Published in final edited form as:

Stat Probab Lett. 2011 July 1; 81(7): 717–723. doi:10.1016/j.spl.2010.11.002.

Testing for Constant Nonparametric Effects in General Semiparametric Regression Models with Interactions

Jiawei Wei,

Department of Statistics, 3143 TAMU, Texas A&M University, College Station, Texas 77843, USA

Raymond J. Carroll, and

Department of Statistics, 3143 TAMU, Texas A&M University, College Station, Texas 77843, USA

Arnab Maity

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A

Jiawei Wei: wjw@stat.tamu.edu; Raymond J. Carroll: carroll@stat.tamu.edu; Arnab Maity: amait@ncsu.edu

Abstract

We consider the problem of testing for a constant nonparametric effect in a general semi-parametric regression model when there is the potential for interaction between the parametrically and nonparametrically modeled variables. The work was originally motivated by a unique testing problem in genetic epidemiology (Chatterjee, et al., 2006) that involved a typical generalized linear model but with an additional term reminiscent of the Tukey one-degree-of-freedom formulation, and their interest was in testing for main effects of the genetic variables, while gaining statistical power by allowing for a possible interaction between genes and the environment. Later work (Maity, et al., 2009) involved the possibility of modeling the environmental variable nonparametrically, but they focused on whether there was a parametric main effect for the genetic variables. In this paper, we consider the complementary problem, where the interest is in testing for the main effect of the nonparametrically modeled environmental variable. We derive a generalized likelihood ratio test for this hypothesis, show how to implement it, and provide evidence that our method can improve statistical power when compared to standard partially linear models with main effects only. We use the method for the primary purpose of analyzing data from a case-control study of colorectal adenoma.

Some Key Words

Function estimation; Generalized likelihood ratio; Interactions; Non-parametric regression; Partially linear logistic model; Partially linear model; Semiparametric models

1 Introduction

The motivating example for this research comes from a case-control study of colorectal adenoma, a recognized precursor of colorectal cancer, involving 628 prevalent advanced adenoma cases and 635 gender-matched controls, selected from the screening arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Cancer Institute, USA (Gohagan et al, 2000; Moslehi et al, 2006). Data were collected on each of the subjects on their smoking history and other demographic variables such as gender, age etc. Each of the subjects were genotyped for six known functional Single Nucleotide Polymorphisms (SNPs) related to NAT2, a candidate gene that is known to play an important role in detoxification of certain aromatic carcinogens in cigarette smoke. The goal of our study is to investigate the effect of cigarette smoking on the risk of colorectal adenoma while accounting for the genetic markers and their interaction with cigarette smoking.

As part of the analysis of the data, more generally we consider the problem of testing for constant nonparametric effect in a general semiparametric regression model when there is the potential for interaction between the parametrically and nonparametrically modeled variables. We develop a generalized likelihood ratio test for testing constant nonparametric effect, show how to implement it, and provide evidence that our testing procedure can improve statistical power compared to standard partially linear models.

Chatterjee, et al. (2006) and Maity, et al (2009) considered the complementary problem of testing the genetic association of a disease with a set of genetic variants, such as tagging SNPs in a candidate gene, that may potentially interact with another set of genetic variants or/and with one or more environmental exposures. They selected six SNPs for genotyping which are known to be informative for reconstructing diplotypes. They investigated the association between colorectal adenoma and NAT2 by considering a logistic regression type problem where the interaction between NAT2 diplotypes and cigarette smoking is modeled using Tukey-like 1-degree of freedom formulation. Specifically, Chatterjee, et al. (2006) considered a binary disease status Y , a set of genetic variables X that might possibly interact with a scalar environmental variable Z , and propose the model

$$\text{pr}(Y=1|X, S, Z)=H(\kappa_0+X^T\beta_0+S^T\eta_0+Z\theta_0+\gamma X^T\beta_0Z\theta_0), \quad (1)$$

where $H(\cdot)$ be the logistic distribution function and S are other demographic variables not thought to interact with Z or X . Here the term $\gamma X^T\beta_0Z\theta_0$ represents the Tukey-like 1-degree of freedom interaction term parameterized by a scalar interaction parameter γ . Their interest was in testing for a possible genetic main effect, $H_0 : \beta_0 = 0$ versus $H_A : \beta_0 \neq 0$. Maity, et al. (2009) generalized models such as (1) to allow the effect of the environmental variable Z to enter nonparametrically, that is, they consider the model

$$\text{pr}(Y=1|X, S, Z)=H\{X^T\beta_0+S^T\eta_0+\theta_0(Z)+\gamma X^T\beta_0\theta_0(Z)\}, \quad (2)$$

where $\theta_0(\cdot)$ is an unknown function quantifying the effect of the environmental factor Z . When $\gamma = 0$, (1) and (2) reduce to ordinary and partially linear main effect only logistic regression models and thus the testing for H_0 is routine. However, Chatterjee, et al. (2006) and Maity, et al. (2009) argue that if there is a possible gene-environment interaction, then capturing it via the Tukey-like 1-degree of freedom interaction term can increase statistical power greatly. It is important to see that under the null hypothesis $H_0 : \beta_0 = 0$, the interaction parameter γ is not identifiable. Chatterjee, et al. (2006) and Maity, et al. (2009) propose to construct maximal score-type test statistic where they construct score statistic for each fixed γ along a range of values $L \leq \gamma \leq R$ and then take the maximum value of these score statistics as the final test statistic. They then develop a simulation-based procedure for computing an overall p-value.

While testing for effect of genetic factors is an important problem by itself, risk of colorectal adenoma may also be modified by certain environmental variables, e.g., individual's addiction to smoking. Thus, testing for a main effect of the environmental variable is also of great interest in our problem, because we wish to understand what smoking variables are affecting colorectal adenoma. Thus, we consider the problem of testing for the effect of the environmental factor, such as cigarette smoking, on the risk of colorectal adenoma. In general, we consider a general semiparametric regression problem with Tukey-type 1 degree of freedom interaction

$$\mathcal{L}\{Y, X^T\beta_0 + S^T\eta_0 + \theta_0(Z) + \gamma X^T\beta_0\theta_0(Z), \zeta_0\},$$

where $\mathcal{L}$ is a loglikelihood function and ζ_0 is a collection of other parameters, e.g., a scale parameter or any nuisance parameters present in the model. We address the problem of testing for constant environmental effect $H_0 : \theta(z) = \text{constant}$ under this model. We develop a generalized likelihood ratio based testing procedure for H_0 and demonstrate that our method can result in substantial gain in power when there is in fact interaction present between X and Z , but with little loss of power if $\gamma = 0$ actually holds.

