate of $\mathrm{MSE}(\widehat{m}_h(\mathbf{x}))$ is given by

$$\mathrm{MSE}_*(\widehat{m}_h(\mathbf{x})) = E_*\left[\left(\widehat{m}_h^*(\mathbf{x}) - \widehat{m}_{h_0}(\mathbf{x})\right)^2\right], \tag{3.1}$$

where E_* denotes the expectation over the resampling distributions. The simplicity of the considered bootstrap mechanism allows us to compute the exact expectations in (3.1). More specifically, since any proposed marginal estimator can be written as a linear combination of the block-independent responses $\widehat{m}_h(\mathbf{x}) = \sum_{i=1}^q \mathbf{w}_i(\mathbf{x})\mathbf{y}_i$, we have that

$$\begin{aligned} &\mathrm{MSE}_*(\widehat{m}_h(\mathbf{x})) = \{B_*(h; \mathbf{x})\}^2 + V_*(h; \mathbf{x}), \\ &B_*(h; \mathbf{x}) = \sum_{i=1}^q \mathbf{w}_i(\mathbf{x})\widehat{\mathbf{m}}_{i,h_0} - \widehat{m}_{h_0}(\mathbf{x}), \\ &V_*(h; \mathbf{x}) = \sum_{i=1}^q \mathbf{w}_i(\mathbf{x})\mathbf{V}_i\mathbf{w}_i^t(\mathbf{x}), \end{aligned} \tag{3.2}$$

using $\mathrm{Var}_*(\mathbf{y}_i^*) = \mathbf{B}_i + \boldsymbol{\Sigma}_i = \mathbf{V}_i$. Therefore it is not required to use any Monte Carlo simulations to calculate the $\mathrm{MSE}_*(\widehat{m}_h(\mathbf{x}))$.

For the mean prediction errors of $\widehat{\eta}_i$ or $\widehat{\Theta}_i$ however, we do not get such a simplified version. Indeed, for each bootstrap sample the prediction error contains the terms $\widehat{v}_i^* - v_i^*$, which is not easy to get without performing a Monte Carlo simulation. In those cases we have to add the following step.

Step 5. Repeat the procedure for $l = 1, \ldots, R$ and denote by $\widehat{m}_h^{*(l)}(\mathbf{x})$ the bootstrap estimator computed for the l-th bootstrap sample.

The Monte Carlo approximation of $\mathrm{MSE}_*(\widehat{\eta}_{i,h}(\mathbf{x}, \mathbf{z}))$ is then given by

$$\mathrm{MSE}_*(\widehat{\eta}_i(\mathbf{x}, \mathbf{z})) = R^{-1} \sum_{l=1}^R \left(\widehat{\eta}_{i,h}^{*(l)}(\mathbf{x}, \mathbf{z}) - \widehat{\eta}_{i,h_0}(\mathbf{x}, \mathbf{z})\right)^2, \tag{3.3}$$

and analogously for $\widehat{\Theta}_i$.

In the case of unknown variances one has to modify Step 1 to the following.

Step 1. Estimate $\gamma(\cdot, \cdot)$ (or $\widehat{\mathbf{B}}_i$) and $\widehat{\boldsymbol{\Sigma}}_i$. Take a pilot estimate $\widehat{m}_{h_0}(x)$, involving these covariance estimates, with a pilot bandwidth h_0.

In our simulations the method with estimated variances performs quite well. In fact, the impact of estimating the variances is not significant for our purposes; see the Appendix A.2 for details and simulation results.

In the bootstrap strategies presented the choice of a pilot bandwidth h_0 is required. Related works – though in a simpler context – propose asymptotic expressions for an "optimal pilot bandwidth"; see for example [8]. This pilot bandwidth must tend to zero at a rate slower than h; see for example [13] or [1] who showed that the optimal bandwidth rate slowed down from $O(n^{-1/5})$ for h to $O(n^{-1/9})$ for h_0 in the one-dimensional case. Note that our plug-in proposal does not involve further iterations like most of the so far published methods do. In other words, we avoid this way two nested iterations and get a quick and easy method instead.

The proof of consistency for the bootstrap approximation of the MSE is given in the Appendix. It is derived analogously to Martínez-Miranda et al. [22], and is based on calculating the imitations done in the bootstrap. The key issue is to know the asymptotic expansion of the MSE which we approximate using the bootstrap method. With respect to the MSPE of both, individual curves η_i and related mean parameters Θ_i, such a proof can only be derived for the semiparametric model.

Table 1
Coverage probabilities for simple bootstrap confidence intervals (3.6).

(n, D)	Size (%)	$-3\pi/4$	$-\pi/2$	$-\pi/4$	0.0	$\pi/4$	$\pi/2$	$3\pi/4$
(120, 30)	90	0.868	0.884	0.881	0.880	0.885	0.874	0.857
	95	0.929	0.939	0.936	0.936	0.938	0.933	0.919
(240, 40)	90	0.901	0.919	0.918	0.913	0.912	0.933	0.924
	95	0.949	0.962	0.960	0.958	0.961	0.968	0.965

Specifically, since in the semiparametric model the nonparametric part, $\widehat{m}$, dominates the asymptotic expressions (due to its slower convergence rate), it is sufficient to study the asymptotics of the MSE of $\widehat{m}_h$, and the consistency of the bandwidth selection method.

Assuming some regularity conditions given in the Appendix, the bootstrap approximation $\text{MSE}_*(\widehat{m}_h(\mathbf{x}))$ in (3.1) is consistent in the interior of the support of f, being f the density of $\mathbf{x}$. In probability we have then

$$\text{MSE}_*(\widehat{m}_h(\mathbf{x})) - \text{MSE}(\widehat{m}_h(\mathbf{x})) \to 0. \tag{3.4}$$

For the sake of presentation we restrict to the semiparametric model for the rest of the paper, i.e. we assume a linear random-effects component from now on. Then, Step 2 is just drawing random effects $\mathbf{b}_i^*$, $i = 1, \dots, q$, from $\mathbf{B}_i^{1/2}\mathbf{W}$, where $\mathbf{W}$ is a random vector with zero mean and covariance matrix being equal to the identity. For Step 4 one has then $v_i^*(\mathbf{z}_{ij}) = \mathbf{z}_{ij}'\mathbf{b}_i^*$.

