INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/22

GOODNESS-OF-FIT TEST IN PARAMETRIC MIXED-EFFECTS MODELS BASED ON THE ESTIMATION OF THE ERROR DISTRIBUTION

GONZÀLEZ-MANTEIGA, W., MARTINEZ MIRANDA, M.D. and I. VAN KEILEGOM

Goodness-of-fit Test in Parametric Mixed-Effects Models based on the Estimation of the Error Distribution

Wenceslao González Manteiga

University of Santiago de Compostela, Spain

email: wenceslao.gonzalez@usc.es

and

María Dolores Martínez Miranda

University of Granada, Spain

email: mmiranda@ugr.es

and

Ingrid Van Keilegom

Université catholique de Louvain, Belgium

email: ingrid.vankeilegom@uclouvain.be

Summary: We address the problem of checking the adequacy of a parametric functional form for the fixed effects function in mixed effects models. We propose a test based on the distance between the empirical distribution of the parametric residuals calculated under the null hypothesis and those calculated under the alternative. The proposed method is an extension of the one introduced by Van Keilegom, González-Manteiga and Sánchez-Sellero (2008) for cross-sectional independent data i.e. when no random effects are present in the regression model. This formal test is combined for exploratory data analysis purposes with the graphical tool SiZer Map, and applied to longitudinal data analysis.

Key words: Bootstrap; Empirical distribution of the residuals; Goodness-of-fit; Kernel estimation; Mixed models; SiZer Map.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Mixed effects models assume a flexible covariance structure which allows for non-constant correlation among the observations. These models have become very popular and suitable for many practical situations. A mixed effects model (or simply mixed model) assumes that the covariates may involve a mix of fixed and random effects. While the fixed effects function describes the relationship between the covariates and the response for all the observations (population model), the random effects are specific to clusters or subjects within a population. Such specification allows to model properly several interesting problems related to e.g. longitudinal data, repeated measurements, clustered data and small area estimation problems.

The most popular parametric mixed effects models are linear mixed models (LMMs) or generalized linear mixed models (GLMMs), which can be described as follows:

$$g\left(\mathrm{E}\left[Y_{ij}|X_{ij}, b_i\right]\right) = m(X_{ij}) + b_i^t Z_{ij}. \tag{1}$$

Here, $g(\cdot)$ is a known link function, X_{ij} is a random vector of covariates of dimension k, Z_{ij} is a subvector of $(1, X_{ij}^t)^t$ of dimension r, the function m represents the fixed effects and b_i is a r-dimensional vector of mean zero corresponding to the random effects (which are introduced in the model in a linear way). When g is the identity and $m = m_\theta$ with $\theta \in \Theta$ (with Θ being a parametric space) we have the LMMs, or the nonlinear mixed effects models (NLMMs). More general NLMMs arise when m_θ is a parametric nonlinear function, but we also allow for a nonlinear random effects component in (1).

The parametric assumption becomes a simplification for both theoretical and computational aspects, but also it provides valuable intuition and interpretation in real data applications. In this sense it is of substantial interest to test the adequacy of simple parametric mixed models. There are different approaches to test the assumptions in model (1) with respect to the fixed effects given by the m function, or with respect to the distribution of the random

effects, typically considered as gaussian in the literature. To understand the recent advances in the specification about m in model (1), we must go back to the early nineties, where we find the beginning of a large amount of contributions for models without random effects. Different methods have been developed in the last twenty years (see González-Manteiga and Crujeiras, 2011, for a review on this topic). These methods are based on a) the comparison between nonparametric and parametric estimations under the assumption of the model; b) generalized likelihood ratio tests; or c) the empirical distribution of the residuals. Contributions for the case with random effects are more recent and scarce. Zhang and Lin (2003) consider a test about a semiparametric additive mixed model (SAMM), where m belongs to the class of additive models. In particular for the case where one additive component is linear and the other one is nonparametric, they designed a goodness-of-fit test for polynomial regression in the nonparametric component. The authors assume clustered gaussian and non-gaussian data and the test is based on nonparametric estimation by smoothing splines. Later, in the same direction, but for generalized semiparametric additive models, where m is a partial linear model, Lombardía and Sperlich (2008) designed a goodness-of-fit test for the nonparametric component, testing the assumption of linearity using kernel smoothing. See also the paper of Sperlich and Lombardía (2010), which is motivated by the small area estimation problem, and also Henderson, Carroll and Lin (2008) for a quite close test. In a different direction, and avoiding smoothing techniques, Lin, Wei and Ying (2002), Pan and Lin (2005) and more recently Sánchez, Houseman and Ryan (2009), provide omnibus tests for checking the adequacy of LMMs and GLMMs, based on cumulative sums of the residuals with respect to the covariates or the predicted values. In all these papers a random effects component is present in the model, and estimation is done by extending methods that were previously introduced in the nineties for the simpler model, in which the regression function is only specified by a fixed effects function. On the other hand, inference about the assumptions

made for the random effects in model (1), for example the assumption about the gaussian distribution of the b_i's, has been considered recently in many papers (see for example the paper of Claeskens and Hart, 2009, for an extensive review, or Meintanis and Portnoy, 2011, for a very recent reference in the topic). In this context the calibration of the distribution of the used statistics is crucial to provide a test of the desired α level. Bootstrap approximations have been widely used for this task, as can be seen in many of the references cited earlier. Many of the resampling techniques used must be adapted to mimic the model which assumes random effects.

In this paper we propose a test based on the empirical distribution of the residuals, under the null and the alternative hypotheses, which is an extension of the test introduced by Van Keilegom, González-Manteiga and Sánchez-Sellero (2008) for testing about the fixed effects function in model (1) with $k = 1$ and g being the identity function, but ignoring the random effects component. This is an option not considered before in the literature related with mixed models and it represents a suitable extension, which corresponds to option c) following the classification of goodness-of-fit tests defined above. This kind of methods is very powerful, because they can detect alternatives to the null hypothesis at the parametric rate $n^{-1/2}$. But they can also be combined in a nice way for exploratory data analysis purposes. In fact we describe in this paper how to perform an exploratory analysis of the residuals using the graphical tool SiZer Map (Chaudhuri and Marron, 1999). The strategy is based on a SiZer analysis of the residuals, which follows the spirit of the methods in González-Manteiga et al. (2008), and is much simpler than plotting conventional residual plots. As was pointed out by Lin et al. (2002), conventional residual diagnostics based on plots of the raw residuals, are highly subjective. In fact these techniques rely on the human mind and eye to decide whether or not a specific pattern exists in the residuals, or whether or not the model fits correctly the data. A major problem arises when dealing with correlated data as in this paper. Diagnostics

for mixed effects models are harder to interpret due to the presence of random effects and different covariance structures. It is well-known that the correlation among observations in the same group tends to hide the actual patterns. Some authors recommended to plot separately residuals for each group, which is a complex and hard task. Such a procedure was followed by Lin et al. (2002) and Pan and Lin (2005) for model checking purposes. In this paper, we use the SiZer Map, which allows a local exploration of these residuals. In fact the SiZer Map is a simpler, faster and very intuitive procedure to diagnose deviations from the assumed parametric structure. Through the SiZer analysis the data analyst is able to extract the necessary information, not only for model checking purposes, but also for the specific task of looking for a proper parametric functional for the fixed effects function. Note that one of the major challenges in modeling longitudinal data is the correct specification of the time trend of the response. In this regard the methods proposed in this paper will be shown to provide a suitable tool both for exploratory analysis, but also for confirmatory purposes. Using two examples from the context of longitudinal data analysis, we show how to perform a powerful residual analysis to explore the data and to look for some suitable parametric fit. After this preliminary and exploratory analysis we are able to confirm the results through formal testing methods proposed in this paper.

The rest of the paper is organized as follows. In Section 2 we present the model and the estimation method. The testing methods are described in Section 3. Section 4 describes the results of some simulation studies and real data analysis. Some general conclusions are given in Section 5. The theoretical developments are deferred to the Appendix.

2. Model and estimation

In this paper we consider the following semiparametric one-way model:

$$Y_{ij} = m(X_{ij}) + b_i^t Z_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, n_i;\ i = 1, \ldots, q, \tag{2}$$

where q is the number of levels in the model and $n = \sum_{i=1}^{q} n_i$ is the total number of observations. The covariate X_{ij} is a k-dimensional random vector, and Z_{ij} is a subvector of $(1, X_{ij}^t)^t$ of dimension r. We assume that all X_{ij}'s are identically distributed, and $X_1, \ldots, X_q$ are mutually independent, with $X_i = (X_{i1}, \ldots, X_{in_i})^t$. Also, we assume that the errors $\epsilon_{11}, \ldots, \epsilon_{qn_q}$ are i.i.d. normal random variables with mean zero and variance σ^2, and that $\mathrm{E}(\epsilon_{ij}|X_{ij}) = 0$. The random effects $b_1, \ldots, b_q$ are i.i.d. r-dimensional normal random variables with mean zero and covariance matrix V_b. The matrix V_b quantifies the within-subject variation. Further, assume that b_i and $X_{i'}$ are independent for all $i, i' = 1, \ldots, q$, and so in particular b_i is independent of X_i. Moreover, $\mathrm{Cov}(b_i, \epsilon_{i'j}|X_i, X_{i'}) = 0$ for all $i, i' = 1, \ldots, q$ and $j = 1, \ldots, n_{i'}$. The gaussian assumption made for the random effects and the errors, could be relaxed (Severini and Staniswalis, 1994, Lin and Carroll, 2000). However we introduce it to develop simpler likelihood-based inferences for the function m, but also for the estimation of the variances, V_b and σ^2.

