Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2009 January 1; 71(1): 75–96. doi:10.1111/j.1467-9868.2008.00671.x.

Testing in semiparametric models with interaction, with applications to gene-environment interactions

Arnab Maity,
Texas A&M University, College Station, USA

Raymond J. Carroll,
Texas A&M University, College Station, USA

Enno Mammen, and
University of Mannheim, Germany

Nilanjan Chatterjee
National Cancer Institute, Rockville, USA

Summary

Motivated from the problem of testing for genetic effects on complex traits in the presence of gene-environment interaction, we develop score tests in general semiparametric regression problems that involves Tukey style 1 degree-of-freedom form of interaction between parametrically and non-parametrically modelled covariates. We find that the score test in this type of model, as recently developed by Chatterjee and co-workers in the fully parametric setting, is biased and requires undersmoothing to be valid in the presence of non-parametric components. Moreover, in the presence of repeated outcomes, the asymptotic distribution of the score test depends on the estimation of functions which are defined as solutions of integral equations, making implementation difficult and computationally taxing. We develop profiled score statistics which are unbiased and asymptotically efficient and can be performed by using standard bandwidth selection methods. In addition, to overcome the difficulty of solving functional equations, we give easy interpretations of the target functions, which in turn allow us to develop estimation procedures that can be easily implemented by using standard computational methods. We present simulation studies to evaluate type I error and power of the method proposed compared with a naive test that does not consider interaction. Finally, we illustrate our methodology by analysing data from a case-control study of colorectal adenoma that was designed to investigate the association between colorectal adenoma and the candidate gene NAT2 in relation to smoking history.

Keywords

Additive models; Diplotypes; Function estimation; Non-parametric regression; Omnibus hypothesis testing; Partially linear model; Repeated measures; Score test; Semiparametric models; Smooth backfitting; Tukey's 1 degree-of-freedom model

Address for correspondence: Raymond J. Carroll, Department of Statistics, Room 447, Blocker Building, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA. E-mail: carroll@stat.tamu.edu.

1. Introduction

Modern genetic association studies often focus on discovery of susceptibility loci, i.e. identification of genetic variants that are associated with the trait under study. The risks of multifactorial traits, such as cancer, however, are determined by complex interactions between genetic and environmental exposures and the chance for discovery of the underlying susceptibility genes can be substantially reduced if the possibility of heterogeneity in genetic effects due to interactions is ignored. Thus, in recent years, there has been increasing attention in omnibus testing of genetic main effects and gene-environment or gene-gene interactions for detection of susceptibility genes for complex traits. Clearly, tests of association incorporating interactions require larger degrees of freedom than those which are based only on main effects. When the number of extra degrees of freedom required is relatively small, recent studies have shown that the omnibus tests can be a robust and powerful approach for detecting genetic association irrespectively of whether certain specific forms of interactions are present or not (Chatterjee *et al.*, 2006; Kraft *et al.*, 2007). However, if the required number of degrees of freedom is large, then the omnibus tests can have poor power. Thus parsimonious modelling of gene-gene and gene-environment interactions should be considered for construction of powerful omnibus tests.

Chatterjee *et al.* (2006) proposed the use of a Tukey style 1 degree-of-freedom model for interaction for testing the genetic association of a disease with a set of genetic variants, such as tagging single nucleotide polymorphisms (SNPs) in a candidage gene, that may potentially interact with another set of genetic variants or/and with one or more environmental exposures. SNPs represent a natural genetic variability at high density in the human genome. A genetic locus corresponding to an SNP has two possible alleles (states), namely the normal and the variant. The SNP-genotype data for a subject can have three possible values and are often coded numerically as the number of variant alleles that the subject carries on the pair of homologous chromosomes that are inherited from his or her parents.

In this paper, we shall consider extending the work of Chatterjee *et al.* (2006) focusing on the problem of gene-environment interaction. Thus, for example, if D denotes the binary indicator of a disease outcome, X denotes a 'design matrix' that is associated with a set of genetic variants G, Z denotes the desig matrix that is associated with an environmental exposure of interest and S denotes a set of additional cofactors, such as age and sex, then the risk of the disease can be modelled by using Tukey's form of gene-environment interaction as

$$\mathrm{pr}\,(D=1|X,S,Z,\gamma)=H\left(X^{\mathrm{T}}\beta_0+S^{\mathrm{T}}\eta_0+Z^{\mathrm{T}}\theta_0+\gamma X^{\mathrm{T}}\beta_0 Z^{\mathrm{T}}\theta_0\right), \tag{1}$$

where $H(\cdot)$ is the logistic distribution function. Unlike in the standard logistic regression model where potentially a separate interaction parameter is allowed between each pair of design elements of the genetic and environmental factors, in model (1), a single parameter (γ) is used to capture interactions. Moreover, in model (1), the omnibus null hypothesis of interest can be simply stated as $\beta_0=0$ under which both genetic main effects and gene-environment interactions disappear from the model. A complication, however, is that, under $\beta_0=0$, the parameter γ also disappears from the model and hence is not identifiable from the data. Nevertheless, Chatterjee *et al.* (2006) noticed that, for each fixed value of γ, model (1) can be used to construct a valid score test for $\beta_0=0$. They proposed to use maxima of such score statistics over a range of the parameter γ as the final test statistics for testing $\beta_0=0$. They observed that the score test has particular computational advantages, because under the null hypothesis model (1) reduces to a standard logistic regression model involving only main effects of Z and S.

In this paper, we extend the work by Chatterjee *et al.* (2006) in two novel ways. First, we consider modelling complex effects of continuous environmental exposures by using non-parametric regression models. The problem is particularly motivated by the fact that modern molecular epidemiologic studies often involve measurement of environmental exposures through continuous biomarkers, the relationships of which with the disease can be highly complex and non-linear. Thus for exmaple in the logistic context, we might consider the model

$$\mathrm{pr}\,(D=1|X,S,Z,\gamma)=H\left\{X^{\mathrm{T}}\beta_0+S^{\mathrm{T}}\eta_0+\theta_0\,(Z)+\gamma X^{\mathrm{T}}\beta_0\,(Z)+\gamma X^{\mathrm{T}}\beta_0\theta_0\,(Z)\right\}, \tag{2}$$

where $\theta_0(\cdot)$ is an unknown function. Second, we consider general semiparametric models with possible repeated measures (Lin and Carroll, 2006), where the effects are given through terms roughly of the form on the right-hand side of model (2). In particular, we assume that, for each subject or cluster *i*, there are $j=1,\ldots,J$ observations (Y_{ij}, X_{ij}, S_{ij}, Z_{ij}). We write $\tilde{Y}_i=(Y_{i1},\ldots,Y_{iJ})$ and work with a criterion function

$$\mathcal{L}\left\{\tilde{Y},\nu_1,\ldots,\nu_J,\zeta_0\right\},\qquad \text{with}\quad \nu_j=X_j^{\mathrm{T}}\beta_0\left\{1+\gamma\theta_0\left(Z_j\right)\right\}+S_j^{\mathrm{T}}\eta_0, \tag{3}$$

where a *criterion function* could mean either an actual likelihood function, a composite likelihood function, i.e. one that is a likelihood function for a reduced set of data, or a working independence likelihood function. In particular, criterion functions have scores in the parameters (β_0, η_0, ζ_0, θ_0) that have mean 0 given appropriate subcomponents of $\left(X_j,S_jZ_j\right)_{j=1}^{J}$. The case of no repeated measures as in model (1) occurs when $J=1$.

Our interest is in testing for the hypothesis of the form $H_0:\beta_0=0$. As in Chatterjee *et al.* (2006), it is natural to use a score testing approach to this problem to avoid numerical difficulty that is associated with parameter estimation under general models of the form (1) and (2). In particular, we note that estimation of γ in these models can be numerically unstable because of lack of identifiability of this parameter under $\beta_0=0$. This also means that γ cannot be consistently estimated at contiguous alternatives. In practice, even in fully parametric models, this lack of identifiability means that estimating γ is numerically instable, leading to non-convergence if its range is not restricted.

Following Chatterjee *et al.* (2006), we propose to perform score-type tests for eadh value of γ and then to maximize these tests over an interval of γ-values, and to use numerical devices to create levels of significance. It is possible to create the score statistic directly, and to apply the asymptotic expansions that were developed by Lin and Carroll (2006) to analyse these statistics. However, two problems arise.

- **a.** The first problem is that the direct score statistic requires undersmoothing for the non-parametric estimation of $\theta_0(\cdot)$ in expression (3). By modifying the directly calculated score statistic in a suitable manner using a profile argument, we shall show how to create test statistics that lose no local power yet allow regular smoothing, such as cross-validation.
- **b.** The second problem to overcome is that, in the repeated measures case that $J>1$, the distribution of the profile score statistic depends on random variables that are formed as solutions to integral equations. Rather than go about this problem directly and solve the integral equation, which would be extremely difficult, we show that the crucial terms can be estimated by using nothing more than the Gaussian repeated measures

algorithm of Wang (2003); see also Lin *et al.* (2004) for a non-iterative solution and Huggins (2006) for another simple computational device.

Thus, we shall develop a test statistic that is straightforward to compute and does not require undersmoothing, and the method also allows a simple implementation when the score test is maximized over a range for γ.

Our methodology is easiest to understand in the non-repeated measures case that $J=1$, and we take this up in Section 2. The repeated measures case is described in Section 3. Section 4 gives the results of a simulation study. Here we find that our maximized tests lose little power when there is no interaction and can gain great power advantages over a main effects test when there are interactions. Section 5 illustrates an application of the proposed method for omnibus testing of the effects genetic variants in the NAT2 gene and their interactions with the number of years since stopping smoking on the risk of colorectal adenoma by using a case-control study that was conducted with the prostate, lung, colorectal and ovaraian cancer screening trial (Hayes *et al.*, 2000).

We close this section with a few remarks about identifiability. The models that we study are examples of a problem where γ is a nuisance parameter and, under the null hypothesis (5) that $\beta_0=0$, the nuisance parameter is unidentified. Model (1) is of course reminiscent of Tukey's 1 degree-of-freedom test for interaction (Tukey, 1949). However, unlike in that context, in our problem the parameter γ is a nuisance parameter and is not of primary interest. The method of Chatterjee *et al.* (2006) is more closely akin to the basic suggestion in Davies (1987), namely to fix the nuisance parameter, to compute an appropriate test statistic and then to maximize that test statistic over a range of values for the nuisance parameter. Thus, one way to think about our testing procedure is as the appropriate, efficient (both computationally and in terms of power) way of implementing the basic approach of Davies (1987) in our context, while taking care to eliminate the concerns of undersmoothing and solution of integral equations that arise from a less targeted approach.

2. Testing without repeated measures

2.1. Data and notation

The data consist of a response Y, parametrically modelled covariates S and X, the latter possibly interacting with a non-parametrically modelled covariate Z. We consider a general log-likelihood or criterion function

$$\mathcal{L}\left[Y, S^{\mathrm{T}}\eta_0+\theta_0(Z)+X^{\mathrm{T}}\beta_0\{1+\gamma\theta_0(Z)\}, \zeta_0\right], \tag{4}$$

where β_0 and η_0 are the main effects, $\theta_0(\cdot)$ is an unknown function, γ is the interaction effect and ζ_0 are nuisance parameters. In this section, we are interested in testing the parametric hypothesis

$$H_0:\beta_0=0. \tag{5}$$

As described in Section 1, Chatterjee *et al.* (2006) addressed a similar problem for a fully parameteric model where Z is also modelled parametrically. They used a score-based testing procedure to test H_0. We generalize their idea for the general semiparametric model that is given in expression (4). We describe below the major steps to derive the test statistic for testing hypothesis (5).