To test the behavior of the MSE approximation in finite samples, we performed a simulation study for which we generated data samples $\{x_{ij}, y_{ij}\}_{i,j=1}^{n_i,q}$ from model

$$y_{ij} = \sin(x_{ij}) + b_i + \varepsilon_{ij} \tag{3.5}$$

with random design $x_{ij} \sim U[-3, 3]$, i.i.d. errors $\varepsilon_{i,d} \sim N(0, 0.3)$, and i.i.d. random effects $b_i \sim N(0, 0.3)$. In our simulations, we always set $n_1 = n_2 = \cdots = n_q$ for simplicity. For the above discussed reasons we consider here only the marginal estimator $M3$ with quartic kernels. We estimated 500 times function $m(\cdot)$ at the points indicated in Fig. 1 with the non-optimal global bandwidth $h = 1.0$. This exercise has been done first for a sample with $(n_i, q) = (4, 30)$, and was then repeated for $(n_i, q) = (6, 40)$. We studied the construction of local confidence intervals based on our bootstrap procedure. As it is known that even with bootstrap, the estimation of the bias is a serious problem, we follow the common spirit of neglecting the bias. That is, we will simply use the variance estimates V_* given in (3.2). Note that due to the neglecting of the bias in the construction of confidence intervals we either should undersmooth or argue that we construct confidence intervals for the expectation of the estimates but not for the true function. In the first case (i.e. undersmoothing), the random confidence intervals should guarantee via undersmoothing that they hold the given coverage probability for the true function. In the second case, we say that the calculated confidence interval covers $(1 - \alpha/2)\%$ of all estimates when repeating the experiment. In both cases, $(1 - \alpha/2)\%$ confidence intervals for $m(x)$ based on the estimates $\hat{m}_h(x)$, V_*, and the asymptotic normality of our estimates, would result in

$$\left[\hat{m}_h(x) - z_{1-\alpha/2}V_*^{1/2}(x); \hat{m}_h(x) + z_{1-\alpha/2}V_*^{1/2}(x)\right]. \tag{3.6}$$

Note that these intervals are constructed without Monte Carlo simulations and therefore they are calculated in a quick and easy way.

We investigated the realized coverage probabilities for 90% and 95% confidence intervals. For the true function we realized always coverage probabilities slightly larger than the nominal size (1%–4% points higher for the 90% confidence intervals, i.e. conservative ones) for each sample size. The coverage of estimates in repeated experiments is shown in Table 1. We see how the coverage probability converges to the expected one for increasing sample size. When we also have to estimate the variance, then the coverage probability diminishes slightly (for the 90% confidence interval in average about 1%–4% points compared to the here reported ones, depending on x and the sample size). For constructing prediction intervals for η_i or Θ_i we cannot avoid the computationally more expensive Monte Carlo approximation. This however has the advantage that for those cases we no longer need to work with the normality approximation in (3.6). Note that, while in parametric mixed effects models bootstrap prediction intervals improved only marginally compared to the typically used linear approximations, in non- and semiparametric mixed effects models they are the only applicable method (at least to our knowledge).

3.1. Local optimal bandwidth selection

In nonparametric estimation the choice of the optimal smoothing parameter for estimation is the counterpart of model selection in parametric regression. When there is no random effect term, the optimality of smoothing parameters and its data-driven estimation have been extensively studied; see [16]. The so-called cross-validation method is quite popular due to its intuitive definition as a simple minimizer of the MSE. But it suffers from high variability, tends to undersmooth in practice, is computationally intensive, and does not allow for finding locally optimal smoothing parameters. The so-called plug-in methods require pre-estimates for all expressions in the MSE, which are not easily available. The bootstrap method can be a remedy here, and it can be used to find locally optimal smoothing parameters as we will see next.

Table 2
Results from 500 simulation runs based on model (3.5) with $n_i = 6$ for all $q = 40$ clusters and known variance components.

x		(σ_e^2, σ_b^2)			
		(0.1, 0.5)	(0.5, 0.1)	(0.6, 0.0)	(0.3, 0.3)
$\frac{-3\pi}{4}$	h_{opt}	2.75	2.75	2.75	2.75
	$mean_{h_b}, std_{h_b}$	2.616(0.2213)	2.608(0.4547)	2.385(0.7325)	2.641(0.2657)
	$\text{mse}(\widehat{m}_{h_{opt}})$	0.0157	0.0122	0.0120	0.0137
	$\text{mse}(\widehat{m}_{h_b}), std$	0.0161(0.0008)	0.0126(0.0011)	0.0125(0.0010)	0.0141(0.0009)
$\frac{-2\pi}{4}$	h_{opt}	0.95	0.95	0.95	0.95
	$mean_{h_b}, std_{h_b}$	1.035(0.1561)	1.021(0.1289)	1.005(0.1151)	1.036(0.1450)
	$\text{mse}(\widehat{m}_{h_{opt}})$	0.0222	0.0168	0.0153	0.0196
	$\text{mse}(\widehat{m}_{h_b}), std$	0.0229(0.0011)	0.0177(0.0017)	0.0162(0.0017)	0.0204(0.0013)
$\frac{-\pi}{4}$	h_{opt}	1.1	1.1	1.1	1.1
	$mean_{h_b}, std_{h_b}$	1.204(0.3265)	1.159(0.1485)	1.128(0.1236)	1.169(0.1741)
	$\text{mse}(\widehat{m}_{h_{opt}})$	0.0196	0.0147	0.0133	0.0173
	$\text{mse}(\widehat{m}_{h_b}), std$	0.0203(0.0015)	0.0153(0.0011)	0.0138(0.0010)	0.0179(0.0010)
0.0	h_{opt}	2.3	2.75	2.75	2.75
	$mean_{h_b}, std_{h_b}$	2.151(0.2586)	2.587(0.3279)	2.548(0.3684)	2.469(0.3347)
	$\text{mse}(\widehat{m}_{h_{opt}})$	0.0147	0.0064	0.0039	0.0111
	$\text{mse}(\widehat{m}_{h_b}), std$	0.0149(0.0003)	0.0067(0.0007)	0.0043(0.0009)	0.0113(0.0004)

The inclusion of a random term in the regression model modifies also the standard optimal values for fixed-effects models. For longitudinal or clustered data the asymptotically optimal (global) bandwidth has been derived by Lin and Carroll [17], Wu and Zhang [31], Park and Wu [24], and Chen and Jin [2]. In all cases the optimal bandwidth has been calculated for the case where the true variance matrices were known and afterward were adjusted to the common situation of unknown variances. Park and Wu [24] proposed a backfitting procedure with nested iterative algorithms, updating all, different model components, variances and bandwidths, based on a cross-validatory technique. Xu and Zhu [32] proposed a generalized cross validation method to simultaneously estimate the bandwidth and the variances. Chen and Jin [2] estimated the global optimal bandwidth by mimicking the rule-of-thumb global bandwidth selector of Fan and Gijbels [7], implemented cross-validation and discussed the empirical-bias bandwidth selector of Ruppert [25]. But neither of them are adjusted to the presence of the new random-effect component and the consequent increase of variance.

