Semiparametric nonlinear regression for detecting gene and environment interactions

Shujie Ma\textsuperscript{a,}\textsuperscript{*}, Shizhong Xu\textsuperscript{b}

\textsuperscript{a} Department of Statistics, University of California at Riverside, Riverside, CA 92521, United States
\textsuperscript{b} Center for Plant Cell Biology, University of California at Riverside, Riverside, CA 92521, United States

\begin{abstract}
It is commonly accepted that gene and environment (G\times E) interactions play a pivotal role in determining the risk of human diseases. In conventional parametric models such as linear models and generalized linear models which are applied frequently to study statistical interactions, effects of covariates are decomposed into main effects and interaction effects (products of two components). Such decomposition, however, may not reflect the true interaction effect of gene and environment. In this paper, we propose a semiparametric regression approach to capture possible nonlinear G\times E interactions. A profile quasi-log-likelihood estimation method is applied with asymptotic consistency and normality established for the profile estimators. Moreover, we develop Rao-score-type test procedures based on the profile estimation for regression parameters and nonparametric coefficient functions, respectively. Our models and methods are illustrated by both simulation studies and analysis of a dataset application.
\end{abstract}

\section{Introduction}

It is commonly accepted that most human diseases result from a complex interaction between genetic and environmental factors, such as obesity (Hebebrand and Hinney, 2009), psychiatric disorders (Tsuang et al., 2004; Caspi and Moffitt, 2006), heart disease (Talmud, 2007), diabetes (Grarup and Andersen, 2007) and cancer (Song et al., 2011). By learning how genetic and environmental factors jointly influence the risk of developing a human disease, it would help scientists to develop new methods for prevention and treatment of illnesses. Despite the enthusiasm for investigation of gene–environment (G\times E) interactions, published works on studying these interactions via statistical modeling are limited. In the literature, linear models as well as generalized linear models such as logistic and log-linear models are used frequently to study statistical interactions. In the conventional parametric models, the effects of covariates are decomposed into main effects and interaction effects (products of two components). Such decomposition, however, may not reflect the true nonlinear interaction between gene and environment. As a result, mis-specification in parametric models could lead to a large estimation bias. To overcome this limitation, different non- and semi-parametric modeling methods have been recently applied to study G\times E interactions. For example, Chatterjee and Carroll (2005) and Chen et al. (2012) studied semiparametric maximum likelihood estimates of logistic regression parameters in case-control studies. Maity et al. (2009) developed a score test for parametric main effects of genetic factors in a semiparametric model with Tukey’s form of interaction, and Wei et al. (2011) derived a generalized likelihood ratio test for nonparametric effects. Lobach et al. (2011) and Ahn et al. (2013) developed a score test for parametric main effects of genetic factors in a semiparametric model with Tukey’s form of interaction, and Wei et al. (2011) derived a generalized likelihood ratio test for nonparametric effects. Lobach et al. (2011) and Ahn et al. (2013) developed a score test for parametric main effects of genetic factors in a semiparametric model with Tukey’s form of interaction, and Wei et al. (2011) derived a generalized likelihood ratio test for nonparametric effects.

\* Corresponding author.

\textsuperscript{*} Corresponding author.

\textsuperscript{*} E-mail addresses: shujie.ma@ucr.edu (S. Ma), shizhong.xu@ucr.edu (S. Xu).

\textsuperscript{*} Corresponding author.

\textsuperscript{*} E-mail addresses: shujie.ma@ucr.edu (S. Ma), shizhong.xu@ucr.edu (S. Xu).
proposed semiparametric Bayesian analysis of $G \times E$ interaction. Ma et al. (2011) applied a varying-coefficient model for $G \times E$ interaction.

In this paper, we apply a generalized partially linear single-index coefficient model (GPLSiCM) to study nonlinear $G \times E$ interactions. Let $Z = (Z_1, Z_2, \ldots, Z_p)^T = (1, Z_2, \ldots, Z_p)^T$ be the $p$-dimensional vector, where $(Z_2, \ldots, Z_p)^T$ are the genetic factors. A parametric model for studying genetic effects on phenotype is given as $E(Y|Z) = \mu(Z) = g^{-1}(\eta)$ with $\eta = \sum_{\ell=1}^{p} \beta_{\ell} Z_{\ell}$, where $Y$ is the response variable which can be continuous or discrete such as binary variables or counts, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the regression coefficient vector, and $g$ is a known monotone link function. By considering effects of genetic factors interacting with environmental factors, we may employ the parametric model with both main and interaction effects given as