Since the observations are only dependent if they come from the same individual, it is suitable to write the previous model using matrix notation. Thus we write model (2) first by stacking the observations at the individual level, i.e. $Y_i = m(X_i) + Z_i b_i + \epsilon_i$, $i = 1, \ldots, q$, where Z_i is the $(n_i \times r)$ matrix with rows $Z_{i1}^t, \ldots, Z_{in_i}^t$, $Y_i = (Y_{i1}, \ldots, Y_{in_i})^t$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^t$. Note that ϵ_i has diagonal covariance matrix $\sigma^2 I_{n_i}$. Also, the variance of Y_i conditionally on X_i is given by $V_i = Z_i V_b Z_i^t + \sigma^2 I_{n_i}$. Finally, the model can be compactly written for the whole set of n observations by $Y = m(X) + Zb + \epsilon$, where $Y = (Y_1^t, \ldots, Y_q^t)^t$, X is the $(n \times k)$ matrix with rows X_i^t, and Z is the $(n \times qr)$ matrix with diagonal blocks Z_i. Here, the variance of Y conditionally on X is given by $V = ZBZ^t + \sigma^2 I_n$, where B is the matrix with diagonal blocks V_b.

Under a parametric mixed effects model the fixed effects function m is commonly estimated by the (global) likelihood method. Under the above conditions the density of Y_i, conditionally

on X_i, is a Normal with mean $m(X_i) = (m(X_{i1}), \ldots, m(X_{in_i}))^t$ and covariance matrix V_i. Then, the global log-likelihood of Y (conditional on X) is given by

$$\ell(m, V_b, \sigma^2) \;\; = -\frac{1}{2}\sum_{i=1}^{q}\left\{[Y_i - m(X_i)]^t\, V_i^{-1}\,[Y_i - m(X_i)] + \log|V_i| + 2n_i\log(2\pi)\right\}. \qquad (3)$$

Here we are interested in estimating $m(x)$, for any fixed x in $\mathcal{R}^k$, by a local constant approach. We could also use local linear estimation, but local constant is simpler and enough for the purpose of this paper. We consider an estimator which is an extension of the common Nadaraya-Watson estimator for i.i.d. data. The estimator is derived using a local likelihood approach. Note that for a given estimation point x and supposing that X_{ij} is close to x, we have that $m(X_{ij}) \approx m(x) := \beta_x$ $(j = 1, \ldots, n_i)$. Model (2) can then be locally approximated by the simple linear mixed model $Y_i = 1_{n_i}\beta_x + b_i Z_i + \epsilon_i$, with 1_{n_i} being the n_i-dimensional vector of ones, for each $i = 1, \ldots, q$. The local log-likelihood can be defined from the global log-likelihood (3), by introducing some kernel weights $K_h(X_{ij} - x) = h^{-1}K((X_{ij} - x)/h)$ for each observation, where h is an appropriate bandwidth sequence. However, in the presence of within-subject correlation in the model, this should be done using blocks, i.e. considering each of the q independent components of the global log-likelihood. As Lin and Carroll (2000) pointed out, the way of introducing the kernel weights into the individual components is problem specific, and different ways provide estimators with different theoretical and practical properties (see for example González-Manteiga, Lombardía, Martínez-Miranda and Sperlich (2012) for a recent discussion on this topic). In this paper we follow the approximation by Park and Wu (2006), which is simpler and has good finite sample properties. In particular, Park and Wu (2006) define the local log-likelihood by

$$\ell_{loc}(\beta_x, V_b, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{q}\left\{[Y_i - 1_{n_i}\beta_x]^t\, W_{ih,x}^{1/2} V_i^{-1} W_{ih,x}^{1/2}\,[Y_i - 1_{n_i}\beta_x] + \log|V_i| + 2n_i\log(2\pi)\right\}, \qquad (4)$$

which involves diagonal matrices, $W_{ih,x}$, with elements $K_h(X_{ij} - x)$ for each independent block $i = 1, \ldots, q$. The derived estimator can be explicitly written as

$$\widehat{m}(x) = \widehat{\beta}_x = \left(\sum_{i=1}^{q} 1_{n_i}^t W_{ih,x}^{1/2} V_i^{-1} W_{ih,x}^{1/2} 1_{n_i} \right)^{-1} \sum_{i=1}^{q} 1_{n_i}^t W_{ih,x}^{1/2} V_i^{-1} W_{ih,x}^{1/2} \, Y_i. \tag{5}$$

Note that if there is no within-subject correlation (or it is ignored) the derived estimator for $m(x)$ becomes the common Nadaraya-Watson estimator for independent data. Hereafter we simplify the notation to $\beta_x = \beta$ and $W_{ih,x} = W_{ih}$ when no confusion is possible.

Now, since the estimator of $m(x)$ in (5) depends on the variances V_b and σ^2, which are unknown in general, the corresponding feasible (or also called empirical) estimator at each x can be derived using a three-step procedure. The procedure accomplishes the estimation of β_x and the variances V_b and σ^2, using simultaneously the local and global scores (4) and (3), and is formulated as follows:

Step 1. For arbitrary values of σ^2 and V_b define the estimator of $m(x)$ for any x by $\widehat{m}_{\sigma,V_b}(x) = \widehat{\beta}_{\sigma,V_b}$, which is the maximizer of $\ell_{loc}(\beta, V_b, \sigma^2)$, and has the explicit expression given in equation (5). Calculate $\widehat{m}_{\sigma,V_b}(X_{ij})$ for all observed X_{ij}'s.

Step 2. Define the estimator of (V_b, σ^2) as the maximizer $(\widehat{V}_b, \widehat{\sigma}^2)$ of $\ell(\widehat{m}_{\sigma,V_b}, V_b, \sigma^2)$, from the global log-likelihood $\ell(m, V_b, \sigma^2)$ given in (3).

Step 3. Finally, define $\widehat{m}(x) = \widehat{\beta}_{\widehat{\sigma}^2,\widehat{V}_b}$.

From $\widehat{m}(X_i) = (\widehat{m}(X_1), \ldots, \widehat{m}(X_q))^t$, the random effects b_i can be predicted using standard methods for linear mixed models by $\widehat{b}_i = \widehat{V}_b Z_i^t \widehat{V}_i^{-1}(Y_i - \widehat{m}(X_i))$, for $i = 1, \ldots, q$.

Note that the above three-step procedure is suitable for model (2), where σ^2 and V_b are global parameters, and β is the only local parameter. However the method can be easily adapted depending on which parameters in the model are global and which are local. Specifically we define the following variations of the basic model: [1] If we would assume that the model is heteroscedastic, i.e. $\mathrm{Var}(\epsilon_{ij}|X_{ij}) = \sigma^2(X_{ij})$, then we could estimate this variance in the first step of the above procedure. [2] We could also consider that m belongs to a parametric class $\mathcal{M} = \{m_\theta : \theta \in \Theta\}$, in which case the estimation of m would move to the

second step of the procedure. Finally, note that the three-step procedure which we propose, can be applied in general to any local log-likelihood and hence to any local estimator of m proposed in the literature. Also, if the error is not supposed to be normal, steps 1 and 2 could be based on a quasi-likelihood and on general estimating equations (GEE), see e.g. Liang and Zeger (1986) and Lin and Carroll (2000).

3. Goodness-of-fit testing

In this section we propose a test for the parametric null hypothesis about the fixed effects function m, formulated as follows:

$$\begin{aligned} H_0 &: m \in \mathcal{M} = \{m_\theta : \ \theta \in \Theta\} \\ H_1 &: m \notin \mathcal{M}. \end{aligned} \tag{6}$$

The proposed test extends the test introduced by Van Keilegom et al. (2008) for cross-sectional independent data. It is based on the distance between the empirical distribution of the parametric residuals (under H_0) and those calculated under the alternative. The first step to define the test is the characterization of the null hypothesis (see Theorem 2.1 in Van Keilegom et al., 2008). When no random effects are present in the regression model, the characterization is based on the comparison between the error distribution calculated under the null and the one calculated under the alternative hypothesis. However, if the regression model involves random effects, two different approaches can be considered. More precisely, we can consider either the errors $U_{ij} := Y_{ij} - m(X_{ij})$, or the errors from the full regression structure, i.e. $\epsilon_{ij} = Y_{ij} - m(X_{ij}) - b_i^t Z_{ij}$. We refer to the errors U_{ij} as the marginal errors, because they arise from the marginal distribution of the response Y_{ij} (conditional on the covariates), and to the errors ϵ_{ij} as the conditional errors, since they arise from the conditional distribution on the random effects b_i and the covariates. Note that inference from conditional errors requires the estimation of both the random and the fixed effects. In the latter case, as Pan and Lin (2005) discussed, the results of the tests can be affected

by a possible misspecification of such random component. Indeed, test procedures based on conditional residuals are testing about the whole regression model, instead of testing about the fixed effects function, which is actually our goal here. For this reason we suggest in this paper a test based on the marginal errors, which is formulated in what follows.

From the true function m, consider the marginal errors $U_{ij} = Y_{ij} - m(X_{ij})$. Note that such errors are not i.i.d. variables, which does not agree with the assumptions in Van Keilegom et al. (2008). In order to remove the within-subject correlation such errors have to be standardized. In this aim we consider the block transformation $V^{-1/2}U$, based on the whole vector $U = (U_1^t, \ldots, U_q^t)^t$, with $U_i^t = (U_{i1}, \ldots, U_{in_i})$. Note that the elements of $V^{-1/2}U$ are then i.i.d. variables and therefore we can follow similar arguments as Van Keilegom et al. (2008) to formulate the test. Let us denote the elements of the transformed vector of errors by U'_{ij}, and note that they all have the same distribution as a generic variable U'. Analogously, consider the transformed errors based on the parametric regression function (under the null hypothesis), i.e. $V^{-1/2}(Y - m_\theta(X))$, with elements denoted by $U'_{ij,0}$, which are also i.i.d. variables with the same distribution as U'_0. From these definitions the characterization of the null hypothesis in problem (6) follows by using arguments from Van Keilegom et al. (2008). In fact we prove in the Appendix that the null hypothesis in (6) holds if and only if the standardized marginal errors U' and U'_0 have the same distribution.

