Semiparametric Inference in Generalized Mixed Effects Models

María José Lombardía
Departamento de Estadística e Investigación Operativa
Universidad de Santiago de Compostela
15782 Santiago de Compostela, SPAIN

Stefan Sperlich
Departamento de Economía, c/Madrid 126-128
Universidad Carlos III de Madrid
28903 Getafe, SPAIN
tel: +34 916249861, fax: +34 916249875
stefan@eco.uc3m.es

Abstract

We consider a so called generalized partially linear model including random effects in the linear part. For these kind of models we first propose an estimator combining likelihood approaches for mixed effects models with kernel methods. Next we introduce different tests that allow to choose between a parametric and the semiparametric mixed effects model, following the methodology of Härdle, Mammen and Müller (1998). To this end we also discuss some bootstrap procedures to simulate the critical values. Various alternatives and extensions to other semiparametric models are discussed. We prove consistency and give asymptotic theory for all our methods. Finally, a simulation study is provided in order to see the performance of our methods, in particular the tests.[1]

Keywords and Phrases: semiparametric inference, mixed effects models, bootstrap, generalized partial linear models, small area statistics.

[1]The authors gratefully acknowledge the financial supported of the Spanish "Dirección General de Investigación del Ministerio de Ciencia y Tecnología", SEJ2004-04583/ECON, BFM2002-03213 and of the Xunta de Galicia PGIDIT03PXIC20702PN. We further thank Jean Opsomer for helpful discussion.

1 Introduction

In the last decade linear random effect models have attracted an increasing attention as an effective tool for either reducing the dimensionality of a high-dimensional regression problem or to increase the efficiency of statistical inference, when the not explained heterogeneity can be partly classified (therefore one often speaks also of cluster-specific intercepts). They are an extension of linear regression models that allow for the incorporation of random effects. In this formulation, the probability distribution for the multiple measurements has the same form for each individual, but the parameters of that distribution vary over individuals. The simplest model includes a single random component or intercept that varies between clusters of observations and induces dependence within these clusters. Random effects models are also useful for modeling panel data or grouped cross-sectional data where the responses for the same person or group cannot be assumed to be independent after conditioning on exogenous variables. In the grouped cross-sectional case the clusters could be households, schools, hospitals, firms, geographical entities (here is included all small-area literature, see below). Also, we can find applications of linear random effect models in the analysis of longitudinal data sets, see e.g. Laird and Ware (1982).

Similarly to the usual linear models, these linear random or also called mixed effects models have then been extended to generalized mixed effect models. They are commonly defined by

$$G\left\{E\left[Y_{dj}|u_d, \boldsymbol{X}_{dj}\right]\right\} \;=\; \boldsymbol{X}_{dj}^t\boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D;\; j = 1, \ldots, n_d,$$

with $G(\cdot)$ being a known link function, Y_{dj} the dependent variable, $\boldsymbol{X}_{dj}$ some observable regressors, and u_d unobservable cluster-specific effects for which we can only estimate a reduced number of parameters. In practice, they are treated like random effects and only its variance will be estimated to improve inference on $\boldsymbol{\beta}$. Due to its above mentioned effectiveness this kind of model today is broadly applied in different fields of statistical analysis. Overviews of this vast topic are provided by Searle, Casella, and McCulloch (1982), Vonesh and Chinchilli (1997), Pinheiro and Bates (2000), Verbeke and Molenberghs (2000), McCulloch and Searle (2001). Further examples and explanations, but in particular different ways of the nontrivial problem of implementation are provided, among others, by Fahrmeir and Tutz (2001), Diggle, Tawn and Moyeed (2002), and most recently by Skrondal and Rabe-Hesketh (2005). We also want to mention the extension to nonlinear parametric models with random effects. They are always straight forward as long as we assume to know the distribution of the random parts and stick to conditional maximum likelihood methods, see e.g. Kuhn and Lavielle (2005).

A particular research area that is of strong public interest but almost unthinkable without the techniques of mixed models is the statistical analysis in small areas. “Small area” may refer to a small geographical region (states, provinces, school districts, health

service areas) or a particular group obtained by a cross-classification of various demographic factors such as age, gender, race, etc. Small area statistics are needed in regional planning and fund allocation in many government programs, so e.g. EUROSTAT is demanding since 2003 from the EU states providing statistics for its so called small areas (provinces, districts, departments, etc.), examples of major small area estimation programs in the United States include Census Bureau's Small Area Income and Poverty Estimates program, the Bureau of Labor Statistics' Local Area Unemployment Statistics program, and the National Agricultural Statistics Service's County Estimates Program. See Ghosh and Rao (1994) and Rao (2003) for a thorough review of different small area estimation techniques. In the exclusively model-based framework there is some interesting research done using Bayes methodology (see e.g. Malec, Sedransk, Moriarity, and LeClere 1997; Ghosh, Natarajan, Stroud, and Carlin, 1998; Butar and Lahiri, 2003) or using frequentist methodology (Prasad and Rao, 1990; Jiang and Lahiri, 2001; Jiang, 2003; Das, Sedransk, Moriarity, and LeClere, 2004; González-Manteiga, Lombardía, Molina, Morales, and Santamaría, 2005).

Still quite recently, mixed effect models have entered the world of non- and semiparametric statistics. A first, rather appealing step was to consider the smoothing parameters of spline or sieve estimators as random effects; further extensions followed immediately, see Ruppert, Wand and Carroll (2003) or Wand (2003) who takes a general look on smoothing in mixed models. So far this research concentrates mainly on the challenging development of feasible algorithms for non- and semiparametric mixed models using spline methods. Kneip, Sickles and Song (2005) provide a series estimator and its asymptotic theory for a partial linear model with time varying individual effects, what is a particular case but falls clearly in the class of semiparametric mixed effects models. However, in the most cases of the so far existing literature, asymptotic theory is missing. The same holds for theory based suggestions for model specification tests in (generalized) mixed models. In particular, at least to our knowledge, mixed models have so far not been combined with kernel smoothing methods. Although often, spline methods are preferred due to its easy handling and implementation in the one dimensional or additively one dimensional case, a major part of the existing asymptotic theory for non- and semiparametric statistics is based on kernel smoothing methods. Notice that the notation "semiparametric mixed effects models" is also used in the literature where only the assumption of having normal distribution of the error terms and / or of the random effects is relaxed.

This article is aimed to show how the combination of kernel based methods and mixed effects models can open a huge variety of statistical methods for the analysis of high dimensional data and enrich the inference e.g. in small areas. We consider a so called generalized partially linear model (see e.g. Severini and Staniswalis, 1994) but including now random effects. This results in a model of the form

$$G\left(E[Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}]\right) = m(\boldsymbol{T}_{dj}) + \boldsymbol{X}_{dj}^t\boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D;\ j = 1, \ldots, n_d, \qquad (1)$$

where D is the number of random factors and $n = \sum_{d=1}^{D} n_d$ is the sample size. As above, for unit j of factor d, let $Y_{dj} \in I\!R$ be the dependent variable, $\boldsymbol{X}_{dj} \in I\!R^p$

and $\boldsymbol{T}_{dj} \in I\!R^q$ being the covariates. We restrict the random effects u_d to be variables with zero mean and constant variance σ_u^2, assuming that $u_1, \ldots, u_D$ are independent. Again, the function $G(\cdot)$ is a (known) link function and $g(\cdot)$ its unique inverse function. An extension to the case where $G(\cdot)$, respectively $g(\cdot)$ depend on unknown parameters will be discussed. Finally, $m : I\!R^q \to I\!R$ is a nonparametric smooth function that can also be modelled additively or multiplicatively. Let us denote by

$$\eta_{dj} \;=\; G\left(E[Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}]\right)$$

the partial linear predictor and by

$$\mu_{dj} \;=\; E[Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}] \;=\; g\left(m(\boldsymbol{T}_{dj}) + \boldsymbol{X}_{dj}^t\boldsymbol{\beta} + u_d\right)$$

the conditional expectation. Under this setup, let $\boldsymbol{Y}$ be the $n \times 1$ vector with elements Y_{dj}, $\boldsymbol{X}$ be the $p \times n$ matrix with rows $\boldsymbol{X}_{dj}$ and $\boldsymbol{u} = (u_1, \ldots, u_D)^t$ be the $D \times 1$ vector of random effects. Let us define the $D \times n$ matrix $\boldsymbol{Z} = \mathrm{diag}\{\boldsymbol{1}_{n_d}^t,\ d = 1, \ldots, D\}$, where $\boldsymbol{1}_a$ denotes a column vector of ones with size a. Let $\boldsymbol{\mu} = E[\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{T}, \boldsymbol{X}]$ be the conditional mean vector with elements μ_{dj}, $\boldsymbol{\Sigma} = Var[\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{T}, \boldsymbol{X}]$ the conditional covariance matrix, which is diagonal with elements σ_{dj}, and $V = Var[\boldsymbol{Y}|\boldsymbol{T}, \boldsymbol{X}]$. In abuse of notation, we denote the vectors and matrices in bold letters. Then, in matrix notation, the linear predictor is $\boldsymbol{\eta} = m(\boldsymbol{T}) + \boldsymbol{X}^t\boldsymbol{\beta} + \boldsymbol{Z}^t\boldsymbol{u}$.

Later, we will discuss alternatives and extensions, i.e. easily available generalizations, of model (1). The estimation of such a model without a nonparametric part is well studied and its literature on it has been discussed already above. When $m(\cdot)$ is included but not u_d, then the estimation and testing is well studied, too. Severini and Wong (1992) and Severini and Staniswalis (1994) studied intensively the asymptotic theory for kernel based quasi and profiled likelihood estimators of these kinds of models. Hastie and Tibshirani (1990) provided useful algorithms, and Müller (2001) a comparative study and survey of existing estimation methods in those models.

For the parametric part, inference can be derived directly from the asymptotic theory (if provided). For the nonparametric part $m(\cdot)$ statistical inference is much more sophisticated in theory and unfortunately also in practice. Therefore, a first step should be to check whether such an effort is justified. This means to test $m(\cdot)$ for significant nonlinearity. An extension to significant deviation of $m(\cdot)$ from a fixed polynomial structure is obvious.

In order to do so, we propose a test of the parametric hypothesis

$$H_0: \quad m(\boldsymbol{T}) = c + \boldsymbol{T}^t\boldsymbol{\gamma} \quad \text{vs} \quad H_1: \quad m(\boldsymbol{T}) \neq c + \boldsymbol{T}^t\boldsymbol{\gamma} \tag{2}$$

for any $\boldsymbol{\gamma}$ and c, i.e. a generalized linear mixed effects model versus the semiparametric alternative (1). For the case of having no random effects, such a test has been introduced by Härdle, Mammen and Müller (1998). See this paper also for further references. It turns out that their theory carries over to our mixed effects model. This

holds also true for the (nonparametric) bootstrap we will use to obtain reasonable critical values for the test statistic. Finally, extensions to related bootstrap tests in these kind of models, like proposed most recently by Härdle, Huet, Mammen, and Sperlich (2004) or Rodríguez-Póo, Sperlich, and Vieu (2005) are obvious then, too.

Our test is also of particular interest in small areas statistics, where u_d is the random factor referring to the small area $d = 1, \cdots, D$. Inferences from model-based estimators refer to the distribution implied by the assumed model. Therefore, model selection and validation play a vital role in this type of estimation. If the assumed models do not provide a good fit to the data, the model-based estimators will be model biased which can lead to (sometimes completely) erroneous inferences. However, the hypothesis testing in the general mixed model framework for the small areas inference has been scarcely investigated. So e.g. Jiang and Lahiri (2001) study a generalization of the Pearson's χ^2 goodness-of-fit test, which is applied to a real data example with geographical small areas; Zhu and Fung (2004) investigate the test for heteroscedasticity under the framework of a semiparametric mixed model, which is illustrated with the analysis of a longitudinal study.