In what follows, we use a simple subscripting convention for derivatives of the log-likelihood. Thus, with (•), we set

$$\begin{aligned}
\mathcal{L}_\theta(\bullet) &= (\partial/\partial\nu)\,\mathcal{L}\left\{Y, S^{\mathrm{T}}\eta+\nu+X^{\mathrm{T}}\beta(1+\gamma\nu), \zeta\right\}|_{\nu=\theta(Z)},\\
\mathcal{L}_{\theta\theta}(\bullet) &= \left(\partial^2/\partial\nu^2\right)\lambda\left\{Y, S^{\mathrm{T}}\zeta+\nu+X^{\mathrm{T}}\beta(1+\gamma\nu), \xi\right\}|_{\nu=\theta(Z)},\\
\mathcal{L}_\zeta(\bullet) &= (\partial/\partial\zeta)\,\mathcal{L}\left\{Y, S^{\mathrm{T}}\eta+\nu+X^{\mathrm{T}}\beta(1+\gamma\nu), \zeta\right\}|_{\nu=\theta(Z)},\\
\mathcal{L}_{\theta\zeta}(\bullet) &= (\partial/\partial\zeta)\,\mathcal{L}_\theta\left\{Y, S^{\mathrm{T}}\eta+\nu+X^{\mathrm{T}}\beta(1+\gamma\nu), \zeta\right\}|_{\nu=\theta(Z)},
\end{aligned}$$

etc. Thus, in an abuse of notation we do not indicate in the notation that these partial derivatives do not depend on the parameters and covariates only via $S^{\mathrm{T}}\eta(Z)+X^{\mathrm{T}}\beta\{1+\gamma\,\theta(Z)\}$.

2.2. Estimation of parameters under the null hypothesis

Here we show how to estimate the parameters and the function at the null hypothesis.

The strength of score tests is that we fit the model under the null hypothesis. Under the null hypothesis, the log-likelihood or criterion function for the model is written as $\mathcal{L}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z), \zeta_0\right\}$, a standard form that is easy to handle. The log-likelihood under the alternative is much more difficult to deal with numerically because of the interaction.

By definition of a log-likelihood or criterion function, at the null hypothesis,

$$0=E\left[\mathcal{L}_\theta\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z), \zeta_0\right\}|X, S, Z\right]. \tag{6}$$

The first step of the process is to estimate the function $\theta_0(\cdot)$ for any fixed value of $\delta=\delta^*=(\eta^*\zeta^*)$. We shall use kernel methods because of their convenient theory, bu this step can be modified in practice by using any smoother. The resulting estimate is denoted as $\theta\hat{}(\cdot\,,\delta^*)$. Let $K(\cdot)$ be a smooth symmetric density function with bounded support, let h be a bandwidth and let $K_h(z)=h^{-1}K(z/h)$. Define $\phi_k=\int Z^k K(z)\,dz$ and $G_h(z)=(1, z/h)^{\mathrm{T}}$. We follow Lin and Carroll (2006) to estimate the parameters under hypothesis H_0: for any fixed value of $\delta=\delta^*$, estimate $\theta_0(z)$ by solving the local likelihood equation

$$0=n^{-1}\sum_{i=1}^{n} K_h(Z_i-z)\,G_h(Z_i-z)\,\mathcal{L}_\theta\left\{Y_i, S_i^{\mathrm{T}}\eta^*+\alpha_0+\alpha_1(Z_i-z), \zeta^*\right\},$$

for $\alpha\hat{}_0$ and set $\theta\hat{}(z, \delta^*)$.

The second step in the process is now smoothing method independent. To estimate $\delta_0=(\eta_0,\zeta_0)$ maximize in δ the function

$$n^{-1}\sum_{i=1}^{n}\mathcal{L}\left\{Y_i, S_i^{\mathrm{T}}\eta+\widehat{\theta}(Z_i,\delta), \zeta\right\},$$

the so-called profile method, which solves

$$0=n^{-1}\sum_{i=1}^{n}\left\{S_i+\widehat{\theta}_\eta(Z_i,\delta)\right\}\mathcal{L}_\theta\left\{Y_i,S_i^{\mathrm{T}}\eta+\widehat{\theta}(Z_i,\delta),\zeta\right\},$$
$$0=n^{-1}\sum_{i=1}^{n}\left[\mathcal{L}_\zeta\left\{Y_i,S_i^{\mathrm{T}}\eta+\widehat{\theta}(Z_i,\delta),\zeta\right\}+\widehat{\theta}_\zeta(Z_i,\delta)\,\mathcal{L}_\theta\left\{Y_i,S_i^{\mathrm{T}}\eta+\widehat{\theta}(Z_i,\delta),\zeta\right\}\right],$$

where $\theta\hat{}\eta(Z_i,\delta)$ and $\theta\hat{}\zeta(Z_i,\delta)$ are the derivatives of $\theta\hat{}(Z_i,\delta)$ with respect to η or ζ respectively. Call the resulting estimate $\delta\hat{}$.

2.3. The score function and asymptotic theory

2.3.1. Derivation—One approach to developing a score statistic is to fix the function $\theta(\cdot)$, to derive the score statistic and then to plug in estimates of nuisance parameters and the function $\theta(\cdot)$. This does not work well because the function estimate itself needs profiling, and indeed this approach requires undersmoothing for its validity.

In contrast, our test statistic is a particular implementation of the profiled log-likelihood or criterion function, which is derived as follows. In general, the log-likelihood function for an observation is $\mathcal{L}\left\{Y,S^{\mathrm{T}}\eta+X^{\mathrm{T}}\beta+\theta(Z)+\gamma X^{\mathrm{T}}\beta\theta(Z),\zeta\right\}$. Recall that $\delta(\eta,\zeta)$. For given (β,δ), let $\theta(Z,\beta,\delta)$ be the profile function that solves

$$E\left[\mathcal{L}_\theta\left\{Y,S^{\mathrm{T}}\eta+X^{\mathrm{T}}\beta+\theta(Z,\beta,\delta)+\gamma X^{\mathrm{T}}\beta\theta(Z,\beta,\delta),\delta\right\}|Z\right]=0. \tag{7}$$

Define $\tilde{X}_{\mathrm{pro}}=X\{1+\gamma\,\theta(Z,0,\delta)\}+\theta_\beta(Z,0,\beta)$, where $\theta_\beta(Z,\beta,\beta)=(\partial/\partial\,\beta)\theta(Z,\beta,\delta)$. The profiled log-likelihood is $\mathcal{L}\left\{Y,S^{\mathrm{T}}\eta+X^{\mathrm{T}}\beta+\theta(Z,\beta,\delta)+\gamma X^{\mathrm{T}}\beta\theta(Z,\beta,\delta),\zeta\right\}$. Differentiating it with respect to β and evaluating at the null hypothesis $\beta=0$, the profiled (efficient) score is easily seen to be $\tilde{X}_{\mathrm{pro}}\mathcal{L}_\theta\left\{Y,S^{\mathrm{T}}\eta+\theta(Z,0,\delta),\zeta\right\}$.

In addition, differentiating equation (7) with respect to $\beta=0$ and evaluating it at $\beta=0$ and $\delta=\delta_0$ shows that $\tilde{X}_{\mathrm{pro}}=\{1+\gamma\theta_0(Z)\}\,\tilde{X}$, where

$$\tilde{X}=X-E\left[X\mathcal{L}_{\theta\theta}\left\{Y,S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}|Z\right]/E\left[\mathcal{L}_{\theta\theta}\left\{Y,S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}|Z\right].$$

We thus propose the following profiled score statistic for β_0:

$$\mathcal{T}_{n,\mathrm{pro}}(\gamma)=n^{-1/2}\sum_{i=1}^{n}\left\{1+\gamma\widehat{\theta}\left(Z_i,\widehat{\delta}\right)\right\}\tilde{X}_{i,\mathrm{est}}\mathcal{L}_\theta\left\{Y_i,S_i^{\mathrm{T}}\widehat{\eta}+\widehat{\theta}\left(Z_i,\widehat{\delta}\right),\widehat{\zeta}\right\}, \tag{8}$$

where $\tilde{X}_{i,\,\mathrm{est}}$ is an estimated version of $\tilde{X}_i$, with the terms to be estimated in $\tilde{X}$ obtained by separate non-parametric regressions in the numerator and denominator. The normalization by $n^{-1/2}$ is conveient for the asymptotic theory.

2.3.2. Theoretical results—Let $\delta_0=\left(\eta_0^{\mathrm{T}},\zeta_0^{\mathrm{T}}\right)^{\mathrm{T}}$ and make the definitions

$$\theta_\delta(z_0,\delta_0) = -E\left[\mathcal{L}_{\theta\delta}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}|Z=z_0\right]/E\left[\mathcal{L}_{\theta\theta}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}|Z=z_0\right],$$
$$\epsilon=\mathcal{L}_\delta\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}+\theta_\delta(Z,\delta_0)\,\mathcal{L}_\theta\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\},$$
$$\mathcal{M}=-E\left(\epsilon\epsilon^{\mathrm{T}}\right),$$
$$\mathcal{N}=E\left(X\{1+\gamma\theta_0(Z)\}\left[\mathcal{L}_{\theta\delta}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}+\mathcal{L}_{\theta\theta}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}\theta_\delta(Z,\delta_0)\right]^{\mathrm{T}}\right),$$
$$\Psi(\gamma) = \{1+\gamma\theta_0(Z)\}\,\tilde{X}\,\mathcal{L}_\theta\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z),\zeta_0\right\}-\mathcal{N}\mathcal{M}^{-1}\epsilon.$$

The main result of this section justifying our methodology is stated below. Technically, a precise argument requires little more than that the linear expansions for the parametric and non-parametric parts that are given in Lin and Carroll (2006) hold to order $o_p(n^{-1/2})$, the latter uniformly.

Result 1: Suppose that we are testing for H_0:β_0=0. Assume that $h\ \alpha\ n^{-\alpha}$ with $1/3 \le \alpha \le 1/5$. Then, for any fixed γ, the score function for β_0 can be written as

$$\mathcal{T}_{n,\mathrm{pro}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\Psi_i(\gamma)+o_p(1).$$

In addition, assume that, for any γ_1 and γ_2, $\nu(\gamma_1,\gamma_2)=E\left\{\Psi(\gamma_1)\,\Psi^{\mathrm{T}}(\gamma_2)\right\}$ is finite. Then, under the hypothesis that β=0 $\mathcal{T}_{n,\mathrm{pro}}(\gamma)$ as a function of $\gamma \in [L, R]$ converges weakly to a Gaussian process $\mathcal{W}(\gamma)$ with mean 0 and covariance function $\nu(\gamma_1,\gamma_2)$.

