Comparison of nonlinear curves and surfaces

Shi Zhao[a,], Giorgos Bakoyannis[a], Spencer Lourens[b], Wanzhu Tu[a]

[a] *Department of Biostatistics, Indiana University Fairbanks School of Public Health and Indiana University School of Medicine, Indianapolis, Indiana 46202, U.S.A.*
[b] *CliftonLarsonAllen LLP*

Abstract

Estimation of nonlinear curves and surfaces has long been the focus of semiparametric and nonparametric regression analysis. What has been less studied is the comparison of nonlinear functions. In lower-dimensional situations, inference typically involves comparisons of curves and surfaces. The existing comparative procedures are subject to various limitations, and few computational tools have been made available for off-the-shelf use. To address these limitations, two modified testing procedures for nonlinear curve and surface comparisons are proposed. The proposed computational tools are implemented in an `R` package, with a syntax similar to that of the commonly used model fitting packages. An `R` Shiny application is provided with an interactive interface for analysts who do not use `R`. The new tests are consistent against fixed alternative hypotheses. Theoretical details are presented in an appendix. Operating characteristics of the proposed tests are assessed against the existing methods. Applications of the methods are illustrated through real data examples.

Keywords: Comparison of nonlinear functions, Nonparametric and semiparametric regression, Resampling methods

1. Introduction

An essential task in nonparametric and semiparametric regression is to estimate nonlinear functions for depiction of relations between independent and

Email address: `wtu1@iu.edu` (Wanzhu Tu)

This is the author's manuscript of the article published in final edited form as:

Zhao, S., Bakoyannis, G., Lourens, S., & Tu, W. (2020). Comparison of nonlinear curves and surfaces. Computational Statistics & Data Analysis, 150, 106987. https://doi.org/10.1016/j.csda.2020.106987

dependent variables. In lower-dimensional situations, the functions are often expressed as smooth curves and surfaces [1]. Various smoothing techniques have been developed for the estimation of nonlinear functions. Commonly used methods include local polynomial models [2], wavelets [3], smoothing splines [4, 5], and various types of penalized regression splines [6, 7, 8, 9]. Most of these estimation methods can be easily implemented in common computational platforms, giving analysts much flexibility for curve and surface estimation. What has been less studied is the inference concerning nonlinear functions, and there is a dearth of computational tools for practical use. A question of general interest is whether a specific nonparametric smoother, when applied to different comparison groups, gives the same function.

To address the question above, we review the existing literature on curve and surface comparisons, and present two L_2-based testing procedures with related theoretical and numerical justifications. Our procedures are based on B-splines, although the formulation of the test statistics could be extended to other smoothers. We put forward an `R` package, which can be accessed either directly from within `R`, or through an interactive R-Shiny interface; the latter allows analysts who do not use `R` to perform the desired comparisons. To illustrate the use of the proposed methods, we present two real data examples.

2. Existing methods for curve and surface comparison

Comparison of smooth curves can be formalized as a test of the following hypothesis $H_0 : g_1(x) = g_2(x) = \cdots = g_I(x), \forall x \in \mathbb{R}$ vs $H_1 : g_i(x) \neq g_j(x)$ for some $i, j \in \{1, \ldots, I\}$, where i and j indicate different comparison groups. In the situation of $\mathbf{x} \in \mathbb{R}^2$, $g_i(\mathbf{x})$ and $g_j(\mathbf{x})$ are surface functions. The concept can be extended to higher-dimensional functions, although visualization of higher-dimensional functions becomes more difficult. In this paper, we restrict the discussion to lower-dimensional situations, and we write the underlying model as $Y = g_i(x) + \epsilon$. We use capital letters Y and X to indicate the random response and independent variables, and their lower-case counterparts

x and y to indicate the observed values of the corresponding random variables.

Early work on this problem started almost three decades ago. One approach is to frame the problem in a regression setting, where a modified version of the Kolmogorov-Smirnov test could be used to compare the regression curves [10]. Another approach is to transform the nonparametric curves to reduce the comparison to a test of limited dimensional parameters within the transformation matrix [11]. Alternatively, wavelet methods have been used to compare density functions [12]; the methods are especially suitable for comparing higher frequency local features. Many of these methods, however, require the curves to have the same design points, i.e., all functions must be evaluated at the same x values. To remedy, Kulasekera (1995) fitted kernel-based regression models and proposed tests based on the weighted average of partial sum of squares of the quasi-residuals and error variances [13]. But simulations suggest that these tests tend to be overly sensitive to bandwidth selection. There are also specialized tests for parallelism among the curves [14, 15].

In this section, we briefly review the methods that are most frequently used in analytical practice.

2.1. Nonparametric Analysis of Covariance (ANCOVA)

Young and Bowman (1995)[16] described a method for testing the equality of two or more smooth curves, under the model $Y_{ij} = g_i(x_{ij}) + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma^2)$, for $i = 1, 2, \ldots, I$, $j = 1, \ldots, n_i$. The test has a homoscedastic assumption, i.e., the error variance remains constant across all I groups.

Young and Bowman used a kernel-based smoothing method to approximate g_i. Assuming that h_i is the bandwidth for the ith regression function, they proposed to estimate g_i with

$$\hat{g}_i(x) = \frac{\sum_{j=1}^{n_i} K((x - x_{ij})/h_i) y_{ij}}{\sum_{j=1}^{n_i} K((x - x_{ij})/h_i)}, \tag{1}$$

which is sometimes referred to as the Nadaraya-Watson estimator of g_i.

Under the null hypothesis, one could obtain a common regression function

by combining data from all groups

$$\hat{g}(x) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} K((x - x_{ij})/h)y_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} K((x - x_{ij})/h)}, \tag{2}$$

where h is the common bandwidth for estimating g.

The test statistic that Young and Bowman proposed is analogous to the one-way ANOVA,

$$T_1 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\hat{g}(x_{ij}) - \hat{g}_i(x_{ij})]^2}{\hat{\sigma}^2}, \tag{3}$$

where $\hat{g}_i$ and $\hat{g}$ are the group-specific and common curve estimators, and $\hat{\sigma}^2$ is the pooled variance. To estimate σ^2, one uses $\hat{\sigma}^2 = \frac{1}{N-I}\sum_{i=1}^{I}(n_i - 1)\hat{\sigma}_i^2$, where $N = \sum_{i=1}^{I} n_i$. Similarly, the group-specific variance is estimated as

$$\hat{\sigma}_i^2 = \frac{1}{2(n_i - 1)}\sum_{j=1}^{n_i-1}(y_{i,[j+1]} - y_{i,[j]})^2.$$

This test has been extended to comparisons of surface functions [17].

The construction of the test is intuitive, and its implementation straightforward. The equal variance assumption, however, can be overly restrictive in some analytical situations. Additionally, when the explanatory variable x_{ij} takes different values in the comparison groups, the power of the test often drops precipitously because the biases can no longer be canceled out under H_0; see Tables 1- 3 in Young and Bowman (1995)[16].

More recently, Park and colleagues considered a similar ANOVA type test statistic for multiple x values with a given bandwidth [18]. They obtained an empirical distribution for the maximum of the pointwise test statistics for controlling multiplicity. A visualization tool has been developed to show differences between curves at multiple locations. But simulation studies suggest that the test has type I error rates far below the nominal level, and very low power; see Table 1 of Park et al [18].

2.2. Kernel-based nonparametric methods

Dette and Neumeyer (2001) proposed another set of tests, all based on kernel smoothing techniques [19]. Expressing the curves as $Y_{ij} = g_i(x_{ij}) + \epsilon_{ij}(x_{ij})$,

where $i = 1, 2, \ldots, I, j = 1, \ldots, n_i$, the authors introduced heteroscedastic errors $\epsilon_{ij}(x_{ij}) \sim N(0, \sigma_i^2)$ into the model. The main hypothesis remains the same, $H_0 : g_1 = g_2 = \cdots = g_I$ vs $H_1 : g_i \neq g_j$ for some $i, j \in \{1, \ldots, I\}$.

The tests are subject to the following conditions: (1) The variances $\sigma_i(\cdot)$ are continuous functions; (2) the design points x_{ij} satisfy $\int_0^{x_{ij}} r_i(x)dx = \frac{j}{n_j}$ for a density function r_i, where $j = 1, \ldots, n_i$, and $i = 1, \ldots, I$; (3) the regression functions $g(\cdot)$ are sufficiently smooth, i.e., ≥ 2 times continuously differentiable in the supporting space. And the Nadaraya-Watson estimators $\hat{g}_i$ and $\hat{g}$ are as previously defined.

One test (T_2) compares the group-specific error variances against that of the combined sample, in a way that is analogous to one-way ANOVA

$$T_2 = \hat{\sigma}^2 - \frac{1}{N} \sum_{i=1}^{I} n_i \hat{\sigma}_i^2. \tag{4}$$

The second test (T_3) directly assesses the distance between the group-specific curves and a common curve at x_{ij}, assuming that x_{ij} remain exactly the same across the groups,

$$T_3 = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [\hat{g}(x_{ij}) - \hat{g}_i(x_{ij})]^2. \tag{5}$$

The third test (T_4) summarizes all pairwise L_2-distances of the estimated individual grooup curves,

$$T_4 = \sum_{i=1}^{I} \sum_{j=1}^{i-1} \int [\hat{g}_i(x) - \hat{g}_j(x)]^2 w_{ij}(x) dx, \tag{6}$$

where $w_{ij}(\cdot)$ are positive weight functions. Asymptotic normality of the test statistics has been established under the null and fixed alternatives.

The above tests have been extended to comparison of two regression curves with different design points and heteroscedastic variances [20]. For comparison of two curves, the authors proved by using the empirical process theory that the two marked empirical processes converged to a centered Gaussian process at a rate of $N^{-1/2}$ under the null; while under the alternative, the means of the two processes do not converge to zero. Hence, tests could be constructed based

on either functions of the integrated squared residuals or the supremum of the absolute residuals of these two processes.

A practically important extension of Neumeyer and Dette's methods is the comparison of multiple curve functions [21]. These tests have also been extended to compare surface functions. Since the rates of convergence of the test statistics are slower [22], and the wild bootstrap procedure is consistent under more relaxed conditions, analysts often calculate p-values from the distribution of the test statistic under H_0 using a wild bootstrap procedure [23].

2.3. Additive model-based tests

Zhang and Lin (2000) [24] considered testing the equivalence of two nonparametric functions in an additive mixed model for longitudinal data

$$Y_{ijk} = g_i(x_{ijk}) + \boldsymbol{s}_{ijk}^T \boldsymbol{\alpha}_i + \mathbf{Z}_{ijk}^T \mathbf{b}_{ij} + \epsilon_{ijk}, \tag{7}$$

where Y_{ijk} is the response from the jth subject in the ith group at the kth assessment, and $\boldsymbol{\alpha}_i$ is a $p \times 1$ vector associated with covariates $\boldsymbol{s}_{ijk}$.

To test hypothesis $H_0 : g_1 = g_2$ vs. $H_1 : g_1 \neq g_2$, the authors suggested the following test statistic

$$G\{\hat{g}_1(x), \hat{g}_2(x)\} = \int_{T_1}^{T_2} \{[\hat{g}_1(x) - \hat{g}_2(x)]^2\} dx, \tag{8}$$

where $\hat{g}_1$ and $\hat{g}_2$ are estimated by maximizing the penalized log-likelihood function.

The penalized likelihood under the semiparametric additive mixed model for an individual group is $l(g_i, \alpha_i; \mathbf{y}) - \frac{\lambda_i}{2} \int [g_i''(x)]^2 dx$, where λ_i is the parameter that controls the smoothness of function $g_i(x)$. Note that the test statistic (8) is also an L_2-based distance measure, as that in (6). Expressing G in Equation (8) as a quadratic function of $\mathbf{y}$, Zhang and Lin approximated the distribution of $G\{\hat{g}_1(x), \hat{g}_2(x)\}$ with a scaled χ^2 distribution using the moment matching technique.

We summarize the tests reviewed in Table 1, which highlights the key features of each method.

[Table 1 about here.]

3. Curve comparison in semiparametric regression

In a low-dimensional regression analysis, comparing nonlinear effects amounts to a comparison of nonlinear curves and surfaces. As described in Table 1, the existing methods are often overly restrictive in their accommodation of heterogeneity and design points. For example, when we compare two nonlinear functions $g_1(x)$ and $g_2(x)$, it is rather unrealistic to expect the two functions be evaluated at the exact same x values. Furthermore, if the functions are indeed different, it would not be reasonable to expect the two functions to have the same variance. These are the features that the existing methods and analytical software have not accommodated adequately.

To remedy, we propose a testing procedure in the usual context of semiparametric regression. For the convenience of discussion, we consider a comparison of curve functions $g_i(x_{ij})$, where x_{ij} denotes the value of independent variable of the jth subjects in the ith group. We are interested in testing hypothesis $g_1 = g_2 = \cdots = g_I$. We note that the test can be extended to compare higher dimensional functions, although visualization may be difficult.

In order to test the above hypothesis, one needs to estimate the functions, $g_1, g_2, \ldots, g_I$, as well as g. Most of the existing methods are developed based on kernel estimates. In practice, however, many analysts prefer various forms of regression splines, with or without smoothness penalty. In this paper, all theoretical results are derived under B-spline estimates. For all practical purposes, choices of spline basis functions are often less consequential. Here for curve comparison, we write the model as $Y_{ij} = g_i(x_{ij}) + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma_i^2)$, and we write the group-specific function g_i as a B-spline function, i.e., $g_i(x) = \sum_{k=1}^{K_n+m} \gamma_k B_k^m(x)$, where K_n is the number of internal knots with $K_n = O(n^v)$, m is the order of the B-spline with $m \geq 1$, $\{\gamma_k\}_{k=1}^{K_n+m}$ is the set of B-spline coefficients, and $\{B_k^m(x) : x \in [a, b]\}$ are B-spline basis functions. For higher dimensional functions $g_i(x_1, x_2, \ldots, x_d)$, one could use tensor products, radial, or thin-plate splines [25][26][27].