The rest of the paper is organized as follows. In Section 2 we introduce the estimators for the semiparametric model (1), i.e. the parametric counterpart of the null hypothesis H_0, together with its asymptotic properties. In Section 3 we first introduce an estimator of the parametric model that is convoluted with a kernel and will be used in the test statistic to account for the bias the semiparametric alternative suffers from. Then, some test statistics and a bootstrap procedure will be introduced and discussed, again together with its asymptotic behaviors. Several alternative tests and bootstrap procedures, mainly modifications, as well as extensions to other models and test problems are discussed in Section 4. A simulation study in Section 5 shows the excellent performance of the test even for moderate sample sizes. The lists of assumptions are deferred to Section 6.

2 Estimating the Semiparametric Model

The aim is to study the relationship between a dependent variable $Y \in I\!R$ and a set of explanatory variables $(\boldsymbol{T}, \boldsymbol{X})$, $\boldsymbol{T} \in I\!R^q$ and $\boldsymbol{X} \in I\!R^p$, taking into account the influence of a random factor u_d that we suppose to have $N(0, \sigma_u^2)$ distribution $(d = 1, \cdots, D)$. In this work we stick to the particular case $\sigma_{dj}^2 := \sigma_e^2$ for all $d = 1, \ldots, D,\ j = 1, \ldots, n_d$; being $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)$ the vector of variance components. Suppose we have a sample of $n = \sum_{d=1}^{D} n_d$ replicates $\{(Y_{dj}, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj})\}_{d=1,\cdots,D;j=1,\cdots,n_d}$ and the conditional distribution of Y given the random effects u but also $\boldsymbol{T}$ and $\boldsymbol{X}$, belongs to the family with density $\{f(Y|u, \boldsymbol{T}, \boldsymbol{X}; m, \boldsymbol{\delta}) : \boldsymbol{\delta} \in \Delta, m(\boldsymbol{T}) \in \mathcal{M}\}$, where $m(\cdot)$ is an unknown smooth function that takes values in $\mathcal{M} \in I\!R$ and $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\theta}) \in \Delta$, both compact; $\boldsymbol{X}_{dj} \in \mathcal{X}$ and $\boldsymbol{T}_{dj} \in \mathcal{T}$ are also assumed to be from compact sets $\mathcal{X} \subset I\!R^p$ and $\mathcal{T} \subset I\!R^q$. Further, $p(u; \sigma_u^2)$ denotes the density of the random effects.

Call $f_y(\cdot)$ the density of Y conditioned on $\boldsymbol{T}, \boldsymbol{X}$ only. Then, one has

$$f_y(Y|\boldsymbol{T}, \boldsymbol{X}; m, \boldsymbol{\delta}) = \int f(Y_{dj}|u, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}; m, \boldsymbol{\delta}) p(u|\sigma_u^2) du, \tag{3}$$

so that taking the logarithm would provide us with a possible likelihood function, see e.g. McCulloch, Searle (2001) how this works out for the fully parametric case. By $f(\cdot|\cdot)$ and $p(\cdot|\cdot)$ we denote the corresponding conditioned density functions relative to $f(\cdot)$ and $p(\cdot)$.

Alternatively, if we wish to predict simultaneously β and the random effects $\boldsymbol{u} = (u_1, \ldots, u_d)^t$ we could look e.g. at the posterior density

$$f_\delta(\beta, \boldsymbol{u}|Y, \boldsymbol{T}, \boldsymbol{X}; m, \sigma_u^2) = \frac{\prod_{d=1}^{D} \prod_{j=1}^{n_d} f(Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}; m, \boldsymbol{\delta}) \prod_{d=1}^{D} p(u_d|\sigma_u^2)}{\int\int \prod_{d=1}^{D} \prod_{j=1}^{n_d} f(Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}; m, \boldsymbol{\delta}) p(u_d|\sigma_u^2) du d\beta} \tag{4}$$

so that the part to maximize is proportional to the numerator, see Fahrmeir, Tutz (2001) for details. Breslow and Clayton (1993) derived a penalized quasi likelihood (PQL) that is based on this criterium and combine it with the idea of profiled likelihood to get simultaneously estimates for the variance components. González-Manteiga, Lombardía, Molina, Morales, and Santamaría (2005) start also with the PQL but estimate the variance components from a linearized version of the generalized linear model, going back to an idea of Schall (1991). For simplification let us denote by $l(\boldsymbol{Y}; m, \boldsymbol{\delta})$ the log density, whatever the density function under consideration is. E.g., following the idea of maximizing the posterior mode, one would consider

$$\varphi_1(\boldsymbol{Y}; m, \boldsymbol{\delta}) = \sum_{d=1}^{D} \sum_{j=1}^{n_d} \log f(Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}; m, \boldsymbol{\delta}) \ , \tag{5}$$

$$\varphi_2(\boldsymbol{u}; \sigma_u^2) = \sum_{d=1}^{D} \log p(u_d; \sigma_u^2) \tag{6}$$

and

$$\varphi(\boldsymbol{Y}, \boldsymbol{u}; m, \boldsymbol{\delta}) = \varphi_1(\boldsymbol{Y}; m, \boldsymbol{\delta}) + \varphi_2(\boldsymbol{u}; \sigma_u^2) \ . \tag{7}$$

For getting an estimator of the nonparametric one first has to fix a point $\boldsymbol{t}_0$ on which we aim to estimate $m(\cdot)$ and then takes the empirical counterpart of

$$E\left[\log f(Y|u, \boldsymbol{T}, \boldsymbol{X}; m, \boldsymbol{\delta}) + \log p(u; \sigma_u^2)|\boldsymbol{T} = \boldsymbol{t}_0\right]$$

what can be written in terms of

$$\varphi_s(\boldsymbol{Y}; m, \boldsymbol{\delta}) = \sum_{d=1}^{D} \sum_{j=1}^{n_d} K_{\boldsymbol{h}}(\boldsymbol{t}_0 - \boldsymbol{T}_{dj}) \log f(Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}; m(\boldsymbol{t}_0), \boldsymbol{\delta}) + \varphi_2(\boldsymbol{u}; \sigma_u^2) \tag{8}$$

where $K_{\boldsymbol{h}}(\cdot)$ is a q-dimensional product kernel function, $\boldsymbol{h} = (h_1, \ldots, h_q)$ the corresponding bandwidth vector and $\boldsymbol{t}_0$ the fixed value. This is also called the smoothed

likelihood function and is only used to incorporate the nonparametric part. Note that the conditioning on T has no impact on φ_2 ; that is why we do not convolute that part with the kernel function. Alternatively, we define also the simplified smoothed likelihood

$$\varphi_{ss}(\boldsymbol{Y}; m, \boldsymbol{\delta}) \;=\; \sum_{d=1}^{D} \sum_{j=1}^{n_d} K_{\boldsymbol{h}}\left(\boldsymbol{t}_0 - \boldsymbol{T}_{dj}\right) \log f(Y_{dj} | u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}; m(\boldsymbol{t}_0), \boldsymbol{\delta}) \;, \qquad (9)$$

skipping $\varphi_2(u_d; \sigma_u^2)$ from φ_s. We will see later why this might be useful. Analogously one defines likelihoods (5) to (9) based on (3). We have recalled here the version of objective functions based on the Breslow, Clayton (1993) approach only because it is, to our knowledge, the most popular one in practice.

We now have to combine the existing fully parametric likelihood approaches for random effect models with the semiparametric regression problem. In the following, we denote the considered log likelihood by $\varphi(\boldsymbol{Y}; m, \boldsymbol{\delta})$. For this we will use the well known method of profiled likelihood to estimate the parameter $\boldsymbol{\delta}$. This is, let $\lambda_{\boldsymbol{\delta}}$ denote a least favorable curve in $\mathcal{M}$ to take into account the nuisance parameter $m(\cdot)$ when estimating $\boldsymbol{\delta}$, and let $\hat{\lambda}_{\boldsymbol{\delta}}$ an estimator of $\lambda_{\boldsymbol{\delta}}$. If $\hat{\lambda}_{\boldsymbol{\delta}}$ is a *valid estimator* for our least favorable curve, then the $\hat{\boldsymbol{\delta}}$ that maximizes $\varphi(\boldsymbol{Y}, u; \hat{\lambda}_{\boldsymbol{\delta}}, \boldsymbol{\delta})$ is asymptotically efficient. To be a valid estimator for $\lambda_{\boldsymbol{\delta}}$, Severini and Wong (1992) [in the following SW92] have given sufficient conditions, i.e. their so called *Conditions NP*, p.1779 - 1780. In the same paper they give also conditions in a rather general context that guarantee that maximizing the log likelihood convoluted with a kernel, i.e. the empirical version of $E[\varphi(\boldsymbol{Y}; m, \boldsymbol{\delta}) | T = t_0]$ is such a valid estimator of the least favorable curve. Let us denote this convoluted or smoothed likelihood by $\varphi_s(\boldsymbol{Y}; m, \boldsymbol{\delta})$, compare with equation (8). Rodríguez-Póo, Sperlich, Vieu (2003) finally proved this for the particular case we need here (but they do it in a different context[2]).

Summarizing, we propose
Procedure A.

1. For a value t_0 and fixed $\boldsymbol{\delta}$ we estimate $m(\boldsymbol{t}_0)$ as the solution of the problem

$$\check{m}_{\boldsymbol{\delta}} \;= \underset{\{m \in \mathcal{M}\}}{argmax}\; \varphi_s(\boldsymbol{Y}; m, \boldsymbol{\delta}).$$

2. We estimate $\boldsymbol{\delta}$ (together with predictors $\hat{u}$ or not) as the solution of the problem

$$\hat{\boldsymbol{\delta}} \;= \underset{\boldsymbol{\delta}}{argmax}\; \varphi(\boldsymbol{Y}, \boldsymbol{u}; \check{m}_{\boldsymbol{\delta}}, \boldsymbol{\delta}), \qquad (10)$$

[2]They consider the problem of estimating arbitrarily separable function with possibly limited dependent variables but do not discuss the inclusion of random effects.