Remark 1: There are two methods that can be used to estimate the covariance matrix of the estimated score.

a. First, suppose as in logistic regression that there are no nuisance parameters ζ_0, and that $\mathcal{L}(\cdot)$ is a log-likelihood function and not a general criterion function. Then we can write $\Psi_i(\gamma)=\Psi_i^*(\gamma)\,\mathcal{L}_\theta\left\{Y_i, S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\right\}$ with
$\Psi_i^*(\gamma) = \{1+\gamma\theta_0(Z_i)\}\,\tilde{X}_i - \mathcal{N}\mathcal{M}^{-1}\tilde{S}_i$, where

$$\tilde{S} = S - E\left[S\,\mathcal{L}_{\theta\theta}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z)\right\}|Z\right]/E\left[\mathcal{L}_{\theta\theta}\left\{Y, S^{\mathrm{T}}\eta_0+\theta_0(Z)\right\}|Z\right].$$

Let $\widehat{\Psi}_i^*(\gamma)$ be the estimated version of $\Psi_i^*(\gamma)$ This estimated version requires the definition of $\tilde{X}_i$, $\tilde{S}_i$ and additional non-parametric regressions, which are easily accomplished via kernel or spline methods. Further, let $\mathcal{I}_{\theta,\mathrm{null}}\left\{S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\right\}$ be the conditional information matrix for θ under the null model. Then we estimate the covariance matrix of $\mathcal{T}_n(\gamma)$

$$\mathcal{I}_{\beta_{0,n}}(\gamma) = n^{-1}\sum_{i=1}^{n}\mathcal{I}_{\theta,\mathrm{null}}\left\{S_i^{\mathrm{T}}\widehat{\eta}+\widehat{\theta}(Z_i,\widehat{\eta})\right\}\widehat{\Psi}_i^*(\gamma)\,\widehat{\Psi}_i^*(\gamma)^{\mathrm{T}}.$$

b. In general, $\mathcal{I}_{\beta_0,n}(\gamma)$ can be estimated as the sample covariance matrix of the terms $\hat{\Psi}_i(\gamma)$, the estimated version of $\Psi_i(\gamma)$. In likelihood problems, simplifications arise becuase we can compute the covariance matrix of $\Psi(\cdot)$ given (X, Z, S) by using Fisher information calculations.

Remark 2: The validity and unbiasedness of the profiled score statistic primarily depend on the use of $\tilde{x}$. In simpler models, such as the Gaussian model, $\tilde{x}=X - E(X|Z)$ is simply the residual of a non-parametric Gaussian regression of each component of X on Z. In general, $\tilde{x}$ can be thought of as the residual of a weighted non-parametric Gaussian regression of each component of X on Z, where the error variance for weighting is taken to be $-1/\mathcal{L}_{\theta\theta}(\cdot)$. This interpretation enables us to construct estimates of $\tilde{x}$ with considerable ease in many cases, especially in the presence of repeated measurements; see Section 3 for details.

2.4. The test statistic and its implementation

Here we define our test statistic and show how to implement it in practice to compute critical values.

The score test statistic, for a fixed value of γ, is then given by $\mathcal{T}_{n,\mathrm{pro}}(\gamma)^{\mathrm{T}}\mathcal{I}^{-1}_{\beta_{0,n}}(\gamma)\mathcal{T}_{n,\mathrm{pro}}(\gamma)$. We compute the final test statistic as

$$\mathcal{T}^{*}_{n}=\max_{L\le\gamma\le R}\left\{\mathcal{T}^{\mathrm{T}}_{n,\mathrm{pro}}(\gamma)\,\mathcal{I}^{-1}_{\beta_{0,n}}(\gamma)\,\mathcal{T}_{n,\mathrm{pro}}(\gamma)\right\},$$

Where L and R are prespecified lower and upper bounds of γ. Our approach is also related to adaptive tests that have been developed for non-parametric alternatives of functions with unknown smoothness; see for example Horowitz and Spokoiny (2001).

To implement the test, we need to simulate the null distribution of $\mathcal{T}^{*}_{n}$ and to obtain the desired critical values. Oure method avoids the need to determine critical values for the maximum of a function of a Gaussian process. Using result 1 we can generate realizations from the limiting distribution of the score statistic as

$$T_0(\gamma)=n^{-1/2}\sum_{i=1}^{n}\widehat{\Psi}_i(\gamma)\,\mathcal{Z}_i,$$

where $\hat{\Psi}(\gamma)$ is $\Psi(\gamma)$ evaluated at $\delta\hat{}$ and $\theta\hat{}(z, \delta\hat{})$ and $\mathcal{Z}_1, \ldots, \mathcal{Z}_n$, are standard normal random variates which are drawn independently of the data. The null distribution of $\mathcal{T}^{*}_{n}$ is then simulated by generating $\mathcal{T}^{*}_{0}=\max_{L\le\gamma\le R}\left\{T_0(\gamma)^{\mathrm{T}}\mathcal{I}^{-1}_{\beta_{0,n}}(\gamma)\,T_0(\gamma)\right\}$ repeatedly. This methods is the semiparametric version of a method that was discussed by Lin and Zou (2004) and Chatterjee *et al.* (2006).

3. General interaction model with repeated measures

3.1. Data and notation

In this section we generalize the ideas that were presented earlier to the case when repeated measures are present in the data. Repeated measures models can arise from various fields of research, e.g. matched case-control studies, finance and epidemiology. The key feature of these models is that the non-parametric function is evaluated for each of the repeated measurements. Lin and Carroll (2006) developed kernel-based estimation procedures and investigated asymptotic properties of the estimators in general semiparametric regression problems. We shall use their results and methodology in our context.

In this section we set out the notation to be used.

For simplicity only, we suppose that there are J repeated measurements for each individual. Only obvious notational changes are requires for the more general case. Specifically, we consider a log-likelihood or criterion function

$$\mathcal{L}\left\{\tilde{Y}, \nu_1(\beta_0,\theta_0,\eta_0),\ldots,\nu_J(\beta_0,\theta_0,\eta_0),\zeta_0\right\},$$

Where $\nu_j(\beta_0,\theta_0,\eta_0)=X_j^{\mathrm{T}}\beta_0\left\{1+\gamma\theta_0\left(Z_j\right)\right\}+\theta_0\left(Z_j\right)+S_j^{\mathrm{T}}\eta_0\gamma$ is the common interaction parameter for each of the repeated measurements and $\zeta 0$ is the collection of all the nuisance parameters. Then, with a slight abuse of notation in the first formula below,

$$E\left[\partial\mathcal{L}\left\{\tilde{Y}, \nu_1(\beta_0,\theta_0,\eta_0),\ldots,\nu_J(\beta_0,\theta_0,\eta_0),\zeta_0\right\}/\partial\theta_0(Z_k)\,|\left(X_j,Z_j,S_j\right)_{j=1}^{J}\right]=0,$$
$$E\left[\partial\mathcal{L}\left\{\tilde{Y}, \nu_1(\beta_0,\theta_0,\eta_0),\ldots,\nu_J(\beta_0,\theta_0,\eta_0),\zeta_0\right\}/\partial(\beta,\eta,\zeta)\,|\left(X_j,Z_j,S_j\right)_{j=1}^{J}\right]=0;$$

see Lin and Carroll (2006) for more discussion. In Section 3.6, we describe methods for the partially linear model when working independence among the errors is used, and hence weaker conditioning assumptions are required.

Letting $\bullet=\{\tilde{Y}, \nu_1(\beta,\theta,\eta),\ldots,\nu_J(\beta,\theta,\nu),\ldots,\nu_J(\beta,\theta,\zeta\}$, we define terms $\mathcal{L}_{j\theta}(\bullet)$, $\mathcal{L}_{jk\theta}(\bullet)$, $\mathcal{L}_{\zeta}(\bullet)$ and $\mathcal{L}_{j\theta\zeta}(\bullet)$ in the same way as described in Section 2.1. Thus, for example,

$$\begin{aligned}\mathcal{L}_{j\theta}(\bullet)=\ & \frac{\partial}{\partial\nu_j}\mathcal{L}\Big[\tilde{Y}, S_1^{\mathrm{T}}\eta+\theta(Z_1)+X_1^{\mathrm{T}}\beta\{1+\gamma\theta(Z_1)\},\ldots,S_j^{\mathrm{T}}\eta+\nu_j+X_j^{\mathrm{T}}\beta\left(1+\gamma\nu_j\right),\ldots,\\ & S_J^{\mathrm{T}}\eta+\theta(Z_J)+X_J^{\mathrm{T}}\beta\{1+\gamma\theta(Z_J)\},\zeta\Big]_{\nu_j=\theta(Z_j)},\\ \mathcal{L}_{jk\theta}(\bullet)=\ & \frac{\partial^2}{\partial\nu_j\partial\nu_k}\mathcal{L}\Big[\tilde{Y}, S_1^{\mathrm{T}}\eta+\theta(Z_1)+X_1^{\mathrm{T}}\beta\{1+\gamma\theta(Z_1)\},\ldots,S_j^{\mathrm{T}}\eta+\nu_j+X_j^{\mathrm{T}}\beta\left(1+\gamma\nu_j\right),\ldots,\\ & S_k^{\mathrm{T}}\eta+\nu_k+X_k^{\mathrm{T}}\beta(1+\gamma\nu_k),\ldots,S_J^{\mathrm{T}}\eta+\theta(Z_J)+X_J^{\mathrm{T}}\beta\{1+\gamma\theta(Z_J)\},\zeta\Big]_{\nu_j=\theta(Z_k)}\ldots.\end{aligned}$$

3.2. Estimation under the null model

In this section, we display the method for estimation of parameters and the function $\theta(\cdot)$, at the null hypothesis.

Under the null hypothesis, the criterion function is given by

$$\mathcal{L}\left\{\tilde{Y}, \theta_0(Z_1)+S_1^{\mathrm{T}}\eta_0,\ldots,\theta_0(Z_J)+S_J^{\mathrm{T}}\eta_0,\zeta_0\right\}.$$

Let $\delta=(\eta,\zeta)$. We estimate $\theta_0\cdot$ and δ_0 under the null model by using methodology that was proposed in Lin and Carroll (2006): for any fixed $\delta=\delta^*=(\eta^*,\zeta^*)$, estimate $\theta_0(z)$ by solving for $(\alpha 0,\alpha 1)$

$$\begin{aligned}0=\ & \sum_{i=1}^{n}\sum_{j=1}^{J}K_h\left(Z_{ij}-z\right)G\left(Z_{ij}-z\right)\mathcal{L}_{j\theta}\Big\{\tilde{Y}_i,\widehat{\theta}(Z_{i1},\delta^*)+S_{i1}^{\mathrm{T}}\eta^*,\ldots,\alpha_0+\alpha_1\left(Z_{ij}-z\right)/h\\ & +S_{ij}^{\mathrm{T}}\eta^*,\ldots,\widehat{\theta}(Z_{iJ},\delta^*)+S_{iJ}^{\mathrm{T}}\eta^*,\zeta^*\Big\},\end{aligned}$$

and setting $\theta\hat{}(z,\delta^*)$ Next, estimate δ by maximizing

$$\sum_{i=1}^{n} \mathcal{L}\left\{\tilde{Y}_i, \widehat{\theta}(Z_{i1}, \delta) + S_{i1}^{\mathrm{T}}, \ldots, \widehat{\theta}(Z_{ij}, \delta) + S_{ij}^{\mathrm{T}}\eta, \zeta\right\}$$

with respect to δ. This can be accomplished by implementing a profiling algorithm as in Lin and Carroll (2006)

3.3. The score function and asymptotic theory

3.3.1. Derivation of the profile score—As we have seen in Section 2.3, our test statistic will be based on the score function of a profiled log-likelihood. In this section, we derive the profiled log-likelihood and the score function, but here the repeated measures aspect makes the calculations less transparent and indeed leads to real issues of implementation. Let $f_j(z)$ be the marginal density of Z_j. Again, for any (β, δ), we define $\theta(z, \beta, \delta)$ by the repeated measures version of equation (7), namely the solution to the equation

$$\begin{aligned} 0= \sum_{j=1}^{J} f_j(z)\, E\Big[\mathcal{L}_{j\theta}\Big\{\tilde{Y}, X_1^{\mathrm{T}}\beta\{1+\gamma\theta(Z_1,\beta,\delta)\} + \theta(Z_1,\beta,\delta) + S_1^{\mathrm{T}}\eta, \ldots, \\ X_J^{\mathrm{T}}\beta\{1+\gamma\theta(Z_J,\beta,\delta)\} + \theta(Z_J,\beta,\delta) + S_J^{\mathrm{T}}\eta, \zeta\Big\} | Z_j = z\Big]. \end{aligned} \tag{9}$$