3.1. Test statistics and comparison procedures

An intuitive way to compare two functions is to measure the distance between them. The L_2 norm is a commonly used distance measure. As described previously, both Zhang and Lin (2000) and Dette and Neumeyer (2003) used the L_2 norm in the construction of their test statistics. Herein, we reexamine the test statistic

$$T_{spline} = \frac{1}{N} \sum_{1 \leq i < m \leq I} \sum_{j=1}^{n_i} [\hat{g}_i(\mathbf{x}_{ij}) - \hat{g}_m(\mathbf{x}_{ij})]^2,$$

under B-spline estimates of $\hat{g}_i$ and $\hat{g}_m$ for testing the hypothesis $H_0 : g_1 = g_2 = \cdots = g_I$ vs $H_1 : g_i \neq g_j$ for some $i, j \in \{1, \ldots, I\}$. Theoretical properties of the test statistic are examined in Section 3.1.2.

In the absence of an asymptotic normal distribution, however, one has to devise a method to approximate the distribution of the test statistic under the null hypothesis. In the following section, we demonstrate how such an approximation can be done through a resampling procedure. Specifically, we show how to ascertain p values for the test statistic from a wild bootstrap procedure.

3.1.1. A wild bootstrap-based comparison method

For the standard linear regression models, it is usually sufficient to draw bootstrap samples from centralized residuals, because the errors are homoscedastic [28]. In the one dimensional case, the underlying model can be written as $Y_{ij} = g_i(x_{ij}) + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma_i^2)$. We present g as a curve function here, although the method can be easily extended to higher dimensional situations.

To accommodate error heteroscedasticity, we consider a wild bootstrap procedure, which assures that the bootstrap error terms possess properties that are similar to those of the actual errors [29]. Another alternative approach is to use pairs bootstrapping, in which the analyst directly resamples from the joint empirical distribution function of $\mathbf{Y}_i$ and $\mathbf{x}_i$, which are the vectors of the response and independent variables respectively. The computational burden of

pairs bootstrap, however, tends to be greater especially if the dimension of $\mathbf{x}_i$ is high [28].

Wild bootstrap has been used to resample the residuals of nonparametric regression models, as suggested by Härdle and Mammen (1993) [30], and Mammen (1993) [31]. The essence of wild bootstrap is to express the regression function as a conditional expectation of the observed response variable, i.e. $E(Y_i^*|X_i = x_i) = g(x_i)$, where Y_i^* is the bootstrap data. Since this method uses a single residual $\hat{\epsilon}_i$ to estimate the conditional distribution $l(Y_i - g(x_i)|X_i = x_i)$ of an arbitrary distribution ($\hat{F}_i$ in the following), it is often referred to as the wild bootstrap.

Let V_i be a random variable following a two-point distribution $\hat{F}_i$ such that $E_{\hat{F}_i}(V_i) = 0$, $E_{\hat{F}_i}(V_i^2) = 1$, and $E_{\hat{F}_i}(V_i^3) = 1$. We define random independent quantities $\epsilon_i^* = V_i\hat{\epsilon}_i \sim \hat{F}_i$, and use $(X_i, Y_i^* = \hat{g}(x_i) + \epsilon_i^*)$ as the bootstrap observations. We then create a new bootstrap test statistic T^*.

With the bootstrap samples, for a test at level α, the null hypothesis will be rejected if T is greater than the corresponding quantile of the bootstrap distribution of the test statistic T^*, i.e. $T > T^*_{[B(1-\alpha)]}$, where $T^*_{[B(1-\alpha)]}$ is the ith order value of the bootstrap statistic T^*. Härdle and Mammen (1993)[30] showed that under the null hypothesis, T^* estimated the distribution of T consistently, since the regression function with bootstrap data $g^*(\cdot)$ had mean $g(\cdot)$ for nonlinear models under the standard regularity conditions.

Using the wild bootstrap method, we propose the following procedure:

Step 1: Estimate function $g_i(\mathbf{x})$ with $\hat{g}_i(\mathbf{x})$, $i = 1, 2, \ldots, I$, and compute the test statistic

$$T_{spline} = \frac{1}{N} \sum_{1 \le i < m \le I} \sum_{j=1}^{n_i} [\hat{g}_i(\mathbf{x}_{ij}) - \hat{g}_m(\mathbf{x}_{ij})]^2.$$

Step 2: Estimate the common function $\hat{g}(\mathbf{x})$ from the combined sample and calculate the residuals $\hat{\epsilon}_{ij} = y_{ij} - \hat{g}(\mathbf{x}_{ij})$.

Step 3: For each $\mathbf{x}_{ij}$, draw a bootstrap residual $\epsilon_{ij}^{(b)}$, for $b = 1, 2, \ldots, B$, where B is the number of bootstrap samples, from the two-point distribution

with probability mass points $\frac{1-\sqrt{5}}{2}\hat{\epsilon}_{ij}$ and $\frac{1+\sqrt{5}}{2}\hat{\epsilon}_{ij}$, occurring with probabilities $\frac{5+\sqrt{5}}{10}$ and $\frac{5-\sqrt{5}}{10}$ respectively, so that $E(\epsilon_{ij}^{(b)}) = 0$, $E(\epsilon_{ij}^{(b)2}) = \hat{\epsilon}_{ij}^2$ and $E(\epsilon_{ij}^{(b)3}) = \hat{\epsilon}_{ij}^3$.

Step 4: Generate a bootstrap sample $(\mathbf{x}_{ij}, Y_{ij}^{(b)})$ from $Y_{ij}^{(b)} = \hat{g}(\mathbf{x}_{ij}) + \epsilon_{ij}^{(b)}$.

Step 5: From this sample, estimate the bth bootstrap regression function $\hat{g}_i^{(b)}$, and calculate the test statistic $T^{(b)}$ as in the original T_{spline} calculation.

Step 6: Repeat Step 3 to 5 B times, and use the B generated values of the test statistics, $T^*_{spline} = (T^{(1)}, T^{(2)}, \ldots, T^{(B)})$, to determine the quantiles of the distribution of the test statistic. For a test at significance level α, the null hypothesis is rejected if T_{spline} is greater than the corresponding $(1-\alpha)$th quantile of the bootstrap distribution of T^*_{spline}.

3.1.2. Consistency of the test against fixed alternatives

In this section we show that the proposed test is consistent against any fixed alternatives. We also provide the optimal number of internal knots for the B-spline that leads to the best possible rate of convergence. We first rewrite the model in a slightly more general form; here we use X instead of x to emphasize the random nature of the independent variable.

We write the true model as follows

$$Y_j = g_0(X_j) + \epsilon_j = g_0(X_j) + \sigma e_j,$$

where g_0 is an unknown function of interest, (Y_j, X_j), $j = 1, \ldots, n$, are i.i.d. random variables independent of the error term $e_j \sim N(0,1)$. For simplicity and without loss of generality, we assume that the covariate $X_j \in \mathbb{X}$ a.s., where $\mathbb{X} = [0,1]$. As defined, $\mathbb{X}$ is a compact subset in $\mathbb{R}$.

As previously described, a B-spline estimate of $g(x) = \sum_{k=1}^{K_n+m} \gamma_k B_k^m(x)$ can be achieved by minimizing the objective function

$$\frac{1}{n}\sum_{j=1}^{n}[Y_j - g(X_j)]^2,$$

or, equivalently, by maximizing

$$\mathbb{M}_n(g) \equiv \mathbb{P}_n m_g = 2\mathbb{P}_n(g - g_0)e - \mathbb{P}_n(g - g_0)^2,$$

where $\mathbb{P}_n m_g$ denotes the empirical process indexed by the function m_g, i.e.

$$\mathbb{P}_n m_g = \frac{1}{n}\sum_{j=1}^{n} m_g(X_j) = -\frac{1}{n}\sum_{j=1}^{n}[Y_j - g(X_j)]^2.$$

Direct maximization of the above objective function over the full infinite-dimensional parameter space $\mathcal{G}$ may lead to inconsistant estimates[32]. Therefore, one has to use the sieve M-estimation framework by considering the spaces of B-spline functions

$$\mathcal{G}_n(D_n, K_n, m) = \left\{ g_n : g_n(x) = \sum_{k=1}^{K_n+m} \gamma_k B_k^m(x) \in S_n(D_n, K_n, m), x \in [0,1] \right\},$$

where $D_n = \{d_1, \ldots, d_{K_n}\}$ is a set of partition points for the set $[0,1]$, K_n is the number of internal knots with $K_n = O(n^v)$, m is the order of the B-spline with $m \geq 1$, $\{\gamma_k\}_{k=1}^{K_n+m}$ is the set of the unknown coefficients or control points for the B-spline, $\{B_k^m(x) : x \in [a,b]\}$ are the basis functions, and $S_n(D_n, K_n, m)$ is the space of polynomial splines on a partition D_n with K_n internal knots and of order m. Then, the sieve estimator $\hat{g}_n$ of g_0 satisfies

$$\mathbb{M}_n(\hat{g}_n) \geq \mathbb{M}_n(g) \quad \text{for all} \quad g \in \mathcal{G}_n,$$

that is $\hat{g}_n$ maximizes $g \mapsto \mathbb{M}_n(g)$ over the sieve space $\mathcal{G}_n(D_n, K_n, m)$.

For simplicity of presentation, we consider the special case of the two-sample comparison. We assume the following regularity conditions:

C1. The error e_j has a distribution with zero mean and sub-exponential tails, i.e., tails bound by the supremum of an empirical process. Also, e and the covariate X are independent.

C2. The parameter space $\mathcal{G}_i \ni g_{0,i}$, $i = 1, 2$, contains functions uniformly bounded by $C \geq 1/2$ on $[0,1]$, with bounded pth derivatives, for fixed $p \geq 1$, with the first derivative being continuous.

C3. The number of internal knots satisfies $K_n = O(n^v)$, such that

$$\max_{1 \leq k \leq K_n+1} \{d_k - d_{k-1}\} = O(n^{-v}).$$

C4. The sample sizes of the two groups satisfy

$$\frac{n_1}{n_1 + n_2} \to \lambda \in (0, 1),$$

as $\min(n_1, n_2) \to \infty$.

Theorem 1. Assuming conditions C1-C4 hold, we consider a wild-bootstrap procedure that of level α asymptotically. Then, the proposed test is consistent against any fixed alternative hypothesis: If $\pi_n(\theta_\delta)$ is the power function of the test under the fixed alternative hypothesis θ_δ, then $\pi_n(\theta_\delta) \to 1$ as $n \to \infty$.

A sketch of the proof is provided in Appendix A. Even though the proposed test is, by Theorem 1, consistent against any fixed alternative hypothesis, the asymptotic distribution of the test statistic is difficult to derive. The difficulty stems from the fact that the convergence rate of the B-spline estimator of g_1 and g_2 is

$$d(\hat{g}_{n_i,i}, g_{0,i}) = O_p\left(n^{\frac{p}{1+2p}}\right), \qquad i = 1, 2,$$

where $d(g_1, g_2) = \{E[g_1(X) - g_2(X)]^2\}^{1/2}$. Note that the above rate is the optimal convergence rate for nonparametric regression and is achieved if one sets $v = 1/(1 + 2p)$. This convergence rate is slower than the usual $\sqrt{n}$ rate for parametric models, even though it is the optimal rate in nonparametric regression.

In the absence of an analytically derived asymptotic distribution, we assessed the performance of the wild-bootstrap procedure through extensive simulations. Results of the simulation experiments are presented in Section 5.

3.1.3. Extending the testing procedure to correlated data

The same testing procedures can be modified and extended to the analysis of correlated data. A key requirement for the modification is the preservation of the correlation structure that exists within each subject. The following algorithm is a natural extension and it performs a Cholesky decomposition on the estimated covariance structure [33]. Combining with the estimated regression functions, we generate the bootstrap sample $Y_{ijk}^{(b)}$.

The algorithm is described below.

Step 1: Obtain group-specific estimates $\hat{g}_i(\mathbf{x})$ of function g_i by using semiparametric mixed effect models, and then compute the test statistic $T_{splcorr}$.

Step 2: Obtain a common regression function estimate $\hat{g}(\mathbf{x})$ for the combined sample.

Step 3: For $i = 1, \ldots, I$, obtain the group-specific covariance matrix estimate $\hat{\mathbf{R}}_i$ from the fitted group-specific model and ascertain the residuals $\hat{\eta}_i = (\hat{\eta}_{i11}, \hat{\eta}_{i12}, \ldots, \hat{\eta}_{inn_i})$.

Step 4: Perform a Cholesky decomposition on $\hat{\mathbf{R}}_i$ so that $\hat{\mathbf{R}}_i = \hat{\mathbf{L}}_i \hat{\mathbf{L}}_i^T$, where $\hat{\mathbf{L}}_i$ is a lower triangular matrix. We then obtain

$$\hat{\mathbf{e}}_i = (\hat{e}_{i11}, \hat{e}_{i12}, \ldots, \hat{e}_{inn_i}) = \hat{\mathbf{L}}_i^{-1} \hat{\eta}_i;$$

and calculate the "whitened" residuals $\tilde{\mathbf{e}}_i = \hat{\mathbf{e}}_i - \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\mathbf{e}}_i$.