3. With the estimators obtained in steps 1 and 2, we set finally $\hat{m} = \check{m}_{\hat{\boldsymbol{\delta}}}$.

This procedure is suitable for the case D tending to ∞ with rate $O(n)$ as it is typically assumed e.g. in small area statistics. Alternatives that are interesting in practice, in particular for D relatively small, are discussed in Section 4. Clearly, in Procedure A as well as in the alternative procedures, the first steps are performed in order to get estimators for the least favorable curve. Suppose the first step provides such a valid estimator $\check{m}_{\boldsymbol{\delta}}$ of the least favorable curve in the sense of SW92 (Conditions NP). Then, a direct consequence of Propositions 1 and 2 of SW92 is:

Corollary 1. *Assume that assumptions [A.1] to [A.3] from the Appendix hold. Let $\hat{\boldsymbol{\delta}}$ be the log likelihood estimate as given in step 2 of Procedures A. Then, as $n = \sum_{d=1}^{D} n_d$ tends to infinity*

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \xrightarrow{d} N\left(0, I_{\boldsymbol{\delta}}^{-1}\right),$$

where $I_{\boldsymbol{\delta}}$ is the (marginal) Fisher information matrix, i.e.

$$\begin{aligned} I_{\boldsymbol{\delta}} &= E_T\left\{ E_{X,u,T}\left[\frac{\partial}{\partial \boldsymbol{\delta}} l\left(Y; m, \boldsymbol{\delta}\right) \frac{\partial}{\partial \boldsymbol{\delta}^t}\, l\left(Y; m, \boldsymbol{\delta}\right)\right] - E_{X,u}\left[\frac{\partial}{\partial \boldsymbol{\delta}} l\left(Y; m, \boldsymbol{\delta}\right) \frac{\partial}{\partial m} l\left(Y; m, \boldsymbol{\delta}\right) | T\right]\right. \\ &\times \left. E_{X,u}\left[\left(\frac{\partial}{\partial m} l(Y; m, \boldsymbol{\delta})\right)^2 | T\right]^{-1} E_{X,u}\left[\frac{\partial}{\partial m} l\left(Y; m, \boldsymbol{\delta}\right) \frac{\partial}{\partial \boldsymbol{\delta}^t} l\left(Y; m, \boldsymbol{\delta}\right) | T\right]\right\}, \end{aligned} \tag{11}$$

where $E_A[\cdot]$ is the expectation with respect to the variable A. Further, $l\left(Y; m, \boldsymbol{\delta}\right) = \log f(Y|u, \boldsymbol{T}, \boldsymbol{X}; m, \boldsymbol{\delta}) + \log p(u; \sigma_u^2)$ and

$$\frac{\partial}{\partial \boldsymbol{\delta}} l\left(Y; m, \boldsymbol{\delta}\right) = \left(\frac{\partial}{\partial \beta_1} l\left(Y; m, \boldsymbol{\delta}\right), \cdots, \frac{\partial}{\partial \beta_p} l\left(Y; m, \boldsymbol{\delta}\right), \frac{\partial}{\partial \sigma_u^2} l\left(Y; m, \boldsymbol{\delta}\right), \frac{\partial}{\partial \sigma_e^2} l\left(Y; m, \boldsymbol{\delta}\right)\right)^t.$$

As can be observed from this result, the semiparametric estimator achieves the semiparametric efficiency bound (see Newey, 1990, 1994). The asymptotic variance could be approximated with the aid of the Hessian matrix that one obtains as a by-product from the maximum likelihood estimation. Note further that our model restrictions do not contain any information about a possible dependence structure between X and T.

Remark 1. *If we assume to have a link $g(\cdot)$ being the identity function, i.e.*

$$Y_{dj} = m(\boldsymbol{T}_{dj}) + \boldsymbol{X}_{dj}^t \boldsymbol{\beta} + u_d + \epsilon_{dj}, \quad d = 1, \ldots, D,\ j = 1, \ldots, n_d;$$

with u_d defined as above and $\epsilon_{dj} \in N(0, \sigma_e^2)$, we would get for the variance of $\hat{\boldsymbol{\beta}}$, the Fisher information

$$\begin{aligned} I_\beta &= E_T\left\{X^t V^{-1} X - E[X|T]^t V^{-1} E[X|T]\right\} \\ &= E_T\left\{(X - E[X|T])^t V^{-1} (X - E[X|T])\right\} \end{aligned}$$

what e.g. equals the variance of Robinson (1988) in the simple partial linear model.

Certainly, our proposal (step 1 of Procedure A) is not the only valid one but is given for simplicity as it corresponds to the ones (each one in a different, particular context) of SW92, Severini, Staniswalis (1994), Rodríguez-Póo, Sperlich, Vieu (2003) and others.

Corollary 2. *Suppose assumptions [A.1] - [A.3], [B.1] - [B.3] and [N.1] - [N.2] from the Appendix hold. Then maximizing the smoothed likelihood as given in step 1 of Procedure A provides a valid estimator of the least favorable curve.*

In many practical examples this proposal yields the Nadaraya Watson smoother. Note however, that Fan, Heckman, Wand (1995) discuss in detail the extension to local polynomial smoother and its advantages. That paper and the above mentioned provide us also the asymptotic distribution of the final nonparametric estimate. Note that, due to the faster rate of convergence, the randomness of the parametric estimators does not affect (asymptotically) the nonparametric estimate. To this aim, define $h_{prod} = \prod_{j=1}^{q} h_j$ and $h_{max} = \max_{1\leq j\leq q} h_j$.

Corollary 3. *With the same conditions as in Corollary 2,* $\boldsymbol{t}_0$ *being from the interior of the support of* $\boldsymbol{T}$*,* $p(\cdot)$ *its density function, and* $n = \sum_{d=1}^{D} n_d$ *going to infinity, we have*

$$\sqrt{nh_{prod}}\,(\hat{m}(\boldsymbol{t}_0) - m(\boldsymbol{t}_0) - B_m(\boldsymbol{t}_0)) \xrightarrow{d} N(0, Var_m(\boldsymbol{t}_0))\,,$$

with

$$Var_m(\boldsymbol{t}_0) = \frac{\int K(\boldsymbol{t})^2 d\boldsymbol{t}}{p(\boldsymbol{t}_0) H'(m(\boldsymbol{t}_0), \boldsymbol{\delta}_0)}\,, \tag{12}$$

$$H'(m(\boldsymbol{t}_0), \boldsymbol{\delta}_0) = E\left[\frac{\partial}{\partial m} l(Y; m, \boldsymbol{\delta}_0)^2 \,|\boldsymbol{T} = \boldsymbol{t}_0\right] \tag{13}$$

and $B_m(\boldsymbol{t}_0) = O(h_{max}^2)$ *being the bias.*

If the conditional distribution of Y belongs to the exponential family, $H'(\cdot)$ simplifies to

$$H'(m(\boldsymbol{t}_0), \boldsymbol{\delta}_0) = E\left[g'(\boldsymbol{X}^t\boldsymbol{\beta}_0 + \boldsymbol{Z}^t\boldsymbol{u} + m(\boldsymbol{t}_0))^2\right] Var\left[Y|u, \boldsymbol{t}_0, \boldsymbol{X}\right]^{-1}$$

what gives an asymptotic variance of the form

$$\int K(\boldsymbol{t})^2 d\boldsymbol{t}\; p^{-1}(\boldsymbol{t}_0)\; Var\left[Y|u, \boldsymbol{t}_0, \boldsymbol{X}\right]\; g'\left(\boldsymbol{X}^t\boldsymbol{\beta}_0 + \boldsymbol{Z}^t\boldsymbol{u} + m(\boldsymbol{t}_0)\right)^{-2}\,.$$

As for the test statistic only the variance is of interest, we discuss here only the bias one gets when using in step 3 of Procedure A a local linear maximum likelihood, see Fan, Heckman, Wand (1995) for details. Then the bias of Corollary 3 is

$$B_m(\boldsymbol{t}_0) = \frac{1}{2}\boldsymbol{h}^t\; \mathrm{diag}\left(\frac{\partial^2}{\partial \boldsymbol{t}^2} m(\boldsymbol{t}_0)\right) \boldsymbol{h}\mu_2(K) + o_p(1)\,,$$

with $\operatorname{diag}\left(\frac{\partial^2}{\partial \boldsymbol{t}^2} m(\boldsymbol{t}_0)\right)$ being a diagonal matrix with the main diagonal of the Hessian matrix of $m(\cdot)$, and $\mu_2(K)$ is implicitly defined by $\int \boldsymbol{t}\boldsymbol{t}^t K(\boldsymbol{t})d\boldsymbol{t} = \mu_2(K)\boldsymbol{I}$ with $\boldsymbol{I}$ being the identity matrix. Unfortunately, as can be seen in the very same paper (p. 144), the bias of the local constant estimator is much more complex.

Finally, let us add two more remarks.

First, often we find in the literature of (generalized) linear mixed models the so called REML and / or Moment estimation methods to get unbiased estimators for the vector of variances θ. For the Moment Methods one needs to know the degrees of freedom one looses by estimating $\boldsymbol{\beta}$ and $m(\cdot)$. Unfortunately, for $m(\cdot)$ this information is not exactly available. The REML does not work here neither, because the idea is based on the possibility of applying a linear mapping L with $L\boldsymbol{X} = 0$ so that the variances can be estimated from the linearly transformed data without getting distorted by $\boldsymbol{X}$, respectively the estimation of $\boldsymbol{\beta}$. In other words, such a linear mapping applied onto the data corrects automatically for the degrees of freedom. However, in our case one would need a linear mapping such that also $m(\boldsymbol{T})$ vanishes. This, in general, is not available or would reduce the degrees of freedom to almost zero. This is why these alternatives are not feasible in this context.

Second, it might be worth to study more in detail the implementation of these estimation procedures. However, as can be seen e.g. in Fahrmeier, Tutz (2001) this is a topic on its own already in the fully parametric case. Often, it depends on the particular situation (model, number of random effect, etc.) what kind of implementation is the less costly and / or most efficient one. It would be for example interesting to study extensions of the EM-algorithm to our context, something that is clearly beyond the scope of this paper.

3 Testing the Parametric versus the Semiparametric Model

We now turn to the testing problem $H_0: \; m(\boldsymbol{t}) = c + \boldsymbol{t}^t\boldsymbol{\gamma}$ vs. $H_1: \; m(\boldsymbol{t}) \neq c + \boldsymbol{t}^t\boldsymbol{\gamma}$. Our test statistic is based on a direct comparison of the semiparametric estimate with the corresponding estimate in the parametric model. First of all note that for this purpose it is certainly enough to have

$$\sup_{\boldsymbol{t}_0 \in \mathcal{T}} |\hat{m}(\boldsymbol{t}_0) - m(\boldsymbol{t}_0)| = O_p\left(\sqrt{\frac{\log n}{n h_{prod}}}\right) \; .$$

It follows from Lemma 1 of Rodríguez-Póo, Sperlich, Vieu (2005) that this holds for our estimator introduced in Section 2.

Under the null hypothesis we have $\eta_{dj} = c + \boldsymbol{T}_{dj}^t\boldsymbol{\gamma} + \boldsymbol{X}_{dj}^t\boldsymbol{\beta} + u_d$, so the estimation problem is purely parametric. It certainly should be based on the same likelihood

functions as in the semiparametric case, i.e. be based on the same objective function. Set in the following $\boldsymbol{\gamma}_c^t = (c, \boldsymbol{\gamma}^t)$. The estimators for this model are denoted by

$$(\tilde{\boldsymbol{\gamma}}_c, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\delta}}) = \underset{\{\boldsymbol{\gamma}_c, \boldsymbol{u}, \boldsymbol{\delta}\}}{argmax} \ \varphi_p(\boldsymbol{Y}; \boldsymbol{\gamma}_c, \boldsymbol{\delta}),$$

where φ_p denotes the fully parametric log likelihood. Then, following the arguments of Härdle, Mammen, Müller (2002), a direct comparison of $\hat{m}(\boldsymbol{T})$ with $\tilde{c} + \boldsymbol{T}^t\tilde{\boldsymbol{\gamma}}$ may be misleading, because $\hat{m}(\cdot)$ has a smoothing bias which is typically non negligible. To avoid this effect, we add a bias to $\tilde{c} + \boldsymbol{T}^t\tilde{\boldsymbol{\gamma}}$ that will compensate for the bias of $\hat{m}(\boldsymbol{T})$:

Procedure B.

1. We build the artificial data set: $\{\tilde{Y}_{dj}, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}\}$ with

   $$\tilde{Y}_{dj} = g(\tilde{c} + \boldsymbol{T}_{dj}^t\tilde{\boldsymbol{\gamma}} + \boldsymbol{X}_{dj}^t\tilde{\boldsymbol{\beta}} + \tilde{u}_d),$$

   the parametric fit of $\mu_{dj} = E[Y_{dj}|u_d, \boldsymbol{T}_{dj}, \boldsymbol{X}_{dj}]$.

2. Repeat only the nonparametric step from Procedure A but replacing all parametric unknowns by their estimates $\tilde{\boldsymbol{\delta}}$ (and eventually $\tilde{\boldsymbol{u}}$). E.g., using the likelihood (8) or (9) one would set

   $$\begin{aligned} \widetilde{m}(\boldsymbol{t}_0) &= \underset{\{m\in\mathcal{M}\}}{argmax}\ \varphi_s(\tilde{\boldsymbol{Y}}; m, \tilde{\boldsymbol{\delta}}), \\ \text{or} \quad \widetilde{m}(\boldsymbol{t}_0) &= \underset{\{m\in\mathcal{M}\}}{argmax}\ \varphi_{ss}(\tilde{\boldsymbol{Y}}; m, \tilde{\boldsymbol{\delta}}) \qquad \text{respectively.} \end{aligned} \tag{14}$$

3. The resulting estimators we use for the direct comparison with its semiparametric analog are therefore $(\widetilde{m}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\delta}})$.