Now the next step is to estimate the distribution of the random variables U' and U'_0. Note that this can be done by considering the estimators of m and the variance V resulting from the three-step estimation methods presented in Section 2. Denote these estimators by $\widehat{m}$ and $\widehat{V}$, where $\widehat{V}$ is the block diagonal matrix with blocks $\widehat{V}_i = Z_i \widehat{V}_b Z_i^t + \widehat{\sigma}^2 I_{n_i}$. Then, we can estimate the distribution of U' by the empirical distribution of the standardized nonparametric marginal residuals, $\widehat{U}'_{ij}$, given by

$$\widehat{F}_{u'}(t) = n^{-1} \sum_{i=1}^{q} \sum_{j=1}^{n_i} I\left(\widehat{U}'_{ij} \leq t\right).$$

Here, the $\widehat{U}'_{ij}$'s denote the elements of the vector $\widehat{V}^{-1/2}(Y - \widehat{m}(X))$. Also, the distribution of U'_0 is estimated by the empirical distribution of the parametric marginal residuals $\widehat{U}'_{ij,0}$, i.e.

$$\widehat{F}_{u'_0}(t) = n^{-1} \sum_{i=1}^{q} \sum_{j=1}^{n_i} I\left(\widehat{U}'_{ij,0} \leq t\right),$$

where the $\widehat{U}'_{ij,0}$'s are defined as the elements of the vector $\widehat{V}^{-1/2}(Y - \widehat{m}_0(X))$. The estimator $\widehat{m}_0(\cdot)$ is defined analogously to $\widehat{m}(\cdot)$, but replacing the observed responses Y_{ij} by the parametric estimations of the fixed effects, $m_{\widehat{\theta}}(X_{ij})$, or also by the whole estimated effects $m_{\widehat{\theta}}(X_{ij}) + \widehat{b}_i Z_{ij}$, for some suitable estimator $\widehat{b}_i$ (for instance the linear predictor which was provided in Section 2).

Finally, we calibrate the distance between the empirical distributions $\widehat{F}_{u'}$ and $\widehat{F}_{u'_0}$, using Kolmogorov-Smirnov (KS) and Cramér-von Mises (CM) type statistics, defined as

$$T_{n,KS} = n^{1/2} \sup_{-\infty<t<\infty} \left|\widehat{F}_{u'}(t) - \widehat{F}_{u'_0}(t)\right| \quad \text{and} \quad T_{n,CM} = n \int \left[\widehat{F}_{u'}(t) - \widehat{F}_{u'_0}(t)\right]^2 d\widehat{F}_{u'_0}(t).$$

In the Appendix we develop the asymptotic distribution of $T_{n,KS}$ and $T_{n,CM}$. Since these limiting distributions are rather complicated, we suggest to use bootstrap methods to approximate the critical values of these test statistics. More precisely, we define a bootstrap algorithm suitable for the assumed mixed model, which can be described along the following four steps:

Step 1. Calculate the estimator $\widehat{m}$ of the fixed effects function m, the estimators $\widehat{V}_b$ and $\widehat{\sigma}^2$ of the variances V_b and σ^2, and also the estimator $m_{\widehat{\theta}}$ of the parametric regression function (under H_0). These estimators are derived using the three-step method in Section 2.

Step 2. Generate bootstrap conditional errors $\epsilon^\star_{ij}$ and bootstrap random effects $b^\star_i$, independently from the standard gaussian distribution.

Step 3. Under the null hypothesis the bootstrap responses are constructed by $Y^\star_{ij} = m_{\widehat{\theta}}(X_{ij}) + \widehat{V}_b^{1/2} b^\star_i + \widehat{\sigma}\epsilon^\star_{ij}$ $(j = 1, \ldots, n_i,\ i = 1, \ldots, q)$. Then, the bootstrap sample is given by $\{(X_{ij}, Z_{ij}, Y^\star_{ij}),\ j = 1, \ldots, n_i\ i = 1, \ldots, q\}$.

Step 4. Calculate the bootstrapped test statistics $T^{\star}_{n,KS}$ and $T^{\star}_{n,CM}$ from the bootstrap sample generated in the previous step.

Finally, the quantiles of the distribution of $T^{\star}_{n,KS}$ and $T^{\star}_{n,CM}$ can be approximated by Monte Carlo simulation repeating steps 2-4 in the above algorithm B times.

Note that the defined resampling scheme could also be defined in a more general way in the absence of the gaussian assumption. In that case both the conditional residuals and the random effects in Step 2 above could be generated from the smoothed empirical distribution of the residuals (see Van Keilegom et al., 2008).

4. Empirical study

4.1 *Simulation experiments*

Here we investigate the finite sample performance of the test proposed in the previous section. We simulate the model $Y_{ij} = m(X_{ij}) + b_i + \epsilon_{ij}$, which is a particular case of model (2) with unidimensional X_{ij} and $Z_{ij} = 1$. Our aim is to test the linear mixed model. To calculate the size of the test we simulated the fixed effect function by $m(X) = 1 + X$. The power of the test is evaluated by considering two different alternatives. The first alternative consists of contaminating the null hypothesis with a sinusoidal functional, namely we simulate $m(X) = 1 + (1 - a)X + a\sin(\pi X)$, with $a = 0.1$ and 0.2. The second alternative is harder to detect by test procedures and allows to check the power against quadratic terms by simulating $m(X) = 1 + (1 - a)X + aX^2$, for $a = 0.1$ and 0.2. We consider two different designs to generate the covariate X, namely a Uniform on $[0, 2]$ and a Normal with zero mean and variance 0.6. The random effects b_i and the errors ϵ_{ij} are generated independently from a Normal distribution with mean 0 and standard deviations σ_b and σ, respectively. We consider three situations. In the three cases $\sigma = 0.3$ but $\sigma_b = 0.3, 0.6, 1$ for cases 1, 2 and 3, respectively. We simulate samples of size $n = 100, 200$ and 500. For $n = 100$ we consider

$q = 10$ groups of sizes n_i equal to $5, 7, 8, 9, 10, 10, 11, 12, 13$ and 15. For $n = 200$ we increase q to 20 and repeat the previous sizes twice. Finally for $n = 500$ the number of groups is $q = 50$ and we repeat the sequence of sizes five times.

The proposed test involves the local constant estimator presented in Section 2. Such estimator has been calculated from the three-step estimation method, using the Epanechnikov kernel and considering different values of the bandwidth h. More precisely, we defined $h = h_0 n^{-3/10}$, and considered $h_0 = 2, 2.5, 3$. For each sample size, design and case (1 to 3) we performed 1000 replications of the model under the null hypothesis of linearity ($a = 0$) and also under the alternatives. At each replication we solved the testing problem considering the Kolmogorov-Smirnov (KS) and also the Cramér-von Mises (CM) statistics defined in Section 3. Also, we calibrated the test statistics using the suggested bootstrap algorithm with $B = 1000$ bootstrap samples.

Table 1 reports the percentages of rejection and the average p-values under the null hypothesis of linearity. The considered nominal level is $\alpha = 0.05$. We considered the randomized rule to determine the rejection levels for the KS statistics. As it is expected the size varies slightly with the bandwidth, but in general the test attains the level quite well for both the CM and the KS statistics. The results are even better when the covariate has been simulated under a Normal design.

[Table 1 about here.]

The power of the test has been evaluated for the two considered alternatives. The observed percentages of rejections under the nominal level of $\alpha = 0.05$ are shown in Table 2. The quadratic alternative is harder to detect under a Uniform design, but the results improve substantially under the Normal design. Again we can see some slight variations for different bandwidths, but they do not change the conclusions. Note that both alternatives are relatively hard to detect for case 3. In fact the total variability of the response is quite big

for case 3, and comes mainly from the within-subject correlation. However, even in such a situation the proposed test achieves reasonable rejection levels for moderate and high sample sizes.

[Table 2 about here.]

For comparison purposes we implemented the omnibus test proposed by Pan and Lin (2005). This test is based on the cumulative marginal residuals (see Pan and Lin, 2005, for more details and also for the explicit expression of the distribution). In the above described scenario this omnibus test does not achieve the level, except for the higher sample size. Note that the approximation which the authors consider to calibrate the test, is valid for q big enough. In fact, the simulation studies by Pan and Lin (2005) consider a model close to our model (with just a random intercept), but where the considered number of groups is at least 50, and the number of correlated observations is always equal to 3. To do a fair comparison with the test by these authors we considered a second scenario where the number of groups equals $q = 50, 100, 200$ and $n_i = 3$ $(i = 1, \ldots, q)$. Table 3 shows the results obtained under this scenario from the Normal design. Columns labeled as PL show the results by Pan and Lin's test. Under the null hypothesis $(a = 0)$ for each sample size, the first row shows the percentage of rejection and the second row the average p-value. Under the alternatives $(a = 0.1, 0.2)$ the percentage of rejection is given. For each value of n the first row shows the results for the sinusoidal alternative $m_s(X) = 1 + (1 - a)X + a\sin(\pi X)$, and the second row shows the results for the quadratic alternative $m_q(X) = 1 + (1 - a)X + aX^2$. The number of replications to approximate the null distribution is 1000 for both the omnibus test and the bootstrap approximation. The smoothing parameter considered for the CM and the KS test is $h = 3n^{-3/10}$ (other bandwidths provided close results). Note that all the compared tests have similar size. However, even in this situation, the tests proposed in this paper clearly outperform the test by Pan and Lin (2005), which has much less power against both

types of alternatives. The results are even more dramatic when the within-subject correlation increases.