We propose a data-driven smoothing parameter selection that is based on the bootstrap method looking for the locally optimal bandwidths

$$h_{opt}(\mathbf{x}) = \arg\min_h \text{MSE}(\widehat{m}_h(\mathbf{x})). \tag{3.7}$$

Making use of the bootstrap MSE estimate, the local bandwidth selector for the marginal estimator $\widehat{m}_h(\cdot)$ is consequently

$$h_b(\mathbf{x}) = \arg\min_h \text{MSE}_*(\widehat{m}_h(\mathbf{x})). \tag{3.8}$$

The consistency of the selection procedure follows from the one of the bootstrap approximation of the MSE; see also Appendix.

To check out the behavior of this bandwidth selector we extended our simulation study from above. Again we generated 500 samples $\{x_{ij}, y_{ij}\}_{i,j=1}^{n_i,q}$ from model (3.5) but $x_{ij} \in [-3, 3]$ taken from a fixed equidistant grid, i.i.d. errors $\varepsilon_{i,d} \sim N(0, \sigma_e^2)$, and i.i.d. random effects $b_i \sim N(0, \sigma_b^2)$ with different combinations of (σ_e, σ_b) fulfilling $\sigma_e^2 + \sigma_b^2 = 0.6$. We varied sample size and the number of clusters q. The rest was as before.

The optimal bandwidth was searched for estimating function $m(\cdot)$ with M3 at $x_k = k\pi/8$ for all integers k from -7 to 7. In this way we verified that the method works equally well for symmetric problems. For all points we searched the optimal local bandwidth out of the following equidistant grid $\{0.65, 0.8, \ldots, 2.6, 2.75\}$ (15 bandwidths). The pilot plug-in bandwidth h_0, see step 1, was calculated as follows: we applied a local quadratic estimator to get an estimate of the second derivative, $\widehat{m''}(\cdot)$ and set then, c.f. [7],

$$h_{PI}^5 = 2.75^{1/5} \wedge \frac{(\sigma_e^2 + \sigma_b^2) \cdot range(x) \cdot \int K^2(u)du}{n(\int u^2 K(u)du)^2 \frac{1}{q} \sum_{i=1}^{q} \sum_{j=1}^{n_i} \frac{1}{n_i} \widehat{m''}^2(\mathbf{x}_{ij})}, \quad n = \sum_{i=1}^{q} n_i \tag{3.9}$$

which for simplicity ignores the dependence structure of the data. Following arguments of [13] or [1] we set $h_0 = h_{PI} n^{\frac{1}{5}-\frac{1}{9}}$, compare also discussion from above. To avoid that the estimation of m'' by local quadratic smoothing fails due to data sparse areas we used a bandwidth of size 2.75. This, however, might lead to oversmoothing and result in $\sum_{j=1}^{n_i} \widehat{m''}^2(\mathbf{x}_{ij}) \approx 0$. Therefore we introduced the upper boundary of 2.75 for h_{PI} in (3.9).

Comparing h_{opt} with the mean and standard deviation of h_b, and the mean squared errors of the corresponding estimates of m respectively (calculated from 500 simulation runs), we see in Table 2 that our bandwidth selection methods works

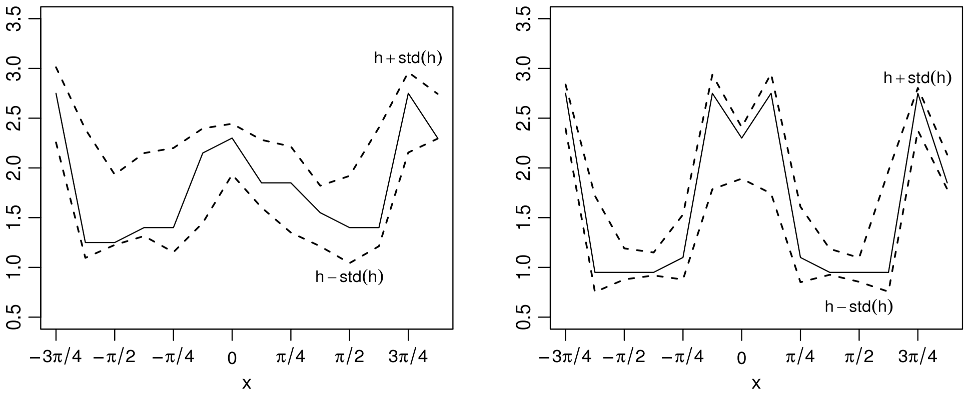

Fig. 2. For $\mathbf{x}_{ij} = k/(8\pi)$, $k = -7, -6, \ldots, 7$ the solid line indicates h_{opt}, the upper and lower dashed line indicate $mean(h_b) \pm std(h_b)$ where *mean*, *std* indicate mean and standard deviation over 500 simulation runs from model (3.5) with $\sigma_b^2 = 0.5$, $\sigma_e^2 = 0.1$. $(n_i, q) = (3, 20)$ on the left plot; $(n_i, q) = (6, 40)$ on the right plot.

Table 3
Results from 500 simulation runs based on model (3.5) with $\sigma_e^2 = \sigma_b^2 = .3$ and known variance components.

x		n			
		60	120	240	480
$\frac{-3\pi}{4}$	h_{opt}	2.75	2.75	2.75	2.75
	$mean_{h_b}, std_{h_b}$	2.612(0.4380)	2.681(0.1864)	2.641(0.2657)	2.635(0.2154)
	$\mathrm{mse}(\hat{m}_{h_{opt}})$	0.0444	0.0247	0.0137	0.0082
	$\mathrm{mse}(\hat{m}_{h_b}), std$	0.0456(0.0040)	0.0251(0.0011)	0.0141(0.0009)	0.0086(0.0007)
$\frac{-2\pi}{4}$	h_{opt}	1.25	1.1	0.95	0.8
	$mean_{h_b}, std_{h_b}$	1.552(0.3655)	1.239(0.2154)	1.036(0.1450)	0.868(0.0935)
	$\mathrm{mse}(\hat{m}_{h_{opt}})$	0.0564	0.0300	0.0196	0.0121
	$\mathrm{mse}(\hat{m}_{h_b}), std$	0.0658(0.0149)	0.0320(0.0048)	0.0204(0.0013)	0.0123(0.0004)
$\frac{-\pi}{4}$	h_{opt}	1.4	1.25	1.1	0.95
	$mean_{h_b}, std_{h_b}$	1.687(0.4186)	1.436(0.2729)	1.169(0.1741)	1.014(0.1155)
	$\mathrm{mse}(\hat{m}_{h_{opt}})$	0.0484	0.0281	0.0173	0.0110
	$\mathrm{mse}(\hat{m}_{h_b}), std$	0.0539(0.0096)	0.0300(0.0036)	0.0179(0.0010)	0.0112(0.0004)
0.0	h_{opt}	2.75	2.75	2.75	2.6
	$mean_{h_b}, std_{h_b}$	2.575(0.3272)	2.572(0.3342)	2.469(0.3347)	2.336(0.3575)
	$\mathrm{mse}(\hat{m}_{h_{opt}})$	0.0274	0.0166	0.0111	0.0070
	$\mathrm{mse}(\hat{m}_{h_b}), std$	0.0284(0.0025)	0.0170(0.0009)	0.0113(0.0004)	0.0071(0.0003)