Then, under $H_0: \ m(\boldsymbol{t}) = c + \boldsymbol{t}^t\boldsymbol{\gamma}$, one will get $|\widetilde{m}(\boldsymbol{t}) - [\tilde{c} + \boldsymbol{t}^t\tilde{\boldsymbol{\gamma}} + B_m(\boldsymbol{t})]| = o_p(1)$, where $B_m(\boldsymbol{t})$ is the bias of $\hat{m}(\boldsymbol{t})$, and therefore $|\hat{m}(\boldsymbol{t}) - \widetilde{m}(\boldsymbol{t})| = o_p(1)$.

A most traditional testing approach would be based on the likelihood ratio. But this test does not work because $\hat{m}$ and $\hat{\boldsymbol{\delta}}$ were calculated with different likelihood functions (smoothed and unsmooothed functions), see Härdle, Mammen, Müller (1998). Therefore we consider weighted and unweighted squared differences.

$$R_{1w} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} \mathrm{H}\left(\hat{m}(\boldsymbol{t}_{dj}), \hat{\boldsymbol{\delta}}\right)\left[\hat{m}(\boldsymbol{t}_{dj}) - \widetilde{m}(\boldsymbol{t}_{dj}) + \boldsymbol{X}_{dj}^t(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\right]^2 \pi(\boldsymbol{t}_{dj}) \ , \tag{15}$$

or just

$$R_1 = \sum_{d=1}^{D}\sum_{j=1}^{n_d}\left[\hat{m}(\boldsymbol{t}_{dj}) - \widetilde{m}(\boldsymbol{t}_{dj}) + \boldsymbol{X}_{dj}^t(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\right]^2 \pi(\boldsymbol{t}_{dj}) \ , \tag{16}$$

with $\pi(\cdot)$ being a weight function chosen by the empirical researcher and

$$\mathrm{H}\left(m(\boldsymbol{t}_{dj}), \boldsymbol{\delta}\right) = \frac{\partial}{\partial m} l\left(Y_{dj}; m, \boldsymbol{\delta}\right)^2 .$$

Note that this test statistic comes close to the likelihood ratio comparing the deterministic parts of the indices weighted by something proportional to the variance of $\{\hat{m}(\boldsymbol{t}_{dj}) - \tilde{m}(\boldsymbol{t}_{dj})\}$. Further, the covariances between the $\{\hat{m}(\boldsymbol{t}_{dj}) - \tilde{m}(\boldsymbol{t}_{dj})\}$ are asymptotically negligible for different observations of $\boldsymbol{t}_{dj}$. Due to this fact, the asymptotic distribution of our test statistic can be concluded from Härdle, Mammen, Müller (1998) and summarized as follows:

Corollary 4. *Under the hypothesis* $m(\boldsymbol{t}) = c + \boldsymbol{t}^t\boldsymbol{\gamma}$ *, the previous assumptions and [A.4], it holds that*

$$v^{-1}(R_{1w} - b) \xrightarrow{d} N(0,1) , \tag{17}$$

with $b = h_{prod}^{-1} \int K(\boldsymbol{t})^2 d\boldsymbol{t} E[\pi(\boldsymbol{T})p^{-1}(\boldsymbol{T})] + o(1)$ *and* $v^2 = 2h_{prod}^{-1} \int K^{(2)}(\boldsymbol{t})^2 d\boldsymbol{t} E[\pi(\boldsymbol{T})^2 p^{-1}(\boldsymbol{T})]$. *Here,* $K^{(2)}$ *refers to a two-fold convolution of kernel* K*.*

If we skip the weighting with $H(\cdot)$ in the test statistic, bias and variance will become

$$b = h_{prod}^{-1} \int K(\boldsymbol{t})^2 d\boldsymbol{t} E\left[E[H^{-1}(m(\boldsymbol{T}), \boldsymbol{\delta}_0)\pi(\boldsymbol{T})p^{-1}(\boldsymbol{T})\right] + o(1)$$

and

$$v^2 = 2h_{prod}^{-1} \int K^{(2)}(\boldsymbol{t})^2 d\boldsymbol{t} E\left[H^{-2}(m(\boldsymbol{T}), \boldsymbol{\delta}_0)\pi(\boldsymbol{T})^2 p^{-1}(\boldsymbol{T})\right] .$$

These asymptotic expressions, however, are unknown expressions in the asymptotic distribution. In practice, even though we can substitute most of the unknowns by estimates, the asymptotics do not come even close to the real finite sample distribution. Therefore we propose to use bootstrap procedures which allow us to simulate the critical value. The first one, Procedure C, is a purely parametric bootstrap. Alternatives are discussed in Section 4.

Procedure C.

1. From the sample, calculate a consistent estimator $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ of $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)$.

2. Generate a vector $\boldsymbol{w}_1$ containing D independent copies of a variable w_1 with $E[w_1] = 0$ and $E[w_1^2] = 1$ with subexponential tails; that is, for a constant C_1 it holds that $E[exp\{|w_1|/C_1\}] \leq C_1$ (c.f. [A.4]). Construct the vector $\boldsymbol{u}^* = \hat{\sigma}_u \boldsymbol{w}_1$ such that the mean vector is $\mathbf{0}_D$ and the variance covariance matrix is $\hat{\Sigma}_u = \hat{\sigma}_u^2 \boldsymbol{I}_D$.

3. Generate a vector $\boldsymbol{w}_2$ containing n independent copies of a variable w_2 with $E[w_2] = 0$ and $E[w_2^2] = 1$ with subexponential tails (c.f. [A.4]). Construct the vector $\boldsymbol{e}^* = \hat{\sigma}_e \boldsymbol{w}_2$, which is independent of $\boldsymbol{u}^*$, such that the mean vector is $\mathbf{0}_n$ and the variance covariance matrix is $\hat{\Sigma} = \hat{\sigma}_e^2 \boldsymbol{I}_n$.

4. Under H_0 true, set

$$Y^*_{dj} = g\left(\boldsymbol{T}^t_{dj}\tilde{\boldsymbol{\gamma}} + \boldsymbol{X}^t_{dj}\tilde{\boldsymbol{\beta}} + u^*_d\right) + e^*_{dj},\ d = 1, \cdots, D,\ j = 1, \cdots, n_d.$$

5. Calculate the test statistic R^*_1 (R^*_{1w} respectively) from the bootstrap sample $(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{T})$.

Again, applying quasi likelihood estimation allows to relax the distribution assumptions up to exponential families. Then, Procedure C is a version of wild bootstrap, which has been introduced by Wu (1986) (see also Beran 1986; Mammen 1992) and was first proposed by Härdle and Mammen (1993) in nonparametric setups. Liu (1988) studied the wild bootstrap under regression models with non-i.i.d. observations (e.g. taking $\Sigma = \sigma^2_e \boldsymbol{A}^{-1}$ with matrix of weights $\boldsymbol{A} = diag(a_{11}, \cdots, a_{Dn_D})$), fulfills in addition to the conditions $E[w_\bullet] = 0$, $E[w^2_\bullet] = 1$ also condition $E[w^3_\bullet] = 1$, in order to obtain the second order properties of Wu's bootstrap.

The computation of quantiles of the distributions of R_l ($l = 1w, 1$) can be done by Monte Carlo: generate B independent sets of bootstrap samples $(\boldsymbol{Y}^{*(b)}, \boldsymbol{X}, \boldsymbol{T})$, $b = 1, \cdots, B$. The $(1-\alpha)$ quantiles of the distributions R_l can be estimated by the $\{[(1-\alpha)B] + 1\}$th order statistic of $R^{*(b)}_l = R^*_l(\boldsymbol{Y}^{*(b)}, \boldsymbol{X}, \boldsymbol{T})$ $(b = 1, \cdots, B)$.

Theorem 1 shows that the bootstrap procedure works.

Theorem 1. *Under the assumptions of Corollary 4, it holds for $l = 1w, 1$, that*

$$d_k\left(F^*_{R^*_l}, F_{R_l}\right) \longrightarrow 0.$$

Where F_{R_l} is the distribution of R_l, $F^*_{R^*_l}$ is the conditional distribution of R^*_l (given the sample), and d_k is the Kolgomorov distance, which is defined as

$$d_k(\nu, \tau) = sup_{\{t \in I\!R\}}|\nu(X \leq t) - \tau(X \leq t)|$$

for two probability measures ν and τ on the real line.

Proof. The consistency of bootstrap methods is proved by imitation (for a general discussion of the validation of bootstrap methods see Shao and Tu (1995, pp.76)). One proceeds as in Härdle, Mammen and Müller (1998, see proof of their Theorem 2 in Appendix), taking into account the asymptotics results of the previous section and that $|Y^*_{dj}|$ has a bounded conditional Laplace transform (in a neighborhood of 0). For more details see Mammen and van de Geer (1997, Section 5); these authors studied the asymptotic distribution of the parametric component of a regression model using wild bootstrap.

Thus, for $l = 1w, 1$, it holds that

$$d_k\left(F^*_{R^*_l}, N(b, v^2)\right) \longrightarrow 0,$$

in probability, with b and v^2 introduced in Corollary 4. □

4 Alternatives and Extensions

4.1 Estimation for "small" D

Even though the following differentiation is not necessary, let us distinguish the case with $D \to \infty$ where $\boldsymbol{u} = (u_1, u_2, \ldots, u_D)$ becomes therefor an infinite dimensional vector, and the case where D is small relatively to n. A particular case is when D is fixed, i.e. $\boldsymbol{u}$ is a finite dimensional vector. As has been seen above, concerning theory Procedure A is valid for any case. In practice, however, one has the problem that one has either to integrate over all u , see equation (3) or to calculate $\hat{u}$ simultaneously (e.g. to get σ_u^2). This can only be avoided in particular situations like in simple linear models with a perfectly known variance matrix. In case of prediction, especially in small area statistics, one is interested in $\hat{u}$ in any case. When we have to (or want to) calculate also the $\hat{u}$, the likelihood functions are commonly based on the PQL approach, and one would replace Procedure A by

Procedure D.

1. For fixed $(\boldsymbol{u}, \boldsymbol{\delta})$ we estimate $m(\boldsymbol{t}_0)$ as the solution of the problem
$$\check{m}_{(\boldsymbol{u},\boldsymbol{\delta})} = \underset{\{m \in \mathcal{M}\}}{argmax}\ \varphi_{ss}(\boldsymbol{Y}; m, \boldsymbol{\delta}) \ .$$
2. We calculate $(\boldsymbol{u}, \boldsymbol{\delta})$ as the solution of the problem
$$(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\delta}}) = \underset{\{\boldsymbol{u},\boldsymbol{\delta}\}}{argmax}\ \varphi(\boldsymbol{Y}, \boldsymbol{u}; \check{m}_{\boldsymbol{\delta}}, \boldsymbol{\delta}). \tag{18}$$
3. With the estimators obtained in steps 1 and 2, we set finally $\hat{m} = \check{m}_{(\hat{\boldsymbol{\delta}},\hat{\boldsymbol{u}})}$.

This can simplify calculations a lot but can be a harder problem than the first one when D converges to infinity without any restrictions. To apply standard theory it is necessary to assume D to be fixed (D of order $o(n)$ might do). In that case the statements of Corollary 1 to 3 hold also for Procedure D. Whether it is preferable to work with Procedure A or D will depend on the particular context.

4.2 Alternative Tests Statistics

Although we think the test statistic R_{1w} is the most natural approach, we would like to add two more statistics that are very much related to the first one.