[Table 3 about here.]

4.2 *Applications to longitudinal data*

Now we illustrate the proposed tests with two datasets. We consider longitudinal data previously used in other papers, which are related to estimation and model-checking techniques about the fixed effects function m. First, we show how to perform a powerful residual analysis to explore the data and to look for some suitable parametric fit. After this preliminary analysis we are able to confirm the results through the formal tests proposed in this paper.

4.2.1 *SiZer analysis of the residuals.* The SiZer Map provides the scale and location space which González-Manteiga et al. (2008) considered to perform SiZer analysis of residuals for model-checking purposes. Such SiZer analysis consists of plotting and interpreting the maps of the estimated residuals under the null hypothesis. Following the spirit of conventional residual diagnostics, if the tested model is adequate then the residuals under the null hypothesis should be small without any significant characteristics. Therefore, we should inspect these residuals and exclude the existence of significant features, otherwise the tested model should be rejected. We can reformulate this approach under the mixed models framework and derive a strategy consisting of the following steps: [1] Define the testing problem and estimate the model under the null hypothesis. Denote by $m_{\widehat{\theta}}$ the resulting parametric estimator. [2] Calculate the estimated marginal residuals under the null hypothesis and remove the existing within-subject correlation. Denote these estimated standardized residuals by $\widehat{U}'_{\widehat{\theta},\widehat{V}} = (\widehat{U}'_{ij;\widehat{\theta},\widehat{V}}; j = 1, \ldots, n_i, i = 1, \ldots, q)^t = \widehat{V}^{-1/2}(Y - m_{\widehat{\theta}}(X))$, with $\widehat{V}$ being a consistent estimator of the marginal covariance matrix of the responses. [3] Construct the SiZer Map from the dataset $\{(X_{ij}, \widehat{U}'_{ij;\widehat{\theta},\widehat{V}}); j = 1, \ldots, n_i, i = 1, \ldots, q\}$. The presence of significant features (peaks,

valleys, inflection points, etc.) or significant increasing or decreasing patterns would mean the non-adequacy of the tested parametric model.

The defined strategy is analogous to the one suggested by González-Manteiga et al. (2008) except for the normalization of the residuals in the second step. SiZer relies strongly on the assumption of uncorrelated observations in order to reveal significant features, i.e. those which are really present in the data. If some correlation is present in the data some of these characteristics could be hidden and other false patterns could also appear due to the dependence structure. The normalization step avoids such problems and agrees with the requirements of the SiZer Map, as originally proposed by Chaudhuri and Marron (1999) for regression models.

4.2.2 *AIDS clinical trial.* The first dataset consists of CD4 counts data from an AIDS clinical trial study to evaluate the efficacy of Zidovudine (AZT) in treating patients with mild symptomatic HIV infection. These data have also been analyzed by Lin et al. (2002) and Pan and Lin (2005) to illustrate their model-checking methods. Specifically they performed a testing analysis about the functional form of the covariates. Our aim here is slightly different, because we aim to test about the whole fixed effects function. The study enrolled 711 patients with 361 randomized to AZT and 350 to placebo. Experts on this type of data suggest that the CD4 counts among the placebo patients tend to decline monotonically over the entire study period, whereas for the AZT patients they tend to rise for the first few weeks and then decline over time. Hence, it seems reasonable to describe the time trend with a linear function for the placebo group and with a higher order polynomial function for the AZT patients. The spaghetti plots for both groups are shown in Figure 1.

[Figure 1 about here.]

From these plots it is difficult to extract any useful information, because the individual CD4 cell counts are quite noisy over time. However, the nonparametric estimator proposed

in Section 2 is able to capture the underlying structure in the data. This kernel estimator is shown by the black solid curve in Figure 1. We calculated this estimator by estimating model (2) with the response, Y_{ij}, being the CD4 cell counts, and with the covariate, X_{ij}, being the time (in weeks). Also, we considered the simple covariance structure with only a random intercept $Z_{ij} = 1$. We tried other more complex covariance structures, but they did not modify substantially our results. We considered a bandwidth of 8 weeks for both groups and used the Epanechnikov kernel. In this example the bandwidth has been chosen by eye taking into account the variability in the data. Such choice is sufficient for our purposes, however, we direct the reader to the recent paper by González-Manteiga et al. (2012) for proper bandwidth selection methods for this type of kernel estimators. Looking at the nonparametric estimator it seems that the underlying structure in the placebo, respectively AZT, group could be modeled by a linear, respectively quadratic or cubic, polynomial. However this discussion is heavily depending on the degree of smoothness considered in the kernel estimator. To perform a proper exploratory analysis we recommend to perform a SiZer analysis of the residuals described in Subsection 4.2.1. By visualizing the problem into a scale and location space we avoid the tricky issue of choosing a "correct" smoothing parameter and provide more objective conclusions. We focus from now on on the AZT group and try to assess the experts' judge about the time trend. We look for deviations from the simple parametric linear mixed model. Figure 2 shows the SiZer Map for the residuals from the fitted linear parametric mixed model. The figure consists of three plots: the family plot in the top panel, the Slope SiZer Map (usually called SiZer Map) in the middle and the Curvature SiZer Map (also called SiCon) in the bottom. The family plot shows the smoothed curves at different smoothing levels. The maps in the middle highlight the characteristics of the target curve defined as significant changes on the sign of the first derivative. The conclusions are given using a color language (see the paper by Chaudhuri and Marron (1999) for a further

explanation about the SiZer Map and González-Manteiga et al. (2008) for its application for testing purposes). Specifically, the color blue (black for monochrome versions) means that the derivative is significantly positive, red (dark gray) if the derivative is significantly negative, and purple (light gray) if the derivative is not significantly non-zero. The SiCon Map indicates significant changes in the sign of the second derivative: cyan (black) if the second derivative is negative and therefore concave, orange (dark gray) if it is positive and so convex and green (light gray) if the derivative is not significantly non-zero. In both maps the gray color (lightest gray) is used to indicate smoothing parameter values that are too small for inference purposes. Having this color language in mind we see from the Slope SiZer Map a significant increase (blue color) for the first 5 weeks, and also a significant decrease for the later weeks (from 15 to the end). For the middle weeks SiZer does not show any significant feature for any of the bandwidths. From this information we can conclude that the linear model is not able to describe the time trend. Besides we can guess that the problem with such a model arises in the first weeks, where an increasing pattern in the data is not captured, but also a decreasing pattern in the later weeks. Note that these conclusions agree with the experts' judge about the time trend of the AZT patients (Lin et al. 2002). Clearly to remove the observed pattern in the residuals a higher order polynomial is required. Such a model should involve the necessary curvature in the target function m. In fact the Curvature SiZer (bottom panel) shows that the underlying structure in these residuals is significantly concave (color cyan) up to the 20th week. The tests proposed in this paper confirm these conclusions. As it is expected both the CM and the KS test provide p-values close to zero. These values have been calculated considering $B = 1000$ bootstrap samples and $h = 8$, but similar results are obtained for other bandwidths.

[Figure 2 about here.]

Now we fit a quadratic mixed model to the same data and repeat the SiZer analysis on the new standardized residuals. Figure 3 shows the resulting maps. The Slope SiZer Map shows a significant decrease in the residual pattern between week 10 and 20. This feature can only be visualized by considering proper scales or bandwidths. The information reported by the bottom panel allows to go further in the conclusions. The Curvature SiZer Map points out that a better model should capture a change in the curvature: concavity (cyan color) in the first weeks and convexity (orange color) in the later weeks. Therefore, the quadratic model could be a reasonable model for most of the periods but it could be improved for the middle weeks. Note that such features can only be visualized considering proper scales and therefore it would be very hard just from conventional plots of the residuals. Again we confirm our conclusions with the test proposed in this paper. The resulting p-values are 0.003 using the CM test and 0.005 using the KS test.

[Figure 3 about here.]

Now we consider a bit more complexity in the parametric model and move to a cubic mixed model. The SiZer Maps for the corresponding residuals are shown in Figure 4. The information reported by the maps suggests that the cubic model is a good model for the AZT patients. In fact, neither of the maps reveals any features or any significant patterns or trends in the residuals. Confirmation of such exploratory analysis is obtained by calculating the CM and KS test statistics, which provide p-values equal to 0.332 and 0.148, respectively.

[Figure 4 about here.]

A very similar analysis for the placebo group allows to confirm again the parametric model suggested by the experts, describing a decreasing time trend.

4.2.3 *Progesterone data.* The progesterone dataset comes from a study on early pregnancy loss. These data were considered by Wu and Zhang (2006), among others, to motivate

the nonparametric estimation of the fixed effects function. The dataset consists of two groups: the conceptive progesterone curves (22 menstrual cycles) and the non-conceptive progesterone curves (69 menstrual cycles). Figure 5 shows the spaghetti plots for each group.

[Figure 5 about here.]

From these graphs we can see that it is difficult to find a suitable parametric model describing the data. We use this example to illustrate how the testing methods in this paper and also the SiZer analysis of the residuals can be conducted to confirm such intuition. Since the analysis is similar for both groups we only include the results for the non-conceptive group. In this aim we consider again a simple mixed model to describe the logarithm of the progesterone (Y_{ij}) along the days in the cycle (X_{ij}), with covariance structure specified by only a random intercept. We calculated the nonparametric estimator presented in Section 2 using a bandwidth of 2 days. The resulting estimator is shown in Figure 5. From visual inspection of the data and from the nonparametric estimator it is clear that it is necessary to use at least a quadratic or cubic polynomial to describe the trend. The exploratory analysis through the SiZer Map, however, reveals that neither of these models allows to capture the underlying structure. Figure 6 shows the resulting maps for a parametric cubic fit. Even from such a complex model, many features still underly in the residuals. The Slope SiZer and the SiCon Maps show at each location i.e. day in cycle, how the tested model should be corrected in order to describe the underlying structure. For example the curvature around the second day should be corrected to describe an inflection point, which is not captured by the cubic fit. Also a more correct model should increase faster around such location. A similar behavior is observed around the seventh day, where such a more correct model should decrease faster than the cubic fit. However, using this valuable information it is still very hard to find a suitable parametric model to describe such a shape. The testing methods proposed in this paper confirm this exploratory analysis and do not accept the cubic model as a good model

to describe the log-progesterone along the days in the cycle. The p-values obtained from the CM and the KS statistics are 0.005 and 0.011 respectively (using $h = 2$). Therefore, we recommend to use nonparametric methods like the one proposed in this paper, or the one suggested by Wu and Zhang (2006), which are more flexible to describe the progesterone curves.