pretty well for different "critical points" $\mathbf{x}_{ij}$ (see Fig. 1) and different combinations of (σ_b, σ_e), including the fixed effects model with $\sigma_b = 0$. We used here $n_i = 6$ for all of the $q = 40$ clusters. This finding is illustrated by the two plots in Fig. 2 where for $\mathbf{x}_{ij} = k/(8\pi)$, $k = -7, -6, \ldots, 7$ the solid line indicates h_{opt}, whereas the upper and the lower dashed line indicate $mean(h_b) \pm std(h_b)$ with *mean*, *std* being the mean and standard deviation over 500 simulation runs from model (3.5). We have considered the model with $\sigma_b^2 = 0.5$, $\sigma_e^2 = 0.1$ and sample sizes $(n_i, q) = (3, 20)$ on the left plot, and $(n_i, q) = (6, 40)$ on the right plot. We see first, that the selection procedure reveals the functional form of $h_{opt}(\mathbf{x})$ with respect to $\mathbf{x}$, even for such a small sample of only $n = 60$. We see further how this method improves for increasing sample size. The latter point is also illustrated in Table 3 for a model with $\sigma_e^2 = \sigma_b^2 = 0.3$. Again, like in Table 2, we see that not just h_b comes close to h_{opt} but, even more important, the optimal mean squared errors for each $\mathbf{x}$ can be reached, indeed.

4. Conclusions

Our aim in this paper has been to compare different kernel estimators, and solve the practical but the crucial problem of doing inference and choosing the bandwidth for nonparametric mixed models. For convenience we considered a simpler semiparametric version with parametric random effect impacts. We have defined several local kernel estimation strategies to estimate the nonparametric function in the model. For all considered smoothers we have solved the problem of constructing confidence intervals, and finding the optimal (local) bandwidth by a simple bootstrap method that does not require expensive Monte Carlo simulations. We have provided the consistency of our methods, and have carried out simulation studies which have revealed the good performance of our methods in practice. For the assumed mixed model not many competitors are available which actually involve the assumed correlation structure, and only cross-validation strategies have been proposed recently in the literature for bandwidth selection. These however, become very slow and even inefficient for big datasets.

Further research is still necessary to extend the results to a fully nonparametric mixed model where inference and bandwidth selection becomes also relevant for the prediction of the random effects. As already indicated in our paper, several additional difficulties will arise then, and the practicability is questionable.

Appendix

A.1. Appendix A: asymptotic results

We first derive the mean squared error of the marginal estimator (2.14) in the semiparametric model (2.16). To simplify notation we will write the estimator as $\widehat{m}_M(\mathbf{x}) = \mathbf{e}_1^t\widehat{\boldsymbol{\beta}}_M$. We will work with two sets of hypothesis, namely cases 1 and 2 below.

- *Case* 1. The number of observations in each group, n_i, tends to ∞ and also the number of groups, $q \to \infty$.
- *Case* 2. The number of observations in each group, n_i, is bounded and only the number of groups, q goes to infinity.

Because the covariance matrices can be estimated at parametric rates we derive the asymptotics assuming that the variance matrices are known without loss of generality. For notational reasons, but without loss of generality, we set the dimension of $\mathbf{x}_{ij} = x_{ij}$ to one ($k = 1$). For each case we consider the following set of assumptions.

For the first case we have the following.

A11. The design points $x_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, q$, are i.i.d. with density $f(\cdot)$.
A12. The point x is in the interior of the support of f where $f(x) > 0$ and $f'(x)$ exists.
A13. The fixed-effect function $m(x)$ has twice-continuous derivatives at x.
A14. The variances of the random effects, B_i are uniformly bounded.
A15. The conditional error variance $\sigma^2(x) = E[\varepsilon(x)^2]$ is continuous at x. Also $\int \left(f(u)/\sigma^4(u)du\right) < \infty$.
A16. The kernel K is a bounded symmetrical probability density function with bounded support ($[-1, 1]$) so that $\int K(u)u^2du < \infty$ and $\int K^2(u)du < \infty$.
A17. As $q \to \infty, h \to 0, n_ih \to \infty$ and $n_ih^3 \to 0$, for $i = 1, \ldots, q$.

For the second case we have the following.

A21. The assumptions A11–A16 are satisfied.
A22. For some $0 < C < \infty, n_i \le C, i = 1, \ldots, q$.
A23. As $q \to \infty, h \to 0, nh \to \infty$ and $nh^3 \to 0$.

Let $B_{(r,s)}(K) = \int u^sK(u)^rdu$ for any adequate kernel K, and $\widetilde{n} = \frac{q}{\sum_{i=1}^q n_i^{-1}}$.

Lemma 1. • *Case* 1. *Under conditions A11–A17the asymptotic mean squared error of $\widehat{m}_M(x)$ is given by*

$$\mathrm{MSE}(\widehat{m}_M(x)) = \frac{h^4}{4}\left(f(x)m''(x)\right)^2 B^2_{(1,2)}(K)\left[1 + O_P\left((\widetilde{n}h)^{-1/2}\right)\right]^2 + \left\{\sum_{i=1}^q \frac{n_i^2B_i}{n^2} + \sum_{i=1}^q \frac{n_i\left(B_i + \sigma^2(x)\right)}{n^2hf(x)}B_{(2,0)}(K)\right\}\left[1 + O_P((\widetilde{n}h)^{1/2}) + O_P((h/\widetilde{n})^{1/2})\right].$$

• *Case* 2. *Under conditions A21–A23 the asymptotic mean squared error of $\widehat{m}_M(x)$ is given by*

$$\mathrm{MSE}(\widehat{m}_M(x)) = \frac{h^4}{4}\left(f(x)m''(x)\right)^2 B^2_{(1,2)}(K)\left[1 + O_P\left(h^{-1/2}\right)\right]^2 + \left\{\sum_{i=1}^q \frac{n_i^2B_i}{n^2} + \sum_{i=1}^q \frac{n_i\left(B_i + \sigma^2(x)\right)}{n^2hf(x)}B_{(2,0)}(K)\right\}\left[1 + O_P(h^{1/2}) + O_P(h^{1/2})\right].$$

Remark 1. Note that for the case 1 when $B_i \equiv \Sigma_B, \forall i = 1, \ldots, q$, since $\frac{n_i^2}{n^2} \approx \frac{1}{q}$, the error can be written as

$$\mathrm{MSE}(\widehat{m}_{M3}(x)) = \frac{h^4}{4}\left(m''(x)\right)^2 B^2_{(1,2)}(K)\left[1 + O_P\left((\widetilde{n}h)^{-1/2}\right)\right]^2 + \left\{\frac{\Sigma_B}{q} + \frac{\Sigma_B + \sigma^2(x)}{nhf(x)}B_{(2,0)}(K)\right\}\left[1 + O_P((\widetilde{n}h)^{1/2}) + O_P((h/\widetilde{n})^{1/2})\right].$$

The above expression is equivalent to that given by Park and Wu [24, Theorem 4.1], but considering the parametric structure for the random effects defined in the semiparametric model (2.16).