Recalling the likelihood ratio comparing the full indices weighted by something proportional to the variance of $\{\hat{m}(\boldsymbol{t}_{dj}) - \tilde{m}(\boldsymbol{t}_{dj})\}$ we propose a first alternative:

$$R_{2w} = \sum_{d=1}^{D} \sum_{j=1}^{n_d} \mathrm{H}\left(\hat{m}(\boldsymbol{t}_{dj}), \hat{\boldsymbol{\delta}}\right) \left[\hat{m}(\boldsymbol{t}_{dj}) - \widetilde{m}(\boldsymbol{t}_{dj}) + \boldsymbol{X}_{dj}^t(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + \hat{u}_d - \tilde{u}_d\right]^2 \pi(\boldsymbol{t}_{dj}) \tag{19}$$

with $\pi(\cdot)$ being a weight function as before. This statistic comes close to the one of Härdle, Mammen, Müller (1998). Obviously, including the random effects yields a quite interesting alternative to R_{1w} when the target of the analysis is to find the better prediction model.

Remember that the random effects are supposed to be independent of $\boldsymbol{X}$ and $\boldsymbol{T}$ but only correct for the mean in each subset indexed by the same d (e.g. being from the same area). Consequently including them now in R_{2w} will not affect systematically the (squared) differences under consideration. Intuitively, the test statistics R_{1w} and R_{2w} should test exactly the same. However, it is an important modification because the handling of the asymptotic distribution is straight forward for R_{1w}, whereas the distribution of $\hat{u}_d$ respectively $\tilde{u}_d$ gets rather cumbersome when D goes to infinity.

A very simple test statistic would be to reduce the prior versions to (weighted) differences of the $m(\cdot)$ estimates only:

$$R_{3w} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} \mathrm{H}\left(\hat{m}(\boldsymbol{t}_{dj}), \hat{\boldsymbol{\delta}}\right) \left[\hat{m}(\boldsymbol{t}_{dj}) - \widetilde{m}(\boldsymbol{t}_{dj})\right]^2 \pi(\boldsymbol{t}_{dj}) \ , \tag{20}$$

$$R_3 = \sum_{d=1}^{D}\sum_{j=1}^{n_d} \left[\hat{m}(\boldsymbol{t}_{dj}) - \widetilde{m}(\boldsymbol{t}_{dj})\right]^2 \pi(\boldsymbol{t}_{dj}) \ . \tag{21}$$

Even though the asymptotic distribution is the same as for R_1 , respectively R_{1w} (i.e. the statements of Corollary 4 hold also for R_3 and R_{3w}), in finite samples we expect a different performance when the impact of the regressors $\boldsymbol{X}$ and $\boldsymbol{T}$ are related (existence of concurvity).

We conclude with the remark that the bootstrap procedure discussed in the last section as well as the ones that will be discussed next is valid for all of these test statistics.

4.3 Alternative Bootstrap Procedures

In case we face a model with additive error terms we could consider the following alternative. In Procedure C we have made use of the homoscedasticity assumption estimating σ_ϵ and using it for the generation of the bootstrap samples. Often, in practice some extreme values make the test quite conservative as they produce rather huge estimates of σ_ϵ. In case of additive errors one can circumvent this problem by using the wild bootstrap typically used when the assumption of homoscedasticity is dropped:

Procedure E.

1. From the sample, calculate a consistent estimator $\hat{\sigma}_u^2$ of σ_u^2.

2. Generate a vector $\boldsymbol{w}_1$ containing D independent copies of a variable w_1 with $E[w_1] = 0$ and $E[w_1^2] = 1$ with subexponential tails. Construct the vector $\boldsymbol{u}^* = \hat{\sigma}_u \boldsymbol{w}_1$ such that the mean vector is $\boldsymbol{0}_D$ and the variance covariance matrix is $\hat{\Sigma}_u = \hat{\sigma}_u^2 \boldsymbol{I}_D$.

3. Generate a vector $\boldsymbol{w}_2$ containing n independent copies of a variable w_2, which is independent of the random variable w_1, with $E[w_2] = 0$ and $E[w_2^2] = 1$, and that fulfills for a constant C_2 that $|w_2| \leq C_2$ (a.s.). Construct the vector $\boldsymbol{e}^* = \hat{\boldsymbol{e}} \boldsymbol{w}_2$, with the residual vector $\hat{\boldsymbol{e}} = \boldsymbol{Y} - g\left(\boldsymbol{T}^t \hat{\boldsymbol{\gamma}} + \boldsymbol{X}^t \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}\right)$.

4. Under H_0 true, set

$$Y_{dj}^* = g\left(\boldsymbol{T}_{dj}^t \tilde{\boldsymbol{\gamma}} + \boldsymbol{X}_{dj}^t \tilde{\boldsymbol{\beta}} + u_d^*\right) + e_{dj}^*,\ d = 1, \cdots, D,\ j = 1, \cdots, n_d.$$

5. Calculate the test statistic under consideration from the bootstrap sample $(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{T})$.

Let us also consider the special situation of the logistic semiparametric mixed model. For this we recommend the following parametric bootstrap, which is a modification of the resampling methods discussed so far:

Procedure F.

1. From the sample, calculate a consistent estimator $\hat{\sigma}_u^2$ of σ_u^2.

2. Generate a vector $\boldsymbol{w}_1$ containing D independent copies of a variable w_1 with $E[w_1] = 0$ and $E[w_1^2] = 1$ with subexponential tails. Construct the vector $\boldsymbol{u}^* = \hat{\sigma}_u \boldsymbol{w}_1$ such that the mean vector is $\boldsymbol{0}_D$ and the variance covariance matrix is $\hat{\Sigma}_u = \hat{\sigma}_u^2 \boldsymbol{I}_D$.

3. Under H_0 true, generate observations by generating values of a binomial distribution with sizes n_{dj} and probabilities

$$p_{dj}^* = \frac{exp\{\boldsymbol{T}_{dj}^t \tilde{\boldsymbol{\gamma}} + \boldsymbol{X}_{dj}^t \tilde{\boldsymbol{\beta}} + u_d^*\}}{1 + exp\{\boldsymbol{T}_{dj}^t \tilde{\boldsymbol{\gamma}} + \boldsymbol{X}_{dj}^t \tilde{\boldsymbol{\beta}} + u_d^*\}},\ d = 1, \cdots, D,\ j = 1, \cdots, n_d.$$

4. Calculate the test statistic under consideration from the bootstrap sample $(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{T})$.

Finally, note that the statement of Theorem 1 holds also for these bootstrap types.

4.4 Extensions to other Models and Hypothesis

So far we have introduced the generalized partial linear model for mixed effects and given some bootstrap based test statistics for testing linearity of the nonparametric part. As the estimation technique is based on the well studied semiparametric profiled likelihood approach, our method can extended, quite straight forwardly, to similar models who's analysis is based on the same methodology. In particular we are thinking of introducing random effects in

- generalized partially linear single-index models

$$E[Y|\boldsymbol{X},\boldsymbol{T}] \;=\; g\left\{\boldsymbol{X}^t\boldsymbol{\beta} + m(\boldsymbol{T}^t\gamma)\right\} \tag{22}$$

  see Carroll, Fan, Gijbels, Wand (1997)

- generalized additive partially linear models

$$E[Y|\boldsymbol{X},\boldsymbol{T}] \;=\; g\left\{\boldsymbol{X}^t\boldsymbol{\beta} + \sum_{k=1}^{p} m_k(T_k)\right\} \;, T_k \in I\!R \tag{23}$$

  see Härdle, Huet, Mammen, Sperlich (2004). They actually also allow for introducing nonparametric interaction terms of the form $m_{kl}(T_k, T_l)$.

- semiparametric separable models

$$E[Y|\boldsymbol{X},\boldsymbol{T}] \;=\; g_\theta\left\{\boldsymbol{X}, m_1(T_1), m_2(T_2), \ldots, m_p(T_p)\right\} \;, T_k \in I\!R \;, \tag{24}$$

  where $g(\cdot)$ is allowed to depend on a vector of unknown parameter θ. The estimation of those models have been introduced by Rodríguez-Póo, Sperlich, Vieu (2003).

In all these models it is obvious now how to estimate extensions where also random effects enter (linearly). Further, Härdle, Huet, Mammen, Sperlich (2004) give a large bunch of bootstrap based tests for analyzing model (23). In particular, they provide statistics for testing $m_k(\cdot)$ for any given parametric structure, testing for interaction, and testing the link specification. They also construct uniform confidence bands for each function $m_k(\cdot)$.

Note that these tests have important application in small area estimation. Consider e.g. the nested-error regression type models of Battese, Harter and Fuller (1988), where Y_{dj} is the target character for the j'th sample unit in the d'th area (domain) sample:

$$E[Y_{dj}|u_d, \boldsymbol{X}_{dj}, \boldsymbol{T}_{dj}] = g\left\{m(\boldsymbol{T}_{dj}) + \boldsymbol{X}_{dj}^t\boldsymbol{\beta} + u_d\right\} \quad d = 1, \ldots, D;\; j = 1, \ldots, n_d$$

or the Fay-Herriot type model (1979), assuming that the sample mean $\bar{y}_d$ is related with the area mean $\mu_d = m(\boldsymbol{T}_d) + \boldsymbol{X}_d^t\boldsymbol{\beta} + u_d$ via

$$E[\bar{y}_d|u_d, \boldsymbol{X}_d, \boldsymbol{T}_d] = g\left\{m(\boldsymbol{T}_d) + \boldsymbol{X}_d^t\boldsymbol{\beta} + u_d\right\} \quad d = 1, \ldots, D.$$

These models have been studied so far only for the case of linearity of $m(\cdot)$.

5 Finite Sample Performance

As discussed in the previous sections, there exist already a large amount of papers considering the estimation of semiparametric mixed effects models for different smoothers, implementations, and likelihoods. Therefore, we concentrated here once more on the testing part and stick to a relatively easy to estimate model. Instead, we will intensively study the effect of using different bootstrap procedures, modified test statistics, (slightly) different smoothers, different bandwidths, etc. as well as on the effect of facing different data generating processes. For all this we studied the first error type and the power having only samples of moderate size.

The data generating process was

$$y_{dj} \;=\; 1 + (1 - \;a\;)t_{dj} + \;a\; sin(\pi t_{dj}) + \beta^t \boldsymbol{x}_{dj} + u_d + \epsilon_{dj}\;, \tag{25}$$

for $d = 1, \ldots, D,\ ,\ j = 1, \ldots, n_d$, where $\beta^t = (2, 1)\ ,\ (t_{dj}, \boldsymbol{x}^t_{dj}) \in I\!R^3$ i.i.d., $u_d \sim N(0, 1)$ i.i.d., and $\epsilon_{dj} \sim N(0, 0.25)$ i.i.d., where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. We simulated the case where a is running from 0 (giving the null hypothesis model) to 0.5 to study the error of the first and second type. Further, for the explanatory variables $(t_{dj}, \boldsymbol{x}^t_{dj})$ we simulated three different (always random) designs; first $U[0, 2]^3$, then normal with mean 1.0, variance 0.6 but uncorrelated, and finally normal with mean 1.0, variance 0.6 but covariance 0.15. This has been done as it is well known that non- and semiparametric inference unfortunately is strongly affected by the experimental design; in our context it obviously is of special interest to see the (expected) loss in power when we change from uncorrelated to correlated designs. Note that they all have the same mean, but in case of normal distribution about 10 to 20% of the observations fall outside of the $[0, 2]^3$ cube. We studied two sample sizes, $n = 100$ and $n = 200$. When $n = 100$ we set $D = 10$ with n_1 to n_D equal to 5, 7, 8, 9, 10, 10 ,11, 12, 13, and 15. For $n = 200$ we set $D = 20$ and each of the n_d from above occurs twice. We used always $B = 500$ bootstrap replications to calculate the critical values of the test statistic.