An outline of this paper is as follows. In Section 2 we will develop a generalized likelihood ratio based testing procedure for more general semiparametric likelihood problems. Section 3 gives a simulation study, while Section 4 gives the data analysis. The technical details justifying the method are described for the logistic case in the Appendix.

2 Methodology

2.1 Basic Framework

Suppose, for the i^{th} sample, $i = 1, \dots, n$, we observe (Y_i, X_i, Z_i, S_i) , where Y_i is an outcome of interest, X_i denotes the parametrically modeled genetic covariate vector, Z_i denotes the nonparametrically modeled environmental factor and S_i denotes other demographic covariates. We assume that the vectors X and Z do not have an entry 1.0 for an intercept. We consider general loglikelihood functions of the form

$$\mathcal{L}\{Y, X^T\beta_0 + S^T\eta_0 + \theta_0(Z) + \gamma X^T\beta_0\theta_0(Z), \zeta_0\}, \quad (3)$$

where β_0 and η_0 are the main effects, $\theta_0(\cdot)$ is an unknown function, ζ_0 is a set of other parameters, e.g., scale parameter or any nuisance parameters present in the model, and γ is the interaction effect. Recall that the interaction parameter is not to be estimated directly since it is unidentifiable when either $\beta_0 = 0$ or $\theta_0(\cdot)$ is a constant. Henceforth, all technical details will be exhibited for the logistic model (2), although as we indicate below, the result holds much more generally. As stated previously, we are interested in testing the null hypothesis $H_0 : \theta_0(\cdot) = \text{constant}$.

2.2 Estimation of Model Components

To test H_0 , we use the concept of a generalized likelihood ratio test (Fan et al., 2001). To implement the testing procedure, we need to estimate the model components under the full and null models.

We use a kernel based profile method to estimate the parameters under the full model. However, since γ is not identifiable when either $\beta_0 = 0$ or $\theta_0(\cdot)$ is a constant, we will perform

the estimation for fixed values of γ . Let $K(\cdot)$ be a symmetric density function and for any bandwidth h , let $K_h(t) = K(t/h)/h$. Then the local linear profile method works as follows: for any given $(\beta, \eta, \zeta) = (\beta^*, \eta^*, \zeta^*)$ and γ , we maximize the local loglikelihood

$$\sum_{i=1}^n K_h(Z_i - z_0) \mathcal{L}(Y_i, \alpha_0 + \alpha_1(Z_i - z_0) + X_i^T \beta^* [1 + \gamma \{\alpha_0 + \alpha_1(Z_i - z_0)\}] + S_i^T \eta^*, \zeta^*)$$

with respect to α_0 and α_1 , and set $\hat{\theta}(z_0, \beta^*, \eta^*, \zeta^*, \gamma) = \hat{\alpha}_0$. Then the profile estimates of (β, η, ζ) is obtained by maximizing

$$\sum_{i=1}^n \mathcal{L} \{Y_i, X_i^T \beta + S_i^T \eta + \hat{\theta}(Z, \beta, \eta, \zeta, \gamma) + \gamma X_i^T \hat{\theta}(Z, \beta, \eta, \zeta, \gamma), \zeta\}. \quad (4)$$

Let the resulting estimator be $(\hat{\beta}_F, \hat{\eta}_F, \hat{\zeta}_F)$, where it is understood that these estimates depend on the value of γ chosen.

Estimation under the null model is a purely parametric problem where one computes the MLE in the reduced model under H_0 . We add an intercept κ so that we are maximizing

$$\sum_{i=1}^n \mathcal{L}(Y_i, \kappa + X_i^T \beta + S_i^T \eta, \zeta).$$

Let $(\hat{\kappa}_R, \hat{\beta}_R, \hat{\eta}_R, \hat{\zeta}_R)$ be the resulting null model estimates.

Remark 1—For specific models, the maximization of (4) is quite simple and can be implemented easily. For example, in logistic regression, the steps are as follows. There is no nuisance parameter ζ . For any given (β, η) , define $\mathcal{U}_i = X_i^T \beta + S_i^T \eta$ and $\mathcal{V}_i = 1 + \gamma X_i^T \beta$. Then $\hat{\theta}(z_0, \beta, \eta, \gamma)$ is the estimated intercept ζ_0 in the linear logistic regression model

$$\text{pr}(Y_i=1) = H\{\mathcal{U}_i + \xi_0 \mathcal{V}_i + \xi_1 \mathcal{V}_i (Z_i - z_0)\}$$

with the weights $K_h(Z_i - z_0)$. This procedure is a weighted logistic regression with no intercept, an offset $\mathcal{U}_i$, and predictors $\mathcal{V}_i$ and $\mathcal{V}_i(Z_i - z_0)$, and is hence easily implemented. Computing $(\hat{\beta}_F, \hat{\eta}_F)$ is then done by performing maximum likelihood under the model

$\text{pr}(Y_i=1) = H\{X_i^T \beta + S_i^T \eta + \hat{\theta}(Z, \beta, \eta, \gamma) + \gamma X_i^T \hat{\theta}(Z, \beta, \eta, \gamma)\}$ based on profile method. In our numerical computations, we used the function `optim()` in R with initial values estimated by backfitting.

Remark 2—The one-degree of freedom model (3) can be relaxed by replacing $\gamma X^T \beta_0$ by a more general $X^T \beta_1$ when the dimension of X is low. In this respect, the problem is not symmetric in the two components X and Z , and the null hypothesis would be $H_0 : \beta_1 = 0$ instead of $H_0 : \theta_0(\cdot) = \text{constant}$. The local linear profile method works as follows: for any given $(\beta_0, \beta_1, \eta, \zeta) = (\beta_0^*, \beta_1^*, \eta^*, \zeta^*)$ and γ , we maximize the local loglikelihood

$$\sum_{i=1}^n K_h(Z_i - z_0) \mathcal{L}[Y_i, \alpha_0 + \alpha_1(Z_i - z_0) + X_i^T \beta_1^* \{\alpha_0 + \alpha_1(Z_i - z_0)\} + X_i^T \beta_0^* + S_i^T \eta^*, \zeta^*]$$

with respect to α_0 and α_1 , and set $\widehat{\theta}(z_0, \beta_0^*, \beta_1^*, \eta^*, \zeta^*) = \widehat{\alpha}_0$. Then the profile estimates of $(\beta_0, \beta_1, \eta, \zeta)$ is obtained by maximizing

$$\sum_{i=1}^n \mathcal{L}\{Y_i, X_i^T \beta_0 + S_i^T \eta + \widehat{\theta}(Z, \beta_0, \beta_1, \eta, \zeta) + X_i^T \beta_1 \widehat{\theta}(Z, \beta_0, \beta_1, \eta, \zeta), \zeta\}.$$

2.3 Properties of Profile Estimates of Parameters and Functions

In order to be able to draw upon the work of Fan and Huang (2005) and Fan, et al. (2001), we require to know the properties of the parameters and function estimates under the null hypothesis of constant $\theta_0(\cdot)$.