Step 5: Draw a random sample from the "whitened" residuals $\tilde{\mathbf{e}}_i$ and calculate $\tilde{\eta}_i^{(b)} = \hat{\mathbf{L}}_i \tilde{\mathbf{e}}_i$. We use $Y_{ijk}^{(b)} = \hat{g}(\mathbf{x}_{ijk}) + \tilde{\eta}_{ijk}^{(b)}$ as the bootstrap sample.

Step 6: Calculate the test statistic $T^{(b)}$ with the bootstrap sample. Repeat Steps 3 to 5 B times to construct an empirical distribution of the statistic with bootstrap samples, and then calculate the p value from the tail area of the empirical distribution.

4. Software development

We published an R package `gamm4.test` in CRAN to make the proposed testing procedures available to practitioners. The two main functions are `gam.grptest` for comparisons of nonlinear functions with cross-sectional data, and `gamm4.grptest` for comparisons involving correlated data. Key features of this package are:

a) It utilizes a syntax that is consistent with `mgcv` and `gamm4`, two packages that are often used for fitting semiparametric regression models. Users familiar with those packages can perform comparisons with `gam.grptest` and `gamm4.grptest`.

b) The package performs parallel computing with an automatic detection of the numbers of available CPU cores for enhanced computational efficiency.

c) The R package includes a data visualization function `plot.gamtest` that allows users to visually examine the fitted curves and surfaces. The graphics are produced by the R package `plotly`. With option `type = plotly.persp`, the users can create 3-dimensional interactive plots. Finally, setting the option `test.statistic = TRUE` generates the empirical distribution of the test statistic under the null hypothesis of equal regression functions.

Computational efficiency of `gamm4.test` in the analysis of the example data in the package was assessed on a computer with Intel(R) Core(TM) i5-3470, CPU @3.20GHz, 64-bit operating system, and 4 CPU cores. The computing time is summarized in Table 2.

To enhance the usability of the testing methods, we also created an interactive R Shiny interface for the `gamm4.test` package. This interface allows analysts that do not use R to access the testing procedure through a web link. See `https://heather.shinyapps.io/shinygamm4/` for the app. See `https://youtu.be/SHqaZXSLaMw` for a related youtube tutorial.

[Table 2 about here.]

5. Simulation Studies

We conducted a series of simulation studies to verify the theoretical results and to examine the operating characteristics of the proposed tests, in comparison with the other L_2-based methods. We additionally investigated the influences of the number of knots on the tests.

5.1. Curve comparisons

For curve comparisons, we considered the following model

$$Y_{ij} = g_{id}(x_{ij}) + \epsilon_{ij}, \tag{9}$$

where $i = 1, 2; j = 1, \ldots, n_i$.

In this simulation, we generated values of the independent variable x_{ij} from $Unif[0, 1]$, with sample sizes n_1 and n_2 for the two comparison groups. The

nonlinear functions for the two groups were specified as $g_1(x) = 2x\exp(2-4x) - 2x + 0.5$ and $g_2(x) = 2x\exp(2-4x) - 0.5$, with $g(x) = (4x\exp(2-4x) - 2x)/2$. More generally, we considered $g_{id}(x_{ij}) = (d/10)g_i(x_{ij}) + (1 - d/10)g(x_{ij})$, with $d = 0, 1, 2, 3$, where d controlled the distance between the two group-specific functions. For example, $d = 0$ corresponded to the situation where the two groups shared the same regression function; as d increased, the functions grew further apart. These functions were plotted in Appendix B; see Supplemental Figure 1.

Values of the dependent variables Y_{ij} were generated from Equation (9) with standard errors σ_1 and σ_2, i.e. $\epsilon_{1j} \sim N(0, \sigma_1^2)$, $\epsilon_{2j} \sim N(0, \sigma_2^2)$. Comparisons of the nonlinear functions were carried out under the following three conditions: (1) a distance parameter $d = 0, 1, 2, 3$; (2) sample sizes (n_1, n_2)=(125, 125), (216, 216), and (512, 512); (3) values of the error standard deviations (σ_1, σ_2)=(0.20, 0.15) and (0.25, 0.20).

With the generated data sets, we comparatively evaluated the performance of five discussed methods:

- Method 1: The proposed testing method with cubic B-spline regression bases for curve estimation, and a wild bootstrap procedure for p-value calculation.

- Method 2: The proposed testing method with penalized cubic spline bases for curve estimation, with a wild bootstrap procedure for p-value calculation by using the `gam` function in the R package `mgcv`. Numbers of knots were set to the default value, which was determined by a generalized cross-validation (GCV) method.

- Method 3: Kernel smoothing based on the L_2 distance test statistic, followed by a wild bootstrap procedure [19].

- Method 4: The testing method based on variance estimator [19].

- Method 5: Young and Bowman's (1995) method [16], which calculates the p value by matching with a scaled chi-square distribution.

For each simulation setting, we generated a total of $1,000$ testing datasets. For each dataset, we performed the proposed test on 200 wild bootstrap samples to calculate the p value. We calculated the rate of rejection in the $1,000$ simulated samples at a significance level of 0.05.

Rejection rates for curve comparison under the null and alternative hypotheses are reported in Appendix B; see supplemental Table 1. When $d = 0$, the rejection rates are Type I error rates; when $d = 1, 2, 3$, the rejection rates represent the power of the tests. In comparison with other testing methods, the proposed tests in general had an excellent control of Type I error rates. As d increased, the power of rejecting the null hypothesis increased as well. The power of the new tests was comparable to, if not slightly better than, that of the existing tests. We noticed that Method 5 showed a slightly higher type I error rates than others. On the other hand, the Method 4 exhibited a tighter type I error control, while having considerably lower power. For the proposed testing methods, B-splines and penalized splines produced similar results. As previously discussed, we set the number of knots to $\sqrt[3]{n_1}$. In the simulation studies we observed that when an unpenalized semiparametric model was used, incorrect number of knots selection could lead to substantial bias and hence inflated the Type I error rates. The penalized semiparametric regression estimates, however, were generally very robust.

To verify the consistency theory in Section 3, we further examined the rejection rates (i.e., power) of the test as the sample size increased, under a fixed alternative hypothesis. We showed that when the distance between the null and alternative hypotheses was set to $d = 1$, the power increased with the sample size. The power approached to 1 when $n_1 = n_2 = 1000$. See Figure 2 in Appendix B.

Besides of the five methods listed above, we also compared the proposed methods with the test proposed by Kulasekera (1995) [13]. The latter's performance has left much to be desired: In the tested settings, the type I error rates were close to zero, and the power was low as well. The suboptimal performance could be due to the simulation settings we chose, where the functional curves

were close and data variability was large. The proposed tests, on the other hand, performed well in such situations. We omitted the results of Kulasekera's test from the summary table.

We further examined the performance of the tests in situations of three group comparison. The nonlinear functions of the three groups were specified respectively as $g_1(x) = 2x\exp(2-4x) - x - (d_1/10)(x-0.5)$, $g_2(x) = 2x\exp(2-4x) - x + (d_2/10)(x-0.5)$ and $g_3(x) = 2x\exp(2-4x) - x$, with $d_i = 0, 1, 2, 3$ and $i = 1, 2$, where d_1, d_2 respectively controlled the distances between g_1 and g_3, and g_2 and g_3. Comparisons of the nonlinear functions were carried out with various distance values of d_1 and d_2, as well as sample sizes. For each setting, we generated a total of 500 testing datasets.

See Supplemental Table 2 in Appendix B for rates of rejection in 500 simulated samples at a significance level of 0.05. Similar to two curve comparisons, when $d_1 = d_2 = 0$, the proposed tests in general had a good control of type I error rates. As d_i increased, regardless the equal or unequal distances between the curves, the power of rejecting the null hypothesis increased with the sample size.

5.2. Surface comparisons

Simulations for surface comparisons were carried out in a similar manner. We used the following surface functions in the simulation:

a. $g_1(\mathbf{x}) = g_2(\mathbf{x}) = \sin(2\pi x_1) + \cos(2\pi x_2)$

b. $g_1(\mathbf{x}) = g_2(\mathbf{x}) = 2x_1^2 + 3x_2^2$

c. $g_1(\mathbf{x}) = g_2(\mathbf{x}) = \exp(-x_1^2 - x_2^2)$

d. $g_1(\mathbf{x}) = \sin(2\pi x_1) + \cos(2\pi x_2) \qquad g_2(\mathbf{x}) = \sin(2\pi x_1) + \cos(2\pi x_2) + x_1$

e. $g_1(\mathbf{x}) = 2x_1^2 + 3x_2^2 \qquad g_2(\mathbf{x}) = 2x_1^2 + 3x_2^2 + \sin(2\pi x_1)$

f. $g_1(\mathbf{x}) = \exp(-x_1^2 - x_2^2) \qquad g_2(\mathbf{x}) = \exp(-x_1^2 - x_2^2) + \sin(2\pi x_1)$

Scenarios a-c represented situations under the null hypothesis, i.e., where the surfaces were the same; Scenarios d-f corresponded to various alternative

hypotheses. The independent variables x_1 and x_2 were simulated from independent $Unif[0,1]$ with sample size n_1 and n_2 for each group. The dependent variables Y_{ij} were generated from the above functions with standard errors σ_1 and σ_2. i.e. $Y_{ij} = g_i(\mathbf{x}_{ij}) + \epsilon_{ij}$, where $i = 1, 2; j = 1, \ldots, n_i$, $\epsilon_{1j} \sim N(0, \sigma_1^2)$, $\epsilon_{2j} \sim N(0, \sigma_2^2)$.

We conducted the simulation under the following parameter settings: (1) Three sample size settings of (n_1, n_2) as (125, 125), and (216, 216), (512,512); (2) Two different values of standard errors (σ_1, σ_2) for each function. For each simulation setting, we generated 500 datasets. For each dataset, we tested the new method with 300 wild bootstrap resamples. We calculated the rejection rate based on the 500 simulated datasets at $\alpha = 0.05$.

Type I error rates for function pairs of a-c are reported in Appendix B; see supplemental Table 4; powers for function pairs in d-f are reported in Table 5 of Appendix B. The Type I error rates of the proposed tests were generally good and power was superior than the existing tests. Similar to the simulation studies for curve comparisons, the penalized semiparametirc model with the default number of knots from 'GCV' method showed a performance similar to the tests using semiparametirc estimating methods and $\sqrt[3]{n_i}$ number of knots. Numbers of knots had relatively minor influences on the testing performance.

In comparison with the existing methods, we found that in general the proposed methods had reasonable Type I error control as expected. The power was either comparable to or superior than that of the other methods.

5.3. Tests with correlated data

We considered the following models for correlated data

$$Y_{ijk} = g_{id}(x_{ijk}) + b_{ij} + \epsilon_{ijk},$$

where $i = 1, 2; j = 1, \ldots, n_i; k = 1, 2, 3$. As previously presented in Equation (9), we used Y_{ijk} to indicate the measure on the kth occasion in subject j from group i. Values of the independent variable x_{ijk} were generated from independent $Unif[0,1]$. Values of the dependent variable Y_{ij} were generated based on the

above functions with random effect $b_i \sim N(0, \sigma_i')$ and the i.i.d. random error $\epsilon_{ijk} \sim N(0, \sigma_i)$. We used the same regression functions in Section 5.1, where the two curves gradually grew apart with an increasing d (see Appendix Figure 1).

We considered the following parameter settings: (1) d=0, 1, 2; (2) three sample size settings (n_1, n_2) = (50, 60), (100, 120), and (150, 160) and all with three repeated measures; (3) three different combinations of the standard deviations of the random intercept and the i.i.d random variable as $(\sigma_1', \sigma_2', \sigma_1, \sigma_2)$= (0.2, 0.15, 0.04, 0.05), (0.2, 0.15, 0.10, 0.12), and (0.25, 0.20, 0.10, 0.12). We used penalized semiparametric mixed regression to estimate the curves.

The simulation results are reported in supplemental Table 6 of Appendix B. Results suggested a relatively tight Type 1 error rate control and good power. Zhang's (2000) scaled chi-square test is the only existing comparative test for correlated data [24]. We compared the performance of the new test with that of the scaled chi-square test. Under sample sizes (n_1, n_2) = (50, 50) and (100, 100), we performed the test using 1) the same x values for two groups; 2) slightly different x ($x_2 = x_1 + Unif(0, 0.05)$); 3) completely random and independent x_1, x_2 for two groups. Simulation was repeated for 200 times under each scenario and the results were shown in Appendix B Table 7.

The proposed test clearly outperformed the scaled χ^2 test. When the two groups had the same values in x, the scaled χ^2 test had Type I error rates that were lower than the nominal level. The power was generally lower as well. The scaled chi-square test was not designed for situations of randomly distributed independent variables so the power deteriorated when we introduced different x values between the two groups.

6. Real data applications

To illustrate the proposed testing procedures, we analyzed two data sets from a large observational study. The original study was designed to examine the factors related to blood pressure development in children. Detailed study protocols were published elsewhere [34][35]. Briefly, healthy children between

5 and 17 years of age were recruited from schools in Indianapolis, Indiana. Blood pressure, height, and weight were measured twice a year from the study participants. Blood and overnight urine samples were collected. The study protocol was approved by a local Internal Review Board. Informed consent was obtained from study participants, or their parents when appropriate.

6.1. Comparisons of weight growth curves

We compared the weight growth curves between blacks and whites within each sex, and between boys and girls within each race. We write the model as

$$Weight_{ij} = g_i(Age_{ij}) + \epsilon_{ij},$$

where i indexes the groups and j the n_i observations within each group. Here we used the baseline assessment data to examine the weight-age relationships in the four sex and race combinations. The sizes of the four groups were: 205 black boys, 311 white boys, 232 black girls, and 289 white girls. We performed comparisons by testing the hypotheses $H_0 : g_1 = g_2$ vs $H_1 : g_1 \neq g_2$, where g_1, g_2 are the weight growth curves between the sexes within each race group, or weight growth functions between the races within each sex group.