The test statistics have been implemented first with a Nadaraya Watson smoother. Even though this smoother suffers from boundary effects, we did neither boundary corrections nor any trimming, i.e. we set $\pi(\boldsymbol{t}) = 1$ throughout; instead, we trusted in the ability of the bootstrap procedures to replicate these effects adequately. The literature on bandwidths selection for nonparametric (kernel) estimation is abundant but it is also well known that the optimal bandwidth for testing has a faster rate, i.e. should be undersmoothing in practice. Although the cross validation bandwidth is asymptotically optimal for estimation rather then for testing, to our experience it behaves pretty well for testing problems with finite samples since it has indeed the tendency to somewhat undersmooth. Alternatively, there exists an increasing literature on adaptive testing, i.e. choosing a bandwidth that maximizes the power of the test. However, these methods, so far only available for some particular testing problems, are rather expensive with respect to implementation and computational time. An approach

that could probably be extended to our testing problem has been recently proposed by Rodríguez-Póo, Sperlich, and Vieu (2005) and is based on the idea of Spokoiny (2001). The results presented in this chapter are calculated with bandwidths $h = h_0/n^{2/9}$ where $h_0 = 1.0$, 1.5, and 2.0 respectively.

Although both bootstrap procedures are implemented (C and E), in the following are given only the results for Procedure C. As we face simulated data without extreme values or outliers it is clear that the procedure making use of the homoscedasticity (i.e. Procedure C) is superior and gives always somewhat better results. The power loss when using Procedure E was in our simulations between 5 to 15%.

Let us start with a comparison of the different proposed test statistics. As in (25) the canonical link function is the identity function, the test statistics R_{jw} and R_j coincide (for $j = 1, 3$). It remains therefore to compare R_1 with R_2 and R_3.

		R_2			R_1			R_3		
	h_0	1.0	1.5	2.0	1.0	1.5	2.0	1.0	1.5	2.0
$U[0,2]^3$	p	.485	.510	.534	.481	.501	.526	.484	.498	.518
	1%	.010	.004	.000	.012	.002	.002	.010	.002	.004
	5%	.070	.050	.022	.068	.052	.024	.066	.058	.026
	10%	.108	.096	.054	.110	.098	.072	.116	.098	.084
$N(0, 0.75)$	p	.495	.560	.607	.491	.554	.600	.488	.557	.599
$Cov = 0.0$	1%	.010	.002	.002	.014	.004	.002	.014	.002	.002
	5%	.064	.034	.014	.066	.034	.014	.068	.034	.016
	10%	.120	.060	.050	.120	.066	.052	.120	.070	.052
$N(0, 0.75)$	p	.503	.594	.647	.502	.588	.639	.502	.585	.631
$Cov = .15$	1%	.014	.002	.002	.014	.002	.002	.016	.004	.004
	5%	.070	.018	.018	.078	.028	.016	.070	.022	.014
	10%	.122	.066	.032	.130	.066	.034	.116	.066	.032

Table 1: The p-values (p) and first error type at 1, 5, and 10% level for the **different tests** using Nad.-Wat., parametric bootstrap (Procedure C), $n = 100$, $D = 10$.

In Table 1 are given the real rejection levels for different nominal levels under the null hypothesis of linearity of $m(\cdot)$, i.e. setting $a = 0$ in (25), calculated from 500 simulation runs. As expected, depending on the bandwidth, the rejection levels vary somewhat but due to the implemented bias reduction (compare discussion in Section 3, Procedure B) this test tends to under-reject for small samples, i.e. to be conservative instead of being too liberal. In any case, we cannot detect any clear differences between R_1, R_2, and R_3 in the sense that one of them would in general be more correct than the other.

To study the power performance of these test statistics we let a in model (25) run from 0 to 0.5 determining the real rejection levels for the different $a \in [0, 0.5]$ based on 100 simulation runs. In Figure 1 are plotted the power functions of the three tests

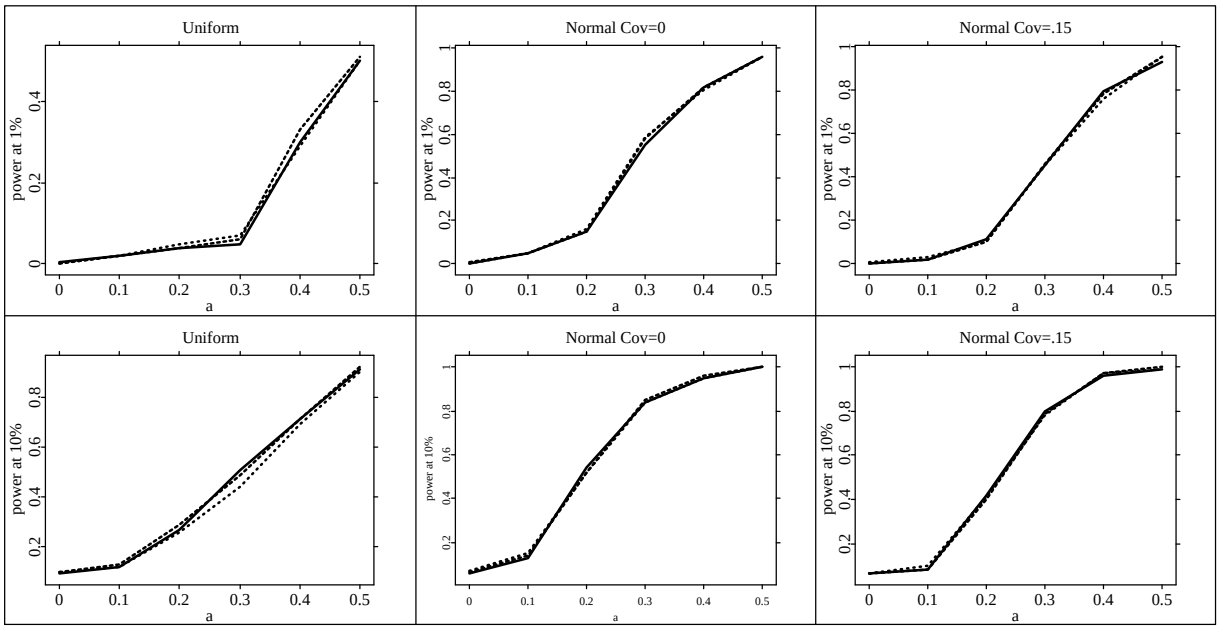

Figure 1: Comparing the powers of R_2 (solid), R_1 (dotted) and R_3 (dashed) with $n = 100$ for different designs using Nadaraya Watson smoother.

(solid line for R_2, dotted for R_1 and dashed for R_3). From these plots we can draw several conclusions. First, the tests do hardly differ, so for further investigations it does not matter whether we consider R_1, R_2 or R_3. Second, obviously, even for this rather small sample size of only 100 ($D = 10$) observations our tests work quite well detecting already moderate deviations from the null hypothesis. Third, the loss of power caused by introducing correlation in the design is moderate but, at lease visually, evident. The case study with uniform design does, maybe surprisingly, not better than the one with normally uncorrelated distributed regressors but even worse. This is probably due to the larger support even though the observations are rather sparse outside the $[0, 2]^3$ cube.

	R_2				R_1			
	$n = 100$		$n = 200$		$n = 100$		$n = 200$	
h_0	1.5	2.0	1.5	2.0	1.5	2.0	1.5	2.0
p	.499	.501	.513	.514	.495	.495	.511	.511
1%	.010	.016	.004	.004	.006	.012	.002	.004
5%	.068	.060	.034	.040	.070	.060	.038	.042
10%	.108	.100	.086	.086	.114	.096	.086	.086

Table 2: R_1 and R_2, Uniform design (uncorrelated), **Local Linear Smoother**, p-values (p) and first error type at level α% for bootstrap (Procedure C).

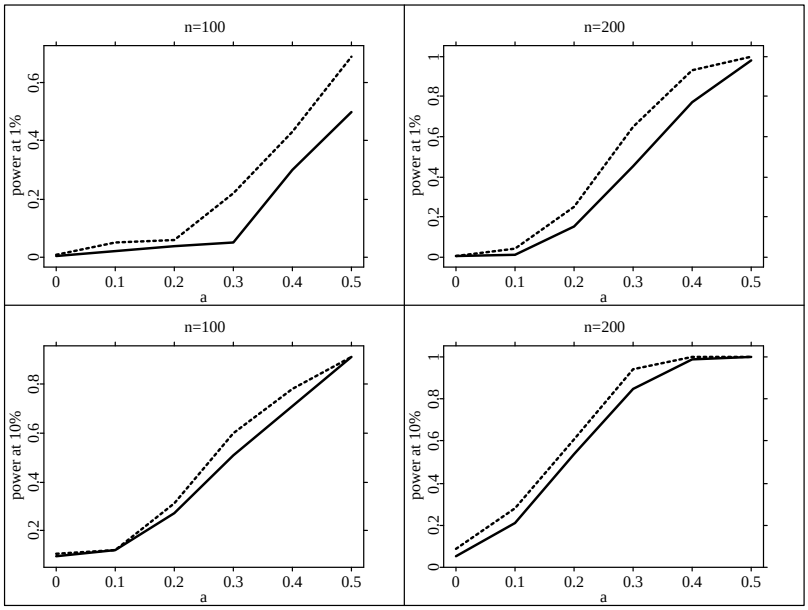

Figure 2: Comparing the power of R_2 when using Nadaraya Watson smoother (solid) vs Local Linear smoother (dotted) for different sample size, always uniform design.

The estimation procedure has been implemented with Nadaraya Watson as well as with Local Linear smoother. We next study whether and how the performance of the test changes, maybe improves. On the one hand the main advantage of the local linear smoother is its strong reduction of boundary effects which are certainly more serious for the uniform design when changing from Nadaraya Watson to Local Linear smoothing. On the other hand a local linear smoother is unbiased under the null hypothesis of linearity what should, at least for increasing sample size, make it clearly superior over the Nadaraya Watson approach. For these two reasons we concentrate on the uniform design and compare their performances also for $n = 200$, $D = 20$. First let us have a look on the error of the first type, see Table 2. As we had serious numerical problems for $h_0 = 1.0$, i.e. got too many zero - weights, we give only results for $h_0 = 1.5$ and 2.0 respectively. As can be seen, the Local Linear based test is much more robust against bandwidths choice although we have to admit that in our small simulation study it is slightly too liberal for the nominal 5% rejection level, compare with Table 1. Turning now to the power study, see Figure 2, we can only state a clear improvement for the case $n = 100$, $D = 10$ at the nominal 1% level. Here, all plots refer to results using bandwidths $h_0 = 1.5$. Certainly, also for the other cases we see the power functions of the Local Linear based test is located above its competitors, but only a little bit and this having already started from a higher (real) rejection level under the null. Nevertheless, together with the results on the first error type we conclude that a Local Linear based test should be given preference if the additional computational cost is low.