Sketch of the Proof. The proof follows steps similar to those of [24, Theorems 4.1 and 4.2] and therefore we only show the main calculations. Let us consider the minus-log-likelihood in (2.13),

$$Q(\boldsymbol{\beta}) = -\mathcal{L}_M(\boldsymbol{\beta};\mathbf{y}) = -\sum_{i=1}^{q} \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{V}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} [\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}].$$

Using a Taylor expansion at the optimizer $\widehat{\boldsymbol{\beta}}_M$, the (conditional) bias can be computed by

$$\mathrm{Bias}(\widehat{\boldsymbol{\beta}}_M) = E\left[\widehat{\boldsymbol{\beta}}_M\right] - \boldsymbol{\beta} = -\left(Q'(\boldsymbol{\beta})\right)^{-1} E\left[Q(\boldsymbol{\beta})\right] \tag{A.1}$$

since $Q'(\boldsymbol{\beta}) = \sum_{i=1}^q \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{V}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{X}_i$ is constant under the conditional distribution. And the (conditional) variance is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_M) = \left(Q'(\boldsymbol{\beta})\right)^{-1} \mathrm{Var}\left(Q(\boldsymbol{\beta})\right) \left(Q'(\boldsymbol{\beta})\right)^{-1}. \tag{A.2}$$

We can write the two factors in (A.1) by

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{q} \left\{ \mathbf{G}_{x1i} - \mathbf{G}_{x2i} \Lambda_i^{-1} \mathbf{G}_{x2i}^t \right\},$$

$$E\left[Q(\boldsymbol{\beta})\right] = -\sum_{i=1}^{q} \left\{ \mathbf{G}_{m1i} - \mathbf{G}_{x2i} \Lambda_i^{-1} G_{m2i} \right\}.$$

Here we have used that $\mathbf{V}_i^{-1} = \boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \Lambda_i^{-1} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1}$, with $\Lambda_i = B_i^{-1} + \sum_{j=1}^{n_i} \sigma^{-2}(x_{ij})$ $(i = 1, \ldots, q)$, and introduced the following notation: $\mathbf{m}_i = (m_{i1}, \ldots, m_{in_i})^t$, $m_{ij} = m(x_{ij}) - \mathbf{X}_i\boldsymbol{\beta} = m(x_{ij}) - m(x) - (x_{ij} - x)m'(x)$, $\mathbf{G}_{m1i} = \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{m}_i$, $\mathbf{G}_{x1i} = \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{X}_i$, $\mathbf{G}_{x2i} = \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}$ and $G_{m2i} = \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{m}_i$ $(i = 1, \ldots, q)$. Now we use standard asymptotic approximations for each of the elements in the defined matrices, based on the Markov property, i.e. for a random variable Z with finite second-order moment it holds $Z = E[Z] + O_P(\sqrt{\mathrm{Var}(Z)})$. The resulting approximations are given by

$$Q'(\boldsymbol{\beta}) = \frac{nf(x)}{\sigma^2(x)} \begin{pmatrix} 1 + O_P((\widetilde{n}h)^{-\frac{1}{2}}) & O(h^2) + O_P((h/\widetilde{n})^{\frac{1}{2}}) \\ O(h^2) + O_P((h/\widetilde{n})^{\frac{1}{2}}) & O(h^2) + O_P((h^3/\widetilde{n})^{\frac{1}{2}}) \end{pmatrix} \tag{A.3}$$

and

$$E\left[Q(\boldsymbol{\beta})\right] = \frac{nh^2}{\sigma^2(x)} \frac{1}{2} m''(x) f(x) B_{(1,2)} \begin{pmatrix} 1 + O_P((\widetilde{n}h)^{-\frac{1}{2}}) \\ O_P((h/\widetilde{n})^{\frac{1}{2}}) \end{pmatrix}. \tag{A.4}$$

If we consider the situation in case 2, the results are similar using analogous approximations and the assumption (A15). In this case we get

$$Q'(\boldsymbol{\beta}) = \frac{nf(x)}{\sigma^2(x)} \begin{pmatrix} 1 + O_P(h^{-\frac{1}{2}}) & O(h^2) + O_P(h^{\frac{1}{2}}) \\ O(h^2) + O_P(h^{\frac{1}{2}}) & O(h^2) + O_P(h^{\frac{3}{2}}) \end{pmatrix}. \tag{A.5}$$

For $E\left[Q(\boldsymbol{\beta})\right]$ we get an expression analogous to (A.4) but dropping $\widetilde{n}$ from the $O_P(\cdot)$ terms. Finally, substituting the above expressions into (A.1) the asymptotic expression for the bias term in the lemma is proved.

For the variance (A.2) we have to calculate the term

$$\mathrm{Var}(Q(\boldsymbol{\beta})) = \sum_{i=1}^{q} \left\{ \boldsymbol{\Omega}_{1i} - 2\mathbf{G}_{x2i} \Lambda_i^{-1} \boldsymbol{\Omega}_{2i} + \mathbf{G}_{x2i} \Lambda_i^{-1} \Omega_{3i} \Lambda_i^{-1} \mathbf{G}_{x2i}^t \right\}, \tag{A.6}$$

where we have defined the matrices: $\boldsymbol{\Omega}_{1i} = \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} (\boldsymbol{\Gamma}_i + \boldsymbol{\Sigma}_i) \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{X}_i$, $\boldsymbol{\Omega}_{2i} = \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} (\boldsymbol{\Gamma}_i + \boldsymbol{\Sigma}_i) \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{X}_i$, and $\Omega_{3i} = \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} (\boldsymbol{\Gamma}_i + \boldsymbol{\Sigma}_i) \mathbf{W}_{ih}^{\frac{1}{2}} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}$, with $\boldsymbol{\Gamma}_i = \mathbf{1}_{n_i} B_i \mathbf{1}_{n_i}^t$ $(i = 1, \ldots, q)$. Since we are interested in $\mathrm{Var}(\widehat{m}_M(x)) = \mathbf{e}_1^t \mathrm{Var}(\widehat{\boldsymbol{\beta}}_M) \mathbf{e}_1$, we only have to provide an asymptotic approximation for the first element of the matrices involved in expression (A.6). More specifically, for $i = 1, \ldots, q$, we get:

$$\mathbf{e}_1^t \mathbf{G}_{x2i} \Lambda_i^{-1} \boldsymbol{\Omega}_{2i} \mathbf{e}_1 = \frac{n_i^2 f^2(x)}{\sigma^4(x)} \left\{ hO(1) + O(n_i^{-1}) + O_P((n_i h)^{-\frac{1}{2}}) \right\},$$

$$\mathbf{e}_1^t \mathbf{G}_{x2i} \Lambda_i^{-1} \Omega_{3i} \Lambda_i^{-1} \mathbf{G}_{x2i}^t \mathbf{e}_1 = n_i^2 h \left\{ O(1) + O_P((n_i h)^{-2}) O(n_i^{-1}) + O_P((n_i h)^{-\frac{1}{2}}) \right\},$$

$$\mathbf{e}_1^t \boldsymbol{\Omega}_{1i} \mathbf{e}_1 = \frac{n_i^2 f^2(x) B_i}{\sigma^4(x)} \left(1 + O_P(h^2)\right) + \frac{n_i^2 f(x) \tau_i^2(x) B_{(2,0)} h^{-1}}{\sigma^4(x)} \left(1 + O_P((n_i h)^{-\frac{1}{2}})\right).$$