We first estimated the weight growth curves of the groups as part of the preliminary analysis. See scatter plots in Figure 1(a). We reported the p values of the four competing testing methods in Table 3. We presented the curve estimates with 95% pointwise confidence intervals from the semiparametric regression model (Generalized Cross-Validation for selecting smoothing parameter and thin-plate penalized basis function) in Figure 1(b). The nonparametric smoothing curves by loess produced curve estimates that were similar to the semiparametric regression estimates.

Testing results from the semiparametric spline-based estimating method were consistent with the curve estimations shown in Figure 1. The tests showed that the weight-for-age curves were significantly different between white and black girls. In our sample, the black girls gained more weight around ages 12 and 13 than their white peers, but the two curves converged gradually at age 14

years. The confidence intervals became wider after age 15, possibly due to the reduced sample sizes. Similar patterns were seen in the height-for-age curves.

For surface comparison, we considered weight as a function of age and height. We wrote the model as

$$Weight_{ij} = g_i(Age_{ij}, Height_{ij}) + \epsilon_{ij},$$

where i is the index for the sex-race group, and j is the index for a specific subject, $j = 1, 2, \ldots, n_i$ within the group. We compared simultaneous effects of height and age on weight, among the four race-sex groups. The p-values of the four types of tests were summarized in Table 3 and the corresponding contour plots were presented in Figure 2 . No statistically significant differences were detected using the four tests.

[Figure 1 about here.]

[Figure 2 about here.]

[Table 3 about here.]

6.2. Hormonal influences on blood pressure

Blood pressure is regulated by hormones in the renin-angiotensin-aldosterone system (RAAS). An essential product of RAAS is aldosterone, a mineralocorticoid hormone. Aldosterone acts on the epithelial sodium channel (ENaC) to help retain sodium. Increased sodium load causes extracellular fluid volume (ECFV) expansion, which in turn leads to blood pressure elevation. Recent biological experiments have shown that in the American population, blacks are more responsive to the stimulation of aldosterone in comparison to whites [35]. As a result, blacks tend to have greater levels of ECFV, which helps to suppress renin secretion. Renin, together with potassium, helps production of aldosterone [36]. This process is essential for the maintenance of blood pressure [37].

In this analysis, we examined the simultaneous influences of plasma renin activity (PRA) and plasma aldosterone concentration (PAC) on systolic blood

pressure (BP):

$$BP_{ij} = g_i(log(PRA)_{ij}, log(PAC)_{ij}) + \epsilon_{ij},$$

where i is the index for the race group, and j is the index for a specific subject, $j = 1, 2, \ldots, n_i$ within the group. We were interested in comparing the surface functions in blacks and in whites. A quick visual examination (Figure 3) showed that blood pressure in whites (n=313) was on average lower than that in blacks (n=184). In whites, PRA and PAC were not significantly correlated with blood pressure, as one would expect in a steady-state sample. But in blacks, lower PRA and higher PAC were associated with a higher systolic blood pressure, which suggested an increased blood pressure sensitivity to aldosterone. We compared the two surface functions and obtained a p-value of 0.054, based on 500 bootstrap resamples. The comparison generated a p value that was close to, but did not reach the commonly accepted threshold of 0.05. Other testing methods, including Young and Bowman's (1995) tests (Method 4 and 5 in Section 5.1) gave p-values of 0.15 and 0.73, while the L_2 distance-based test using kernel smoothing (Method 3 in Section 5.1) provided a p-value of 0.03. Indication for the racial difference in this analysis is generally consistent with the experimental evidence from drug-induced hyperaldosteronism in human subjects [35]; the actual power of the tests, however, is likely influenced by the sample size and data variability.

[Figure 3 about here.]

7. Discussion

Statistical analysis of biomedical data is never complete without a proper test. In analysis of lower-dimensional nonlinear functions, inference typically involves comparisons of curves and surfaces. In parametric analysis where the functions are fully specified, inference is generally straightforward and can be carried out in the usual likelihood-based framework. In nonparametric or semi-parametric analyses, due to lack of knowledge of the true functional forms of the

relationships, hypotheses cannot be formulated solely on prespecified parameters. Analysts, therefore, can no longer rely on likelihood-based tests. Standard software packages or functions typically do not produce comparison of interest. In practice, estimation and inference of the functional curves and surfaces are often done separately, in part due to the lack of integration of estimation and inference tools and common programming syntax. In the present paper, we propose new testing procedures based on the L_2 distance. We show that the proposed tests are consistent against any fixed alternative hypothesis. To evaluate the level of statistical significance we provide a set of bootstrap testing methods. We have developed an `R` package to assist analysts who are interested in using the tests. For those who do not use `R`, we present an R Shiny interface to wrap around the software package so that analysts could directly upload their data to a web server and implement the tests through interactive web-based operations.

Extensive simulation studies show that, in comparison to the existing methods, the proposed tests have good control of type I error rate and excellent power. Despite our use of computer intensive methods such as the wild bootstrap, the procedures are generally quite efficient. Our own testing of the method with real data shows that the software package is easy to operate and it is flexible for accommodating covariates and repeated measures. We showed that the testing procedure possesses the property of consistency, a necessary condition for bootstrap to work. The rate of convergence, however, has not reached the optimum of $n^{1/2}$. The empirical evidence from our simulation has nonetheless supported the good performance in finite sample situations. Notwithstanding this limitation, we put forward a computational tool for the comparison of curves and surfaces in nonparametric or semiparametric analyses.

References

References

[1] P. Green, B. Silverman, Kernel nonparametric regression and generalized linear models: A roughness penalty approach (1994).

[2] J. Fan, I. gijbels, Local Polynomial Modelling and Its Applications.

[3] T. Ogden, Essential wavelets for statistical applications and data analysis, Springer Science & Business Media, 2012.

[4] G. Wahba, Spline models for observational data, Vol. 59, Siam, 1990.

[5] C. Gu, Smoothing spline ANOVA models, Vol. 297, Springer Science & Business Media, 2013.

[6] B. D. Marx, P. H. Eilers, Generalized linear regression on sampled signals and curves: a p-spline approach, Technometrics 41 (1) (1999) 1–13.

[7] R. L. Eubank, Nonparametric regression and spline smoothing, CRC press, 1999.

[8] C. Bde Boor, A practical guide to splines, revised edition (2001).

[9] T. Hastie, R. Tibshirani, Generalized additive models, London: Chapman and Hall (1990) 137–173.

[10] M. A. Delgado, Testing the equality of nonparametric regression curves, Statistics & probability letters 17 (3) (1993) 199–204.

[11] W. Härdle, J. S. Marron, et al., Semiparametric comparison of regression curves, The Annals of Statistics 18 (1) (1990) 63–89.

[12] J. Fan, S.-K. Lin, Test of significance when data are curves, Journal of the American Statistical Association 93 (443) (1998) 1007–1021.

[13] K. Kulasekera, Comparison of regression curves using quasi-residuals, Journal of the American Statistical Association 90 (431) (1995) 1085–1093.

[14] R. Eubank, C. Li, A diagnostic test for parallelism, Journal of Statistical Sciences (2008) 13–29.

[15] H. Dette, S. S. Dhar, W. Wu, Identifying shifts between two regression curves, arXiv preprint arXiv:1908.04328.

[16] S. G. Young, A. W. Bowman, Non-parametric analysis of covariance, Biometrics (1995) 920–931.

[17] A. W. Bowman, Comparing nonparametric surfaces, Statistical Modelling 6 (4) (2006) 279–299.

[18] C. Park, J. Hannig, K.-H. Kang, Nonparametric comparison of multiple regression curves in scale-space, Journal of Computational and Graphical Statistics 23 (3) (2014) 657–677.

[19] H. Dette, N. Neumeyer, et al., Nonparametric analysis of covariance, the Annals of Statistics 29 (5) (2001) 1361–1400.

[20] N. Neumeyer, H. Dette, et al., Nonparametric comparison of regression curves: an empirical process approach, The Annals of Statistics 31 (3) (2003) 880–920.

[21] J. C. Pardo-Fernández, I. Van Keilegom, W. González-Manteiga, Testing for the equality of k regression curves, Statistica Sinica (2007) 1115–1137.

[22] P. Hall, J. D. Hart, Bootstrap test for difference between means in nonparametric regression, Journal of the American Statistical Association 85 (412) (1990) 1039–1049.

[23] X.-F. Wang, D. Ye, On nonparametric comparison of images and regression surfaces, Journal of statistical planning and inference 140 (10) (2010) 2875–2884.

[24] D. Zhang, X. Lin, M. Sowers, Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles, Biometrics 56 (1) (2000) 31–39.

[25] S. N. Wood, Thin plate regression splines, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 (1) (2003) 95–114.

[26] S. N. Wood, Low-rank scale-invariant tensor product smooths for generalized additive mixed models, Biometrics 62 (4) (2006) 1025–1036.

[27] H. Liu, W. Tu, et al., A semiparametric regression model for paired longitudinal outcomes with application in childhood blood pressure development, The Annals of Applied Statistics 6 (4) (2012) 1861–1882.

[28] D. A. Freedman, et al., Bootstrapping regression models, The Annals of Statistics 9 (6) (1981) 1218–1228.

[29] R. Y. Liu, et al., Bootstrap procedures under some non-iid models, The Annals of Statistics 16 (4) (1988) 1696–1708.

[30] W. Härdle, E. Mammen, et al., Comparing nonparametric versus parametric regression fits, The Annals of Statistics 21 (4) (1993) 1926–1947.

[31] E. Mammen, et al., Bootstrap and wild bootstrap for high dimensional linear models, The Annals of Statistics 21 (1) (1993) 255–285.

[32] X. Shen, W. H. Wong, Convergence rate of sieve estimates, The Annals of Statistics (1994) 580–615.

[33] T. L. McMurry, D. N. Politis, Banded and tapered estimates for autocovariance matrices and the linear process bootstrap, Journal of Time Series Analysis 31 (6) (2010) 471–482.

[34] W. Tu, G. J. Eckert, L. A. DiMeglio, Z. Yu, J. Jung, J. H. Pratt, Intensified effect of adiposity on blood pressure in overweight and obese children, Hypertension 58 (5) (2011) 818–824.

[35] W. Tu, G. J. Eckert, T. S. Hannon, H. Liu, L. M. Pratt, M. A. Wagner, L. A. DiMeglio, J. Jung, J. H. Pratt, Racial differences in sensitivity of blood pressure to aldosterone, Hypertension 63 (6) (2014) 1212–1218.

[36] W. Tu, G. J. Eckert, J. H. Pratt, A. Jan Danser, Plasma levels of prorenin and renin in blacks and whites: their relative abundance and associations with plasma aldosterone concentration, American Journal of Hypertension 25 (9) (2012) 1030–1034.

[37] W. Tu, G. J. Eckert, B. S. Decker, J. Howard Pratt, Varying influences of aldosterone on the plasma potassium concentration in blacks and whites, American Journal of Hypertension 30 (5) (2017) 490–494.

Table 1: Summary of the existing methods

Author(s), Methods	**Same x(s)**	**Correlation**	**>2 Groups**	**Curve/Surface**	**Additional Comments**
Bowman (2006)[17]:	Y	N	N	Curve/Surface	(+) Simple to implement and understand as a derivation from ANOVA test; (-) Assume equal variance across groups.
Dette & Neumeyer (2001)[19], Pardo-Fernandez & Van Keilegom (2007)[21], Wang & Ye (2010)[23]:	N	N	Y	Curve/Surface	(+) Demonstrated asymptotic normality of all three kernel-based statistics under H_0; Recommended wild bootstrap when studying finite samples.
Zhang & Lin (1998)[24]:	Y	Y	N	Curve	(+) Spline-based semiparametric additive model; (-) χ^2 approximation can be biased with different covariate values.
Wang & Ye (2010)[23]:	N	Y	Y	Curve/Surface	(+) Able to adjust for spatial correlation; (-) Larger bias in estimating regression surface hence decreased power.
Kulasekera (1995)[13]:	N	N	N	Curve	(+) Low computational demand; (-) Low power when curve functions are similar.
Park, Hannig,& Kang (2014)[18]:	N	N	N	Curve	(+) A visualization tool to present differences between curves across multiple locations and scales; (-) Type 1 error rate below nominal level; relatively low power

Table 2: Average run times and standard errors in seconds over 20 runs of the proposed methods for each of the examples in the package, with and without using the parallel computing

Data	**Functions**	**Obs per group**	**Time (parallel)**	**Time (no parallel)**
cross-sectional	curve	(474,465)	9.6(0.2)	9.1(0.2)
	surface	(474,465)	14.8 (0.2)	23.5(0.6)
correlated	curve	(1873,1713)	86.4(0.6)	204.5(1.6)
	surface	(1873,1713)	664.6(13.6)	1749.0 (30.3)

Table 3: P-values for weight growth curves and weight growth surfaces for different race and sex groups

Endpoints vs predictor(s)	**Group effect**	**Subset data**	T_{spline}	T_4	T_3	T_2
Weight vs. Age	Sex	Black	0.08	0.16	0.55	<0.01
		White	0.48	0.18	0.69	0.28
	Race	Boys	0.41	0.09	0.69	0.2
		Girls	<0.01	0.01	0.96	<0.01
Weight vs. Age & Height	Sex	Black	0.16	0.65	0.66	0.49
		White	0.49	0.55	0.55	0.77
	Race	Boys	0.42	0.42	0.46	0.35
		Girls	0.34	0.19	0.91	0.06