Defining $\omega_j(\beta,\theta,\delta) = X_j^{\mathrm{T}}\beta\left\{1+\gamma\theta\left(Z_j,\beta,\delta\right)\right\} + \theta\left(Z_j,\beta,\delta\right) + S_j^{\mathrm{T}}\eta$, the profiled log-likelihood function is $\mathcal{L}\left\{\tilde{Y}, \omega_1(\beta,\theta,\delta), \ldots, \omega_J(\beta,\theta,\delta), \zeta\right\}$. Let $\mathcal{L}_{j\theta\beta}\left\{\tilde{Y}, \omega_1(\beta,\theta,\delta), \ldots, \omega_J(\beta,\theta,\delta), \zeta\right\}$ and $\mathcal{L}_{jk\theta}\left\{\tilde{Y}, \omega_1(\beta,\theta,\delta), \ldots, \omega_J(\beta,\theta,\delta), \zeta\right\}$ be the derivatives of $\mathcal{L}_{j\theta}\left\{\tilde{Y}, \omega_1(\beta,\theta,\delta), \ldots, \omega_J(\beta,\theta,\delta), \zeta\right\}$ with respect to β and $\theta(Z_k, \beta, \delta)$ respectively. Differentiating and setting β=0, the profiled score becomes

$$\sum_{j=1}^{J}\left[\left\{1+\gamma\theta\left(Z_j,0,\delta\right)\right\}X_j + \theta_\beta\left(Z_j,0,\delta,\gamma\right)\right]\mathcal{L}_{j\theta}\left\{\tilde{Y}, \omega_1(0,\theta,\delta), \ldots, \omega_J(0,\theta,\delta), \zeta\right\},$$

where, by differentiating equation (9) with respect to β and solving $\theta_\beta(z, \beta, \delta, \gamma)$ is the solution of the functional integral equation

$$\begin{aligned} 0 = \sum_{j=1}^{J} f_j(z)\, E\Big[\mathcal{L}_{j\theta\beta}\left\{\tilde{Y}, \omega_1(\beta,\theta,\delta), \ldots, \omega_J(\beta,\theta,\delta), \zeta\right\} + \sum_{k=1}^{J}\mathcal{L}_{jk\theta}\left\{\tilde{Y}, \omega_1(\beta,\theta,\delta), \ldots, \omega_J(\beta,\theta,\delta), \zeta_0\right\} \\ \times\theta_\beta(Z_k,\beta,\delta,\gamma) | Z_j = z\Big]. \end{aligned} \tag{10}$$

Then, for any fixed value of γ, the profiled score function for β_0 evaluated at β_0=0, β_0=β^ and $\theta(z)$=θ^(z,δ^) is given by

$$\begin{aligned} \mathcal{T}_{n\mathrm{pro}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\left[\left\{1+\gamma\widehat{\theta}\left(Z_{ij},\widehat{\delta}\right)\right\}X_{ij} + \widehat{\theta}_\beta\left(Z_{ij},0,\widehat{\delta},\gamma\right)\right]\mathcal{L}_{j\theta}\left\{\tilde{Y}, \tilde{\theta}\left(Z_{i1},\widehat{\delta}\right) \right. \\ \left. + S_{i1}^{T}\widehat{\eta}, \ldots, \widehat{\theta}\left(Z_{ij},\widehat{\delta}\right) + S_{ij}^{\mathrm{T}}\widehat{\eta}, \widehat{\zeta}\right\} \end{aligned}$$

3.3.2. Asymptotic theory—Denote $(\bullet)=\{\tilde{Y}, \omega_1(\beta_0, \theta_0, \delta_0),\ldots, \omega_J(\beta_0, \theta_0, \delta_0)$ and denote $(\bullet_i)$ to be $(\bullet)$ evaluated at the ith observation. Do all calculations at the null model $\beta_0=0$. Define $\theta_\delta(z,\delta_0)$ such that

$$0=\sum_{j=1}^{J} f_j(z)\, E\left\{\mathcal{L}_{j\theta\delta}(\bullet)+\sum_{k=1}^{J}\theta_\delta(Z_k,\delta_0)\,\mathcal{L}_{jk\theta}(\bullet)\,|Z_j=z\right\}.$$

Further define

$$\mathcal{M}_1=-\operatorname{cov}\left\{\mathcal{L}_\delta(\bullet)+\sum_{j=1}^{J}\mathcal{L}_{j\theta}(\bullet)\,\theta_\delta\left(Z_j,\delta_0\right)\right\},$$

$$\mathcal{M}_2 E\left[\sum_{j=1}^{J}\left\{1+\gamma\theta_0\left(Z_j\right)\right\}X_j\left\{\mathcal{L}_{j\theta\delta}(\bullet)+\sum_{k=1}^{J}\theta_\delta(Z_k,\delta_0)\,\mathcal{L}_{jk\theta}(\bullet)\right\}^{\mathrm{T}}\right],$$

$$\begin{aligned}\Psi_i(\gamma)=\;&\sum_{j=1}^{J}\left[X_{ij}\left\{1+\gamma\theta_0\left(Z_{ij}\right)\right\}+\theta_\beta\left(Z_{ij},0,\delta_0,\gamma\right)\right]\mathcal{L}_{j\theta}(\bullet_i)\\&-\mathcal{M}_2\mathcal{M}_1^{-1}\left\{\mathcal{L}_\theta(\bullet_i)+\sum_{j=1}^{J}\mathcal{L}_{j\theta}(\bullet_i)\,\theta_\delta\left(Z_{ij},\delta_0\right)\right\}.\end{aligned}$$

Then we have the following result.

Result 2: Suppose that we are interested in testing $H_0: \beta_0=0$. Assume that $h\ \alpha\ n^{-\alpha}$ where $1/3 \le \alpha \le 1/5$. Then, for any fixed γ, the score function for β_0 can be written as

$$\mathcal{T}_{n,\mathrm{pro}}(\gamma)=n^{-1/2}\sum_{i=1}^{n}\Psi_i(\gamma)+o_p\left(n^{-1/2}\right).$$

In addition, assume that, for any γ_1 and γ_2 $\mathcal{V}(\gamma_1,\gamma_2)=E\left\{\Psi(\gamma_2)^{\mathrm{T}}\right\}$ is finite. Then, under the hypothesis that $\beta_0=0$, $\mathcal{T}_{n,\mathrm{pro}}(\gamma)$ as a function of $\gamma \in [L, R]$ converges weakly to a Gaussian process $\mathcal{W}(\gamma)$ with mean 0 and covariance function $V(\gamma 1, \gamma 2)$.

Using result 2, we construct the test statistic and the critical values in the obvious analogy with Sections 2.3 and 2.4. To implement this in practice though, we must solve the integral equations for $\theta_\beta(\cdot\_$ and $\theta_\delta(\cdot)$, which is very difficult to do. In the next section, we show how to estimate these quantities without directly solving the integral equations.

3.4. Computation of $\theta_\beta(\cdot)$ and $\theta_\delta(\cdot)$

The main difficulty in performing the score test is that, for each γ, we must compute $\theta\hat{}_\beta(z, 0,\delta_0, \gamma)$ and $\theta\hat{}_\beta(z, 0, \delta_0)$, the former of which is the solution of integral equation (10), making implementation difficult. In this section we show that $\theta_\beta(z, 0, \delta_0, \gamma)$ can be viewed as a regression function and hence can be computed via a non-parametric Gaussian repeated measures regression, which is easily computed and for which the exact solution is known; see Huggins (2006) and Lin *et al.* (2004). The result can be stated as follows: details are in Appendix A.

Result 3—Define Q_{ij} =-$X_{ij}\{1+\gamma\theta_0(Z_{ij})\}$. Let V_i be the $J \times J$ matrix with elements v^{ijk}. Then θ_β (z, 0, δ_0, γ) is identified as the formal solution of the Gaussian repeated measures problem that was solved by Wang (2003) and Huggins (2006) with 'responses' being the components of Q_{ij} and the inverse of the covariance matrix being V_i

The algorithm for estimating $\theta_\beta(\cdot)$ now is quite simple. Define $\hat{Q}_{ij}$=-{1+$\gamma\theta$^(Z_{ij},δ^). Then we construct each component of θ^$_\beta$(z,0,β,γ) by performing a non-parametric repeated measures regression under the null model with β=0, with the response being the appropriate component of Q^$_{ij}$ and the inverse of the covariance matrix being $\hat{V}_i$=(v^^{ijk}), where $\widehat{v}^{jk} = -\mathcal{L}_{jk\theta}\left\{\widetilde{Y}_i, \widehat{\theta}\left(Z_{i1}, \widehat{\delta}\right) + S_{il}^{\mathrm{T}}\widehat{\eta}, \ldots, \widehat{\theta}\left(Z_{iJ}, \widehat{\delta}\right) + S_{iJ}^{\mathrm{T}}\widehat{\eta}, \widehat{\zeta}\right\}$ and θ^(z,δ^) is computed under the null model with β_0=0.

We can estimate $\theta_\delta(\cdot)$ in a similar manner. We do this componentwise. Let $\mathcal{L}_{j\theta\delta,l}(\bullet)$ denote the lth component of $\widehat{\theta}_\beta\left(z, 0, \widehat{\delta}, \gamma\right)$, and similarly for $\theta_{\delta,l}(\cdot)$. Define $\left(R_l^{i1}, \ldots, R_{iJ}^{l}\right)^T = -V_i^{-1}\left\{\mathcal{L}_{i1\theta\delta,l}(\bullet), \ldots, \mathcal{L}_{iJ\theta,l(\bullet)}\right\}^{\mathrm{T}}$. Then $\theta_{\delta,l}(\cdot)$ can be thought of as the Gaussian repeated measures regression of R_{ij}^l on Z_{ij} pretending that the inverse of the covariance matrix for the ith cluster is V_i. In practice, we construct θ^$_{\delta,l}(\cdot)$ by using $\widehat{R}_{ij}^l$ and $\widehat{V}_i^{-1}$.

3.5. Special case: partially linear repeated measurement model

In this section we consider the partially linear Gaussian model as an example to demonstrate our methodology. Specifically, we consider the model

$$Y_{ij} = X_{ij}^{\mathrm{T}}\beta_0\left\{1+\gamma\theta_0\left(Z_{ij}\right)\right\} + \theta_0\left(Z_{ij}\right) + S_{ij}^T\eta_0 + \varepsilon_{ij},$$

where $\widetilde{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$ has an $N(0, \Sigma)$ distribution. We want to test for H_0: β_0=0. The asymptotic theory is not affected by estimation of Σ, so here we assume that it is known.