The properties of profile estimates of parameters and function estimates have been well-studied in fairly general contexts, see for example Claeskens and Van Keilegom (2003), Claeskens and Carroll (2007), Van Keilegom and Carroll (2007) and Apanasovich, et al. (2009), among many, many others. Specifically, the parameter estimates are $n^{1/2}$ -consistent and the function estimates have uniform linear expansions to order $o_p(n^{-1/2})$. Conditions, summarized in Apanasovich, et al. (2009) and translated to our context, are as follows. All assumptions are meant to apply to the null hypothesis, since our asymptotic results pertain only to the null hypothesis of constant $\theta_0(\cdot)$. This means that there are simplifications to the calculations of Apanasovich, et al. (2009), who also study misspecified models, a topic not of relevance in this paper.

2.4 Generalized Likelihood Ratio Test

Given any fixed γ , the generalized likelihood ratio test statistic is given by

$$\Lambda_n(\gamma) = \sum_{i=1}^n \left[\mathcal{L}\{Y_i, X_i^T \widehat{\beta}_F + S_i^T \widehat{\eta}_F + \widehat{\theta}(Z, \widehat{\beta}_F, \widehat{\eta}_F, \widehat{\zeta}_F, \gamma) + \gamma X_i^T \widehat{\beta}_F \widehat{\theta}(Z, \widehat{\beta}_F, \widehat{\eta}_F, \widehat{\zeta}_F, \gamma), \widehat{\zeta}_F\} - \mathcal{L}\{Y_i, \tilde{\kappa}_R + X_i^T \widehat{\beta}_R + S_i^T \widehat{\eta}_R, \widehat{\zeta}_R\} \right].$$

Under H_0 , from Section 2.3 we have that the parameters are estimated $n^{1/2}$ -consistently. As in Fan and Huang (2005), this means that the likelihood ratio statistic behaves asymptotically as if the parameters are known. As they note, it is easy to show that the likelihood ratio statistic is $\Lambda_n(\gamma) = \Lambda_n^*(\gamma) + O_p(1)$, where

$$\Lambda_n^*(\gamma) = \sum_{i=1}^n \left[\mathcal{L}\{Y_i, X_i^T \beta_0 + S_i^T \eta_0 + \widehat{\theta}(Z, \beta_0, \eta_0, \zeta_0, \gamma) + \gamma X_i^T \beta_0 \widehat{\theta}(Z, \beta_0, \eta_0, \zeta_0, \gamma), \zeta_0\} - \mathcal{L}\{Y_i, \kappa_0 + X_i^T \beta_0 + S_i^T \eta_0, \zeta_0\} \right].$$

This statistic is easily analyzed because of the expansion (A.1), and indeed that expansion allows us to show the following result by using similar arguments as in Fan, et al. (2001), see also Fan and Jiang (2005).

Theorem 1—Assume conditions (C.1)–(C.6) in A.1. There is a constant r_K depending on the kernel function and a deterministic sequence $\mu_n(h) \propto h^{-1} \rightarrow \infty$ depending on the bandwidth h such that

$$r_K \{\Lambda_n^*(\gamma) - \mu_n(h)\} / \{2r_K \mu_n(h)\}^{1/2} \Rightarrow \text{Normal}(0, 1). \quad (5)$$

A consequence of Theorem 1 is that because $\mu_n(h) \rightarrow \infty$,

$$r_K \{\Lambda_n(\gamma) - \mu_n(h)\} / \{2r_K \mu_n(h)\}^{1/2} \Rightarrow \text{Normal}(0, 1). \quad (6)$$

Result (6) is the so-called Wilks-phenomenon, namely that the semiparametric likelihood ratio statistic has a common limiting distribution under the null hypothesis independent of the problem.

2.5 Test Statistic and Implementation

While (6) holds, it is not very useful in practice for decision making because it depends upon the bandwidth. This fact motivated Fan and Jiang (2005) to use a bootstrap-type test. Here we propose a parametric bootstrap-type test to overcome this problem as follows.

Since the true value of γ is unknown, we follow the idea of Davies (1987), Chatterjee, et al. (2006) and Maity, et al. (2009) and propose to use as the test statistic

$$\mathcal{T}_n^* = \max_{L \leq \gamma \leq R} \Lambda_n(\gamma),$$

where L and R are pre-specified lower and upper bounds for γ . A normalized version of $\mathcal{T}_n^*$ as in (6) converges to the maximum of a Gaussian process, see the Appendix. However, this is not very useful in terms of setting a critical level due to the dependence upon the bandwidth. We propose instead a simulation based approach to compute p-values as in the papers of J. Fan and colleagues.

- Let B be a large number, and for $b = 1, \dots, B$, generate response data Y_{ib} from the null model fits.
- For each of the $b = 1, \dots, B$ generated data sets, compute the test statistic $T_{n,b}^*$.
- The p-value is then computed as $B^{-1} \sum_{b=1}^B I(T_{n,b}^* > \mathcal{T}_n^*)$.

3 Simulation Study

We investigate the performance of our approach using a simulation study. We simulate data using the partially linear logistic model

$$\text{pr}(Y|X, Z) = H\{X^T \beta_0 + \theta(z) + \gamma X^T \beta_0 \theta(z)\},$$

where $H(\cdot)$ denotes the logistic distribution function. The sample size was $n = 1, 200$, X is generated from a standard bivariate normal distribution and Z is generated from a uniform

distribution on $[-2, 2]$. We set true values of $\beta_0 = (1, -1)^T$ and vary $\gamma_{\text{true}} = 0, 1, 2$. To calculate the Type I error, we set $\theta(z) = 0$ and the alternative values of the function were given as $\theta(z) = c \sin(2z)$ for $c = 0.125, 0.250, 0.375$ for power calculations. We repeat the simulation 1,000 times. For each simulated data set, we fit the null model, namely logistic in X , then simulated from this null model $B = 1,000$ times to obtain a p-value. The values of γ used to construct our test statistic were 11 equally spaced values on the interval $[-2, 2]$. We used the Epanechnikov kernel to estimate the function $\theta(\cdot)$ and used bandwidth $h = \hat{\sigma}_Z n^{-1/5}$, where $\hat{\sigma}_Z$ is the standard deviation of Z . The results were not sensitive to varying h by factors of 3.0 in each direction.