$$\eta = \beta_1 + \sum_{k=1}^{d_1} \beta_{k1} X_k + \sum_{\ell=2}^{p} \beta_{\ell} Z_{\ell} + \sum_{\ell=2}^{p} \sum_{k=1}^{d} \beta_{\ell k} X_k Z_{\ell},$$

where $(\beta_1, \beta_{k1}, \beta_{\ell}, \beta_{\ell k})$ are the regression coefficients and $X = (X_1, \ldots, X_{d_1})^T$ is the $d_1$-dimensional vector of the environmental factors. By simple calculation, (1) can be written as

$$\eta = \sum_{\ell=1}^{p} \left( \beta_{\ell} + \sum_{k=1}^{d_1} \beta_{k\ell} X_k \right) Z_{\ell}.$$  

Thus the coefficient for the $\ell$th genetic factor, which is $\beta_{\ell} + \sum_{k=1}^{d_1} \beta_{k\ell} X_k$, is indeed a linear function of $(X_1, \ldots, X_{d_1})$.

The linearity assumption can be easily violated due to the underlying nonlinear mechanism of the relationship between the response and explanatory variables. For example, research on causes of obesity has brought tremendous attention due to its high prevalence and the associated medical and psychosocial risks. It is known that obesity is related to not only genetic factors but also some environmental factors such as sleeping hours (Knutson, 2012) and physical activity (Wareham et al., 2005). Thus, people may wonder how genetic and environmental factors together influence people’s weight. Using data from the Framingham Heart Study (Dawber et al., 1951), we let $X_1 =$ hours of light activity, $X_2 =$ hours of moderate activity, $X_3 =$ hours of moderate activity, be the environmental factors. The body mass index (BMI) is used as the response variable. After deleting missing data, 299 subjects remain in our study. To illustrate possible $G \times E$ interactions, we divide people in the study into three groups based on the three genotypes of SNP ss66155100. For each group, a linear model $E(Y|X) = \beta^T X$ can be fitted. However, in order to study possible nonlinear relationship between $Y$ and $X$, we let $E(Y|X)$ be a nonlinear function of $\beta^T X$, such that $E(Y|X) = m(\beta^T X)$, where $m$ is an unknown but smooth function, and $U = \beta^T X$ is the index. This model is the semiparametric single-index coefficient model, estimation of which has been substantially studied, see Carroll et al. (1997) for the backfitting method, Liang et al. (2010) for the profile method and Xia and Härdle (2006) for the minimum average variance estimation (MAVE) method. Fig. 1 shows the plots of the estimates of $m(\cdot)$ against the index value by using the np package in R for the three groups with genotypes aa (thick line), Aa (thin line) and AA (dashed line). Here A is the minor allele. We can clearly observe different nonlinear change patterns of the estimated BMI mean functions among the three groups. For example, for people with genotype aa, their BMI shows a decreasing pattern as index value increases. However, for people
with genotypes $Aa$ and $aa$, the estimated BMI mean function has a roughly bell-shaped curve. The different estimated BMI curves provide us a visual evidence that there exist strong nonlinear $G \times E$ interactions.

Motivated by this example, by letting the coefficient for each $Z_\ell$ be a nonlinear function $m_\ell(\cdot)$ of a linear combination of $(X_1, \ldots, X_{d_\ell})$, the following semiparametric model can be considered

$$E(Y|Z, X) = \mu(Z, X) = g^{-1}(\eta), \quad \text{with} \quad \eta = \sum_{\ell=1}^{p} m_\ell(X^T \beta)Z_\ell, \quad (2)$$

where $\beta = (\beta_1, \ldots, \beta_{d_\ell})^T$ are coefficient parameters and $m_\ell(\cdot)$ is an unknown smooth nonparametric function. We impose no specific functional form for each $m_\ell(\cdot)$, so that model (2) can capture dynamic change patterns of the coefficient functions in a flexible way. It has been studied that in addition to genetic factors, sleeping hours and physical activities, BMI may be related to other factors such as health condition, blood pressure, etc. These factors can be either continuous or discrete. Assuming that these factors do not interact with genetic factors, we model them as a linear part. Let $T = (T_1, \ldots, T_{d_\ell})^T$ be the $d_\ell$-dimensional vector of covariates. Then we propose the class of generalized partially linear single-index coefficient models (GPLSiCM) given as