Let Σ=(σ_{jk})$_{j,k=1,\ldots,J}$ and Σ^{-1}=V=(v^{jk}). Then the log-likelihood function is given by

$$\mathcal{L} = -\frac{1}{2}\sum_{q=1}^{J}\sum_{l=1}^{J} v^{q1}\left(Y_q - \mu_q\right)\left(Y_l - \mu_l\right),$$

where $\mu_j = X_j^{\mathrm{T}}\beta_0\left\{1+\gamma\theta_0\left(Z_j\right)\right\} + \theta_0\left(Z_j\right) + S_j^T\eta_0$. Now we observe that, when β_0=0,

$$\mathcal{L}_{j\theta}(\bullet) = \sum_{l=1}^{J} v^{jl}\left(Y_l - \mu_l\right),$$

$$\mathcal{L}_{j\theta\beta}(\bullet) = \gamma X_j \sum_{l=1}^{J} v_{jl}\left(Y_l - \mu l\right) - \sum_{l=1}^{J} v^{jl} X_l\left\{1+\gamma\theta_0\left(Z_l\right)\right\},$$

$$\mathcal{L}_{jk\theta}(\bullet) = -v^{jk}.$$

For β_0=0, θ_β (z, 0, η0, γ) solves

$$0 = \sum_{j=1}^{J} f_j(z)\, E\left(\sum_{k=1}^{J} v^{jk}\left[X_k\left\{1+\gamma\theta_0\left(Z_k\right)\right\} + \theta_\beta\left(Z_k\right), 0, \eta_0, \gamma\right] | Z_j = z\right). \tag{11}$$

Hence the profiled score function is given by

$$\mathcal{T}_{n,\mathrm{pro}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{J}\nu^{jk}\left[\left\{1+\gamma\widehat{\theta}\left(Z_{ij},\widehat{\eta}\right)\right\}X_{ij}+\widehat{\theta}_{\beta}\left(Z_{ij},0,\widehat{\eta},\gamma\right)\right] \times\left\{Y_{ik}-\widehat{\theta}\left(Z_{ik},\widehat{\eta}\right)-S_{ik}^{\mathrm{T}}\widehat{\eta}\right\}.$$

Now we can construct the score test by using result 2.

Remark 3—Referring to Section 3.4, we observe that estimation of $\theta_{\beta}(\cdot)$ becomes much simpler in this case. Using the fact that $\mathcal{L}_{jk\theta}(\bullet) = -\nu^{jk}$, we can construct $\theta\hat{}_{\beta}(\cdot)$ by performing a non-parametric *componentwise* Gaussian repeated measures regression of $\hat{Q}_k$=-{1+γ $\theta\hat{}$(Z_k,$\nu\hat{}$) on Z_k pretending that the error covariance matrix is Σ, where β_0=0 is computed under the null model with β_0=0. Similarly, we can estimate $\theta_{\eta}(\cdot)$ by performing a non-parametric Gaussian repeated measures regression of -S_{ij} on Z_{ij} by using Σ as the error covariance matrix.

3.6. Testing under working independence

In practice, often working independence is used to simplify the computations in the presence of repeated measures. In this set-up, we pretend that there is no correlation among the data. In our context, this leads to the assumption that σ_{jk}=0 for $j \neq k$, and we work with the criterion function

$$\mathcal{L}^{WI} = -\frac{1}{2}\sum_{j=1}^{J}\sigma_{jj}^{-1}\left(Y_j-\mu_j\right)^2,$$

where $\mu_j = X_j^{\mathrm{T}}\beta_0\left\{1+\gamma\theta_0\left(Z_j\right)\right\}+\theta_0\left(Z_j\right)+S_j^{\mathrm{T}}\eta 0$. The use of this criterion function simplifies the calculations to a great extent. For any generic random variable W, define $\tilde{W}_j = W_j - m_z^W\left(Z_j\right)$ with

$$m_z^W(z) = \sum_{j=1}^{J}\sigma_{jj}^{-1}f_j(z)\,E\left(W_j|Z_j=z\right)/\sum_{j=1}^{J}\sigma_{jj}^{-1}f_j(z).$$

Under the hypothesis that H_0: β_0=0, we then observe that now $\theta_{\beta}(\cdot)$ and $\theta_{\eta}(\cdot)$ have closed form expressions:

$$\theta_{\beta}(z,0,\eta_0,\gamma) = -\{1+\gamma\theta_0(z)\}m_z^X(z),$$
$$\theta_{\eta}(z,\eta_0) = -m_z^S(z).$$

The profiled score statistic is given by

$$\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\sigma_{jj}^{-1}\left\{1+\gamma\widehat{\theta}\left(Z_{ij},\widehat{n}\right)\right\}\tilde{X}_{ij,\mathrm{est}}\left\{Y_{ij}-\widehat{\theta}\left(Z_{ij},\widehat{\eta}\right)-S_{ij}^{\mathrm{T}}\widehat{\eta}\right\},$$

where $\tilde{X}_{ij,\mathrm{est}} = X_{ij} - \widehat{m}_z^X\left(Z_{ij}\right)$. We can compute $\widehat{m}_z^X(z)$ by running a componentwise Gaussian repeated measures regression on X_{ij} and Z_{ij} by using the working independence set-up.

Further define

$$\mathcal{M}_1 = -\operatorname{cov}\left[\sum_{j=1}^{J}\sigma_{jj}^{-1}\tilde{S}_j\left\{Y_j - \theta_0\left(Z_j\right) - S_j^{\mathrm{T}}\eta_0\right\}\right],$$
$$\mathcal{M}_2 = -E\left[\sum_{j=1}^{J}\sigma_{jj}^{-1}\left\{1+\gamma\theta_0\left(Z_j\right)\right\}X_j\tilde{S}_j^{\mathrm{T}}\right].$$

Result 2 then translates to the following result.

Result 4—Assume that $h \alpha n^{-\alpha}$ where $1/3 \le \alpha \le 1/5$. Then, under the assumption of working independence,

$$\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\sigma_{jj}^{-1}\left[\left\{1+\gamma\theta_0\left(Z_{ij}\right)\right\}\tilde{X}_{ij} + \mathcal{M}_2\mathcal{M}_1^{-1}\tilde{S}_{ij}\right]\left\{Y_{ij} - \theta_0\left(Z_{ij}\right) - S_{ij}^{\mathrm{T}}\eta_0\right\} + o_p(1).$$

Define $\Psi_{ij}^{*}(\gamma) = \left\{1+\gamma\theta_0\left(Z_{ij}\right)\right\}\tilde{X}_{ij} + \mathcal{M}_2\mathcal{M}_1^{-1}\tilde{S}_{ij}$ and let $\widehat{\Psi}_{ij}^{*}(\gamma)$ (γ) be the sample version. Under the null hypothesis, we estimate the covariance matrix of $\mathcal{T}_{n,\mathrm{pro}}^{WI}$ by

$$\mathcal{I}_{\beta_0,n}^{\mathrm{WI}} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\sigma_{jj}^{-1}\widehat{\Psi}_{ij}^{*}(\gamma)\,\widehat{\Psi}_{ij}^{*}(\gamma)^{\mathrm{T}}.$$

The score statistic, maximized over γ, is then given by

$$\mathcal{T}_n^{*} = \max_{\gamma\in[L,R]}\left\{\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma)^{\mathrm{T}}\left(\mathcal{I}_{\beta_0,n}^{\mathrm{WI}}\right)^{-1}\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma)\right\}.$$

Using lemma 4 in Appendix A, we can now implement the score test by using the technique that was described in Section 2.4. We start by generating

$$\mathcal{T}_0^{\mathrm{WI}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\sigma_{jj}^{-1}\widehat{\Psi}_{ij}^{*}(\gamma)\,\mathcal{Z}_{ij},$$

where $\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{iJ})^{\mathrm{T}}$, $i=1, \ldots, n$, are independent random vectors that are generated from an $N(0,\hat{\Sigma})$ distribution. We can form $\hat{\Sigma}$ as the sample covariance matrix of the residuals $\left\{Y_{ij} - \widehat{\theta}\left(Z_{ij},\widehat{\eta}\right) - S_{ij}^{\mathrm{T}}\widehat{\eta}\right\}$. The null distribution of $\mathcal{T}_n^{*}$ is then simulated by repeatedly generating

$$\mathcal{T}_0^{*} = \max_{\gamma\in[L,R]}\left\{\mathcal{T}_0^{\mathrm{WI}}(\gamma)^{\mathrm{T}}\left(\mathcal{I}_{\beta_0,n}^{\mathrm{WI}}\right)^{-1}\mathcal{T}_0^{\mathrm{WI}}(\gamma)\right\}.$$

Remark 4—We reiterate that one needs to estimate $\widehat{m}_Z^X\left(Z_{ij}\right)$ and $\widehat{m}_Z^S\left(Z_{ij}\right)$ to implement the score test. These quantities can be easily estimated by performing componentwise Gaussian repeated measures regressions of X_{ij} and S_{ij} on Z_{ij} by using the working independence set-up.

4. Simulations

4.1. Testing without repeated measures

For the simulation for the test for $\beta_0=0$, we used the following conventions. We used 31 values of γ in the range [-3, 3]. The variable Z=Uniform[-2, 2], whereas the function $\theta_0(z)=\sin(2z)$ is distinctly non-linear. In keeping with our data example, the sample size was n=1400.

We generated X in three ways:

- **a.** as a bivariate standard normal random variable;
- **b.** $X=(X_1, X_2)$ where X_1=Bernoulli(0.6) and $X_2=N(0, 1)$;
- **c.** as two dummy variables. Thus, we first generated a standard normal random variable r, and $X_1=I(r < -0.4)$ and $X_2=I(r>0.4)$.

We set $\beta_0=c(1, 1)^{\mathrm{T}}$, where we set c=0.0, 0.01, . . . ,, 0.15 for power calculations. The true value of γ was varied: γ_{true}=0,1,2. We ran simulations both with and without additional covariates S: in the former case, we set S to be generated from a univariate $N(0, 1)$ distribution and used $\eta 0=1$.

For each scenario, we ran 1000 simulated data sets. To estimate the level of significance, we applied the method in Section 2.4 with 1500 replications. The Epanechnikov kernel was used to carry out the computation. We used different bandwidths of the form $h=\kappa\ \text{std(Z)}n^{-1/5}$ with various values of κ ranging from 0.5 to 2. The results are very similar in each of those cases and hence we report the results for κ=1 only. The results are displayed in Figs 1-3. There three main conclusions are clear.

- **a.** The test level of our method is near nominal, being 0.051 without S and 0.057 with S in the model.
- **b.** For the main effects model with γ_{true}=0, our maximized score-type test loses only modest power compared with the efficient (in this case) main effects score test.
- **c.** When there are interactions, our methods greatly dominate the main effects score test as γ_{true} increases.

For comparison, we repeated the simulation by using penalized B-spline regression, using a second-order B-spline with 10 basis functions and with a second-order difference penalty. The smoothing parameter was chosen by generalized cross-validation. The results were very similar to those obtained for kenrnel methods. The near equivalence of kernel and spline methods here is no surprise, since there is evidence in Gaussian cases that smoothing splines are equivalent to kernel methods (Silverman, 1984; Lin *et al.*, 2004). Recently, Li and Ruppert (2008) showed that penalized B-spline regression is also asymptotically equivalent to kernel regression methods in the Gaussian case.

4.2. Testing with repeated measures

We use the following set-up for our simulations for testing $\beta_0=0$. We generate samples from the partially linear gaussian repeated measures model: for i=1, . . . , n and j=1, . . . , J,

$$Y_{ij}=X_{ij}^{\mathrm{T}}\beta_0+\theta_0\left(Z_{ij}\right)\left(1+\gamma X_{ij}^{\mathrm{T}}\beta_0\right)+\varepsilon_{ij},$$

with n=200 and J=3, where we take the true value of the parameter to be $\beta_0=c(1, -1)^{\mathrm{T}}$ and set c=0.001, . . . , 0.06 for power calculation. We set $\tau_0(z)=\sin(2z)$ to be the true function. We generated X from the standard bivariate normal distribution and Z from the Uniform[-2, 2]

distribution. The error vectors $(\varepsilon 1, \ldots, \varepsilon J)^{\mathrm{T}}$ are generated from a multivariate normal distribution with covariance matrix $\sigma = I + 0.6\,(\mathbf{1}\mathbf{1}^{\mathrm{T}} - I)$.