In the null case, for nominal 5% tests, the actual significance level of our test was 3.9%, while the actual significance levels of the main effects test that set $\gamma = 0$ was 5.2%. We compare power of our method to the model which ignores interaction altogether ($\gamma = 0$). The results are given in Figure 1. It is evident that our method shows considerable power gain over the naive method of ignoring interaction when there is in fact interaction present ($\gamma = 1$ and 2). However, when the true value of $\gamma = 0$, our method has very little power loss.

Because of the Wilks phenomenon and the similarity between kernel regression and penalized spline regression, we also implemented the tests using penalized 2nd-order B-splines with equally spaced knots and 10 basis functions. Because our theory for kernel regression assumes that the same bandwidth is used for all values of γ , the penalty parameter was chosen with $\gamma = 0$ using GCV (Ruppert, et al., 2003). In fitting the non-null method, for any given γ we obtained estimates of β and $\theta(\cdot)$ by maximizing the loglikelihood function penalized by $-(\lambda/2)\zeta^T \kappa \zeta$, where λ is the penalty parameter chosen as above, $B(z)$ are the basis functions, $\theta(z) = B^T(z)\zeta$, and κ is the penalty matrix. The results were almost identical to the kernel method and hence not displayed here.

4 Data Example

Here we provide an analysis of the study described in the Introduction. We use the data from the PLCO Cancer Screening Trial case-control study in Chatterjee et al. (2006) to demonstrate our method. In our data set, we removed the nonsmokers, leaving 328 cases and 372 controls who were genotyped for six known functional polymorphisms related to NAT2 acetylation activity. As in Maity et al. (2009), we consider the three most common NAT2 diplotypes, which in our notation is X , in comparison to the rest. The demographic variables S include gender and three indicator dummy variables for age level: between 60 and 65 years, between 65 and 70 years and more than 70 years. We explore three different environmental variables Z , namely “CIG STOP”, the number of years since stopping smoking; “PhIP”, 2-Amino-1-methyl-6-phenylimidazo[4,5-b]pyridine, which has been demonstrated to produce adenocarcinomas in mice, and “Red Meat”, daily grams of red meat intake. We test for constant effect for each of these three environmental factors using our method and compare our results to the main effects only model.

The results are displayed in Table 1. The number of years since stopping smoking is a highly statistically significant predictor of colorectal adenoma, using our model and method or simply ignoring any potential interaction. There appears to be little indication of an effect due to red meat consumption, but of course food consumption in general is measured with considerable error, so that this negative result is perhaps not so surprising. For PhIP, when we include the 3 most common diplotypes in the model, the p-value for our method is 0.162 versus a p-value of 0.938 when the interaction is ignored. While of course not statistically significant, this vast difference in p-values, coupled with the results of the simulation, suggest that there may be an interaction going on, and that PhIP might be a significant predictor of colorectal adenoma, although larger sample sizes would be necessary. Since

CIG STOP is a highly statistically significant predictor of colorectal adenoma, we redid the analysis for PhIP while including CIG STOP as a linear covariate in the model. The last three rows in Table 1 show the results. The vast difference in p-values between our method and the naive method disappear. This is not surprising since most of the effect are due to CIG STOP in this model.

Somewhat more generally, we see that in all cases, the p-values using our method are smaller than that when γ is fixed to = 0. This is not a theorem of course, but it does show support with the results of the simulations, which indicate that if there is an interaction, our method will have greater statistical power.

To visualize the environmental effect $\theta(Z)$, we plot the estimated function $\hat{\theta}(\cdot)$ for different values of $\gamma = 0, 1$ and 2 in Figure 2(a) for the model where $Z = \text{PhIP}$ and X represents three most common diplotypes. It is evident that the function is nonlinear and the shape differs as γ changes. We also plot in figures 2(b) – (d) estimates of the total effect $X^T\beta + S^T\eta + \theta(Z) + \gamma X^T\beta\theta(Z)$ (z-axis) versus estimated effect of the environmental covariate PhIP $\theta(Z)$ (x-axis) and estimated effect of the 3 most common diplotypes $X^T\beta$ (y-axis) for γ is 0, 1 and 2, respectively. The distinct patterns of these plots for different values of γ seem to suggest the presence of nontrivial interaction and hence the gain in power from our method.

5 Discussion

We have developed a testing procedure to test for a constant environmental effect in general semiparametric regression models with Tukey-type 1 degree of freedom interaction, where the environmental factor is modeled nonparametrically. The testing procedure is based on maximal generalized likelihood ratio based test statistic and the computation of the p-values are done using a parametric bootstrap type procedure. The methodology was described for kernel regression methods and justified in the important logistic regression case. Numerically, we have found that regression spline approaches are very close to being the same as kernel methods and much faster to compute, although their theory remains an open question in this context.

Acknowledgments

Our research was supported by a grant from the National Cancer Institute (CA57030). Carroll's research was also supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

- Apanasovich TV, Carroll RJ, Maity A. SIMEX and variance estimation in semiparametric measurement error models. *Electronic Journal of Statistics*. 2009; 3:318348.
- Billingsley, P. *Convergence of Probability Measures*. Wiley; New York: 1968.
- Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multi-locus tests for genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics*. 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. *Annals of Statistics*. 2003; 31:1852–1884.
- Davies RB. Hypothesis testing when a nuisance parameter is present only under the null hypothesis. *Biometrika*. 1987; 74:3343.
- de Jong P. A Central Limit Theorem for generalized quadratic forms. *Probability Theory and Related Fields*. 1987; 75:261–277.

- Fan J, Huang T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*. 2005; 11:1031–1057.
- Fan J, Jiang J. Nonparametric inferences for additive models. *Journal of the American Statistical Association*. 2005; 100:890–907.
- Fan J, Zhang C, Zhang J. Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*. 2001; 29:153–193.
- Gohagan JK, Prorok PC, Hayes RB, et al. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Controlled Clinical Trials*. 2000; 21(6 Suppl):251S–272S. [PubMed: 11189683]
- Maity A, Carroll RJ, Mammen E, Chatterjee N. Testing in semiparametric models with interaction, with applications to gene-environment interactions. *Journal of the Royal Statistical Society, Series B*. 2009; 71:75–96.
- Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfield J, Hein DW, Hayes RB. Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics*. 2006; 7:819–829. [PubMed: 16981843]
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Cambridge University Press; 2003.
- Van Keilegom I, Carroll RJ. Backfitting versus profiling in general criterion functions. *Statistica Sinica*. 2007; 17:797–816.