[Figure 6 about here.]

5. Conclusions

In this paper we proposed, studied and illustrated formal testing methods to check the adequacy of parametric specifications of the fixed effects in a mixed effects model. The test is an extension of the method by Van Keilegom et al. (2008), it exhibits very good theoretical properties and has an excellent performance in practice. Not many competitors can deal with the models in this paper apart from the omnibus test by Pan and Lin (2005). The simulation experiments described in this paper show that our test clearly outperforms this omnibus test. The good performance and the applicability of our methods are also demonstrated through two data analyses involving biological and biomedical applications. The SiZer Map considered in this paper is a powerful and meaningful tool for testing purposes. The SiZer analysis of the residuals provides a valuable intuition and a nice visualization of the underlying structures, which are complicated to assess in common longitudinal datasets. In this sense we expect that the methods in this paper will be very useful for data analysts and practitioners in many disciplines.

Acknowledgements

The authors thank Dr. Sánchez-Sellero for helpful discussions about the computational issues, Dr. Pan for supplying the AIDS dataset, and the Centro de Servicios de Informática y

Redes de Comunicaciones (CSIRC), Universidad de Granada, for providing the computing time. The maps in this paper were generated using Dr. Marron's MatLab software (`http://www.stat.unc.edu/ faculty/marron/marron_software.html`). This research has been supported by the Spanish "Ministerio de Ciencia e Innovación" MTM2008-03010. Ingrid Van Keilegom also acknowledges support by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy), by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

References

Akritas, M.G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, 28 (3), 549-567.

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94 (447), 807-823.

Claeskens, G. and Hart, J.D. (2009). Goodness-of-fit tests in mixed models. *Test*, 18 (2), 213-239.

González-Manteiga, W. and Crujeiras, R.M. (2011). A general view of the goodness-of-fit tests for statistical models. In: *Modern Mathematical Tools and Techniques in Capturing Complexity* (pp. 3–17). Springer Series in Synergetics.

González-Manteiga, W., Lombardía, M.J., Martínez-Miranda, M.D. and Sperlich, S. (2012). Kernel smoothers and bootstrapping for semiparametric mixed effects models (submitted).

González-Manteiga, W., Martínez-Miranda, M.D. and Raya-Miranda, R. (2008). SiZer Map for inference with additive models. *Statistics and Computing*, 18, 297-312.

Henderson, D.J., Carroll, R.J. and Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, 144, 257-275.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lin, X. and Carroll, R.J. (2000). Nonparametric function estimation for clustered data when predictor is measured without/with error. *Journal of the American Statistical Association*, 95, 520-534.

Lin, D.Y., Wei, L.J. and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, 58, 1-12.

Lombardía, M.J. and Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *Journal of the Royal Statistical Society - Series B*, 70, 913-930.

Meintanis, S.G. and Portnoy, S. (2011). Specification tests in mixed effects models. *Journal of Statistical Planning and Inference*, 141 (8), 2545-2555.

Pan, Z. and Lin, D.Y. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61, 1000-1009.

Park, J.G. and Wu, H. (2006). Backfitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data. *Journal of Statistical Planning and Inference*, 136, 3760-3782.

Sánchez, B.N., Houseman, E.A. and Ryan, L.M. (2009). Residual-based diagnostic for structural equation models. *Biometrics*, 65, 104-115.

Severini, T.A. and Staniswalis, J.G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89, 501-511.

Sperlich, S. and Lombardía, M.J. (2010). Local polynomial inference for small area statistics: estimation, validation and prediction. *Journal of Nonparametric Statistics*, 22, 633-648.

Van Keilegom, I., González Manteiga, W. and Sánchez Sellero, C. (2008). Goodness-of-fit

tests in parametric regression based on the estimation of the error distribution. *Test*, 17, 401-415.

Wu, H. and Zhang, J.T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley Series in Probability and Statistics, USA.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4, 57-74.

Appendix : Theoretical developments

Characterization of H_0

We will show that the null hypothesis in (6) holds if and only if the standardized marginal errors U'_{ij} and $U'_{ij,0}$ have the same distribution ($i = 1, \ldots, q; j = 1, \ldots, n_i$). Indeed, note that the components $U'_{i1}, \ldots, U'_{in_i}$ are i.i.d. (and similarly for $U'_{i10}, \ldots, U'_{in_i0}$), since they are uncorrelated (by construction) and multivariate normal. Now, suppose $U'_{i1} \stackrel{d}{=} U'_{i10}, \ldots, U'_{in_i} \stackrel{d}{=} U'_{in_i0}$. Then, $U'_i \stackrel{d}{=} U'_{i0}$ because of independence between the components of both vectors. Hence, $E(U'_i) = E(U'_{i0})$ and $\mathrm{Var}(U'_i) = \mathrm{Var}(U'_{i0})$. Write $E(U'_i) = E(U'_{i0}) + V_i^{-1/2} E(m(X_i) - m_0(X_i))$. Hence, $E(m(X_i) - m_0(X_i)) = 0$. Next,

$$\begin{aligned}\mathrm{Var}(U'_{i0}) = \mathrm{Var}(U'_i) &+ V_i^{-1/2}\mathrm{Var}(m(X_i) - m_0(X_i))V_i^{-1/2} \\ &+V_i^{-1/2}\mathrm{Cov}(U_i, m(X_i) - m_0(X_i))V_i^{-1/2} + V_i^{-1/2}\mathrm{Cov}(m(X_i) - m_0(X_i), U_i)V_i^{-1/2}.\end{aligned}$$

The above covariances equal zero, since $E(U_i|X_i) = E(\epsilon_i|X_i) + E(b_i|X_i)^t X_i = 0$. It follows that $\mathrm{Var}(m(X_i) - m_0(X_i)) = 0$, and hence $m \equiv m_0$. This shows our claim, since the reverse statement is trivial.

Asymptotic limit of $T_{n,KS}$ and $T_{n,CM}$

The asymptotic framework under which we work is the following one : we let q tend to infinity, and suppose that $n_i \leq C$ $(i = 1, \ldots, q)$ for some $C < \infty$.

Similarly to the proof of Lemma 1 in Akritas and Van Keilegom (2001), it can be shown that under H_0,

$$\widehat{F}_{u'}(t) = n^{-1}\sum_{i=1}^{q}\sum_{j=1}^{n_i}[I(U'_{ij}\le t) + P(\widehat{U}'_{ij}\le t|X_i) - P(U'_{ij}\le t|X_i)] + R_n(t), \qquad (A.1)$$

and analogously for $\widehat{F}_{u'_0}(t)$, where $\sup_t|R_n(t)| = o_P(n^{-1/2})$, and where $P(\widehat{U}'_{ij}\le t|X_i)$ is calculated with respect to the law of Y_i, conditional on X_i, $\widehat{m}$ and $\widehat{V}_i$.

Write $V_i^{-1/2} = (s_{ij\ell})_{j,\ell=1}^{n_i}$ and $\widehat{V}_i^{-1/2} = (\widehat{s}_{ij\ell})_{j,\ell=1}^{n_i}$. Then, $\widehat{U}'_{ij} = \widehat{U}'_{ij,0} - \sum_{\ell=1}^{n_i}\widehat{s}_{ij\ell}(\widehat{m}(X_{i\ell}) - \widehat{m}_0(X_{i\ell}))$. Therefore, using (A.1) for the first equality below, we have that under H_0,

$$\begin{aligned}
&\widehat{F}_{u'}(t) - \widehat{F}_{u'_0}(t)\\
&= n^{-1}\sum_{i=1}^{q}\sum_{j=1}^{n_i}[P(\widehat{U}'_{ij}\le t|X_i) - P(\widehat{U}'_{ij,0}\le t|X_i)] + o_P(n^{-1/2})\\
&= n^{-1}f_{u'}(t)\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}\widehat{S}_{i\ell}(\widehat{m}(X_{i\ell}) - \widehat{m}_0(X_{i\ell})) + o_P(n^{-1/2})\\
&= n^{-1}f_{u'}(t)\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}S_{i\ell}(\widehat{m}(X_{i\ell}) - \widehat{m}_0(X_{i\ell})) + o_P(n^{-1/2}),
\end{aligned}$$

where $f_{u'}(t)$ is the standard normal density, $\widehat{S}_{i\ell} = \sum_{j=1}^{n_i}\widehat{s}_{ij\ell}$ and $S_{i\ell} = \sum_{j=1}^{n_i}s_{ij\ell}$. Now, let $Q_n(x) = \sum_{i=1}^{q}1_{n_i}^t W_{ih}^{1/2}(x)V_i^{-1}W_{ih}^{1/2}(x)1_{n_i}$, and consider