The result for the variance in the Lemma are obtained by using the above approximations. Again, the situation in case 2 yields analogous results when dropping $\widetilde{n}$ from the $O_P(\cdot)$ terms.

Consistency of the bootstrap approximation

Assuming the same hypotheses as before together with the following ones (depending on the case 1 or 2), we have the following.

- *Case* 1: A18. As $q \to \infty$, $h_0 \to 0$, $n_i h_0 \to \infty$ and $n_i h_0^3 \to 0$. Also $h_0^{-1} h \to 0$.
- *Case* 2: A24. As $q \to \infty$, $h_0 \to 0$,$nh_0 \to \infty$ and $nh_0^3 \to 0$. Also $h_0^{-1} h \to 0$.

The proof of (3.4) is obtained by imitation. We consider for simplicity the semiparametric model (2.16) and the marginal estimator (2.14), $\widehat{m}_M(x) = \mathbf{e}_1^t \widehat{\boldsymbol{\beta}}_M$. Assuming known variance matrices, the marginal estimator is linear in the responses and we obtain the decomposition of the bootstrap MSE into the squared bootstrap bias term

$$\mathrm{Bias}_*(\widehat{m}_h(x)) = \sum_{i=1}^{q} \mathbf{w}_i(x)\widehat{\mathbf{m}}_{i,h_0} - \widehat{m}_{h_0}(x),$$

and the bootstrap variance,

$$\mathrm{Var}_*(\widehat{m}_h(x)) = \sum_{i=1}^{q} \mathbf{w}_i(x)\mathbf{V}_i\mathbf{w}_i^t(x).$$

In a similar way we have a squared-bias and variance decomposition for the true MSE with: $\mathrm{Bias}(\widehat{m}_h(x)) = \sum_{i=1}^{q} \mathbf{w}_i(x)\mathbf{m}_i - m(x)$ and $\mathrm{Var}(\widehat{m}_h(x)) = \sum_{i=1}^{q} \mathbf{w}_i(x)\mathbf{V}_i\mathbf{w}_i^t(x)$. Taking differences we only have to assess that the difference between bias terms goes to zero. Note that with straightforward calculations on the expressions derived above, the difference becomes in fact

$$\mathrm{Bias}_*(\widehat{m}_h(x)) - \mathrm{Bias}(\widehat{m}_h(x)) = \mathbf{e}_1^t \left(Q'(\boldsymbol{\beta})\right)^{-1} \left\{ \sum_{i=1}^{q} \mathbf{X}_i^t \mathbf{W}_{ih}^{\frac{1}{2}} \mathbf{V}_i^{-1} \mathbf{W}_{ih}^{\frac{1}{2}} \left(\widehat{\mathbf{m}}_{i,h_0} - \mathbf{m}_i\right) \right\}$$

which goes to zero as $n \to \infty$ because of the consistency of the pilot estimator $\widehat{\mathbf{m}}_{i,h_0}$ and the assumed oversmoothing by the pilot bandwidth i.e. ($h_0^{-1} h \to 0$).

A.2. Appendix B: estimation of the covariance matrices

All estimators presented depend on the covariance matrices $\mathbf{B}$ and $\boldsymbol{\Sigma}$. Where these matrices are unknown, consistent estimators are used as substitutes. Those are easily available for example when using joint likelihood estimation. Note that the estimators for the fixed effects are not longer linear functions of the responses which makes it harder to deal with. If the conditional variance functions, $\sigma^2(\cdot)$, and $\gamma(\cdot, \cdot)$ or $\mathbf{B}_i$ respectively, are parametrically specified, the covariance matrices can be estimated at parametric rates. When a general linear mixed model is assumed, [14] provides asymptotic results for all estimators. The generalization to a semiparametric setting can be found e.g. in [21], whereas the generalization to a fully nonparametric setting is still an open problem, even in the simpler longitudinal data case of [31,30]. Most frequently used methods are the maximum likelihood estimators (ML) and restricted maximum likelihood estimators (REML). For the above local polynomial estimators, [31] calculate these estimators by the EM algorithm or the Newton–Raphson algorithm. We adapted their method to our setting. In order to do so it is helpful to specify the distribution of ($\mathbf{b}$, $\mathbf{y}$). Taking the joint normal distribution, the ML method for estimating $\mathbf{B}$ and $\boldsymbol{\Sigma}$ is based on the following kernel-weighted log-likelihood

$$\log \mathcal{L}(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Sigma}|\mathbf{y}) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log|\widetilde{\mathbf{V}}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t \boldsymbol{\Omega}_h (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

with $\widetilde{\mathbf{V}} = \mathbf{W}_h^{1/2}\mathbf{Z}\mathbf{B}\mathbf{Z}^t\mathbf{W}_h^{1/2} + \boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}_h = \mathbf{W}_h^{1/2}\widetilde{\mathbf{V}}^{-1}\mathbf{W}_h^{1/2}$. Note that this looks like the ML method for standard linear mixed effects models but adding some kernel-transformations. This likelihood also gives estimates for $\boldsymbol{\beta}$. In contrast, the REML method integrates out these parameters in order to adjust for the loss of degrees of freedom. Then it maximizes the log of $\int \mathcal{L}(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Sigma}|\mathbf{y}) d\boldsymbol{\beta}$ given by

$$\log \mathcal{L}(\mathbf{B}, \boldsymbol{\Sigma}|\mathbf{y}) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log|\widetilde{\mathbf{V}}| - \frac{1}{2} \log\left|\mathbf{X}^t \boldsymbol{\Omega}_h \mathbf{X}\right| - \frac{1}{2}\mathbf{y}^t \mathbf{W}_h \mathbf{P}_V \mathbf{W}_h \mathbf{y}$$

with $\mathbf{P}_V = \widetilde{\mathbf{V}}^{-1} - \widetilde{\mathbf{V}}^{-1}\mathbf{W}_h^{1/2}\mathbf{X}(\mathbf{X}^t\boldsymbol{\Omega}_h\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}_h^{1/2}\widetilde{\mathbf{V}}^{-1}$.