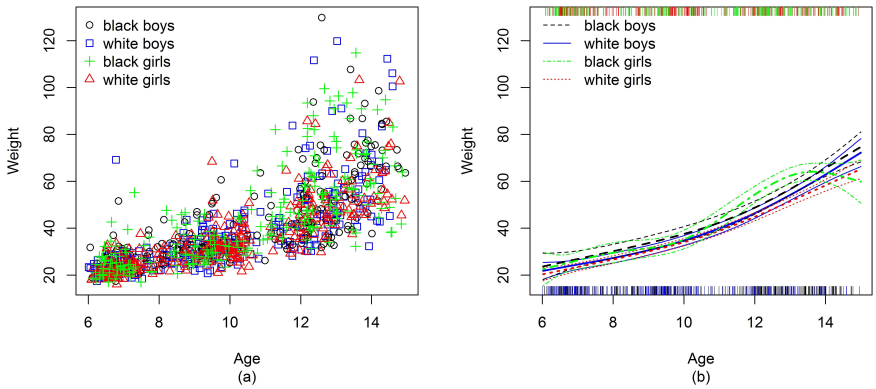

Figure 1: (a) Weight growth by race and sex; (b) Estimated weight growth curves with pointwise 95% CI, by race and sex

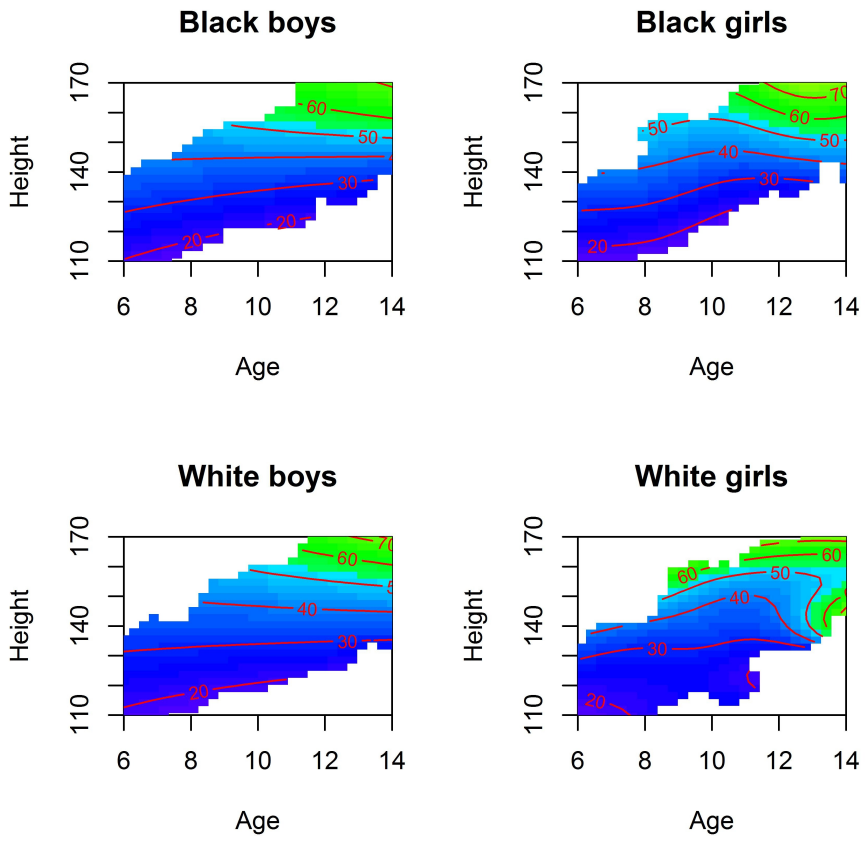

Figure 2: Estimated contour plots of weight as a function of height and age, by race and sex

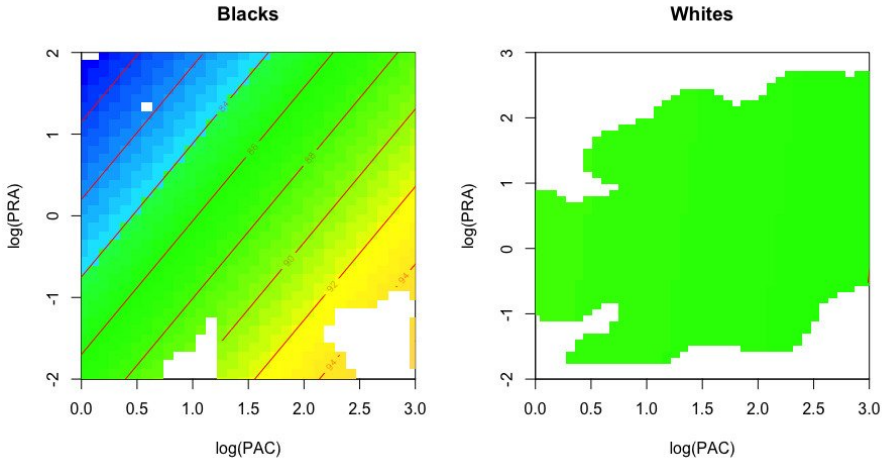

Figure 3: Estimated contour plots of systolic blood pressure as a function of logarithmic transformed plasma renin activity (PRA) and plasma aldosterone concentration (PAC), by race

Appendices

A Proof of Theorem 1

We consider the L_2-metric

$$d(g_1, g_2) = \|g_1 - g_2\|_{L_2(P)} = \left[P(g_1 - g_2)^2\right]^{\frac{1}{2}},$$

where Pf denotes the expectation of a measurable, data-dependent function f. Now, by Condition C1, it is easy to see that the expected objective function is

$$M(g) = Pm_g = -P(g - g_{0,i})^2, \quad i = 1, 2,$$

where $g_{0,i}$ is the true curve for the ith population. This implies the identifiability condition

$$\sup_{g: g \notin G} M(g) < M(g_{0,j})$$

for any neighborhood G_i of $g_{0,i}$.

For simplicity of presentation, we omit the subscript i in the remainder of the proof, and note that the same arguments regarding the consistency of the B-spline estimator hold for both groups. Then, following Kim et al. (2017)[1], we define a linear operator map $\mathcal{Q}$ from $\mathcal{G}$ to the sieve space $\mathcal{G}_n$, as:

$$\mathcal{Q}(\psi) = \sum_{k=1}^{K_n+m} \phi_k(\psi) B^m(x),$$

for any $\psi \in \mathcal{G}$, where $\{\phi_k\}_{k=1}^{K_n+m}$ are linear functionals in $L_\infty(\mathbb{X})$. Then we define $g_n(x) = \mathcal{Q}(g_0)$. Arguments similar to those used by Kim et al. (2017)[1] lead to the following inequality

$$\|g_n - g_0\|_{L_\infty(\mathbb{X})} \leq O(K_n^{-p}),$$

which also implies that

$$\|g_n - g_0\|_{L_2(P)} \leq O(K_n^{-p}). \tag{1}$$

It is straightforward to see that

$$d(\hat{g}_n, g_0) \leq d(\hat{g}_n, g_n) + d(g_n, g_0). \tag{2}$$

To show the convergence of the first term in the right side of (2) to 0, we let $\mathbb{M}_n(g)$ be the empirical objective function based on the data and $\mathbb{P}_n f(X) = n^{-1} \sum_{i=1}^n f(X_i)$. It follows that

$$\begin{aligned} \sup_{g \in \mathcal{G}_n} |\mathbb{M}_n(g) - M(g)| &\equiv \|\mathbb{M}_n(g) - M(g)\|_{\mathcal{G}_n} \\ &\lesssim \|\mathbb{P}_n(g - g_0)e\|_{\mathcal{G}_n} + \|(\mathbb{P}_n - P)(g - g_0)^2\|_{\mathcal{G}_n}. \end{aligned}$$

From Conditions C1, C2, and the Law of Large Numbers, we have $\|\mathbb{P}_n(g - g_0)e\|_{\mathcal{G}_n} = o_p(1)$.

For the second term we consider the class of functions $\mathcal{F}_n = \{(g - g_0)^2 : g \in \mathcal{G}_n\}$. A calculation by Shen and Wang (1994)[2] implies that $N_{[]}(\epsilon, \mathcal{G}_n, L_1(P)) \leq (1/\epsilon)^{c(K_n+m)}$. Based on the set of ϵ-brackets $\{[l_j, u_j] : j = 1, \ldots, (1/\epsilon)^{c(K_n+m)}\}$ in $L_1(P)$ for $\mathcal{G}_n$, we can construct a set of ϵ'-brackets $\{[l'_j, u'_j] : j = 1, \ldots, (1/\epsilon)^{c(K_n+m)}\}$, where

$$l'_j = \left[I(l_j \geq g_0)(l_j - g_0)^2 + I(u_j < g_0)(u_j - g_0)^2\right]\left[1 - I(l_j < g_0, u_j \geq g_0)\right]$$

and

$$\begin{aligned} u'_j &= I(l_j \geq g_0)(u_j - g_0)^2 + I(u_j < g_0)(l_j - g_0)^2 \\ &\quad + I(l_j < g_0, u_j \geq g_0)\max((u_j - g_0)^2, (l_j - g_0)^2) \end{aligned}$$

in $L_1(P)$ for $\mathcal{F}_n$. Therefore, $N_{[]}(\epsilon, \mathcal{F}_n, L_1(P)) < \infty$ for any $\epsilon > 0$. This implies that $\|(\mathbb{P} - P)(g - g_0)^2\|_{\mathcal{G}_n} \overset{as*}{\to} 0$, and thus $\|\mathbb{M}_n(g) - M(g)\|_{\mathcal{G}_n} = o_p(1)$. This fact along with the inequality

$$M(g) - M(g_n) \leq -\frac{1}{4}P(g - g_n)^2,$$

for any g with $P(g - g_n)^2 \geq 4P(g_n - g_0)^2$ (van der Vaart and Wellner, 1996 [3]) implies that $d(\hat{g}_n, g_n) = o_p(1)$. The second term on the right side of (2) is $o(1)$ by condition C3 and inequality (1) and this leads to $d(g_n, g_0) = o(1)$ and, therefore, $d(\hat{g}_n, g_0) \overset{p}{\to} 0$.

For the rate of convergence we consider the key inequality

$$E^\star \sup_{P(g-g_n)^2 \leq \delta^2, g \in \mathcal{G}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (g - g_n)(X_i)e_i \right| \lesssim \tilde{J}_{[]}(\delta, \mathcal{G}_n(\delta), L_2(P)) \left[1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{G}_n(\delta), L_2(P))}{\delta^2\sqrt{n}}\right]$$

holds with $\mathcal{G}_n(\delta) = \{g : g \in \mathcal{G}_n, d(g, g_n) < \delta\}$, given in p. 335 in van der Vaart and Wellner (1996) [3]. The calculation by Shen and Wang (1994)[2] implies that the ϵ-bracketing number for the class $\mathcal{G}_n(\delta)$ is bounded by $(\delta/\epsilon)^{c(K_n+m)}$. Therefore,

$$\tilde{J}_{[]}(\delta, \mathcal{G}_n(\delta), L_2(P)) = \int_0^\delta \sqrt{1 + c(K_n + m)\log\left(\frac{\delta}{\epsilon}\right)}d\epsilon \leq c(K_n + m)^{1/2}\delta.$$

Thus, the key function $\phi_n(\delta)$ given in Theorem 3.4.1. in van der Vaart and Wellner (1996) [3] is

$$\phi_n(\delta) = (K_n + m)^{1/2}\delta + \frac{(K_n + m)}{\sqrt{n}}.$$

After some algebra we conclude that

$$n^{2pv}\phi_n\left(\frac{1}{n^{pv}}\right) \leq \sqrt{n}$$

if $pv \leq (1-v)/2$. Thus, if the rate $r_n = \min(pv, (1-v)/2)$ then

$$r_n^2\phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}.$$

Also, it can be argued that $\mathbb{M}_n(\hat{g}_n) - \mathbb{M}_n(g_0) \geq -O_p(r_n^{-2})$. If $v = 1/(1+2p)$, Theorem 3.4.1 in van der Vaart and Wellener (1996) [3] and (1) and (2) imply that

$$d(\hat{g}_n, g_0) = O_p(n^{-\frac{p}{1+2p}}).$$

Next, consider the test statistic based on two independent samples of size n_1 and n_2

$$\frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{n_i}[\hat{g}_1(\mathbf{x}_{ij}) - \hat{g}_2(\mathbf{x}_{ij})]^2 \equiv \mathbb{P}_N(\hat{g}_1 - \hat{g}_2)^2,$$

where $N = n_1 + n_2$.