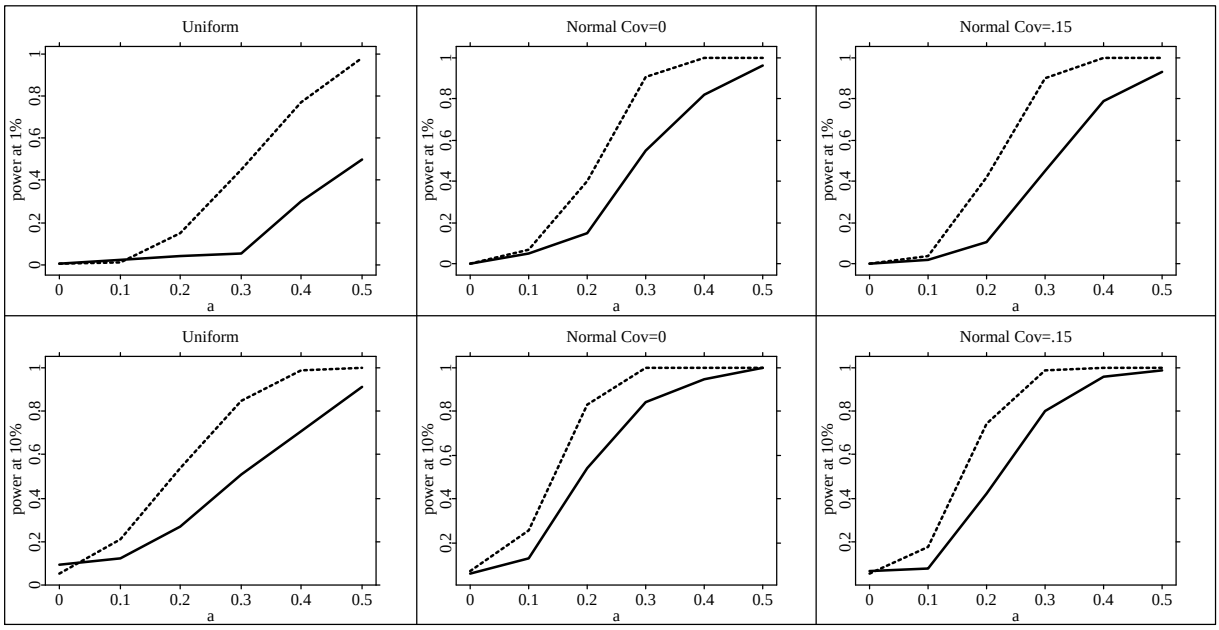

Figure 3: Comparing the power of R_2 when $n = 100$, $D = 10$ (solid) vs $n = 200$, $D = 20$ (dotted) for different designs using Nadaraya Watson smoother.

In the above mentioned study on Local Linear smoother we have also seen that doubling the sample size already improves a lot the power performance. Although this is expected from asymptotic theory, have in mind that the results are of asymptotic nature and especially for very small samples the real improvement is often stronger than the theoretical convergence rate would make us expect. In Figure 3 we directly compare the power functions given $n = 100$, $D = 10$ vs $n = 200$, $D = 20$ for all designs but concentrate again on the test R_2 with Nadaraya Watson based smoother and bandwidths $h_0 = 1.5/n^{2/9}$. We see a quite strong improvement, most evident for the uniform design. Having in mind the complexity of the model, the small deviation from the null hypothesis, and the moderate sample size, the power performance is amazing.

As can be suspected already from the section where we have introduced the estimator, but also thinking of the bootstrap procedure, the computational cost of our procedure is high. Even though it could be argued that spline smoothers can be implemented more efficiently than kernel based smoother (in the one dimensional case), the reason why our method is so expensive is not the smoothing but the different iterations. Therefore, when we come to the bootstrap, in the original (i.e. correct) algorithm there does not exist one "final" smoothing matrix that simply could be applied to all bootstrap responses Y^*. It is clear that the variance as well as the bias of the test statistic is dominated by the nonparametric part, and only in higher order terms by $\hat{u}$ and $\hat{\beta}$ (respectively $\hat{\gamma}$) whereas $\hat{\sigma}_u$, $\hat{\sigma}_\epsilon$ only influence the weighting (for $\hat{\beta}$, $\hat{\gamma}$ and for the test statistic). We therefore implemented a simplified version of the bootstrap where for

		Nadaraya Watson		Local Linear	
	$n =$	100	200	100	200
$U[0,2]^3$	p	.503	.564	.516	.550
	1%	.022	.008	.034	.002
	5%	.072	.026	.078	.026
	10%	.108	.054	.118	.070
$N(0, 0.75)$	p	.576	.572	.501	.530
$Cov = 0.0$	1%	.004	.004	.028	.014
	5%	.032	.028	.080	.052
	10%	.054	.060	.128	.092
$N(0, 0.75)$	p	.612	.610	.480	.524
$Cov = .15$	1%	.002	.004	.020	.014
	5%	.018	.026	.078	.060
	10%	.052	.054	.126	.102

Table 3: **Simplified Version**, R_2 only, p-values (p) and first error type at level α% for parametric bootstrap (Procedure C), always $h_0 = 1.5$.

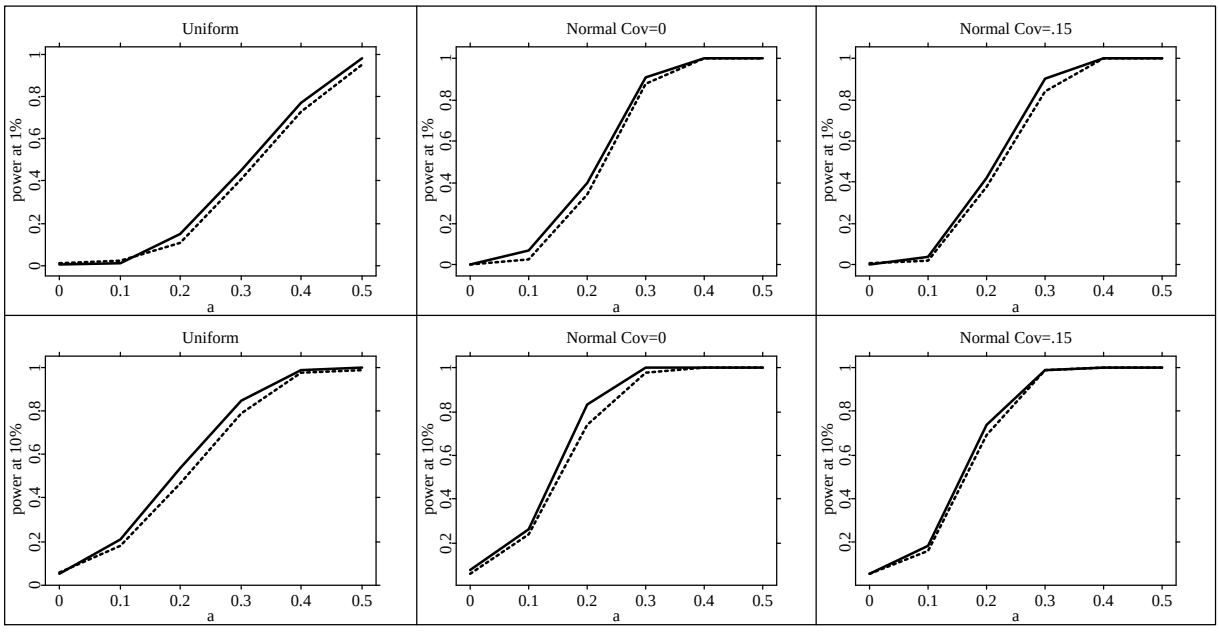

Figure 4: Comparing the power of R_2 calculated by original (solid) vs calculated by simplified algorithm (dotted), always using Nadaraya Watson smoother, $n = 200$ with different designs.

the estimation of the parameter only the last iteration is repeated inside the Newton-Raphson method, using the variance estimates obtained from the original sample. We

only iterate between the estimation of β and $m(\cdot)$. We found that the computational time decreased almost exponentially (depending on the sample size). However, not surprisingly, there is some loss in exactness and power. Our final simulation study is dedicated to this simplified version. In Table 3 is given the error of the first type for R_2, both smoothers (Nadaraya Watson and Local Linear), different sample sizes and all designs. We see that when we use this simplified version for sample size $n = 100$ ($D = 10$), to hold the level, a somewhat larger bandwidths is needed than for the original algorithm, compare with Table 1. However, for $n = 200$, $D = 20$ the simplified version works as good as the original one. This is underpinned also in the power study, see Figure 4. The power functions hardly differ, i.e. the power loss due to simplification is clearly negligible. This weapons us with a quite potential approximation that is rather useful for large samples, higher dimensional nonparametric parts, or if one wishes to do many bootstrap replications. Why extensions to higher dimensional nonparametric parts can be rather interesting has been seen in the previous section.

6 Technical Assumptions

First, let us introduce some more notation. We define

$$D^{\boldsymbol{\mu}_x} a(x) = \frac{\partial^{|\boldsymbol{\mu}_x|}}{\partial {x_1}^{\mu_1}, \cdots, \partial {x_k}^{\mu_k}} \boldsymbol{a}(x),$$

being $\boldsymbol{\mu}_x$ a k-vector of nonnegative integer constants, $|\boldsymbol{\mu}_x| = \sum_{j=1}^{k} \mu_j$ and $\boldsymbol{a}(x) \in I\!R^k$ any function. We denote by $\mathcal{D}_{m,\delta}^{r_m,s_\delta}(Y) = D^{r_m} D^{s_\delta} l(Y; m, \boldsymbol{\delta})$ and by $f_{m\delta}^{(r_m,s_\delta)}(y, u, \boldsymbol{x}|\boldsymbol{t})$ the conditional density of $\mathcal{D}_{m,\delta}^{r_m,s_\delta}(Y)$ given $\boldsymbol{T} = \boldsymbol{t}$. Let us define for each $\boldsymbol{\delta} \in \Delta$ and $\boldsymbol{t} \in \mathcal{T}$

$$h(\boldsymbol{\delta}, m, \boldsymbol{t}) = E[l(Y; m, \boldsymbol{\delta})|\boldsymbol{T} = \boldsymbol{t}] \tag{26}$$

and $\bar{m}_{\boldsymbol{\delta}}(\boldsymbol{t})$ the solution to

$$\frac{\partial}{\partial m} h(\boldsymbol{\delta}, m, \boldsymbol{t}) = 0 \ , \tag{27}$$

with respect to m for each fixed $\boldsymbol{\delta}$ and $\boldsymbol{t}$. Remember that $l\,(Y; m, \boldsymbol{\delta}) = \log f(Y|u, \boldsymbol{T}, \boldsymbol{X}; m, \boldsymbol{\delta}) + \log p(u; \sigma_u^2)$.

Let us write down the conditions for the case when we consider the family of density functions $\{f(\cdot|u, \boldsymbol{t}, \boldsymbol{x}; m, \boldsymbol{\delta}) : \boldsymbol{\delta} \in \Delta, m \in \mathcal{M}\}$. Then, we assume that they satisfy the following conditions:

A.1 For fixed but arbitrary $(m_1, \boldsymbol{\delta}_1) \in \mathcal{M} \times \Delta$, let

$$\rho(m, \boldsymbol{\delta}) = \int l(y; m, \boldsymbol{\delta}) f(y|u, \boldsymbol{t}, \boldsymbol{x}; m_1, \boldsymbol{\delta}_1) dy \ ,$$

with $(m, \boldsymbol{\delta}) \in \mathcal{M} \times \Delta$. If $\boldsymbol{\delta} \neq \boldsymbol{\delta}_1$ then $\rho(m, \boldsymbol{\delta}) < \rho(m_1, \boldsymbol{\delta}_1)$.

A.2 The matrix $I_{\boldsymbol{\delta}}$ is positive definite for all $\boldsymbol{\delta} \in \Delta$ and $m \in \mathcal{M}$.

A.3 Assume that for vectors $|r_m| \leq 4$ and $|s_\delta| \leq 4$ such that $|r_m| + |s_\delta| \leq 4$ the function $D^{r_m} D^{s_\delta} l(Y; m, \boldsymbol{\delta})$ exists for almost all Y. Further, assume that

$$E\left\{sup_{\boldsymbol{\delta}} sup_m |D^{r_m} D^{s_\delta} l(Y; m, \boldsymbol{\delta})|^2\right\} < \infty \ .$$

A.4 The Laplace transform $E[exp\{t|Y_{dj}|\}]$ is finite for $t > 0$ small enough.