Alternative methods include the method of moments. For instance, [32] proposed to use it to estimate the covariances of the random effects starting from a pilot estimation of the fixed-effects function. When checking the behavior of our bandwidth selection, for the case where variances are unknown, we used this moment approach.

Again we generated 500 samples $\{x_{ij}, y_{ij}\}_{i,j=1}^{n_i,q}$ from model (3.5), $x_{ij} \in [-3, 3]$ taken from a fixed equidistant grid, i.i.d. errors $\varepsilon_{i,d} \sim N(0, \sigma_e^2)$, and i.i.d. random effects $b_i \sim N(0, \sigma_b^2)$ with different combinations of (σ_e, σ_b) but keeping

Table 4
Results from 500 simulation runs based on model (3.5) with different sample sizes and combinations of (σ_e, σ_b) comparing the results of our bandwidth selection under known versus unknown variances.

(q, n_i, n)		(30, 4, 120)				(40, 6, 240)			
(σ_e^2, σ_b^2)		(0.1, 0.5)		(0.5, 0.1)		(0.1, 0.5)		(0.5, 0.1)	
x, V		Known	Unkn.	Known	Unkn.	Known	Unkn.	Known	Unkn.
$\frac{-2\pi}{4}$	$mean_{h_b}$	1.323	1.298	1.259	1.262	1.068	1.060	1.050	1.061
	std_{h_b}	0.2733	0.2715	0.2270	0.2316	0.1740	0.1817	0.1437	0.1469
$\frac{-\pi}{4}$	$mean_{h_b}$	1.486	1.444	1.399	1.403	1.218	1.207	1.170	1.181
	std_{h_b}	0.3165	0.3148	0.2272	0.2387	0.2509	0.2496	0.1625	0.1666
0.0	$mean_{h_b}$	1.873	1.898	1.964	1.961	1.874	1.890	1.958	1.945
	std_{h_b}	0.1778	0.1752	0.1202	0.1281	0.1755	0.1710	0.1320	0.1517

always $\sigma_e^2 + \sigma_b^2 = 0.6$. We also varied sample size and the number of clusters q. We did the calculations first under the assumption of known variances, and then repeated the procedure under the assumption of unknown variances where these were estimated by the method of moments. The bandwidth grid was reduced to the range of [0.6; 2.0] with step size 0.1. The results are given for the three critical points $\mathbf{x}$ in Table 4. In can be seen that the outcome is hardly affected by the additional uncertainty of unknown variances.

A.3. Appendix C: prediction of random effects

The nonparametric estimation of fixed effects and prediction of random effects is of particular interest for prediction and small area issues. Again it can be realized for example by making use of a likelihood. Actually, [31] and later [24] proposed such joint estimation in a longitudinal data model. However, though they started out from the same likelihood, they ended up with different predictors. We checked the expressions of [31] and will follow here their lines. The generalization to our more general context can be made as follows.

Under a local modeling approach, the model (1.1) is approximated by a linear mixed effects model such as (2.2) in each neighborhood. Looking at the local linear model, there is no need to restrict to the semiparametric model; we just consider $\mathbf{b}_i$ as a local linear version of υ_i. Assuming that the random effects $\mathbf{b}_i$, and the residuals ε_{ij} are normally distributed, the estimators and predictors for $\boldsymbol{\beta}$ and $\mathbf{b}_i$ arise from the maximization of the local kernel-weighted joint log-likelihood of $(\mathbf{y}_i, \mathbf{b}_i)$, given by

$$\mathcal{L}_J(\boldsymbol{\beta}, \mathbf{b}; \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{q}\left\{[\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i]^t\, \mathbf{W}_{ih}^{1/2}\boldsymbol{\Sigma}_i^{-1}\mathbf{W}_{ih}^{1/2}\,[\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i] + \mathbf{b}_i^t\mathbf{B}_i^{-1}\mathbf{b}_i + C_{iJ}\right\}$$

with $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_q)^t$ and a constant $C_{iJ} = \log|\boldsymbol{\Sigma}_i| + \log|\mathbf{B}_i| + 2n_i\log(2\pi)$ $(i = 1, \ldots, q)$. Note that this matches with the likelihood standard linear mixed effects model estimation but considering the transformed model

$$\widetilde{\mathbf{y}} = \mathbf{W}_h^{1/2}\mathbf{y} = \mathbf{W}_h^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}_h^{1/2}\mathbf{Z}\mathbf{b} + \mathbf{W}_h^{1/2}\boldsymbol{\varepsilon} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\mathbf{Z}}\mathbf{b} + \widetilde{\boldsymbol{\varepsilon}}. \tag{A.7}$$

Under the assumption of known $\mathbf{B}_i$ and $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, q)$, the resultant estimator is given by

$$\widehat{\boldsymbol{\beta}}_J = \left(\sum_{i=1}^{q}\mathbf{X}_i^t\boldsymbol{\Omega}_{ih}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{q}\mathbf{X}_i^t\boldsymbol{\Omega}_{ih}\,\mathbf{y}_i$$

where $\boldsymbol{\Omega}_{ih} = \mathbf{W}_{ih}^{1/2}\widetilde{\mathbf{V}}_i^{-1}\mathbf{W}_{ih}^{1/2}$, $\widetilde{\mathbf{V}}_i = \mathbf{W}_{ih}^{1/2}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^t\mathbf{W}_{ih}^{1/2} + \boldsymbol{\Sigma}_i$. Therefore the (joint) local linear estimator of the fixed-effect function is obtained by

$$\widehat{m}_J(t) = \mathbf{e}_1^t\left(\sum_{i=1}^{q}\mathbf{X}_i^t\boldsymbol{\Omega}_{ih}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{q}\mathbf{X}_i^t\boldsymbol{\Omega}_{ih}\,\mathbf{y}_i \tag{A.8}$$

with $\mathbf{e}_1^t = (1, 0)$. For the prediction of local linear random effects we get

$$\widehat{\mathbf{b}}_i = (\mathbf{Z}_i^t\mathbf{W}_{ih}^{1/2}\boldsymbol{\Sigma}_i^{-1}\mathbf{W}_{ih}^{1/2}\mathbf{Z}_i + \mathbf{B}_i^{-1})^{-1}\mathbf{Z}_i^t\mathbf{W}_{ih}^{1/2}\boldsymbol{\Sigma}_i^{-1}\mathbf{W}_{ih}^{1/2}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_J) \tag{A.9}$$

for $i = 1, \ldots, q$, which is equivalent to

$$\widehat{\mathbf{b}}_i = \mathbf{B}_i\mathbf{Z}_i^t\mathbf{W}_{ih}^{1/2}\widetilde{\mathbf{V}}_i^{-1}\mathbf{W}_{ih}^{1/2}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_J), \tag{A.10}$$

where $\widetilde{\mathbf{V}}_i = \mathbf{W}_{ih}^{1/2}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^t\mathbf{W}_{ih}^{1/2} + \boldsymbol{\Sigma}_i$, for $i = 1, \ldots, q$. The predicted random-effects function is then given by

$$\widehat{\upsilon}_i(\mathbf{z}) = \mathbf{e}_1^t\widehat{\mathbf{b}}_i. \tag{A.11}$$