Appendix A Partially Linear Logistic Model

To illustrate the general result (6), we consider the partially linear logistic model (2). However, it is clear from the expansions described in Section 2.3, which are quite general, that the Wilks phenomenon result holds more generally.

Define $W(\gamma) = 1 + \gamma X^T \beta_0$.

A.1 Assumptions and Preliminary Results

We now state the assumptions and conditions required for our proposed method.

- (C.1) The kernel function K is a symmetric, continuously differentiable density function on $[-1, 1]$ taking on the value zero at the boundaries.
- (C.2) The bandwidth is $h \propto n^{-1/5}$.
- (C.3) The random variables (X, S, Z) have compact support. The design density $f_Z(\cdot)$ of Z is strictly positive and twice continuously differentiable on its support.
- (C.4) The parameter space, here denoted by $\mathcal{B}$, is compact. For any $(\beta^*, \eta^*, \zeta^*)$, let $\theta(z_0, \beta^*, \eta^*, \zeta^*, \gamma)$ be the maximizer in v of $E[\mathcal{L}\{Y, v + X^T \beta^* (1 + \gamma v) + S^T \eta^*, \zeta^*\} | Z = z_0]$, which is assumed to exist. The function $\theta(\cdot, \beta, \eta, \zeta, \gamma)$ has 3 continuous derivatives in its arguments. We also assume that the same calculations done by Claeskens and Carroll (2007) can be applied to our context.
- (C.5) For each γ , and under the null hypothesis, $\theta(z, \beta, \eta, \zeta, \gamma)$ is constant in z , $(\beta_0, \eta_0, \zeta_0)$ is the unique maximizer of $E(\mathcal{L}[Y, \theta(Z, \beta, \eta, \zeta, \gamma) + X^T \beta \{1 + \gamma \theta(Z, \beta, \eta, \zeta, \gamma)\} + S^T \eta, \zeta])$. In addition, the second total derivative of this function is uniformly negative definite in a neighborhood of $(\beta_0, \eta_0, \zeta_0)$.
- (C.6) We can apply the results of Claeskens and Van Keilegom (2003) as needed. In particular, their assumptions imply that uniformly in z_0 , for random variables R_i possessing sufficient moments, then if superscript (1) means first derivative, if

$$C_n = n^{-1} \sum_{i=1}^n K_h(Z_i - z_0)(Z_i - z_0)^j R_i \times \{\theta(z_0) + (Z_i - z_0)\theta^{(1)}(z_0)\};$$

$$D_n = n^{-1} \sum_{i=1}^n K_h(Z_i - z_0)(Z_i - z_0)^j R_i \times \{\theta(Z_i)\},$$

then

$$\sup_{z_0} |C_n - E(C_n)| = O_p[h^j \{\log(n)/(nh)\}^{1/2}];$$

$$\sup_{z_0} |D_n - E(D_n)| = O_p[h^j \{\log(n)/(nh)\}^{1/2}].$$

Under these assumptions, at the null hypothesis, their work can be easily extended to show that uniformly on compact sets of γ , $(\hat{\beta}_F, \hat{\eta}_F, \hat{\zeta}_F)$ are $n^{1/2}$ -consistent estimates of $(\beta_0, \eta_0, \zeta_0)$. In addition, let $\mathcal{L}_\theta(\cdot)$ and $\mathcal{L}_{\theta\theta}(\cdot)$ denote the first and second derivatives with respect to θ , respectively, and define

$$\Omega(z_0, \mathcal{B}_0) = E(\mathcal{L}_{\theta\theta}[Y, \theta(z_0, \beta_0, \eta_0, \zeta_0)] + X^T \beta_0 \{1 + \gamma \theta(z_0, \beta_0, \eta_0, \zeta_0)\} + S^T \eta_0, \zeta_0 | Z = z_0),$$

where $\mathcal{B}_0$ is the true parameter value. Then we also have the result that under the null hypothesis $\theta_0(z_0) \equiv \theta_0$,

$$\widehat{\theta}(z_0, \beta_0, \eta_0, \zeta_0, \gamma) = \theta_0 - n^{-1} \sum_{i=1}^n K_h(Z_i - z_0) R_i + o_p(n^{-1/2});$$

$$R_i = - \frac{\mathcal{L}_{\theta\theta}\{Y_i, \theta_0 + X_i^T \beta_0 (1 + \gamma \theta_0) + S_i^T \eta_0, \zeta_0\}}{f_{Z'}(z_0) \Omega(z_0, \mathcal{B}_0)}, \quad (\text{A.1})$$

uniformly in compact sets of γ .

A.2 Testing Theory

For the null model, we have that

$$\text{pr}(Y=1|X, Z) = H(\kappa_0^* + X^T \beta_0^* + S^T \eta_0^*)$$

independent of γ . Define $\theta(\gamma)$ so that $\kappa_0^* + X^T \beta_0^* + S^T \eta_0^* = X^T \beta_0 + S^T \eta_0 + \theta(\gamma) + \gamma X^T \beta_0 \theta(\gamma)$. Then, under the null model,

$$\text{pr}(Y=1|X, Z) = H\{X^T \beta_0 + S^T \eta_0 + W(\gamma) \theta(\gamma)\}, \quad (\text{A.2})$$

while under the full model,

$$\text{pr}(Y=1|X, Z) = H\{X^T \beta_0 + S^T \eta_0 + W(\gamma) \theta(Z, \beta_0, \eta_0, \gamma)\}, \quad (\text{A.3})$$

and the null hypothesis is that

$$H_0: \theta(z, \beta_0, \eta_0, \gamma) \equiv \theta(\gamma). \quad (\text{A.4})$$

We have already described in Section 2.2 that we can treat (β_0^*, η_0^*) and (β_0, η_0) as if they were known. In this case, with the exception of the minor difference of the offset $X^T \beta_0 + S^T \eta_0$ in (A.2) and (A.3), for any fixed γ the problem is exactly the same as a special case of that studied by Fan, et al. (2001) in their Theorem 10, which applies to generalized linear models. There are, however, huge simplifications because the null hypothesis (A.4) takes a particularly simple form.