The condition [A.2] and [A.3] are usual in likelihood related problems. E.g. [A.3] allows differentiation and integration to be interchanged when differentiating

$$\rho(m, \boldsymbol{\delta}) = \int l(y; m, \boldsymbol{\delta}) f(y|u, \boldsymbol{t}, \boldsymbol{x}; m_1, \boldsymbol{\delta}_1) dy \ .$$

The condition [A.4] is essential for the asymptotic expansions of Corollary 4, see Mammen and van de Geer (1997).

Next, we need to include some smoothness assumptions that are necessary because of the use of nonparametric smoothing methods:

B.1 For each $\boldsymbol{\delta} \in \Delta$ and $\boldsymbol{t} \in \mathcal{T}$,

$$sup_{\{\boldsymbol{\delta}, m, \boldsymbol{t}\}} |D^{r_m} D^{s_\delta} D^{z_t} h(\boldsymbol{\delta}, m, \boldsymbol{t})| < \infty$$

for $2 \leq |r_m| \leq 4$, $|s_\delta| \leq 2$, $|x_t| \leq 1$, and $|r_m| + |s_\delta| + |x_t| \leq 4$.

B.2 The solution to (27), $\bar{m}_{\boldsymbol{\delta}}(\boldsymbol{t})$, is unique and for any constant $\epsilon > 0$ there exists another $\nu > 0$ such that

$$sup_{\boldsymbol{\delta}} sup_{\boldsymbol{t}} \left| \frac{\partial}{\partial m} h(\boldsymbol{\delta}, \bar{m}_{\boldsymbol{\delta}}(\boldsymbol{t}), \boldsymbol{t}) \right| \leq \nu$$

implies that

$$sup_{\boldsymbol{\delta}} sup_{\boldsymbol{t}} \left| \bar{m}_{\boldsymbol{\delta}}(\boldsymbol{t}) - m_{\boldsymbol{\delta}}(\boldsymbol{t}) \right| \leq \epsilon \ .$$

B.3 Assume that

(a) $E[sup_m sup_{\boldsymbol{\delta}} |\mathcal{D}_{m,\delta}^{r_m, s_\delta}(Y)|] < \infty$ for $|r_m| \leq 5$ and $|s_\delta| \leq 3$,

(b) for some even integer $\xi \geq 10$ it holds that $sup_m sup_{\boldsymbol{\delta}} E[|\mathcal{D}_{m,\delta}^{r_m, s_\delta}(Y)|^q] < \infty$ for $|r_m| \leq 3$ and $|s_\delta| \leq 4$,

(c) $sup_m sup_{\boldsymbol{\delta}} sup_{\{y,u,\boldsymbol{t},\boldsymbol{x}\}} |f_{m\delta}^{(r_m, s_\delta)}(y, u, \boldsymbol{x}|\boldsymbol{t})| < \infty$ for $|r_m| \leq 4$ and $|s_\delta| \leq 3$,

(d) $sup_{\boldsymbol{t}} |D^{x_t} p(\boldsymbol{t})| < \infty$ and $sup_m sup_{\boldsymbol{\delta}} sup_{\{y,u,\boldsymbol{t},\boldsymbol{x}\}} |D^{x_t} f_{m\delta}(y, u, \boldsymbol{x}|\boldsymbol{t})| < \infty$ for $|x_t| \leq a + 2$,

(e) and $0 < inf_{\{\boldsymbol{t} \in \mathcal{T}\}} p(\boldsymbol{t}) < sup_{\{\boldsymbol{t} \in \mathcal{T}\}} p(\boldsymbol{t}) < \infty$.

The assumptions [B.1] - [B.3] are sufficient to guarantee that the nonparametric estimator from step 1 of Procedures A and B fulfill $sup_{\{\boldsymbol{t}_0 \in \mathcal{T}\}} |\hat{m}(\boldsymbol{t}_0) - m(\boldsymbol{t}_0)| = o_p(n^{-1/4})$, and thus it is estimator of least favorable curves.

Finally, we also need to impose some conditions on the kernel function $K(\cdot)$ and the bandwidth $\boldsymbol{h}$:

N.1 Function $K(\cdot)$ is a bounded kernel of order a with compact support, and $sup_{\boldsymbol{z}} |D^{t_z} K(\boldsymbol{z})| < \infty$ for $|t_z| \leq a + 2$.

N.2 The bandwidth vector $\boldsymbol{h}$ is of order $O(n^{-\alpha})$, $1/(4a) < \alpha < (\xi - 3)/4q(\xi + 6)$ such that $a/q > (\xi - 3)/(\xi + 6)$ with ξ from [B.3] b).

Note that as Rodríguez-Póo, Sperlich, Vieu (2003) we consider here the use of higher order kernels to allow for higher dimensions of $\boldsymbol{t}$. Else, one could substitute conditions [N.1], [N.2] by the ones of SW92 in Lemma 8 and 9.

References

Battese, G.E., R.M. Harter and W.A. Fuller. (1988) An Error Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**: 28-36.

Beran, R. (1986) Comment on "Jackknife, Bootstrap and Resampling Methods in Regression Analysis" by C.F.J. Wu. *The Annals of Statistics*, **14**: 1295-1298.

Breslow, N.E. and D.G. Clayton (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of American Statistical Association*, **88**: 9-25.

Butar, F.B. and P. Lahiri (2003) On Measures of Uncertainty of Empirical Bayes Small Area Estimators. *Journal of Statistical Planning and Inference*, **112**: 63-76.

Carroll, R.J., J. Fan, I. Gijbels, and M.P. Wand (1997) Generalized Partially Linear Single-index Models. *Journal of the American Statistical Association*, **92**: 477-489.

Das, K., J. Jiang and J.N.K. Rao (2004) Mean Squared Error of Empirical Predictor. *The Annals of Statistics*, **32**: 818-840.

Diggle, P.J., J.A. Tawn and R.A. Moyeed (1998) Model-based Geostatistics (with discussion). *Applied Statistics*, **47**: 299-350.

Fahrmeir, L. and G. Tutz (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics, 2nd edition, Springer-Verlag: New York.

Fan, J., N. Heckman, and M.P. Wand (1995) Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association*, **90**: 141-150.

Fay, R.E. and R.A. Herriot (1979) Estimates of Income for Small Places. An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**: 269-277.

Ghosh, M. and J.N.K. Rao (1994) Small Area Estimation: an Appraisal. *Statistical Science*, **9**: 55-93 (with discussion).

Ghosh, M., D. Natarajan, T.W.F. Stroud and B.P. Carlin (1998) Generalized Linear Models for Small-Area Estimation. *Journal of the American Statistical Association*, **93**: 273-282.

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales and L. Santamaría (2005) Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters Under a Logistic Mixed Model. *Submited*.

Härdle, W., S. Huet, E. Mammen, and S. Sperlich (2004) Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory*, **20**: 265-300.

Härdle, W. and E. Mammen (1993) Testing Parametric Versus Nonparametric Regression. *The Annals of Statistics*, **21**: 1926-1947.

Härdle, W., E. Mammen, and M. Müller (1998) Testing Parametric Versus Semiparametric Modeling in Generalized Linear Models. *Journal of the American Statistical Association*, **93**: 1461-1473.

Jiang, J. (2003) Empirical Best Prediction for Small Area Inference Based on Generalize Linear Mixed Models *Journal Statistical Planning and Inference*, **111**: 117-127.

Jiang, J. and P. Lahiri (2001) Empirical Best Prediction for Small Area Inference with Binary Data *Ann. Inst. Statist. Math.*, **53**: 217-243.

Jiang, J., P. Lahiri and C. Wu (2001) A Generalization of the Pearson χ^2 Goodnes-Of-Fit Test with Estimated Cell Frequencies *The Indian Journal of Statistics*, **63**: 260-276.

Kuhn, E. and M. Lavielle (2005) Maximum Likelihood Estimation in Nonlinear Mixed Effects Models. *Computational Statistics and Data Analysis*, in press

Laird, N.M. and J.H. Ware (1982) Random-Effects Models for Longitudinal Data. *Biometrics*, **38**: 963-974.

Lejeune, M. (1985) Estimation non-paramétrique par noyaux: régression polynomiale mobile. *Rev. Statist. Appl.*, **33**: 43-67.

Liu, R.Y. (1988) Bootstrap Procedures under some Non-I.I.D. Models. *The Annals of Statistics*, **16**: 1696-1708.

Malec, D., J. Sedransk, C.L. Moriarity and F.B. LeClere (1997) Small Area Inference for Binary Variables in the National Health Interview Survey. *Journal of the American Statistical Association*, **92**: 815-826.

Mammen, E. (1992) When Does Bootstrap Work: Asymptotic Results and Simulations. *Lecture Notes in Statistics*, **77**, Berlin: Springer-Verlag.

Mammen, E., S. van de Geer (1997) Penalized Quasi-Likelihood Estimation in Partial Linear Models. *The Annals of Statistics*, **25**: 1014-1035.

McCullagh, P. and J.A. Nelder (1989) *Generalized Linear Models*, London: Chapman and Hall.

McCulloch, C.E. and S.R. Searle (2001) *Generalized, Linear, and Mixed Models*, John Wiley & Sons, Inc.

Newey, W.K. (1990) Semiparametric Efficiency Bounds, *Journal of Applied Econometrics*, **5**: 99-135.

Newey, W.K. (1994) The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, **62**: 1349-1382.

Prasad, N.G.N. and J.N.K. Rao (1990) The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, **85**: 163-171.

Pinheiro, J.C. and D.M. Bates (2000) *Mixed-Effects Models in S and S-Plus*. New-York. Springer-Verlag.

Rao, J.N.K. (2003) *Small Area Estimation*. John Wiley and Sons, Inc., New-York.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2005) Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects. *Journal of Econometrics*, in press

Robinson, P. (1988) Root-N-Consistent Semiparametric Regression. *Econometrica*, **56**: 931-954.

Rodríguez-Póo, J.M., S. Sperlich, and P. Vieu (2003) Semiparametric Estimation of Weak and Strong Separable Models. *Econometric Theory*, **19**: 1008-1039

Rodríguez-Póo, J.M., S. Sperlich, and P. Vieu (2005) An Adaptive Specification Test for Semiparametric Models. *Working Paper, University Carlos III de Madrid, Spain.*

Ruppert, D., M.P. Wand, and R.J. Carroll (2003) *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Searle, S.R., G. Casella and C.E. McCulloch (1982) *Variance Components*. New-York. Wiley.

Severini, T.A. and J.G. Staniswalis (1994) Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association*, **89**: 501-511.

Severini, T.A. and W.W. Wong (1992) Profile Likelihood and Conditionally Parametric Models. *The Annals of Statistics*, **4**: 1768-1802.

Shao, J. and D. Tu (1995) *The Jackknife and Bootstrap*. New-York. Springer.

Skrondal, A. and S. Rabe-Hesketh (2005) Generalized Linear Latent and Mixed Models with Composite Links and Exploded Likelihoods. *http://www.gllamm.org/composite_conf.pdf*

Spokoiny, V. (2001) Data-driven Testing the Fit of Linear Models. *Mathematical methods of statistics*, **10**, 465-497.

Verbeke, G. and G. Molenberghs (2000) *Linear Mixed Models for Longitudinal Data*. New-York. Springer-Verlag.

Vonesh, E.F. and V.M. Chinchilli (1997) *Linear and Nonlinear Models for the Analysis of Repeated Measures*. New-York. Marcel Dekker.

Wand, M.P. (2003) Smoothing and Mixed Models. *Computational Statistics*, **18**: 223-249.

Wu, C.F.J. (1986) Comment on "Jackknife, Bootstrap and Resampling Methods in Regression Analysis". *The Annals of Statistics*, **14**: 1261-1350.

Zhu, Z. and W.K. Fung (2004) Variance Component Testing in Semiparametric Mixed Models. *Journal of Multivariate Analysis*, **91**: 107-118.