We use 11 values of γ in [0, 2] to compute the test statistic. The true values of γ that are used to generate the data are taken to be $\gamma_{\mathrm{true}}=0, 1, 2$. As in the previous simulation, we use the Epanechnikov kernel with bandwidth $h=\kappa$ std $(Z)n^{-1/5}$ where the value of κ ranged from 0.5 to 2. In this case also, we observe that the results are very similar for each of the choices of bandwidth and hence we report the results for $\kappa=1$. We generate 1000 data sets for each case and for each data set we apply our method by using 1000 replications. The results are given in Fig. 4. The level of our test is 0.051, which is very close to the nominal level of 0.05. It is evident that, although our test loses very little power when $\gamma_{\mathrm{true}}=0$, it achieves great power gain in the presence of interaction as seen in cases where $\gamma_{\mathrm{true}}=1, 2$.

We redid the simulation by using *B*-splines with 10 basis functions where the penalty parameter is estimated at the null model by using generalized cross-validation. The results are nearly identical of Fig. 4, as we would expect in the Gaussian case.

5. Data analysis

Chatterjee *et al.* (2006) illustrated application of their methodology by using a case-control study for investigation of association between colorectal adenoma, a precursor of colorectal cancer, and NAT2, a candidate gene that is known to play an important role in detoxification of certain aromatic carcinogens in cigarette smoke. The study involved about 700 cases and 700 controls who were genotyped for six known functional polymorphisms related to NAT2 acetylation activity. The genotype data were used to construct diploytpe information, i.e. the pair of haplotypes that the subjects carried along their pair of homologous chromosomes. The frequency distribution of these diplotypes and associated acetylation phenotypes are shown in Table 4 of Chatterjee *et al.* (2006). In principle, the diplotypes are not observed directly and we can only assign diplotypes on the basis of the unphased genotype data. However, in many instances such as this example, when we have tightly linked SNPs, the phase ambiguity is often minimal, i.e. we can assign a very large proportion (greater than 95%) of the subjects a specific diplotype with a very high probability (greater than 0.95). In such cases, it is esier just to remove those few people for whom the diplotypes are more uncertain and to assume that for the rest of the people the diplotypes are known. In our data set. we removed a small number of people whose haplotypes were quite uncertain.

Chatterjee *et al.* (2006) considered an omnibus test that can account for interaction of NAT2 history with smoling history, defined as ever, former or never smokers. We consider a similar application involving NAT2 diplotypes but model the effect of CIG_STOP (years since stopping smoking) in a continuous fashion with non-parametric regression among smokers. Because of a few high leverage values, we censored CIG_STOP at 45. In our analysis, the cofactor *S* included gender and three indicator dummy variables for age level: between 60 and 65 years, between 65 and 70 years and more than 70 years. For modelling the effect of NAT2 diplotypes, we considered a series of 14 different analyses where in the *k*th analysis we compare the risk that is associated with the *k* ($k=1, \ldots,, 14$) most common diplotypes in reference to the rest, with the associated design matrix X_k being defined by *k* corresponding dummy variables. To account for non-smokers in this analysis, we defined δ to be the indicator of smoking (ever *versus* never) and considered the following model:

$$\mathrm{pr}\,(D=1|X,S,Z)=H\left\{(1-\delta)\beta_0+S^{\mathrm{T}}\beta_1+X^{\mathrm{T}}\beta_2+\delta\theta\,(Z)+\gamma\delta X^{\mathrm{T}}\beta_2\theta\,(Z)\right\}. \tag{12}$$

Modifying our methods to handle this slightly more complex model is straightforward: details are available from the authors.

Table 1 compares results of the proposed method for testing $\beta_2=0$ on the basis of model (12) with those for a test for only the corresponding main effects of the diplotypes, ignoring NAT2-smoking interaction, i.e. assuming $\gamma=0$. We observe that, in each analysis, stronger evidence of association is seen in our new test. For example, when the 12 most common diplotypes were used, our method had a level of significance of 0.036 *versus* a level of significance of 0.214 for the main-effect-based test. Interestingly, when all 14 common diplotypes are used, the level of significance of the test proposed was 0.066, which is quite close to that for the test that was used by Chatterjee *et al.* (2006), also using all the 14 diplotypes, but accounting for interaction with the categorical smoking history variable defined as never, former or current smoker.

6. Discussion

We have developed methodology for an efficient score test for genetic effect in general semiparametric models that can account for gene-environment interaction with non-parametrically specified environmental effects. The procedure proposed allows for repeated measurements.

We proposed a profiled score statistic which can be performed by using standard bandwidth selection procedures. We also found that these profiled score tests are efficient.

The main difficulty of performing the score test is that one must estimate a function which itself is a solution of an integral equation that is difficult to solve. In the case of repeatedly measured data, the solution generally does not a closed form expression and hence some sort of numerical procedure is required for estimation. In this paper, we overcome this problem by developing an easily implementable estimation procedure which does not involve solving integral equations and can be performed easily via standard software. The key idea lies in the fact that the target functions, based on their estimating equations, can be interpreted as Gaussian repeated measures regressions.

Simulations that were presented in the paper show that the score tests proposed maintain the desired type I error level, indicating that the asymptotic approximations work well for studies such as ours. Moreover, both simulation studies and the data example indicate that the score test proposed taking account of the interaction can achieve higher statistical power than naive tests which ignore interaction altogether. Future research areas of interest include extension of the score test to account for the interaction of the genetic factors with several different, but biologically related, environmental factors, such as different biomarkers for a nutrient, simultaneously. In principle, the score test can be extended by using generalized additive models to account for the effect of several different continuous exposures. Further theoretical development, however, is needed to establish the asymptotic theory for such procedures.

Acknowledgements

Maity and Carroll's research was supported by grants from the National Cancer Institute (CA-57030 and CA104620). Mammen's research was supported by the Deutsche Forschungs-gemeinschaft project MA 1026/7-3. Chatterjee's research was supported by a 'Gene-environment initiative' grant from the National Heart, Lung and Blood Institute and by the intramural research programme of the National Cancer Institute.

Appendix A: Argument for result 1

For simplicity of notation, here we consider only the case that there are no nuisance parameters ζ_0. The more general case is a simple extension.

To prove the results, we rely on several technical conditions that for brevity we do not state here explicitly. These conditions are well known and standard in smoothing theory. Refer to Claeskens and Van Keilegom (2003), Claeskens and Carroll (2007) and Lin and Carroll (2006) among many others for the details of these assumptions. As stated just before result 1, we require that the linear expansions for the parametric and non-parametric parts that were given in Lin and Carroll (2006) hold to order o_p ($n^{-1/2}$); the latter uniformly.

A.1. Expansion of $\mathcal{T}_n(\gamma)$

Let $\theta^{(j)}(\cdot)$ be the jth derivative of $\theta(\cdot)$ with respect to z_0. Let $f_z(z_0)$ be the density function of Z. Make the definitions

$$\Omega(z_0) = E\left[\mathcal{L}_{\theta\theta}\left\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z)\right\} | Z = z_0\right],$$
$$\theta_\eta(z_0, \eta_0) = -E\left[S\,\mathcal{L}_{\theta\theta}\left\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z)\right\} | Z = z_0\right] / \Omega(z_0).$$

Note that $S_i + \theta_\eta(Z_i, \eta_0) = \tilde{S}_i$, and recall that

$$\mathcal{M} = -\operatorname{cov}\left[\left\{S + \theta_\eta(Z, \eta_0)\right\}\mathcal{L}_\theta\left\{Y, S^{\mathrm{T}}\eta_0 + \theta(Z)\right\}\right].$$

Then using Lin and Carroll (2006) we have that, uniformly in z_0,

$$\begin{aligned}\widehat{\theta}(Z_0, \eta_0) - \theta_0(Z_0, \eta_0) = {} & -n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\,\mathcal{L}_\theta\left\{Y_i S_i^{T}\eta_0 + \theta_0(Z_i)\right\} / f_z(z_0)\,\Omega(z_0) + \left(\phi_2 h^2/2\right)\theta_0^{(2)}(z_0) \\ & + O_p\left\{h^4 + \log(n)(n)/nh\right\},\end{aligned} \tag{13}$$

$$\widehat{\eta} - \eta_0 = -\mathcal{M}^{-1} n^{-1}\sum_{i=1}^{n}\tilde{S}_i\mathcal{L}_\theta\left\{Y_i S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\right\} + o_p\left(n^{-1/2}\right). \tag{14}$$

The score statistic for β is, via Taylor series,

$$\begin{aligned}\mathcal{T}_{n,\mathrm{pro}}(\gamma) = {} & n^{-1/2}\sum_{i=1}^{n}\left\{1 + \gamma^{\theta_0}(Z_i)\right\}\tilde{X}_i\mathcal{L}\left\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\right\} + n^{-1/2}\sum_{i=1}^{n}\mathcal{S}_{1i}(\gamma)\left\{\widehat{\theta}(Z_i - \theta_0(Z_i))\right\} \\ & + n^{-1/2}\sum_{i=1}^{n}\mathcal{S}_{2i}(\gamma)(\widehat{\eta} - \eta_0) + o_p(1) \\ = {} & A_{1n} + A_{2n} + A_{3n} + o_p(1)\end{aligned}$$

where

$$\mathcal{S}_{1i}(\gamma) = \tilde{X}_i\left[\gamma\mathcal{L}_\theta\left\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\right\} + \left\{1 + \gamma\theta_0(Z_i)\right\}\mathcal{L}_{\theta\theta}\left\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\right\}\right]$$
$$\mathcal{S}_{2i}(\gamma) = \gamma\theta_\eta(Z_i)\tilde{X}_i\mathcal{L}_\theta\left\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\right\} + \left\{1 + \gamma\theta_0(Z_i)\right\}\tilde{X}_i\tilde{S}_i^{T}\mathcal{L}_{\theta\theta}\left\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\right\}.$$

By definition of $\tilde{X}$, it is easy to see that, to order $o_p(1)$,

$$A_{2n} = -n^{-1/2}\sum_{i=1}^{n}\mathcal{L}_\theta\left\{Y_i, S_i^T\eta_0+\theta_0(Z_i)\right\}E\{\mathcal{S}_{1i}(\gamma)|Z_i\}/\Omega(Z_i)=0,$$

where we have used equations (6) and (13). Also, using equation (14) and the definition of $\mathcal{N}$ we obtain

$$A_{3n} = -\mathcal{N}\mathcal{M}^{-1}n^{-1/2}\sum_{i=1}^{n}\tilde{S}_i\mathcal{L}_\theta\left\{Y_i S^{\mathrm{T}}\eta_0+\theta_0(Z_i)\right\}+o_p(1).$$

The result now follows by collecting all ther terms. It is readily seen that the expansion is uniform in $\gamma\in$[L, R].

A.2. Weak convergence

Weak convergence is trivial. Examining the form of the test statistic $\mathcal{T}_{n,\mathrm{pro}}(\gamma)$ in equation (8), we see that it is linear in γ and can be written as $U_n + \gamma V_n$, where (U_n, V_n) are jointly asymptotically normally distributed.