Let $K_c(v) = K * K(v)$, the convolution function, and let $K_{ch}(v) = h^{-1} K_c(v/h)$. Define the volume of the support of Z to be $\nu = E\{1/f_Z(Z)\}$, and also define $\mu_n(h) = h^{-1} \nu \{K(0) - (1/2) \int K^2(t) dt\}$ and $\sigma_n^2(h) = 2\mu_n(h)/r_K$, where

$$r_K = \frac{K(0) - (1/2) \int K^2(t) dt}{\int \{K(t) - (1/2) K_c(t)\}^2 dt}.$$

Let $\Lambda_n(\gamma)$ be the likelihood ratio test statistic. Then by Fan, et al. (2001), under the null hypothesis, independent of the value of γ , $\sigma_n^{-1}(h) \{\Lambda_n(\gamma) - \mu_n(h)\} \Rightarrow \text{Normal}(0, 1)$. This implies that the mean of $r_K \Lambda_n(\gamma)$ is $r_K \mu_n(h)$ and the variance is $2\mu_n(h) r_K$, as one would have with a chi-squared random variable with $r_K \mu_n(h)$ degrees of freedom, see their Theorem 5 on page 165 and Theorem 10 on page 174.

Define

$$\begin{aligned} \Omega(z_0, \gamma) &= f_Z(z_0) E[W^2(\gamma) H^{(1)}\{X^T \beta_0 + S^T \eta_0 + W(\gamma) \theta(Z, \gamma)\} | Z = z_0]; \\ \varepsilon &= Y - H\{X^T \beta_0 + S^T \eta_0 + W(\gamma) \theta(Z, \gamma)\}. \end{aligned}$$

Using Fan, et al. (2001) (see their page 191), which applies to generalized linear models and allows heteroscedastic ε_i , we obtain

$$\begin{aligned} \Lambda_n(\gamma) &= n^{-1} \sum_{k=1}^n \sum_{i=1}^n K_h(Z_k - Z_i) \varepsilon_k \varepsilon_i W_i(\gamma) W_k(\gamma) / \Omega(Z_k, \gamma) - (1/2) n^{-2} \sum_{k=1}^n \sum_{i=1}^n \sum_{j=1}^n H^{(1)}\{X_k^T \beta_0 + S_k^T \eta_0 + W_k(\gamma) \theta(\gamma)\} \varepsilon_i \varepsilon_j W_i(\gamma) W_j(\gamma) \times W_k^2(\gamma) / \Omega(Z_k, \gamma) \\ &= \{T_n(\gamma) - S_n(\gamma)\} \times \{1 + o_p(1)\}. \end{aligned}$$

Also, modulo higher order terms, it follows that

$$\begin{aligned} T_n(\gamma) &= h^{-1} K(0) E\{1/f_Z(Z)\} + n^{-1} \sum_{k \neq i} K_h(Z_k - Z_i) \varepsilon_k \varepsilon_i W_k(\gamma) W_i(\gamma) / \Omega(Z_k, \gamma); \\ S_n(\gamma) &= (1/2) h^{-1} E\{1/f_Z(Z)\} \int K^2(t) dt + (2n)^{-1} \sum_{k \neq i} K_{ch}(Z_i - Z_k) \varepsilon_k \varepsilon_i W_k(\gamma) W_i(\gamma) / \Omega(Z_k, \gamma). \end{aligned}$$

Make the definition

$$\mathcal{U}_n(\gamma) = n^{-1} \sum_{k \neq i}^n \{K_h(Z_k - Z_i) - (1/2)K_{ch}(Z_k - Z_i)\} \varepsilon_k \varepsilon_i W_k(\gamma) W_i(\gamma) / \Omega(Z_k, \gamma).$$

Then we have that

$$\Lambda_n(\gamma) = \{\mu_n(h) + \mathcal{U}_n(\gamma)\} \{1 + o_p(1)\}.$$

It is obvious that these results are uniform in compact sets for γ . Since the terms $\{1 + o_p(1)\}$ are actually $\{1 + o_p(h^{-1/2})\}$, see Fan et al (2001), page 183, if $\mathcal{D}$ is such a compact set, then uniformly in $\gamma \in \mathcal{D}$, $\sup_{\gamma \in \mathcal{D}} |h^{1/2}\{\Lambda_n(\gamma) - \mu_n(h)\} - h^{1/2}\mathcal{U}_n(\gamma)| = o_p(1)$. In Appendix A.3 we show that $h^{1/2}\mathcal{U}_n(\gamma)$, as a process in γ , converges to a Gaussian process.

A.3 Weak Convergence of $\mathcal{U}_n$

To prove that the process $h^{1/2}\mathcal{U}_n(\gamma)$ converges to a Gaussian process in $\gamma \in \mathcal{D}$, we have to show that the finite dimensional distributions converge to normality, and that the process is tight.

A.3.1 Convergence of finite dimensional distributions

Make the definitions

$$\begin{aligned} a_{ii}(\gamma) &= 0; \\ a_{ij}(\gamma) &= h^{1/2} \{K_h(Z_i - Z_j) - (1/2)K_{ch}(Z_i - Z_j)\} W_i(\gamma) W_j(\gamma) / \Omega(Z_i, \gamma); \\ c_{ij}(\gamma) &= n^{-1} \{a_{ij}(\gamma) + a_{ji}(\gamma)\} \varepsilon_i \varepsilon_j, \end{aligned}$$

the latter two are defined when $i \neq j$. Then we have that

$$h^{1/2}\mathcal{U}_n(\gamma) = \sum_{1 \leq i < j \leq n} c_{ij}(\gamma),$$

where once again we note that $(\varepsilon_i, \varepsilon_j)$ are independent of γ under the null hypothesis.

We use Proposition 3.2 in de Jong (1987) to show that $h^{1/2}\mathcal{U}_n(\gamma)$ converges to a Gaussian distribution for any fixed γ . Define

$$\begin{aligned} G_I &= \sum_{1 \leq i < j \leq n} E\{c_{ij}(\gamma)^4\}; \\ G_{II} &= \sum_{1 \leq i < j < k \leq n} \left[E\{c_{ij}^2(\gamma)c_{ik}^2(\gamma)\} + E\{c_{ji}^2(\gamma)c_{jk}^2(\gamma)\} + E\{c_{ki}^2(\gamma)c_{kj}^2(\gamma)\} \right]; \\ G_{IV} &= \sum_{1 \leq i < j < k < l \leq n} \left[E\{c_{ij}(\gamma)c_{ik}(\gamma)c_{lj}(\gamma)c_{lk}(\gamma)\} + E\{c_{ij}(\gamma)c_{il}(\gamma)c_{kj}(\gamma)c_{kl}(\gamma)\} + E\{c_{ik}(\gamma)c_{il}(\gamma)c_{jk}(\gamma)c_{jl}(\gamma)\} \right]; \end{aligned}$$

To apply this proposition, we need to check the following conditions:

C1 $h^{1/2}\mathcal{U}_n(\gamma)$ is clean in the sense of de Jong (1987).