Let us consider the consistency of the test statistic to the true L^2 distance between the two curves under the probability measure P underlying X_{ij}, $i = 1, 2$, $j = 1, \ldots, n_i$. It is straightforward to show that

$$\begin{aligned}\left|\mathbb{P}_N(\hat{g}_1 - \hat{g}_2)^2 - P(g_1 - g_2)^2\right| \quad &\leq \quad \left|\mathbb{P}_N[(\hat{g}_1 - \hat{g}_2)^2 - (g_1 - g_2)^2]\right| \\ &\quad + \left|(\mathbb{P}_N - P)(g_1 - g_2)^2)\right|. \qquad (3)\end{aligned}$$

The second term in (3) is $o_p(1)$, as a consequence of the condition C2 and the law of large numbers. Now see that

$$\begin{aligned}\mathbb{P}_N(\hat{g}_i - g_i)^2 \quad &= \quad \mathbb{P}_N\left\{\sum_{k=1}^{K_{n_i}+m}[\hat{\phi}_{k,i}(g_i) - \phi_{k,i}(g_i)]B_k^m + \left[\sum_{k=1}^{K_n+m}\phi_{k,i}(g_i)B_k^m - g_i\right]\right\}^2 \\ &\equiv \quad \mathbb{P}_N\left\{\sum_{k=1}^{K_{n_i}+m}[\hat{\phi}_{k,i}(g_i) - \phi_{k,i}(g_i)]B_k^m + (g_{n_i,i} - g_i)\right\}^2 \\ &\leq \quad 2\left\{\sum_{k=1}^{K_{n_i}+m}[\hat{\phi}_{k,i}(g_i) - \phi_{k,i}(g_i)]\right\}^2 \mathbb{P}_N\left(\max_{k=1,\ldots,K_{n_i}+m} B_k^m\right)^2 \\ &\quad +2P(g_{n_i,i} - g_i)^2 + o_p(1).\end{aligned}$$

By the uniform boundedness of the B-spline basis functions and the consistency of the estimator of the control points $\hat{\phi}_{k,i}(g_i)$ from the fact that $d(\hat{g}_{i,n}, g_{i,n}) = o_p(1)$ shown above, it follows that the first term in the right side of the above

inequality is $o_p(1)$. Also, the second term in the right side of the above inequality is $o_p(1)$, as it was argued above. Therefore,

$$\mathbb{P}_N(\hat{g}_i - g_i)^2 = o_p(1), \qquad i = 1, 2. \tag{4}$$

(3) can be written into

$$\begin{aligned} \left|\mathbb{P}_N[(\hat{g}_1 - \hat{g}_2)^2 - (g_1 - g_2)^2]\right| &\leq \sum_{i=1}^{2} \left|\mathbb{P}_N(\hat{g}_i^2 - g_i^2)\right| + \left|\mathbb{P}_N \hat{g}_1(\hat{g}_2 - g_2)\right| \\ &\quad + \left|\mathbb{P}_N g_2(\hat{g}_1 - g_1)\right| \\ &\equiv \sum_{i=1}^{2} A_{N,i} + B_N + C_N \end{aligned} \tag{5}$$

It is not hard to see that the first term of (5) becomes

$$\begin{aligned} \sum_{i=1}^{2} A_{N,i} &\leq \sum_{i=1}^{2} \mathbb{P}_N\left(\hat{g}_i - g_i\right)^2 + \sum_{i=1}^{2} \mathbb{P}_N\left|2g_i(\hat{g}_i - g_i)\right| \\ &\leq \sum_{i=1}^{2} \mathbb{P}_N(\hat{g}_i - g_i)^2 + K \sum_{i=1}^{2} \mathbb{P}_N\left|\hat{g}_i - g_i\right| \\ &\leq \sum_{i=1}^{2} \mathbb{P}_N(\hat{g}_i - g_i)^2 + K \sum_{i=1}^{2} [\mathbb{P}_N(\hat{g}_i - g_i)^2]^{1/2} \\ &= o_p(1), \end{aligned}$$

where K represents a generic constant with $K \in (0, \infty)$. By C2 and (5), $i = 1, 2$, we have that for a sufficiently large N

$$(B_N + C_N) \leq K \sum_{i=1}^{2} [\mathbb{P}_N(\hat{g}_i - g_i)^2]^{1/2} = o_p(1).$$

Therefore,

$$\mathbb{P}_N(\hat{g}_1 - \hat{g}_2)^2 \xrightarrow{p} P(g_1 - g_2)^2.$$

This result along with Lemma 14.15 of van der Vaart (2000)[4] leads to the consistency of the proposed test against every fixed alternative hypothesis with $g_1 \neq g_2$. Thus, the proof of Theorem 1 is complete.

B Simulation Results

In Part B of the Appendix, we summarize the results from the simulation studies. Details of the simulation were presented in Section 5 of the manuscript.

B.1 Comparisons of curve functions

We report the average type I error rates and power for parameter settings considered in the simulation. Simulations for curve comparisons were based on 1000 generated samples. Tests were performed at the 0.05 significance level.

Notation for each testing method in the Table 1 is consistent with that presented in the main manuscript. Method 1 $T_{B-spline}$: L^2 distance of pointwise B-spline regression functions with $k = \sqrt[3]{n_i}$; Method 2 $T_{P-spline}$: P-spline regression functions with the default number of knots from GCV ($k \approx 6$ in this example); Method 3 T_4: Kernel-based regression function $\tilde{g}_i(x)$, $T_4 = \sum_{i=1}^{I}\sum_{j=1}^{i-1}[\tilde{g}_i(x) - \tilde{g}_j(x)]^2$; Method 4 T_2: Variance estimating method, $T_2 = \hat{\sigma}^2 - \frac{1}{N}\sum_{i=1}^{I} n_i\hat{\sigma}_i^2$; Method 5 T_1: The scaled chi-square test, $T_1 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\tilde{g}(x_{ij}) - \tilde{g}_i(x_{ij})]^2$.

[Figure 1 about here.]

[Table 1 about here.]

We report the average type I error rates and power for parameter settings considered in the simulation.

Simulations for three-group curve comparisons were based on 500 generated samples. Tests were performed at the 0.05 significance level. Simulation results were presented in Table 2.

[Table 2 about here.]

B.2 Exploration of asymptotic property

To confirm the theoretical results that power increases with the sample size, we randomly chose a scenario that $d = 1$, $(\sigma_1, \sigma_2) = (0.2, 0.15)$ in two curve comparisons based on 1000 simulated samples at the significance level of 0.05. Numerical results are presented in Table 3, and power curve in Figure 2.

[Table 3 about here.]

[Figure 2 about here.]

B.3 Comparisons of surface functions

The average type I error rate and power for various parameter settings in surface comparisons based on 500 simulated samples at the significance level of 0.05 are presented in Tables 4 and 5, where "TP-Spline", "TP-Spline$_+$" and "TP-Spline$_-$" indicate tests using thin-plate splines with $\sqrt[3]{n_i}$, $\sqrt[3]{n_i}+1$, and $\sqrt[3]{n_i}-1$

knots. "TE" indicates testing method using tensor-product basis functions, "TP-Spline.p" and "TE-Spline.p" are tests using penalized splines. T_4(Method 3): Kernel based estimating regression function $\tilde{g}_i(\mathbf{x})$, $T_4 = \sum_{i=1}^{I}\sum_{j=1}^{i-1}[\tilde{g}_i(\mathbf{x}) - \tilde{g}_j(\mathbf{x})]^2$; T_2(Method 4): variance estimating method, $T_2 = \hat{\sigma}^2 - \frac{1}{N}\sum_{i=1}^{I} n_i\hat{\sigma}_i^2$; T_3(revised Method 5): $T_3 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\hat{g}(\mathbf{x}_{ij}) - \hat{g}_i(\mathbf{x}_{ij})]^2$; T_1(Method 5): matching with a scaled chi-square distribution, $T_1 = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\tilde{g}(\mathbf{x}_{ij}) - \tilde{g}_i(\mathbf{x}_{ij})]^2$.

[Table 4 about here.]

[Table 5 about here.]

B.4 Models with correlated data

We present the average type I error rates and power for simulation settings in curve comparisons based on 1000 simulated samples with certain correlations at the significance level of 0.05; see Tables 6. We generated 200 simulated samples for comparison of the rejection probabilities between the proposed methods and Zhang's scaled χ^2 testing method. See Table 7.

[Table 6 about here.]

[Table 7 about here.]

C Package gamm4.test and its R Shiny interface

As described in the manuscript, we developed an R package, gamm4.test, together with an interactive interface for the implementation of the proposed tests. In Part C of the Appendix, we briefly illustrate the use of the package.

C.1 Analysis of cross-sectional data: An example

Using the pubertal growth study data as an example, we compared the weight-for-age curves between boys and girls, using only the baseline measurements. The data are cross-sectional and uncorrelated.

```
R> library("gamm4.test")
R> data("outchild")
R> child <- outchild[order(outchild$SID,outchild$age),]
R> bs <- aggregate(.~SID, child, FUN=head, 1)

R> childcur <- bs[,c("SEX","WEIGHT","age")]
R> test.grpsex1 <- gam.grptest(WEIGHT~s(age, bs="cr"),
      test=~SEX,data=childcur)
R> test.grpsex1
```

The output from the program thus far is as follows

```
Test the equality of curves based on L2 distance

Comparing 2 semiparametric regression curves
Penalized semiparametric regression is used for curve fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 2 curves.
T =  71.92     p-value =  0.01493
```

The density function of the test statistic under the null hypothesis estimated from the bootstrap samples can be obtained by using function

```
plot(test.grpsex1, test.statistic=TRUE).
```

The function provides the flexibility of displaying either a histogram or a density curve through option test.stat.type. In Figure 3 we request a histogram by using

```
plot(test.grpsex1, test.statistic=TRUE, test.stat.type="hist")
```

[Figure 3 about here.]

Function `gam.grptest` is the main function for comparing the curves. Options `bs=` and `k=` can be used in the model formula for changing the basis functions and for specifying the number of knots. If no value for `k` is provided, a penalized semiparametric model estimation with the default number of knots will be used. The following are some sample commands:

```
R> test.grpsex1 <- gam.grptest(WEIGHT~s(age, bs="tp"),test=~SEX,
      data=childcur)
      #penalized thin-plate spline basis
R> test.grpsex1 <- gam.grptest(WEIGHT~s(age, bs="tp", k=5), test=~SEX,
       data=childcur)
      #thin-plate spline basis with five equally spaced knots
      #over the range of variable age
```

Option `N.boot` specifies the number of bootstrap samples. The default value is `N.boot = 200`. Option `parallel=TRUE` calls for parallel computing and it distributes the computational burden to all available CPU cores.

```
R> test.grpsex1 <- gam.grptest(WEIGHT~s(age, bs="cr"),test=~SEX,
      data=childcur, N.boot=300, parallel= TRUE)
R> test.grpsex1
```

The following code produces a plot of the estimated curves with a 95% pointwise confidence interval. The plot is shown in Figure 4.

```
R> plot(test.grpsex1)
R> plot(test.grpsex1, se.est=TRUE)
```

[Figure 4 about here.]

Similarly, one could use the function `gam.grptest` for comparison of surface functions. In the pubertal growth example, we express the body weight as a function of age and height, i.e., $WEIGHT = f(HEIGHT, age)$. The following code produces a comparison of the function f between boys and girls.

```
R> childsurf <- bs[,c("SEX","HEIGHT","WEIGHT","age")]
R> test.grpsex2 <- gam.grptest(WEIGHT~s(HEIGHT,age),test=~SEX,
      data=childsurf)
R> test.grpsex2
R> plot(test.grpsex2)
R> plot(test.grpsex2,type="persp",theta=-35,phi=40)
R> plot(test.grpsex2,type="plotly.persp")
```

with the following output:

```
Test the equality of surfaces based on L2 distance

Comparing 2 semiparametric regression surfaces
```

```
Penalized semiparametric regression is used for surface fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 2 surfaces.
T =  20.92     p-value =  0.4179
```

The option `type=plotly.persp` generates an interactive 3-D plot using the R package `plotly`. The generated plots are shown in Figure 5 and plotly output is uploaded on https://zhaoshi169.github.io/chap5plotlysuf.html.

[Figure 5 about here.]

The R package `fANCOVA` provides functions for kernel-based testing methods. We made a minor correction for the function `T.L2` in our package with a new function name `T.L2c`, which helps to conduct two group comparison.

```
R> n1 <- 200
R> x1 <- runif(n1,min=0, max=3)
R> sd1 <- 0.2
R> e1 <- rnorm(n1,sd=sd1)
R> y1 <- sin(2*x1) + cos(2*x1) + e1
R>
R> n2 <- 120
R> x2 <- runif(n2, min=0, max=3)
R> sd2 <- 0.25
R> e2 <- rnorm(n2, sd=sd2)
R> y2 <- sin(2*x2) + cos(2*x2) + x2 + e2
R>
R> dat <- data.frame(rbind(cbind(x1,y1,1), cbind(x2,y2,2)))
R> colnames(dat)=c('x','y','group')
R>
R> T.L2c(formula=y~x,test=~group,data=dat)
R> gam.grptest(y~s(x,bs="cr"), test=~group, data=dat,
      parallel=TRUE)
R> library(fANCOVA)
R> T.aov(dat$x, dat$y, dat$group)
R> T.var(dat$x, dat$y, dat$group)
```

In this simple example, all testing methods correctly reject the null hypothesis. But as we have shown in the simulation studies, different testing methods do have different operating characteristics.

C.2 Analysis of correlated data

We use the same example to illustrate the analysis of correlated data. Here, we included data from all the visits for the analysis. The goal is to compare the growth functions over age between boys and girls.

The R code for data preparation and testing is presented as below.

```
R> data("outchild")
R> child.rep <- outchild[(outchild$age<16 &outchild$age>10),]
R> child.reptest1 <-  gamm4.grptest(HEIGHT~s(age,bs="cr"),
      random=~(1|SID),test=~SEX,data=child.rep)
R> child.reptest1
R> plot(child.reptest1,test.statistic=TRUE)
R> plot(child.reptest1)
```

The output is as follows:

```
Test the equality of curves based on L2 distance

Comparing 2 semiparametric regression curves
Penalized semiparametric regression mixed model is used for curve fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 2 curves.
T =  33.91     p-value =  0.004975
```

As expected, the height-for-age growth curves are significantly different between the sexes (p<0.001). The empirical distribution of the test statistic under the null hypothesis shows that the value of the test statistic, $T = 33.91$, is located to the far right of the plotted range. At the same time, the 95% pointwise confidence bands were relatively narrow, showing diverging growth patterns around the time of puberty. See Figure 6.

[Figure 6 about here.]