Appendix B: Argument for result 2

Define

$$\Omega(z)=\sum_{j=1}^{J}f_j(z)\,E\left\{\mathcal{L}_{jj\theta}(\cdot)|Z_j=z\right\}$$

and

$$\mathcal{A}(B,Z_1,Z_2)=\sum_{j=1}^{J}\sum_{k\neq j=1}^{J}f_j(z_1)\,E\left\{\mathcal{L}_{jk\theta}(\cdot)\,B(Z_k,z_2)/\Omega(Z_k)|Z_j=z_1\right\},$$
$$Q(Z_1,Z_2)=\sum_{j=1}^{J}\sum_{k\neq j=1}^{J}f_{jk}(z_1,z_2)\,E\left\{\mathcal{L}_{jk\theta}(\cdot)|Z_j=z_1,Z_k=z_2\right\}/\Omega(Z_2),$$

where $f_j(z)$ is the density of Z_j and $f_{jk}(z_1, z_2)$ is the bivariate density of (Z_j, Z_k), which are assumed to have bounded support and are positive on the support. Let $\mathcal{G}(z_1,z_2)$ be the solution to

$$\mathcal{G}(z_1,z_2)=Q(z_1,z_2)-\mathcal{A}(\mathcal{G},z_1,z_2).$$

Using the results of Lin and Carroll (2006) we obtain that, uniformly in z,

$$\begin{aligned}\widehat{\theta}(z,\eta_0)-\theta_0(z)= &\left(\phi h^2/2\right)b(z)-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h\left(Z_{ij}-z\right)\mathcal{L}_{ij\theta}(\bullet)/\Omega(z)+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{L}_{ij\theta}(\bullet)\mathcal{G}\left(z,Z_{ij}\right)/\Omega(z)\\ &+o_p\left\{h^4+\log(n)nh\right\},\end{aligned} \tag{15}$$

$$\widehat{\eta}-\eta_0=-\mathcal{M}_1^{-1}n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{S_{ij}+\theta_n\left(Z_{ij},\eta_0\right)\right\}\mathcal{L}_{ij}(\bullet)+o_p\left(n^{-1/2}\right). \tag{16}$$

Define

$$\mathcal{T}_{k,n}(\gamma)=\sum_{j=1}^{J}\left[X_j\left\{1+\gamma\theta_0\left(Z_j\right)\right\}+\theta_\beta\left(Z_j,0,\eta_0,\gamma\right)\right]\mathcal{L}_{jk\theta}(\cdot),$$
$$\mathcal{T}_{\eta,n}(\gamma)=\sum_{j=1}^{J}\sum_{k=1}^{J}\left[X_j\left\{1+\gamma\theta_0\left(Z_j\right)\right\}+\theta_\beta\left(Z_j,0,\eta_0,\gamma\right)\right]\left(S_k+\theta_\eta\left(Z_k,\eta_0\right)\right)^{\mathrm{T}}\mathcal{L}_{jk\theta}(\cdot).$$

It is easily shown that

$$\begin{aligned}\mathcal{T}_{n,\mathrm{pro}}(\gamma)=\;& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\left[X_{ij}\left\{1+\gamma\theta_0\left(Z_{ij}\right)\right\}+\theta_\beta\left(Z_{ij},0,\eta_0,\gamma\right)\right]\mathcal{L}_{ij\theta}(\cdot)+n^{-1/2}\sum_{i=1}^{n}\mathcal{T}_{i\eta,n}(\gamma)\left(\widehat{\eta}-\eta_0\right)\\ &+\; n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathcal{T}_{ik,n}(\gamma)\left\{\widehat{\theta}\left(Z_{ij},\eta_0\right)-\theta_0\left(Z_{ik}\right)\right\}+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{ij\theta}(\cdot)\left\{\widehat{\theta}_\beta\left(Z_{ij},0,\widehat{\eta},\gamma\right)-\theta_\beta\left(Z_{ij},0,\eta_0,\gamma\right)\right\}\\ &+\; o_p(1).\end{aligned}$$

Using equation (16) and the fact that $E\left\{\mathcal{T}_{\eta,n}(\gamma)\right\}=\mathcal{M}_2$, it is easy to see that

$$n^{-1/2}\sum_{i=1}^{n}\mathcal{T}_{i\eta,n}^{\mathrm{T}}(\gamma)\left(\widehat{\eta}-\eta_0\right)=-\mathcal{M}_2\mathcal{M}_1^{-1}n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\left\{S_{ij}+\theta_\eta\left(Z_{ij},\eta_0\right)\right\}\mathcal{L}_{ij\theta}(\bullet)+o_p(1).$$

Next, using equation (15), we now derive that, up to terms of $o_p(1)$,

$$\begin{aligned}n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\left\{\widehat{\theta}\left(Z_{ik},\eta_0\right)-\theta_0\left(Z_{ik}\right)\right\}\;&=-n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\left\{n^{-1}\sum_{r=1}^{n}\sum_{j=1}^{J}K_h\left(Z_{rj}-Z_{ik}\right)\mathcal{L}_{rj\theta}(\cdot)/\Omega\left(Z_{ik}\right)\right\}\\ &\quad+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{T}_{ik,n}(\gamma)\left\{n^{-1}\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\cdot)\mathcal{G}\left(Z_{ik},Z_{rj}\right)/\Omega\left(Z_{ik}\right)\right\}\\ &=n^{-1/2}\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\cdot)\left\{C_1\left(Z_{rj}\right)+C_2\left(Z_{rj}\right)\right\},\end{aligned}$$

where we define

$$C_1(z,\gamma)=-\sum_{k=1}^{J}f_k(z)\,E\left\{\mathcal{T}_{ik,n}(\gamma)\,|Z_k=z\right\}\Omega(z),$$
$$C_2(z,\gamma)=E\left[\sum_{k=1}^{J}E\left\{\mathcal{T}_{ik,n}(\gamma)\,|Z_k\right\}\mathcal{G}\left(Z_k,z\right)/\Omega\left(Z_k\right)\right].$$

We now note that

$$\sum_{k=1}^{J}f_k(z)\,E\left\{\mathcal{T}_{ik,n}(\gamma)\,|Z_k=z\right\}=0$$

by deifnition of $\theta_\beta(\cdot)$ with $\beta_0=0$ and hence $C_1(z,\gamma)=C_2(z,\gamma)=0$.

Finally, we recognize that $\theta\hat{}_\beta(\cdot)$ is the repeated measures regression of Q_{ij} on Z_{ij} and hence yields an asymptotic expansion similar to equation (15). Together with the fact that $E\left\{\mathcal{L}_{j\theta}(\cdot)|X,S,Z\right\}=0$, it is now straightforward to show that the fourth term in the expansion of $\mathcal{T}_{n,\text{pro}}(\gamma)=o_p(1)$ completing the proof.

Appendix C: Argument for result 3

Under the null hypothesis, $\theta_\beta(z,0,\delta_0,\gamma)$ solves

$$0=\sum_{j=1}^{J} f_j(z)\,E\left(\sum_{k=1}^{J}\left[X_k\{1+\gamma\theta_0(z_k)\}+\theta_\beta(Z_k,0,\delta_0,\gamma)\right]\mathcal{L}_{jk\theta}(\bullet)|Z_j=z\right). \tag{17}$$

Recall that $K_h(z)=h^{-1}K(z/h)$ and $G_h(z)=(1,z/h)^{\mathrm{T}}$. Consider the problem of solving, for $|\{m(z), m^{(1)}(z)|\}$,

$$0=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}K_h\left(Z_{ij}-z\right)G\left(Z_{ij}-z\right)\left[\sum_{k\neq j=1}^{J}\nu^{ijk}\{Q_{ik}-m(Z_{ik})\}+\nu^{ijj}Q_{ij}-\nu^{ijj}G\left(Z_{ij}-z\right)^{\mathrm{T}}\left(m(z),m^{(1)}(z)\right)^{\mathrm{T}}\right]$$

where $\nu^{ijk}=-\mathcal{L}_{ijk\theta}(\bullet)$ Define

$$F_n(z)=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\nu^{ijj}K_h\left(z_{ij}-z\right)G\left(z_{ij}-z\right)^{\mathrm{T}}.$$

The solution then satisfies

$$F_n(z)\left(m(z),m^{(1)}(z)\right)^{\mathrm{T}}=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h\left(Z_{ij}-z\right)\left[\sum_{k\neq j=1}^{J}\nu^{ijk}\{Q_{ik}-m(Z_{ik})\}+\nu^{ijj}Q_{ij}\right].$$

Note that

$$F_n(z)=\sum_{j=1}^{J}E\left(\nu^{ij}|Z_j=z\right)\begin{pmatrix} f_j(z) & 0\\ 0 & \phi_2\end{pmatrix}+o_p(1),$$

Where $\phi_2=\int z^2k(z)\,dz$. Hence, taking the limit of both sides we obtain that $m(z)$ satisfies

$$\sum_{j=1}^{J}E\left(\nu^{jj}|Z_j=z\right)f_j(z)\,m(z)=\sum_{j=1}^{J}f_j(z)\sum_{k\neq j=1}^{J}E\left[\nu^{jk}\{Q_k-m(Z_k)\}|Z_j=z\right]+\sum_{j=1}^{J}f_j(z)\,E\left(\nu^{jj}Q_j|Z_j=z\right),$$

which is identical to equation (17) with $m(z)=\theta_\beta(z,0,\delta_0,\gamma)$. This completes the argument.

References

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multi-locus tests for genetic assocation in the presence of gene-gene and gene-environment interactions. Am. J. Hum. Genet 2006;79:1002–1016. [PubMed: 17186459]

Claeskens G, Carroll RJ. An asymptotic theory for model selection inference in general semiparametric problems. Biometrika 2007;94:249–265.

Claeskens G, Van Keilegom I. Bootstrap confidence bands for regression curves and their derivatives. Ann. Statist 2003;31:1852–1884.

Davies RB. Hypothesis testing when a nuisance parameter is present only under the null hypothesis. Biometrika 1987;74:33–43.

Hayes RB, Reding D, Kopp W, Subar AF, Bhat N, Rothman N, Caporaso N, Ziegler RG, Johnson CC, Weissfeld JL, Hoover RN, Hartge P, Palace C, Gohagan JK. Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (plco) cancer screening trial. Contr. Clin. Trials 2000;21(suppl 6):349S–355S.

Horowitz JT, Spokoiny VG. An adaptive, rate-optimal test of a parametric model against a non-parametric alternative. Econometrica 2001;69:599–631.

Huggins R. Understanding nonparametric estimation for clustered data. Biometrika 2006;93:486–489.

Kraft P, Yen Y-C, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum. Hered 2007;63:111–119. [PubMed: 17283440]

Li Y, Ruppert D. On the asymptotics of penalized splines. Biometrika 2008;95:415–437.

Lin DY, Zou F. Assessing genomewide statistical significance in linkage studies. Genet. Epidem 2004;27:202–214.

Lin X, Carroll RJ. Semiparametric estimation in general repeated measures problems. J. R. Statist. Soc. B 2006;68:69–88.

Lin X, Wang N, Welsh A, Carroll RJ. Equivalent kernels of smoothing splines in nonparametric regression for clustered data. Biometrika 2004;91:177–193.

Silverman B. Spline smoothing: the equivalent variable kernel method. Ann. Statist 1984;12:898–916.

Tukey JW. One degree of freedom for non-additivity. Biometrics 1949;5:232–242.

Wang N. Marginal nonparametric kernel regression accounting for within-subject correlation. Biometrika 2003;90:43–52.