Using Definition 2.1 in de Jong (1987), we call $h^{1/2}\mathcal{U}_n(\gamma)$ is clean if the conditional expectations of c_{ij} vanish, and this is obviously true since $E(\varepsilon|X, S, Z) = 0$.

- C2** $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$ converges to a finite quantity as $n \rightarrow \infty$.
- C3** G_I is of smaller order than $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$.
- C4** G_{II} is of smaller order than $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$.
- C5** G_{IV} is of smaller order than $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$.

In what follows, we check conditions C2 – C5 as condition C1 follows directly from the fact that $E(\varepsilon|X, S, Z) = 0$.

We use the following result:

Lemma 1—Let Z_i and Z_j are independent and identically distributed random variables with a strictly positive density and compact support and let $K(\cdot)$ be a symmetric kernel. Define $K_m(c)$ to be the m -fold convolution of $K(c)$. Then

$$\begin{aligned} E[K_h^2(Z_i - Z_j)\Omega(Z_j, \gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i, \gamma)\}] &= h^{-1}K_2(0)E\{1/f(Z_i)\}\{1+O(h)\} \\ E[K_{2h}^2(Z_i - Z_j)\Omega(Z_j, \gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i, \gamma)\}] &= h^{-1}K_4(0)E\{1/f(Z_i)\}\{1+O(h)\} \\ E[K_{2h}(Z_i - Z_j)K_h(Z_i - Z_j)\Omega(Z_j, \gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i, \gamma)\}] &= h^{-1}K_3(0)E\{1/f(Z_i)\}\{1+O(h)\} \end{aligned}$$

To check condition C2, first observe that

$$\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\} = \sum_{i < j} E\{c_{ij}^2(\gamma)\}$$

Then we derive

$$\begin{aligned} E\{c_{ij}^2(\gamma)\} &= (n^{-2}/4)E[\{a_{ij}(\gamma) + a_{ji}(\gamma)\}^2 \varepsilon_i^2 \varepsilon_j^2] \\ &= (n^{-2}/4)E[\{a_{ij}^2(\gamma) + a_{ji}^2(\gamma) + 2a_{ij}(\gamma)a_{ji}(\gamma)\} \varepsilon_i^2 \varepsilon_j^2] \\ &= A_1 + A_2 + A_3. \end{aligned}$$

Now we see that, using Lemma 1,

$$\begin{aligned} A_1 &= hn^{-2}E\left([\{K_h(Z_i - Z_j) - (1/2)K_{ch}(Z_i - Z_j)\}W_i(\gamma)W_j(\gamma)/\Omega(Z_i, \gamma)]^2 \varepsilon_i^2 \varepsilon_j^2\right) \\ &= hn^{-2}E\left([\{K_h(Z_i - Z_j) - (1/2)K_{ch}(Z_i - Z_j)\}]^2 \Omega(Z_j, \gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i, \gamma)\}\right) \\ &= n^{-2}[K_2(0) - K_4(0) + (1/4)K_4(0)]E\{1/f_Z(Z)\}\{1+O(h)\}. \end{aligned}$$

Similarly,

$$\begin{aligned} A_2 &= n^{-2}[K_2(0) - K_3(0) + (1/4)K_2(0)]E\{1/f_Z(Z)\}\{1+O(h)\}; \\ A_3 &= n^{-2}[2K_2(0) - 2K_3(0) + (1/2)K_4(0)]E\{1/f_Z(Z)\}\{1+O(h)\}. \end{aligned}$$

Hence we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\} &= \lim_{n \rightarrow \infty} \sum_{i < j} E\{c_{ij}^2(\gamma)\} \\ &= [4K_2(0) - 4K_3(0) + K_4(0)]E\{1/f_Z(Z)\}, \end{aligned}$$

and hence condition C2 is satisfied.

Next, similar calculations as in Lemma 1 show that

$$\begin{aligned} E\{a_{ij}^4(\gamma)\varepsilon_i^4\varepsilon_j^4\} &= O(h^{-1}); \\ E\{a_{ij}^2(\gamma)a_{ji}^2(\gamma)\varepsilon_i^4\varepsilon_j^4\} &= O(h^{-1}), \end{aligned}$$

and it follows that $E\{c_{ij}^4(\gamma)\} = O(n^{-4}h^{-1})$. Hence we have that $G_I = O(n^{-2}h^{-1}) = o(1)$.

To check condition C4, we observe that,

$$\begin{aligned} E\{a_{ij}^2(\gamma)a_{ik}^2(\gamma)\varepsilon_i^4\varepsilon_j^2\varepsilon_k^2\} &= O(h^{-1}); \\ E\{a_{ij}(\gamma)a_{ji}(\gamma)a_{ik}(\gamma)a_{ki}(\gamma)\varepsilon_i^4\varepsilon_j^2\varepsilon_k^2\} &= O(h^{-1}), \end{aligned}$$

and similarly for other terms in the expansion of $c_{ij}^2(\gamma)c_{ik}^2(\gamma)$. It follows that $E\{c_{ij}^2(\gamma)c_{ik}^2(\gamma)\} = O(n^{-4}h^{-1})$. Hence we have that $G_{II} = O(n^{-1}h^{-1}) = o(1)$.

Finally, to check condition C5, we note that

$$E\{a_{ij}(\gamma)a_{jk}(\gamma)a_{kl}(\gamma)a_{li}(\gamma)\varepsilon_i^2\varepsilon_j^2\varepsilon_k^2\varepsilon_l^2\} = O(h),$$

and similarly for other cross product terms, and thus implying $E\{c_{ij}(\gamma)c_{jk}(\gamma)c_{kl}(\gamma)c_{li}(\gamma)\} = O(n^{-4}h)$. Hence we get $G_{IV} = O(h) = o(1)$.

We have thus shown that conditions C1–C5 are satisfied and hence the proof is complete.

A.3.2 Tightness

We have to show that there exists $\zeta > 0$, $\eta > 1$, such that, for any $\gamma_1 < \gamma < \gamma_2$,

$$h^\zeta E\{|\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)|^\zeta |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)|^\zeta\} \leq |\gamma_1 - \gamma_2|^\eta, \quad (\text{A.5})$$

see Billingsley (1968, page 128). We show below that (A.5) holds for $\zeta = 1$.