It can be shown that the asymptotic properties of the marginal estimator (2.14) and the joint estimator (A.8) are quite similar. The asymptotic properties of the random effect component is still not worked out. Finally, the individual curves can be estimated by

$$\widehat{\eta}_i(\mathbf{x}, \mathbf{z}) = \widehat{m}_J(\mathbf{x}) + \widehat{v}_i(\mathbf{z}), \tag{A.12}$$

for each group $i = 1, \ldots, q$, and the mean parameter by

$$\widehat{\Theta}_i = \widehat{\dot{m}}_J(\mathbf{x}_i) + \widehat{\dot{v}}_i(\mathbf{z}_i). \tag{A.13}$$

References

[1] R. Cao-Abad, W. González-Manteiga, Bootstrap methods in regression smoothing, Nonparametric Statistics 2 (1993) 379–388.
[2] K. Chen, Z. Jin, Local polynomial regression analysis of clustered data, Biometrika 92 (2005) 59–74.
[3] G. Claeskens, J.D. Hart, Goodness-of-fit tests in mixed models, Test 18 (2009) 265–270.
[4] M. Davidian, D.M. Gilinan, Nonlinear Models for Repeated Measurement Data, Chapman and Hall, London, 1995.
[5] P.J. Diggle, P. Heagerty, K.-Y. Liang, S.L. Zeger, Analysis of Longitudinal Data, second ed., Oxford University Press, Oxford, 2002.
[6] C. Elbers, J.O. Lanjouw, P. Lanjouw, Micro-level estimation of poverty and inequality, Econometrica 71 (1) (2003) 355–364.
[7] J. Fan, I. Gijbels, Variable bandwidth and local regression smoothers, Annals of Statistics 20 (1992) 2008–2036.
[8] W. González-Manteiga, M.D. Martínez-Miranda, A. Pérez-González, The choice of smoothing parameter in nonparametric regression through wild bootstrap, Computational Statistics and Data Analysis 47 (2004) 487–515.
[9] W. González-Manteiga, M.J. Lombardía, I. Molina, D. Morales, L. Santamaría, Small area estimation under Fay–Herriot models with nonparametric estimation of heteroscedasticity, Statistical Modelling 10 (2010) 197–214.
[10] C. Gu, P. Ma, Generalized nonparametric mixed-effect models: computation and smoothing parameter selection, Journal of Computational and Graphical Statistics 14 (2005) 485–504.
[11] C. Gu, P. Ma, Optimal smoothing in nonparametric mixed-effect models, Annals of Statistics 33 (2005) 1357–1379.
[12] P. Hall, T. Maiti, Nonparametric estimation of mean-squared prediction error in nested-error regression models, Annals of Statistics 34 (2008) 1133–1750.
[13] W. Härdle, J.S. Marron, Bootstrap simultaneous bars for nonparametric regression, Annals of Statistics 19 (1991) 778–796.
[14] J. Jiang, Asymptotic properties of the empirical BLUP and BLUE in mixed linear models, Statistica Sinica 8 (1998) 861–863.
[15] J. Jiang, P. Lahiri, Mixed model prediction and small area estimation, Test 15 (2006) 1–96.
[16] M. Köhler, S. Schindler, S. Sperlich, A review and comparison of bandwidth selection methods for kernel regression, CRC-PEG Discussion Papers Nr. 95, 2011.
[17] X. Lin, R.J. Carroll, Nonparametric function estimation for clustered data when predictor is measured without/with error, Journal of the American Statistical Association 95 (2000) 520–534.
[18] X. Lin, R.J. Carroll, Semiparametric estimation in general repeated measures problems, Journal of the Royal Statistical Society B 68 (2006) 69–88.
[19] O. Linton, E. Mammen, X. Lin, R.J. Carroll, Accounting for correlation in marginal longitudinal nonparametric regression, in: Lin, Heagerty (Eds.), Proceedings of the Second Seattle Symposium in Biostatistics, Springer, 2003.
[20] X. Lin, D. Zhang, Inference in generalized additive mixed models by using smoothing splines, Journal of the Royal Statistical Society B 61 (1990) 381–400.
[21] M.J. Lombardía, S. Sperlich, Semiparametric inference in generalized mixed effects models, Journal of the Royal Statistical Society B 70 (2008) 913–930.
[22] M.D. Martínez-Miranda, R. Raya-Miranda, W. González-Manteiga, A. González-Carmona, A bootstrap local bandwidth selector for additive models, Journal of Computational and Graphical Statistics 17 (2008) 38–55.
[23] J. Opsomer, G. Claeskens, M.G. Ranalli, G. Kauermann, F.J. Breidt, Nonparametric small area estimation using penalized spline regression, Journal of the Royal Statistical Society B 70 (2008) 265–286.
[24] J.G. Park, H. Wu, Backfitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data, Journal of Statistical Planning and Inference 136 (2006) 3760–3782.
[25] D. Ruppert, Empirical-bias bandwidths for local polinomial nonparametric regression and density estimation, Journal of the American Statistical Association 92 (1997) 1049–1062.
[26] S. Sperlich, M.J. Lombardía, Local polynomial inference for small area statistics: estimation, validation and prediction, Journal of Nonparametric Statistics 22 (2011) 633–648.
[27] G. Verbeke, G. Molenberghs, Linear Mixed Models for Longitudinal Data, Springer-Verlag, New York, Inc., 2000.
[28] J.M. Vilar, M. Francisco, Local polynomial regression smoothers with AR-error structure, Test 11 (2002) 439–464.
[29] N. Wang, Marginal nonparametric kernel regression accounting for within-subject correlation, Biometrika 90 (2003) 43–52.
[30] H. Wu, J.T. Zhang, Nonparametric regression methods for longitudinal data analysis, in: Wiley Series in Probability and Statistics, USA, 2006.
[31] H. Wu, J.T. Zhang, Local polynomial mixed-effects models for longitudinal data, Journal of the American Statistical Association 97 (2002) 883–897.
[32] W. Xu, L. Zhu, Kernel-based generalized cross-validation in non-parametric mixed-effect models, Scandinavian Journal of Statistics 36 (2009) 229–247.
[33] D. Zhang, X. Lin, J. Raz, M. Sower, Semiparametric stochastic mixed models for longitudinal data, Journal of the American Statistical Association 93 (1998) 710–719.