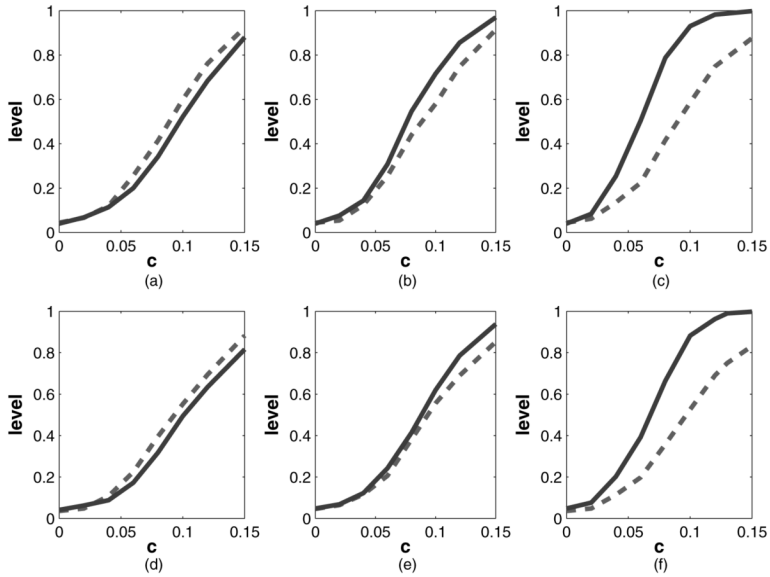

Fig. 1.

Results of the simulation for testing whether $\beta=0$ as described in Section 4.1 by using kernel-based calculations (here X is a bivariate standard normal random variable; ———, our method; - - -, naive test, value of c and the vertical axis plots the corresponding power): (a) $\gamma_{\text{true}}=0$, without S; (b) $\gamma_{\text{true}}=1$, without S; (c) $\gamma_{\text{true}}=2$, without S; (d) $\gamma_{\text{true}}=0$, with S; (e) $\gamma_{\text{true}}=1$, with S; (f) $\gamma_{\text{true}}=2$, with S

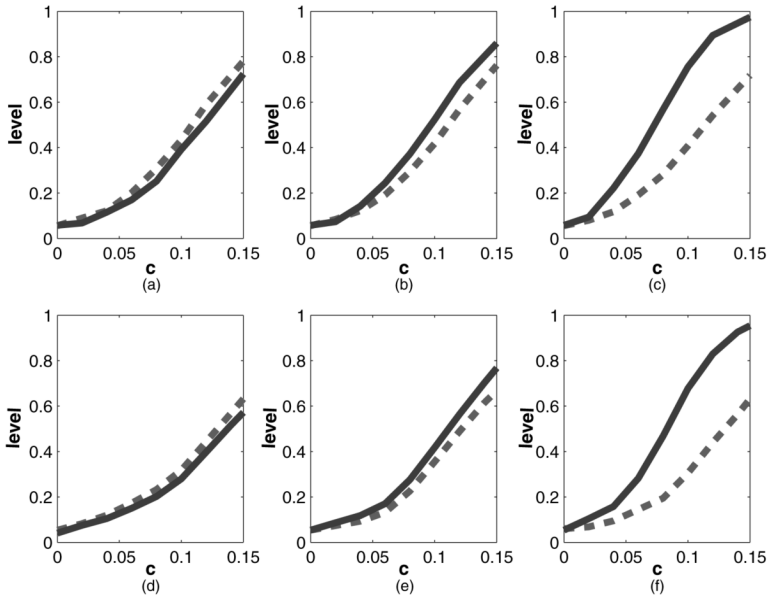

Fig. 2.

Results of the simulation for testing whether $\beta=0$ as described in Section 4.1 by using kernel-based calculations (here $X=(X_1, X_2)$ where X_1=Bernoulli (0.6) and $X_2=N(0, 1)$;————, our method;-------, naive test, which assumes $\gamma=0$; the true value that was used was $\beta=c(1, 1)^T$; the horizontal axis plots the value of c and the vertical axis plots the corresponding power): (a) $\gamma_{true}=0$, without S; (b) $\gamma_{true}=1$, without S; (c) $\gamma_{true}=2$, without S; (d) $\gamma_{true}=0$, with S; (e) $\gamma_{true}=1$, with S; (f) $\gamma_{true}=2$, with S

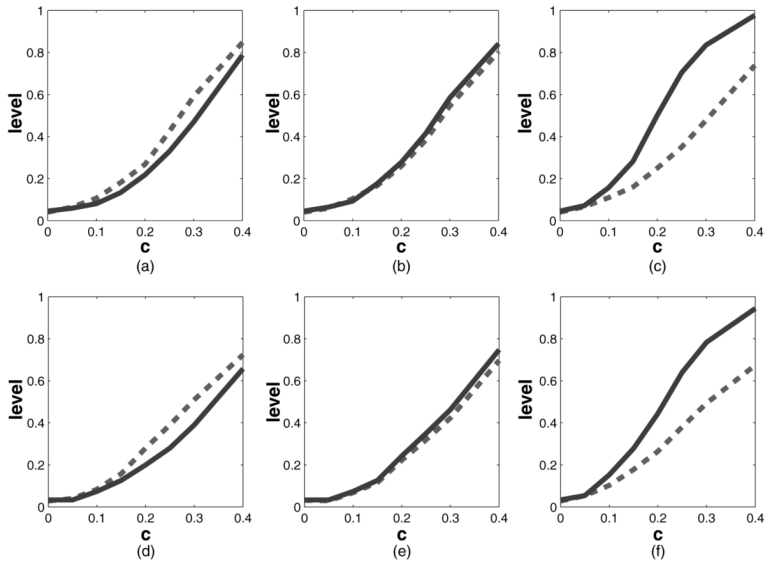

Fig. 3.

Results of the simulation for testing whether $\beta=0$ as described in Section 4.1 by using kernel-based calculations (here $X=(X_1, X_2)$ is two dummy variables; thus, we first generated a standard normal random variable r, and $X_1=I(r<-0.4)$ and $X_2=I(r<0.4)$; ————, our method;-------, naive test, which assumes $\gamma=0$; the true value that was used was $\beta=c(1, 1)^T$; the horizontal axis plots the value of c and the vertical axis plots the corresponding power): (a) $\gamma_{true}=0$, without S; (b) $\gamma_{true}=1$, without S; (c) $\gamma_{true}=2$, without S; (d) $\gamma_{true}=0$, with S; (e) $\gamma_{true}=1$, with S; (f) $\gamma_{true}=2$, with S

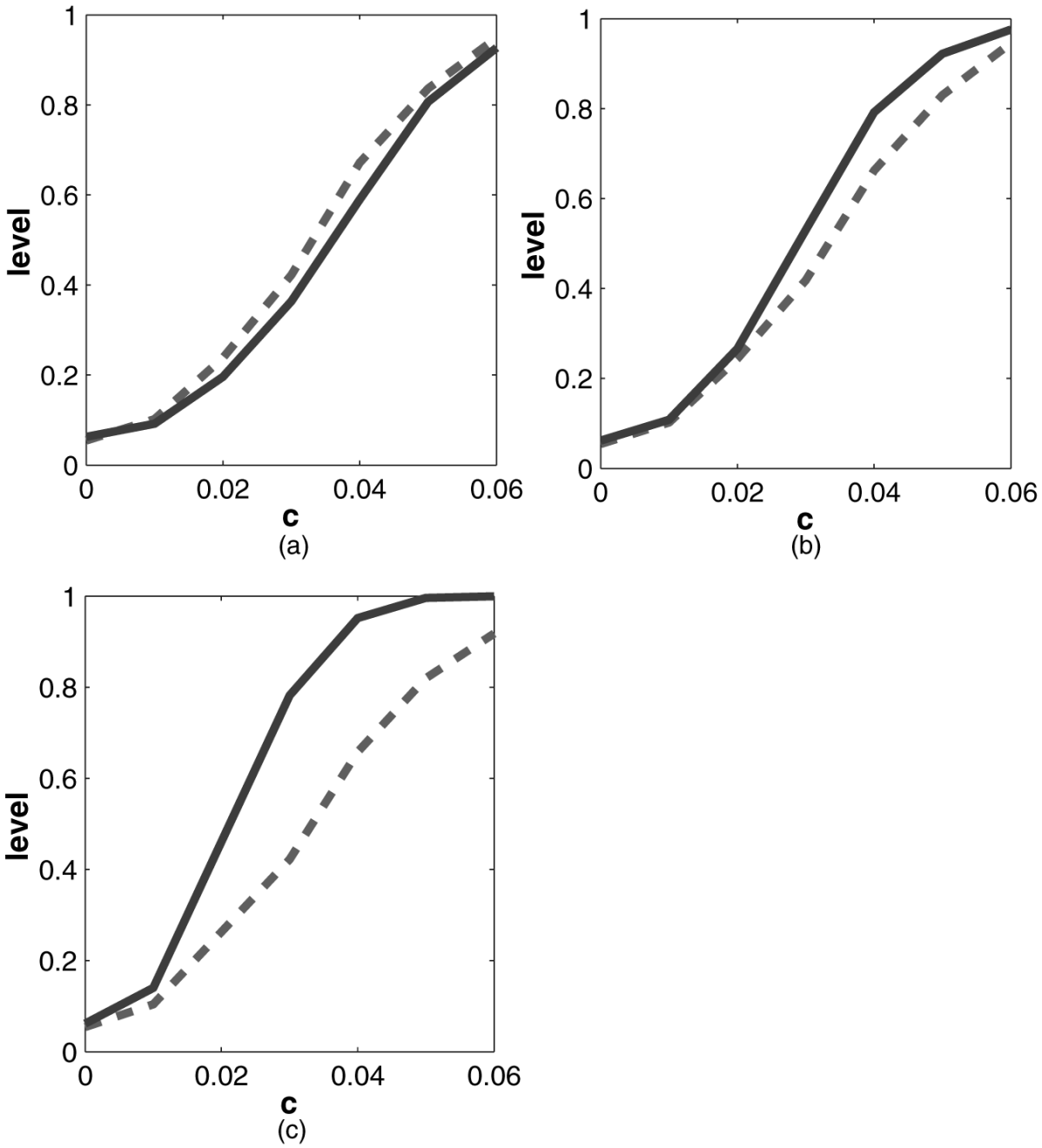

Fig. 4.

Results of the simulation for testing whether $\beta_0=0$, as described in Section 4.2 by using kernel-based calculations (_____, our method; - - -, naive test which assumes $\gamma=0$; the true value that was used was $\beta=c(1, -1)T$; the horizontal axis plots the value of c and the veritcal axis plots the corresponding power): (a) $\gamma_{true}=0$; (b) $\gamma_{true}=1$; (c) $\gamma_{true}=2$

Table 1

Levels of significance (*p*-values) of the test for genetic effects in a regression model in which *Z* is years since stopping smoking[†]

Diplotype	*Results for our method*		*Results for γ=0*	
	Test	***p-value***	***Test***	***p-value***
1	11.4	0.001	3.3	0.066
2	13.9	0.003	5.7	0.055
3	16.6	0.002	9.8	0.016
4	16.7	0.007	9.8	0.041
5	19.5	0.007	11.3	0.045
6	19.7	0.017	11.4	0.087
7	20.0	0.021	12.3	0.098
8	21.3	0.025	13.1	0.111
9	24.1	0.015	14.2	0.116
10	25.2	0.016	15.3	0.120
11	25.2	0.027	15.4	0.180
12	25.6	0.036	15.4	0.214
13	25.9	0.055	15.8	0.262
14	26.7	0.066	16.6	0.279

[†]Age category and gender were modelled additively and parametrically. The analysis is done for the most common diplotype, the most common two diplotypes, and so on. The non-parametric regression was done by using penalized order 2 *B*-splines with 10 segments, with penalization done via generalized cross-validation.