Using the Cauchy-Schwarz inequality we observe

$$h^2 E^2 \{ |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)| |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)| \} \leq E \{ h^{1/2} |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)| \}^2 E \{ h^{1/2} |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)| \}^2.$$

Recall that $h^{1/2} \mathcal{U}_n(\gamma) = \sum_{1 \leq i < j \leq n} c_{ij}(\gamma)$. Let $c_{ij}^{(1)}(\gamma)$ denote the first derivative of $c_{ij}(\gamma)$ with respect to γ and similarly for $a_{ij}(\gamma)$. Now we see that

$$\begin{aligned} E \{ h^{1/2} |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)| \}^2 &= E \left[\sum_{1 \leq i < j \leq n} \{ c_{ij}(\gamma) - c_{ij}(\gamma_1) \} \right]^2 \\ &= E \left[\sum_{1 \leq i < j \leq n} \{ c_{ij}(\gamma) - c_{ij}(\gamma_1) \}^2 \right] \\ &\leq (\gamma_1 - \gamma_2)^2 C, \end{aligned}$$

where $C = n^2 \sup_{\gamma, \gamma_1 \in [L, R]} E \{ \{ c_{ij}(\gamma) - c_{ij}(\gamma_1) \}^2 \}$. Similar calculations can be done for $E \{ h^{1/2} |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)| \}^2$, and hence $h E \{ |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)| |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)| \} \leq |\gamma_1 - \gamma_2|^2 C$. The only thing remaining is to show that C is finite, which follows immediately from condition C2.

A.3.3 Proof of Lemma 1

Note that $K(s) = K(-s)$ and since Z has a positive density function on a compact support, we have $E \{ 1/f_Z(Z) \} = \int_{\mathcal{Z}} dz$. Then

$$\begin{aligned} E \{ K_h^2(Z_i - Z_j) \Omega(Z_j, \gamma) / \{ f_Z(Z_j) f_Z(Z_i) \Omega(Z_i, \gamma) \} \} &= h^{-2} \int K^2 \{ (z_j - z_i)/h \} \Omega(z_j, \gamma) / \{ \Omega(z_i, \gamma) \} dz_i dz_j \\ &= h^{-1} \int K^2(t) \Omega(z_i + th, \gamma) / \{ \Omega(z_i, \gamma) \} dz_i dt \\ &= h^{-1} \int K^2(t) \{ \Omega(z_i, \gamma) + O(h) \} / \{ \Omega(z_i, \gamma) \} dZ_i dt \\ &= h^{-1} \int K^2(t) dz_i dt \{ 1 + O(h) \} \\ &= h^{-1} \int K^2(t) dt E \{ 1/f(Z_i) \} \{ 1 + O(h) \} \\ &= h^{-1} K_2(0) E \{ 1/f(Z_i) \} \{ 1 + O(h) \}. \end{aligned}$$

Also,

$$\begin{aligned} E \{ K_{2h}^2(Z_i - Z_j) \Omega(Z_j, \gamma) / \{ f_Z(Z_j) f_Z(Z_i) \Omega(Z_i, \gamma) \} \} &= h^{-2} \int K_2^2 \{ (z_j - z_i)/h \} \Omega(z_j, \gamma) / \{ \Omega(z_i, \gamma) \} dz_i dz_j \\ &= h^{-1} \int K_2^2(t) \Omega(z_i + th, \gamma) / \{ \Omega(z_i, \gamma) \} dz_i dt \\ &= h^{-1} \int K_2^2(t) \{ \Omega(z_i, \gamma) + O(h) \} / \{ \Omega(z_i, \gamma) \} dz_i dt \\ &= h^{-1} \int K_2^2(t) dz_i dt \{ 1 + O(h) \} \\ &= h^{-1} \int K_2^2(t) dt E \{ 1/f(Z_i) \} \{ 1 + O(h) \} \\ &= h^{-1} K_4(0) E \{ 1/f(Z_i) \} \{ 1 + O(h) \}. \end{aligned}$$

Similarly, we derive

$$\begin{aligned}
E\{K_{2h}(Z_i - Z_j)K_h(Z_i - Z_j)\Omega(Z_j, \gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i, \gamma)\}\} &= h^{-2} \int K_2\{(z_j - z_i)/h\}K\{(z_j - z_i)/h\}\Omega(z_j, \gamma)/\{\Omega(z_i, \gamma)\} dz_i dz_j \\
&= h^{-1} \int K_2(t)K(t)\Omega(z_i + th, \gamma)/\{\Omega(z_i, \gamma)\} dz_i dt \\
&= h^{-1} \int K_2(t)K(t)\{\Omega(z_i, \gamma) + O(h)\}/\{\Omega(z_i, \gamma)\} dz_i dt \\
&= h^{-1} \int K_2(t)K(t) dz_i dt \{1 + O(h)\} \\
&= h^{-1} \int K_2(t)K(t) dt E\{1/f(Z_i)\}\{1 + O(h)\} \\
&= h^{-1} K_3(0) E\{1/f(Z_i)\}\{1 + O(h)\},
\end{aligned}$$

completing the proof.

Figure 1.

Results on power and level in the simulation for testing whether $\theta(\cdot)$ is constant, as described in Section 3, using the kernel method. Solid line is our method, while the dashed line is the usual test. The left panel is for $\gamma = 0$, the middle panel is for $\gamma = 1$ and the right panel is for $\gamma = 2$. The function used was $\theta(z) = c \times \sin(2z)$ on $z \in [0, 2]$; the horizontal axis plots the value of c : $c = 0$ is the null model, and $c > 0$ gives power.

Figure 2. Results from data analysis. Presented are (a) the estimated effect of the environmental covariate PhIP for $\gamma = 0$ (solid line), 1 (dashed line) and 2 (dot-dashed line), and (b) – (d) plots of estimates of the total effect $X^T\beta + S^T\eta + \theta(Z) + \gamma X^T\beta\theta(Z)$ (z-axis) versus the estimated effect of the environmental covariate PhIP $\theta(Z)$ (x-axis) and the estimated effect of the 3 most common diplotypes $X^T\beta$ (y-axis) when $\gamma = 0, 1$ and 2, respectively.

Table 1

Significance levels in the NAT2 example of Section 4. Displayed are the significance levels for testing whether the environmental variables Z listed have no impact on the development of colorectal adenoma, when X consists for the 1, 2 or 3 most common NAT2 diplotypes compared to the rest. PhIP (CIG STOP) denotes the results for PhIP with CIG STOP being included in the model as a linear covariate.

Environment	Number of Diplotypes	Our Method	Fixing $\gamma = 0$
CIG STOP	1	0.000	0.000
	2	0.000	0.000
	3	0.000	0.001
Red Meat	1	0.464	0.639
	2	0.381	0.595
	3	0.470	0.623
PhIP	1	0.984	0.935
	2	0.227	0.939
	3	0.162	0.938
PhIP (CIG STOP)	1	0.922	0.947
	2	0.933	0.937
	3	0.931	0.936