The height-for-age scatter plot is presented in Figure 6. The corresponding regression curves and the 95% pointwise confidence intervals from the semiparametric mixed model analysis were presented in Figure 6.

To implement surface comparisons, we conducted hypothesis testing on the simultaneous nonlinear effects of height and age on weight between boys and girls among blacks.

```
R> child.repw <- child.rep[(child.rep$RACE==1),]
R> child.reptest2 <- gamm4.grptest(WEIGHT~t2(age,HEIGHT),
      random=~(1|SID),test=~SEX,data=child.repw)
R> child.reptest2
R> plot(child.reptest2,type="contour")
R> plot(child.reptest2,type="persp",theta=-35,phi=40)
```

which produces the following output:

```
Test the equality of surfaces based on L2 distance

Comparing 2 semiparametric regression surfaces
Penalized semiparametric regression mixed model is used for surface fitting.
```

```
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 2 surfaces.
T =  10.88     p-value =  0.0995
```

Plots are shown in Figure 7. The option `type=plotly.persp` generates an interactive 3-D plot using the R package `plotly`.

[Figure 7 about here.]

C.3 The R Shiny Interface

To enhance the usability of the testing methods, we created an interactive R Shiny interface for the `gamm4.test` package. This interface allows analysts that do not use R to access the testing procedure through a web link. The interface can be access at https://heather.shinyapps.io/shinygamm4/ and a youtube tutorial is available at https://youtu.be/SHqaZXSLaMw.

To illustrate, we used the `outchild` data from the "gamm4.test" package as an example.

We first put the observed data in the pre-specified format.

```
colnumes(childcur) <- c("grp", "y", "age")
```

Then click "Enter Data" to upload the dataset. To compare curves and plot the estimated regression functions, click "Test summary and plots".

For surface comparison, we first put the raw data in the specified format.

```
R> colnumes(childsurf) <- c("grp", "y", "x1", "x2")
```

We then compare surfaces by clicking "Enter Data" and "Test summary and plots".

References

[1] Kim S, Zeng D, and Taylor J. Joint Partially Linear Model for Longitudinal Data with Informative Drop-Outs. *Biometrics* 2017; 73: 72-82.

[2] Shen X and Wong WH. Convergence rate of sieve estimates. *The Annals of Statistics* 1994; 22: 580-615.

[3] Van der Vaart AW and Wellner JA. Weak convergence. *In Weak convergence and empirical processes*. Springer, New York, NY, 1996. 16-28p.

[4] Van der Vaart AW. *Asymptotic statistics (Vol. 3)*. Cambridge university press, 2000.

Table 1: Two group curve comparison: Power and Type I error rates.

d	(n_1, n_2)	(σ_1, σ_2)	$T_{B-spline}$	$T_{P-spline}$	T_4	T_2	T_1
0	(125, 125)	(0.2, 0.15)	0.051	0.049	0.070	0.043	0.060
	(216, 216)	(0.2, 0.15)	0.048	0.057	0.071	0.046	0.067
	(512, 512)	(0.2, 0.15)	0.055	0.060	0.066	0.049	0.063
	(125, 125)	(0.25, 0.2)	0.049	0.051	0.065	0.043	0.060
	(216, 216)	(0.25, 0.2)	0.051	0.052	0.071	0.058	0.072
	(512, 512)	(0.25, 0.2)	0.060	0.047	0.065	0.057	0.066
1	(125, 125)	(0.2, 0.15)	0.416	0.349	0.406	0.309	0.415
	(216, 216)	(0.2, 0.15)	0.688	0.621	0.655	0.529	0.666
	(512, 512)	(0.2, 0.15)	0.969	0.967	0.973	0.926	0.972
	(125, 125)	(0.25, 0.2)	0.274	0.226	0.321	0.208	0.302
	(216, 216)	(0.25, 0.2)	0.434	0.379	0.446	0.353	0.469
	(512, 512)	(0.25, 0.2)	0.824	0.802	0.850	0.735	0.848
2	(125, 125)	(0.2, 0.15)	0.974	0.941	0.956	0.920	0.958
	(216, 216)	(0.2, 0.15)	1.000	1.000	1.000	0.995	1.000
	(512, 512)	(0.2, 0.15)	1.000	1.000	1.000	1.000	1.000
	(125, 125)	(0.25, 0.2)	0.844	0.764	0.830	0.725	0.821
	(216, 216)	(0.25, 0.2)	0.985	0.970	0.982	0.952	0.983
	(512, 512)	(0.25, 0.2)	1.000	1.000	1.000	1.000	1.000
3	(125, 125)	(0.2, 0.15)	1.000	1.000	0.999	1.000	0.999
	(216, 216)	(0.2, 0.15)	1.000	1.000	1.000	1.000	1.000
	(512, 512)	(0.2, 0.15)	1.000	1.000	1.000	1.000	1.000
	(125, 125)	(0.25, 0.2)	0.995	0.990	0.992	0.981	0.995
	(216, 216)	(0.25, 0.2)	1.000	1.000	1.000	1.000	1.000
	(512, 512)	(0.25, 0.2)	1.000	1.000	1.000	1.000	1.000

Table 2: Three-group curve comparison: Power and Type I error rates.

(d_1, d_2)	(n_1, n_2, n_3)	$T_{B-spline}$
(0, 0)	(100, 50, 75)	0.044
(0, 1)	(100, 50, 75)	0.074
(0, 2)	(100, 50, 75)	0.166
(0, 3)	(100, 50, 75)	0.356
(1, 1)	(100, 50, 75)	0.134
(1, 2)	(100, 50, 75)	0.296
(1, 3)	(100, 50, 75)	0.512
(2, 2)	(100, 50, 75)	0.482
(2, 3)	(100, 50, 75)	0.698
(3, 3)	(100, 50, 75)	0.864
(0, 0)	(200, 100, 150)	0.054
(0, 1)	(200, 100, 150)	0.110
(0, 2)	(200, 100, 150)	0.346
(0, 3)	(200, 100, 150)	0.760
(1, 1)	(200, 100, 150)	0.256
(1, 2)	(200, 100, 150)	0.618
(1, 3)	(200, 100, 150)	0.906
(2, 2)	(200, 100, 150)	0.888
(2, 3)	(200, 100, 150)	0.986
(3, 3)	(200, 100, 150)	1.000
(0, 0)	(300, 150, 225)	0.064
(0, 1)	(300, 150, 225)	0.148
(0, 2)	(300, 150, 225)	0.494
(0, 3)	(300, 150, 225)	0.894
(1, 1)	(300, 150, 225)	0.402
(1, 2)	(300, 150, 225)	0.790
(1, 3)	(300, 150, 225)	0.970
(2, 2)	(300, 150, 225)	0.960
(2, 3)	(300, 150, 225)	1.000
(3, 3)	(300, 150, 225)	1.000

Table 3: Changes of power with sample size.

(n_1, n_2)	(σ_1, σ_2)	$T_{B-spline}$	$T_{P-spline}$
(1000, 1000)	(0.2, 0.15)	1.000	1.000
(729, 729)	(0.2, 0.15)	0.995	0.993
(512, 512)	(0.2, 0.15)	0.978	0.968
(343, 343)	(0.2, 0.15)	0.871	0.837
(216, 216)	(0.2, 0.15)	0.667	0.613
(125, 125)	(0.2, 0.15)	0.408	0.345

Table 4: Surface comparison: Type I error rates.

Func	(n_1, n_2)	(σ_1, σ_2)	**TP-Spline**	**TP-Spline**$_+$	**TP-Spline**$_-$	**TE-Spline**	**TP-Spline.p**	**TE-Spline.p**	T_4	T_1	T_3	T_2
a	(125, 125)	(0.5, 0.3)	0.088	0.10	0.088	0.10	0.106	0.094	0.09	0.038	0.15	0.002
	(216, 216)	(0.5, 0.3)	0.076	0.080	0.094	0.07	0.07	0.068	0.088	0.02	0.110	0.012
	(512, 512)	(0.5, 0.3)	0.074	0.068	0.058	0.080	0.046	0.06	0.064	0.026	0.064	0.012
	(125, 125)	(0.6, 0.4)	0.098	0.086	0.09	0.070	0.080	0.080	0.08	0.02	0.114	0.002
	(216, 216)	(0.6, 0.4)	0.078	0.080	0.100	0.07	0.078	0.068	0.078	0.02	0.120	0.010
	(512, 512)	(0.6, 0.4)	0.05	0.06	0.07	0.060	0.06	0.056	0.056	0.016	0.08	0.012
b	(125, 125)	(0.6, 0.4)	0.06	0.05	0.044	0.068	0.048	0.060	0.058	0.066	0.060	0.052
	(216, 216)	(0.6, 0.4)	0.044	0.038	0.038	0.034	0.044	0.038	0.070	0.080	0.088	0.026
	(512, 512)	(0.6, 0.4)	0.054	0.048	0.05	0.040	0.05	0.05	0.080	0.07	0.076	0.038
	(125, 125)	(0.8, 0.6)	0.060	0.058	0.06	0.060	0.05	0.064	0.074	0.080	0.078	0.068
	(216, 216)	(0.8, 0.6)	0.066	0.07	0.064	0.066	0.05	0.05	0.074	0.076	0.084	0.038
	(512, 512)	(0.8, 0.6)	0.064	0.058	0.06	0.050	0.058	0.050	0.060	0.060	0.06	0.026
c	(125, 125)	(0.8, 0.6)	0.038	0.040	0.04	0.056	0.048	0.040	0.040	0.046	0.04	0.056
	(216, 216)	(0.8, 0.6)	0.05	0.05	0.054	0.048	0.046	0.03	0.046	0.050	0.048	0.056
	(512, 512)	(0.8, 0.6)	0.046	0.050	0.04	0.024	0.04	0.040	0.05	0.044	0.046	0.054
	(125, 125)	(1, 0.8)	0.046	0.044	0.04	0.040	0.046	0.044	0.048	0.05	0.060	0.056
	(216, 216)	(1, 0.8)	0.066	0.05	0.054	0.050	0.058	0.038	0.050	0.046	0.050	0.052
	(512, 512)	(1, 0.8)	0.04	0.044	0.044	0.034	0.038	0.040	0.054	0.064	0.054	0.068

Table 5: Surface comparison (Cont.): Power.

Func	(n_1, n_2)	(σ_1, σ_2)	TP-Spline	TP-Spline$_+$	TP-Spline$_-$	TE-Spline	TP-Spline.p	TE-Spline.p	T_4	T_1	T_3	T_2
d	(125, 125)	(0.5, 0.3)	0.940	0.944	0.952	0.806	0.944	0.826	0.870	0.784	0.940	0.592
	(216, 216)	(0.5, 0.3)	0.998	0.998	1.000	0.984	0.998	0.990	0.988	0.976	0.994	0.960
	(512, 512)	(0.5, 0.3)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(125, 125)	(0.6, 0.4)	0.792	0.780	0.786	0.656	0.804	0.660	0.742	0.604	0.834	0.400
	(216, 216)	(0.6, 0.4)	0.968	0.958	0.968	0.930	0.972	0.942	0.942	0.902	0.964	0.800
	(512, 512)	(0.6, 0.4)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
e	(125, 125)	(0.6, 0.4)	0.960	0.962	0.956	0.932	0.968	0.928	0.894	0.918	0.932	0.934
	(216, 216)	(0.6, 0.4)	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.998	0.998
	(512, 512)	(0.6, 0.4)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(125, 125)	(0.8, 0.6)	0.730	0.724	0.716	0.618	0.730	0.640	0.650	0.646	0.662	0.706
	(216, 216)	(0.8, 0.6)	0.950	0.946	0.944	0.908	0.944	0.936	0.874	0.880	0.890	0.906
	(512, 512)	(0.8, 0.6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000
f	(125, 125)	(0.8, 0.6)	0.742	0.742	0.734	0.632	0.736	0.650	0.696	0.684	0.698	0.734
	(216, 216)	(0.8, 0.6)	0.960	0.958	0.962	0.908	0.960	0.926	0.860	0.882	0.894	0.960
	(512, 512)	(0.8, 0.6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(125, 125)	(1, 0.8)	0.478	0.500	0.490	0.386	0.482	0.388	0.484	0.502	0.500	0.490
	(216, 216)	(1, 0.8)	0.772	0.784	0.784	0.680	0.776	0.736	0.680	0.714	0.722	0.796
	(512, 512)	(1, 0.8)	0.998	0.998	0.998	0.992	0.998	0.996	0.972	0.992	0.990	1.000

Table 6: Type I error rates and power of comparisons with correlated data.

d	(n_1, n_2)	σ_1'	σ_2'	σ_1	σ_2	$T_{splinecorr}$
0	(50,60)	0.20	0.15	0.04	0.05	0.057
	(50,60)	0.20	0.15	0.10	0.12	0.058
	(50,60)	0.25	0.20	0.10	0.12	0.062
	(100,120)	0.20	0.15	0.04	0.05	0.066
	(100,120)	0.20	0.15	0.10	0.12	0.072
	(100,120)	0.25	0.20	0.10	0.12	0.068
	(150,160)	0.20	0.15	0.04	0.05	0.048
	(150,160)	0.20	0.15	0.10	0.12	0.041
	(150,160)	0.25	0.20	0.10	0.12	0.053
1	(50,60)	0.20	0.15	0.04	0.05	0.426
	(50,60)	0.20	0.15	0.10	0.12	0.383
	(50,60)	0.25	0.20	0.10	0.12	0.212
	(100,120)	0.20	0.15	0.04	0.05	0.968
	(100,120)	0.20	0.15	0.10	0.12	0.784
	(100,120)	0.25	0.20	0.10	0.12	0.516
	(150,160)	0.20	0.15	0.04	0.05	1.000
	(150,160)	0.20	0.15	0.10	0.12	0.941
	(150,160)	0.25	0.20	0.10	0.12	0.804
2	(50,60)	0.20	0.15	0.04	0.05	1.000
	(50,60)	0.20	0.15	0.10	0.12	0.994
	(50,60)	0.25	0.20	0.10	0.12	0.956
	(100,120)	0.20	0.15	0.04	0.05	1.000
	(100,120)	0.20	0.15	0.10	0.12	1.000
	(100,120)	0.25	0.20	0.10	0.12	1.000
	(150,160)	0.20	0.15	0.04	0.05	1.000
	(150,160)	0.20	0.15	0.10	0.12	1.000
	(150,160)	0.25	0.20	0.10	0.12	1.000