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}S_{i\ell}(\widehat{m}(X_{i\ell}) - \widehat{m}_0(X_{i\ell}))\\
&= n^{-1}\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}S_{i\ell}Q_n(X_{i\ell})^{-1}\sum_{i'=1}^{q}1_{n_{i'}}^t W_{i'h}^{1/2}(X_{i\ell})V_{i'}^{-1}W_{i'h}^{1/2}(X_{i\ell})(Y_{i'} - m_{\widehat{\theta}}(X_{i'})) + o_P(n^{-1/2})\\
&= \sigma^2 n^{-2}\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}S_{i\ell}f^{-1}(X_{i\ell})\sum_{i'=1}^{q}\sum_{j,j'=1}^{n_{i'}}K_h^{1/2}(X_{i\ell} - X_{i'j})K_h^{1/2}(X_{i\ell} - X_{i'j'})(V_{i'}^{-1})_{jj'}\\
&\qquad\times(Y_{i'j'} - m_{\widehat{\theta}}(X_{i'j'})) + o_P(n^{-1/2})\\
&= \sigma^2 n^{-2}\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}S_{i\ell}f^{-1}(X_{i\ell})\sum_{i'=1}^{q}\sum_{j=1}^{n_{i'}}K_h(X_{i\ell} - X_{i'j})(V_{i'}^{-1})_{jj}(Y_{i'j} - m(X_{i'j}))\\
&\quad -\sigma^2 n^{-2}\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}S_{i\ell}f^{-1}(X_{i\ell})\sum_{i'=1}^{q}\sum_{j=1}^{n_{i'}}K_h(X_{i\ell} - X_{i'j})(V_{i'}^{-1})_{jj}\frac{\partial m_\theta(X_{i'j})}{\partial\theta}|_{\theta=\theta_0}(\widehat{\theta} - \theta_0)\\
&\quad +o_P(n^{-1/2})\\
&:= \sigma^2 n^{-2}\sum_{i=1}^{q}\sum_{\ell=1}^{n_i}\sum_{i'=1}^{q}\sum_{j=1}^{n_{i'}}\Big[\gamma_1(X_{i\ell},Y_{i\ell},X_{i'j},Y_{i'j}) - \gamma_2(X_{i\ell},Y_{i\ell},X_{i'j},Y_{i'j})\Big] + o_P(n^{-1/2}), \qquad (A.2)
\end{aligned}$$

since $nQ_n(X_{i\ell})^{-1}\xrightarrow{P}\sigma^2 f^{-1}(X_{i\ell})$ uniformly in i and ℓ. Using the Hajek projection for U-

statistics with kernel depending on n, we can write the first term of (A.2) as

$$\sigma^2 n^{-1} \sum_{i=1}^{q} \sum_{\ell=1}^{n_i} E[\gamma_1(X_{i\ell}, Y_{i\ell}, X_{i'j}, Y_{i'j})|X_{i\ell}, Y_{i\ell}]$$
$$+\sigma^2 n^{-1} \sum_{i'=1}^{q} \sum_{j=1}^{n_{i'}} E[\gamma_1(X_{i\ell}, Y_{i\ell}, X_{i'j}, Y_{i'j})|X_{i'j}, Y_{i'j}] + o_P(n^{-1/2}).$$

The first conditional expectation equals zero, whereas the second one can be written as

$$\sigma^2 n^{-1} \sum_{i'=1}^{q} \sum_{j=1}^{n_{i'}} W_{i'j}(Y_{i'j} - m(X_{i'j})) + o_P(n^{-1/2}),$$

where $W_{i'j} = plim_{n\to\infty} n^{-1} \sum_{i=1}^{q} \sum_{\ell=1}^{n_i} E[S_{i\ell} f^{-1}(X_{i\ell}) K_h(X_{i\ell} - X_{i'j})|X_{i'j}](V_{i'}^{-1})_{jj}$.

For the second term of (A.2), define

$$\beta = \sigma^2 \, plim_{n\to\infty} \Big[n^{-2} \sum_{i=1}^{q} \sum_{\ell=1}^{n_i} S_{i\ell} f^{-1}(X_{i\ell}) \sum_{i'=1}^{q} \sum_{j=1}^{n_{i'}} K_h(X_{i\ell} - X_{i'j})(V_{i'}^{-1})_{jj} \frac{\partial m_\theta(X_{i'j})}{\partial \theta}|_{\theta=\theta_0} \Big].$$

Then, this second term equals

$$\beta(\widehat{\theta} - \theta_0) + o_P(n^{-1/2}) = \beta n^{-1} \sum_{i=1}^{q} \sum_{j=1}^{n_i} \xi(X_{ij}, Y_{ij}) + o_P(n^{-1/2}),$$

with the function ξ depending on the definition of the estimator $\widehat{\theta}$. This shows that

$$\widehat{F}_{u'}(t) - \widehat{F}_{u'_0}(t) = f_{u'}(t)\sigma^2 n^{-1} \sum_{i=1}^{q} \sum_{j=1}^{n_i} [W_{ij}(Y_{ij} - m(X_{ij})) - \beta\xi(X_{ij}, Y_{ij})] + o_P(n^{-1/2}),$$

under H_0 and uniformly in t.

It now follows that the process $n^{1/2}[\widehat{F}_{u'}(t) - \widehat{F}_{u'_0}(t)]$ $(-\infty < t < \infty)$ converges weakly to $f_{u'}(t)W$, where W is a zero-mean normal random variable with variance

$$\begin{aligned} \mathrm{Var}(W) &= \sigma^4 \, plim_{n\to\infty} n^{-1} \sum_{i=1}^{q} \sum_{j,\ell=1}^{n_i} [W_{ij}(Y_{ij} - m(X_{ij})) - \beta\xi(X_{ij}, Y_{ij})] \\ &\quad \times [W_{i\ell}(Y_{i\ell} - m(X_{i\ell})) - \beta\xi(X_{i\ell}, Y_{i\ell})]. \end{aligned}$$

Finally, we find the following limiting distributions of the test statistics under H_0:

$$T_{n,KS} \xrightarrow{d} \sup_{-\infty<t<\infty} |f_{u'}(t)| \, |W| \quad \text{and} \quad T_{n,CM} \xrightarrow{d} \int f_{u'}^2(t) \, dF_{u'}(t) \, W^2.$$

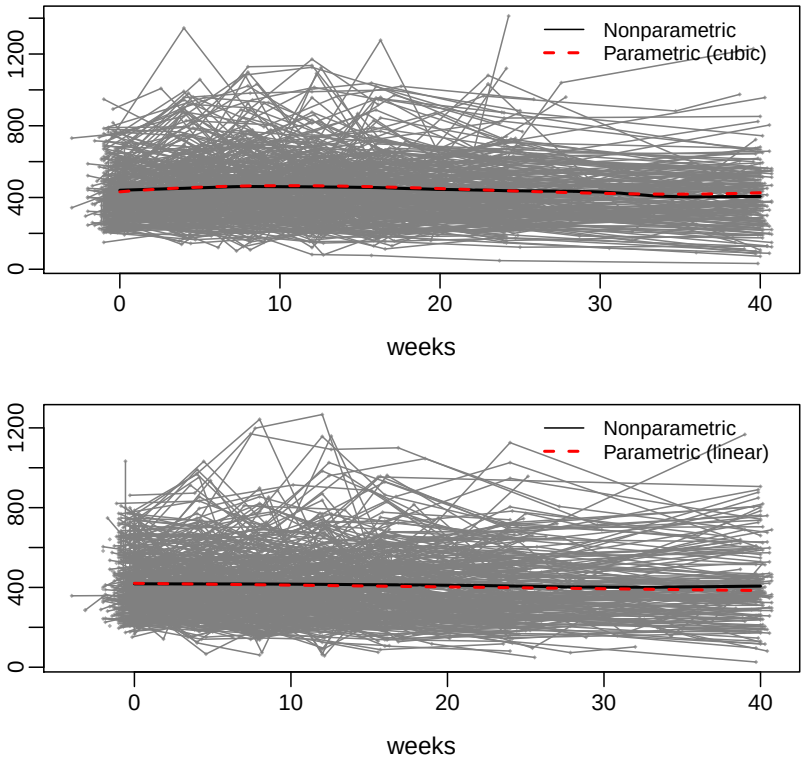

Figure 1. CD4 counts data. Observed individual curves (gray lines) for patients treated with AZT are plotted in the top panel and those for the placebo group in the bottom panel. The estimated fixed effects function using the local constant kernel estimator is shown by a black solid curve using a bandwidth of 8 weeks. Two parametric functions are suggested for each group and plotted by red dashed curves.

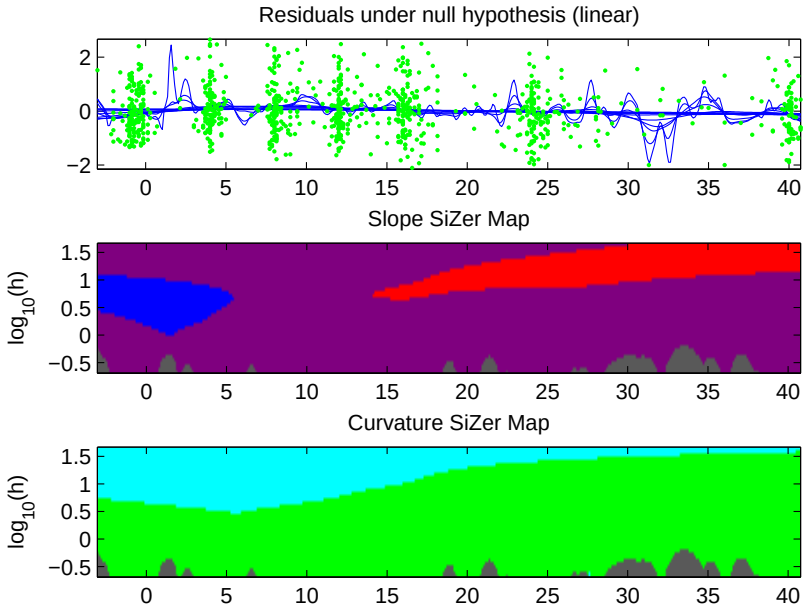

Figure 2. SiZer analysis of the residuals under a parametric linear mixed model $(m_{\widehat{\theta}_1}(X_{ij}) = 452.317 - 0.362X_{ij})$, for the AZT group in the CD4 counts data.

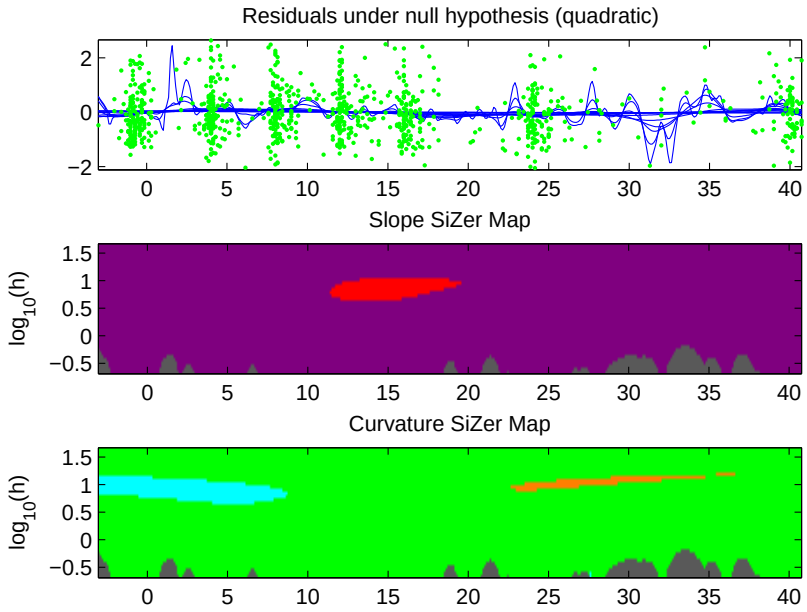

Figure 3. SiZer analysis of the residuals under a parametric quadratic mixed model ($m_{\widehat{\theta}_2}(X_{ij}) = 439.352 + 2.433X_{ij} - 0.077X_{ij}^2$), for the AZT group in the CD4 counts data.

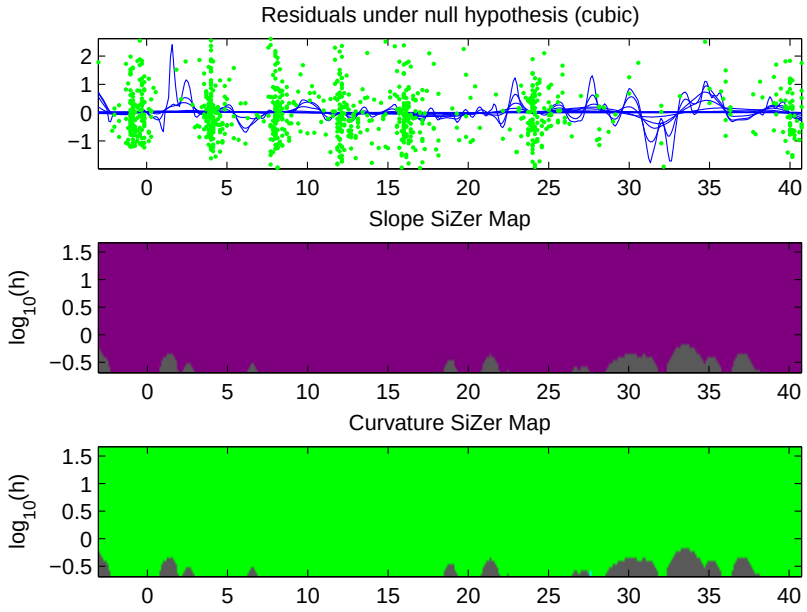

Figure 4. SiZer analysis of the residuals under a parametric cubic mixed model ($m_{\widehat{\theta}_3}(X_{ij}) = 431.617 + 7.273X_{ij} - 0.451X_{ij}^2 + 0.007X_{ij}^3$), for the AZT group in the CD4 counts data.

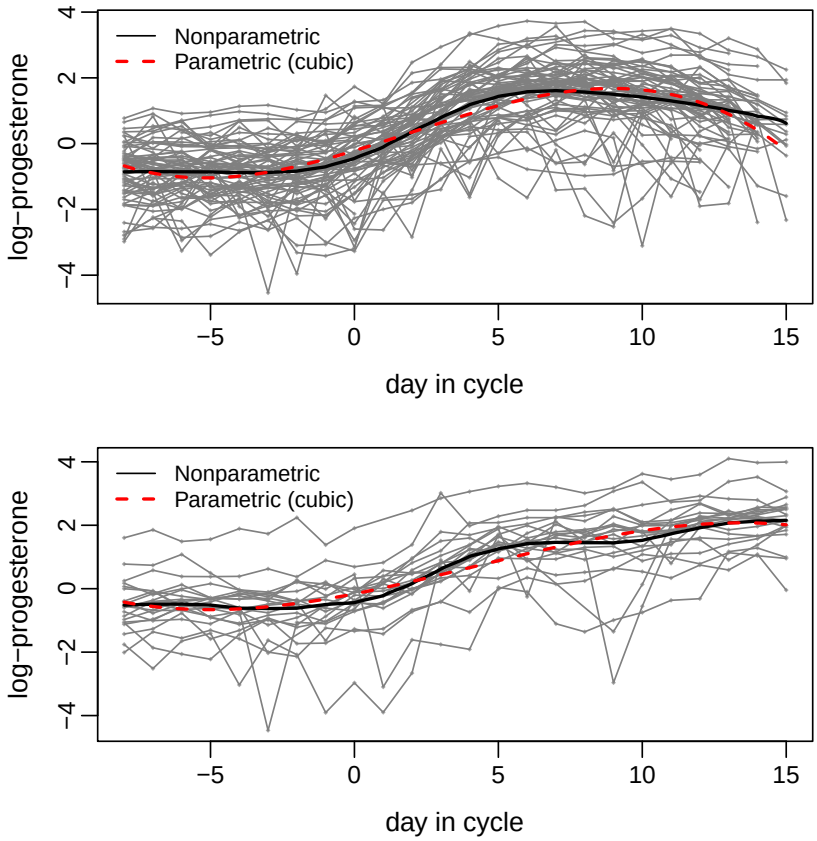

Figure 5. Progesterone data (Wu and Zhang, 2006). The 69 observed raw non-conceptive progesterone curves (gray lines) are plotted in the top panel, and the 22 raw conceptive progesterone curves are shown in the bottom panel. The estimated fixed effects function using the local constant kernel estimator is shown by a black solid curve using a bandwidth of 2 days. Two parametric (cubic) functions are suggested for each group and plotted by red dashed curves.

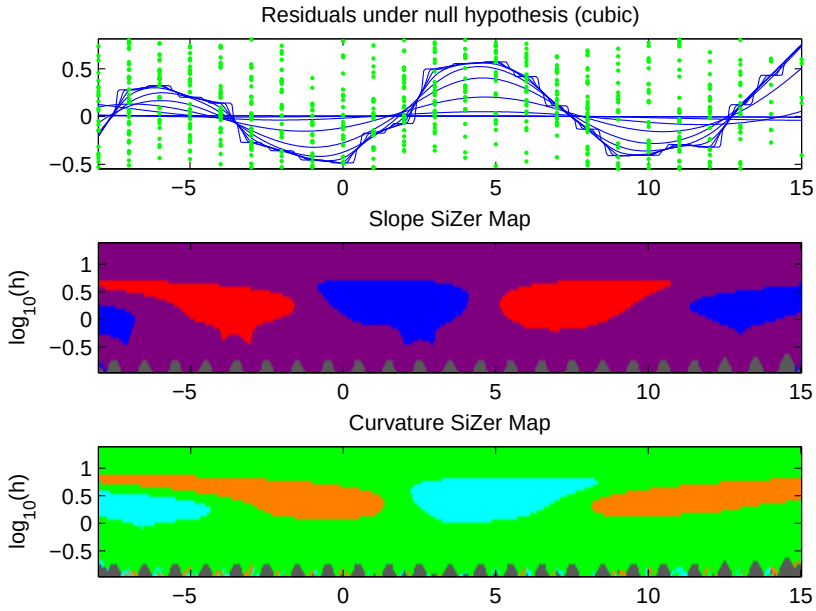

Figure 6. SiZer analysis of the residuals under a parametric cubic mixed model ($m_{\widehat{\theta}_1}(X_{ij}) = -0.215 + 0.267X_{ij} + 0.011X_{ij}^2 - 0.002X_{ij}^3$), for the non-conceptive group in the progesterone data (Wu and Zhang, 2006).

Table 1

Simulation results. Percentage of rejection and average p-value (between brackets) under the null hypothesis of linearity. The considered nominal level is $\alpha = 0.05$.