Table 7: Comparison of the rejection probabilities between the proposed method and Zhang et al's scaled χ^2 testing method

			$\mathbf{x_2} = \mathbf{x_1}$		$\mathbf{x_2} = \mathbf{x_1} + U(0, 0.05)$		Random $\mathbf{x_1}, \mathbf{x_2} \sim U(0, 1)$	
d	(n_1, n_2)	$(\sigma_1', \sigma_2', \sigma_1, \sigma_2)$	Scaled χ^2	$T_{splinecorr}$	Scaled χ^2	$T_{splinecorr}$	Scaled χ^2	$T_{splinecorr}$
0	(50,50)	(0.20, 0.15, 0.04, 0.05)	0.000	0.045	0.010	0.075	0	0.070
	(50,50)	(0.25, 0.20, 0.10, 0.12)	0.005	0.070	0.005	0.070	0	0.065
	(100,100)	(0.20, 0.15, 0.04, 0.05)	0.005	0.035	0.025	0.070	0	0.065
	(100,100)	(0.25, 0.20, 0.10, 0.12)	0.005	0.050	0.010	0.035	0	0.030
1	(50,50)	(0.20, 0.15, 0.04, 0.05)	0.035	0.360	0.005	0.330	0	0.405
	(50,50)	(0.25, 0.20, 0.10, 0.12)	0.005	0.180	0.000	0.165	0	0.185
	(100,100)	(0.20, 0.15, 0.04, 0.05)	0.250	0.955	0.000	0.900	0	0.965
	(100,100)	(0.25, 0.20, 0.10, 0.12)	0.070	0.535	0.000	0.400	0	0.430
2	(50,50)	(0.20, 0.15, 0.04, 0.05)	0.765	1.000	0.020	1.000	0	1.000
	(50,50)	(0.25, 0.20, 0.10, 0.12)	0.270	0.890	0.010	0.885	0	0.935
	(100,100)	(0.20, 0.15, 0.04, 0.05)	1.000	1.000	0.025	1.000	0	1.000
	(100,100)	(0.25, 0.20, 0.10, 0.12)	0.840	1.000	0.005	1.000	0	1.000
3	(50,50)	(0.20, 0.15, 0.04, 0.05)	1.000	1.000	0.460	1.000	0	1.000
	(50,50)	(0.25, 0.20, 0.10, 0.12)	0.915	1.000	0.185	1.000	0	1.000
	(100,100)	(0.20, 0.15, 0.04, 0.05)	1.000	1.000	0.905	1.000	0	1.000
	(100,100)	(0.25, 0.20, 0.10, 0.12)	1.000	1.000	0.565	1.000	0	1.000

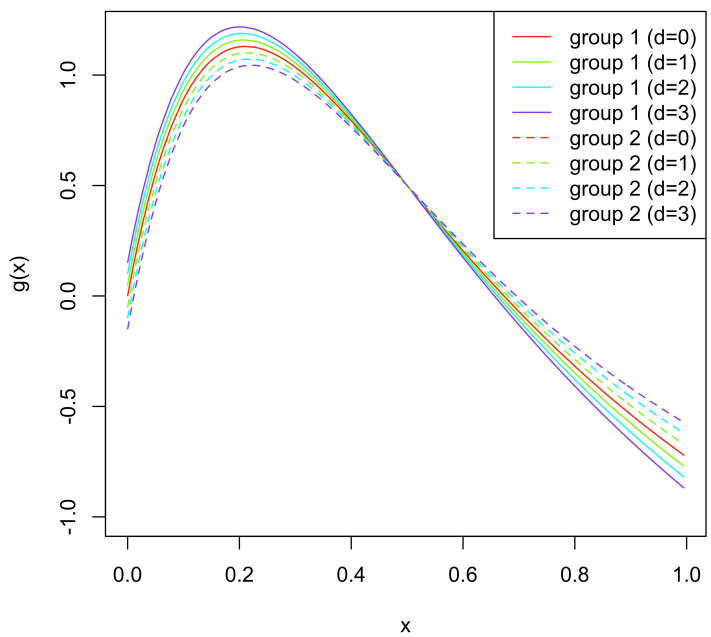

Figure 1: Functions $g_{1d}(x)$ and $g_{2d}(x)$ with $d = 0, 1, 2, 3$ used in the simulation studies

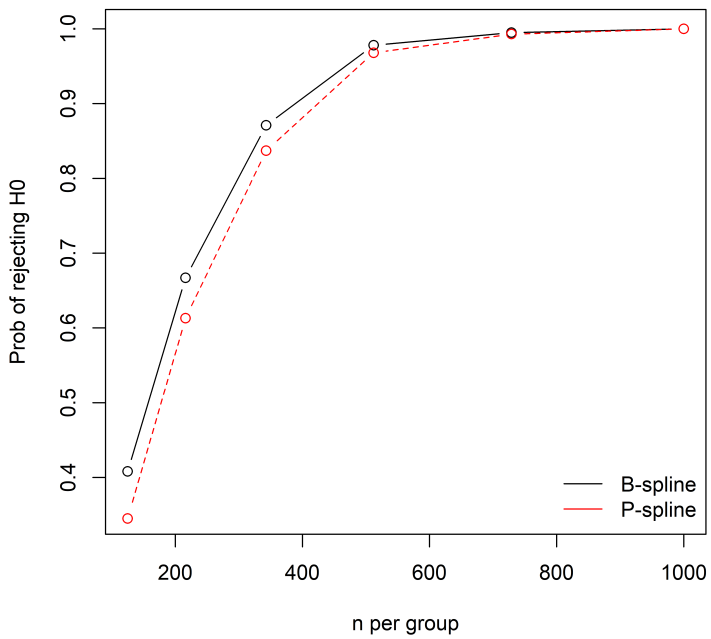

Figure 2: Probability of rejecting H_0 with $d = 1$, $(\sigma_1, \sigma_2) = (0.2, 0.15)$

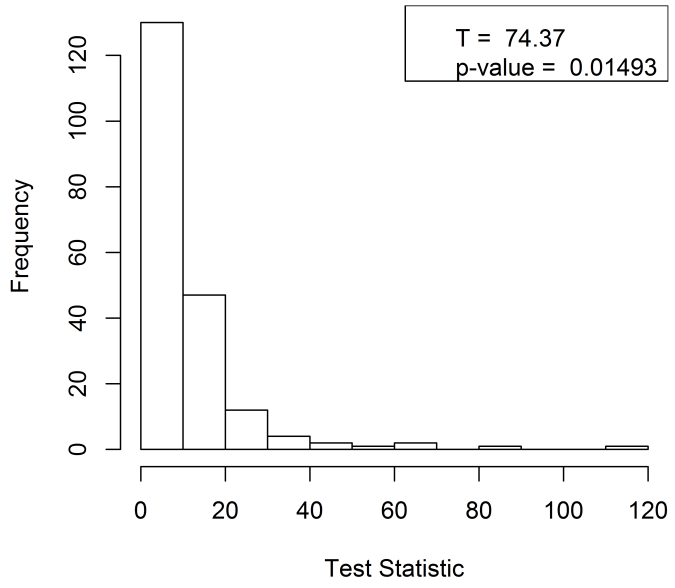

Figure 3: Empirical distribution of the test statistic under the null hypothesis

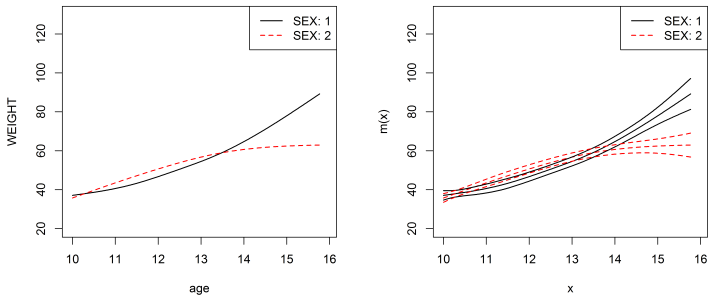

Figure 4: Estimated age effects on weight and associated pointwise 95% CIs in boys and girls

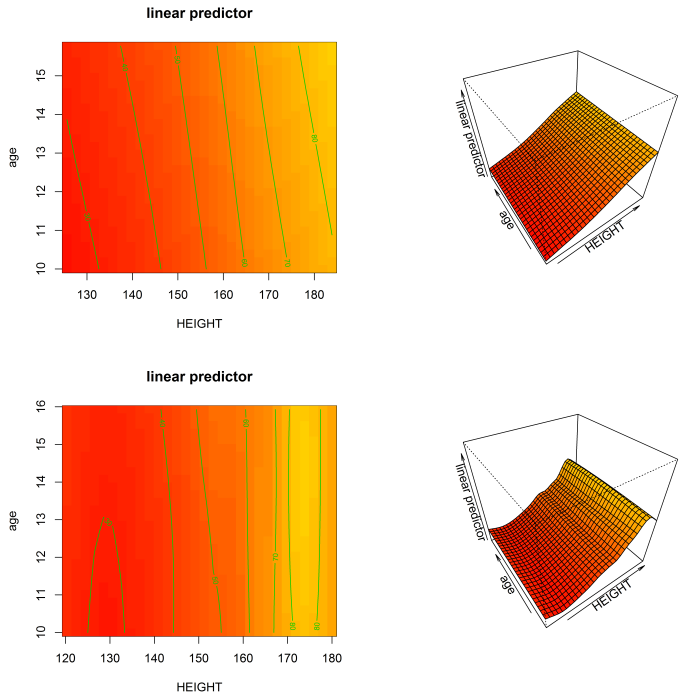

Figure 5: Estimated contour and 3D plots of height and age effects on weight by gender

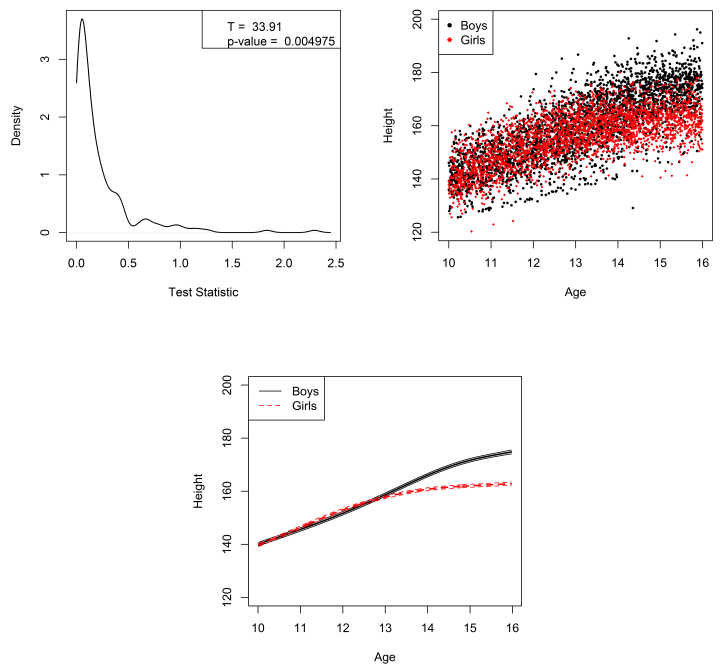

Figure 6: Empirical distribution of the test statistic under the null hypothesis and height over age by gender; height-for-age scatter plot; pointwise 95% CI by gender

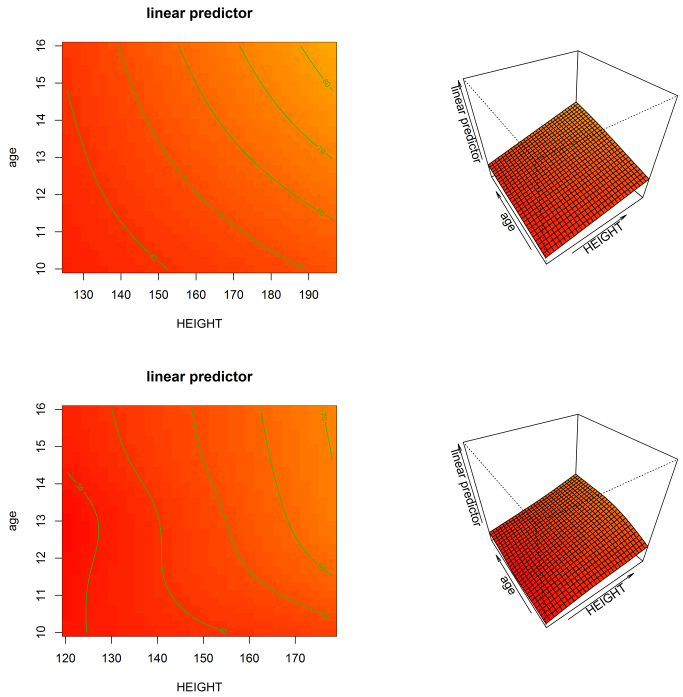

Figure 7: Estimated contour and 3D plots of weight on height and age by sex using correlated data