		Uniform						Normal					
		Case 1		Case 2		Case 3		Case 1		Case 2		Case 3	
n	h_0	CM	KS	CM	KS	CM	KS	CM	KS	CM	KS	CM	KS
100	2	4.5	4.3	4.6	3.8	4.3	4.6	3.8	4.5	4.5	5.4	3.0	3.4
		(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.54)	(.53)
	2.5	4.8	5.1	4.0	4.5	5.0	5.2	4.2	4.8	3.2	4.4	3.3	3.6
		(.48)	(.50)	(.50)	(.50)	(.50)	(.50)	(.49)	(.50)	(.51)	(.50)	(.52)	(.51)
	3	4.1	6.4	5.0	5.3	4.2	4.7	4.5	6.0	3.9	5.0	3.8	4.6
		(.48)	(.49)	(.49)	(.51)	(.49)	(.50)	(.49)	(.49)	(.50)	(.49)	(.51)	(.50)
200	2	5.6	4.5	6.4	4.7	5.2	5.0	5.6	5.8	4.7	4.4	4.5	4.2
		(.49)	(.50)	(.50)	(.50)	(.51)	(.51)	(.50)	(.51)	(.51)	(.51)	(.52)	(.52)
	2.5	5.5	6.0	5.1	5.1	5.9	5.0	4.4	4.1	5.5	3.8	3.9	4.3
		(.49)	(.50)	(.50)	(.49)	(.51)	(.51)	(.50)	(.51)	(.50)	(.50)	(.50)	(.50)
	3	4.4	5.2	5.1	5.2	5.1	5.0	5.0	5.8	4.5	5.3	4.2	4.8
		(.49)	(.50)	(.49)	(.49)	(.50)	(.51)	(.49)	(.50)	(.50)	(.49)	(.51)	(.50)
500	2	5.0	5.2	5.3	6.1	6.2	6.6	4.0	3.7	4.5	4.5	4.7	4.6
		(.50)	(.49)	(.49)	(.49)	(.50)	(.50)	(.50)	(.50)	(.50)	(.50)	(.51)	(.50)
	2.5	5.4	5.1	6.0	5.7	5.5	5.4	4.6	5.2	3.9	5.1	4.3	5.8
		(.49)	(.50)	(.49)	(.50)	(.50)	(.49)	(.50)	(.49)	(.51)	(.50)	(.50)	(.49)
	3	5.4	5.3	6.5	6.0	5.7	5.9	5.0	5.8	3.6	4.0	4.1	4.6
		(.49)	(.51)	(.49)	(.49)	(.49)	(.49)	(.50)	(.49)	(.50)	(.50)	(.51)	(.51)

Table 2

Simulation results: Evaluation of the power against two types of alternatives. The considered nominal level is $\alpha = 0.05$*. For each value of* a*,* n *and* h_0*, the first row shows the percentage of rejection for the sinusoidal alternative* $m_s(X) = 1 + (1-a)X + a\sin(\pi X)$*, and the second row for the quadratic alternative* $m_q(X) = 1 + (1-a)X + aX^2$*.*

			Uniform						Normal					
			Case 1		Case 2		Case 3		Case 1		Case 2		Case 3	
$(a,\ n)$	h_0	H_1	CM	KS	CM	KS	CM	KS	CM	KS	CM	KS	CM	KS
(0.1,100)	2	m_s	12.7	9.6	11.5	9.3	11.6	10.3	20.5	17.0	20.1	14.6	17.0	12.5
		m_q	9.8	7.1	9.2	8.1	9.3	8.0	22.4	15.7	20.9	15.3	18.6	14.1
	2.5	m_s	12.4	10.1	11.5	11.0	11.3	10.1	24.8	18.2	21.7	16.8	20.8	15.9
		m_q	10.6	9.3	9.4	8.6	9.8	8.1	25.3	16.9	23.9	15.2	21.2	15.7
	3	m_s	10.3	8.0	10.4	8.7	10.0	7.9	25.5	19.1	23.5	16.5	22.7	17.1
		m_q	11.2	9.0	10.3	8.6	10.1	8.7	30.3	19.7	27.8	18.6	26.5	16.9
(0.1,200)	2	m_s	23.7	17.2	20.6	14.9	19.2	16.8	34.3	22.3	33.8	23.4	31.7	24.2
		m_q	16.2	12.6	15.1	12.3	13.0	10.3	39.0	22.8	36.9	24.2	35.4	23.1
	2.5	m_s	23.9	16.1	21.9	17.0	21.4	17.6	39.8	26.2	36.8	26.8	36.4	25.3
		m_q	15.8	12.2	15.8	13.3	16.3	10.8	43.8	26.9	42.2	27.0	40.9	25.5
	3	m_s	22.9	17.1	22.4	17.0	20.8	15.7	44.8	29.1	42.9	29.2	41.2	29.2
		m_q	18.1	12.9	18.0	12.4	16.6	13.3	49.4	30.8	45.5	29.2	45.9	29.6
(0.1,500)	2	m_s	48.4	32.4	47.5	31.1	45.9	30.8	70.6	45.1	67.1	46.9	65.3	47.1
		m_q	25.2	17.0	26.5	18.0	25.9	18.1	75.5	50.9	70.2	43.9	67.8	43.9
	2.5	m_s	52.3	37.1	52.8	34.0	53.0	35.3	78.8	53.9	75.2	52.9	71.7	53.6
		m_q	29.4	21.1	30.4	20.7	30.3	21.2	82.5	54.8	75.0	51.0	75.7	50.9
	3	m_s	55.1	39.1	55.4	37.2	53.9	38.7	83.0	56.9	80.6	58.0	78.5	56.9
		m_q	33.3	22.5	32.4	22.2	32.5	25.2	86.2	60.6	78.9	50.6	80.3	56.4
(0.2,100)	2	m_s	38.1	27.0	35.6	24.6	34.2	25.1	61.4	46.2	58.2	43.6	52.3	38.7
		m_q	24.0	17.1	21.7	16.2	21.4	17.1	69.0	48.0	61.8	42.7	55.7	37.8
	2.5	m_s	36.8	23.6	35.3	25.3	35.2	27.2	68.1	50.7	63.7	48.3	60.6	44.5
		m_q	26.3	17.5	25.0	19.3	23.8	18.0	74.1	53.8	70.0	49.4	64.0	46.4
	3	m_s	33.0	22.5	32.4	22.8	32.3	22.7	70.6	51.9	68.2	51.2	65.6	50.1
		m_q	28.9	19.2	26.3	19.0	25.0	18.2	78.6	56.4	73.8	54.1	71.3	49.7
(0.2,200)	2	m_s	69.7	47.3	66.3	47.7	64.2	48.8	89.5	71.5	84.7	68.3	83.3	66.5
		m_q	42.7	28.2	40.7	28.3	40.1	26.4	92.8	74.9	86.5	67.3	82.9	63.4
	2.5	m_s	70.5	49.0	69.3	46.4	67.9	49.4	93.4	77.0	89.2	74.5	86.0	73.4
		m_q	47.3	33.7	46.9	31.5	43.0	32.6	95.7	78.9	91.4	74.9	87.0	70.4
	3	m_s	71.3	48.1	69.5	48.4	68.1	47.4	93.5	80.8	92.6	77.7	89.6	76.5
		m_q	48.6	35.8	48.0	34.4	48.5	36.0	96.2	84.5	93.3	79.2	91.6	77.7
(0.2,500)	2	m_s	96.6	84.3	96.5	82.2	95.0	82.8	99.8	97.2	99.4	94.8	98.0	92.3
		m_q	75.2	51.5	74.6	51.7	73.8	52.1	99.9	98.6	99.5	95.8	97.9	92.3
	2.5	m_s	98.3	87.6	98.5	87.8	97.6	87.6	100	98.5	99.7	97.4	99.3	96.3
		m_q	81.7	60.5	80.9	59.3	82.0	60.9	100	99.2	99.9	98.0	99.0	95.9
	3	m_s	98.6	90.2	99.0	89.5	98.1	89.0	100	99.0	100	98.3	99.7	97.2
		m_q	85.9	64.4	84.4	62.1	85.1	64.4	100	99.5	100	98.5	99.3	96.5

Table 3

Simulation results: Size of the test and power against two types of alternatives. The omnibus test (PL) by Pan and Lin (2005) is compared with the CM and KS tests proposed in this paper. Under the null hypothesis ($a = 0$) for each sample size the first row shows the percentage of rejection and the second row the average p-value. Under the alternatives ($a = 0.1, 0.2$) the percentage of rejection is given. The considered nominal level is $\alpha = 0.05$.

		Case 1			Case 2			Case 3		
$(a,\ n)$	H_1	CM	KS	PL	CM	KS	PL	CM	KS	PL
(0,150)		5.0	4.9	3.6	4.2	4.4	4.0	4.5	4.7	4.3
	(pval)	(.49)	(.50)	(.46)	(.52)	(.51)	(.46)	(.54)	(.53)	(.46)
(0,300)		5.2	5.5	4.2	5.1	5.0	4.4	4.3	4.6	4.6
	(pval)	(.49)	(.49)	(.48)	(.50)	(.49)	(.48)	(.52)	(.51)	(.48)
(0,600)		6.0	6.0	4.2	4.6	4.3	4.1	4.3	5.2	4.2
	(pval)	(.48)	(.47)	(.49)	(.50)	(.50)	(.49)	(.42)	(.59)	(.49)
(0.1,150)	m_s	31.5	21.1	14.9	23.5	17.1	5.4	25.0	17.9	4.5
	m_q	34.3	20.7	15.6	25.0	16.0	7.3	25.0	14.1	5.1
(0.1,300)	m_s	51.6	34.4	32.9	45.4	29.0	12.3	45.8	31.1	7.0
	m_q	57.3	32.2	37.1	42.3	25.2	11.2	44.9	24.9	7.4
(0.1,600)	m_s	83.2	57.2	61.7	72.5	47.0	18.9	76.7	53.1	9.1
	m_q	83.7	55.6	73.5	69.8	42.6	24.0	70.0	49.5	9.6
(0.2,150)	m_s	78.5	59.1	56.3	70.3	52.5	16.1	70.0	51.5	6.9
	m_q	85.4	61.4	59.3	69.4	46.3	19.0	67.9	46.3	8.0
(0.2,300)	m_s	97.5	83.3	91.0	93.3	72.6	36.1	94.2	79.5	15.0
	m_q	98.1	87.4	95.7	91.5	67.6	42.9	91.2	73.4	15.9
(0.2,600)	m_s	100	97.9	99.8	99.5	93.6	68.8	99.6	94.4	27.2
	m_q	100	98.8	100	99.6	94.0	82.1	98.8	90.6	34.0