$$E(Y|Z, X, T) = \mu(Z, X, T) = g^{-1}(\eta), \quad \text{with} \quad \eta = \sum_{\ell=1}^{p} m_\ell(X^T \beta)Z_\ell + T^T \alpha, \quad (3)$$

where $\alpha = (\alpha_1, \ldots, \alpha_{d_\ell})^T$ are regression coefficients for $T$. The conditional variance of $Y$ is typically a function of the mean, i.e., $\text{var}(Y|Z, X) = V(\mu(Z, X, T)) = \sigma^2(Y|Z, X, T)$.

Model (3) is very flexible which provides a way of unifying various other statistical models. Some examples include: (1) when $p = 1$ and $Z_1 = 1$, it is a generalized partially linear single-index model (Carroll et al., 1997); (2) when $p = 1, Z_1 = 1$ and $d_1 = 1$, it is a generalized partially linear model (Hardle et al., 2000); (3) when $d_1 = 1$ and $\alpha = 0$, it is a generalized varying coefficient model (Cai et al., 2000); and (4) without predictors $T$, when the link function $g^{-1}(\cdot)$ is the identity function, it is a single-index varying coefficient model (SiVCM) studied in Xue and Wang (2012) and Xue and Pang (2013), in which the nonparametric functions are estimated by local linear fitting and the index parameters are estimated by maximizing empirical likelihood and estimating equations, respectively.

Clearly, model (3) is not identifiable without any constraints on the parameters. For example, by letting $\tilde{\alpha} = a\beta$ and $\tilde{m}_\ell = a^{-1}m_\ell$, where $a$ is a nonzero constant, we have $\tilde{m}_\ell(X^T \tilde{\beta}) = m_\ell(X^T \beta)$ for each $\ell = 1, \ldots, p$. In the following, we discuss the model identifiability for two different cases. When $X$ and $T$ are two different sets of covariates, to ensure identifiability, we let $\|\beta\|_2 = 1$ and $\beta_1 > 0$, where $\|a\|_2 = (a_1^2 + \cdots + a_d^2)^{1/2}$ denotes the $l_2$ norm for any vector $a = (a_1, \ldots, a_d)^T$. When $X$ and $T$ are the same set of covariates such that $X = T$ and $d_1 = d_2 = d, \alpha \in \mathbb{R}^d$ is assumed to be perpendicular to $\beta$.

Otherwise by taking $\tilde{\alpha} = \alpha - (\alpha^T \beta)\beta$, we would have

$$\eta = m_1(\beta^T X) + \sum_{\ell=2}^{p} m_\ell(\beta^T X)Z_\ell + \tilde{\alpha}^T X$$

$$= \tilde{m}_1(\beta^T X) + \sum_{\ell=2}^{p} m_\ell(\beta^T X)Z_\ell + \tilde{\alpha}^T X.$$

where $\tilde{m}_1(\beta^T X) = m_1(\beta^T X) + (\alpha^T \beta)\beta^T X$. In this paper, we consider the first case that $X$ and $T$ are different sets of covariates, so that model (3) is identifiable when $\|\beta\|_2 = 1$ and $\beta_1 > 0$.

For estimation, we first approximate the nonparametric functions $m_\ell(\cdot)$ by B-spline functions and then estimate parameters $\beta$ and $\alpha$ by a profile method. The spline approximation is known as computationally more efficient than the kernel-based method. More importantly, the proposed spline and profile estimation procedure enables us to develop score tests for inferences on both of the coefficient functions $m_\ell(\cdot)$ and the parameters $\beta$ and $\alpha$, so that we can identify important genetic factors by testing $m_\ell(\cdot) = 0$ in the context of nonlinear $G \times E$ interactions. The proposed testing procedure is easy to implement and fast to compute with the $p$-values or critical values obtained from the asymptotic distributions of the test statistics, and it provides a useful inferential tool to identify genetic risk factors in a more flexible model other than parametric linear models in a dataset with a large number of genes by conducting the test for each gene combined with multiple testing procedures. The proposed estimation procedure in model (3) is briefly described as follows. For given $\beta$ and $\alpha$, we approximate the coefficient functions $m_\ell(\cdot)$ by B-spline functions (de Boor, 2001) with coefficients obtained by maximizing a quasi-log-likelihood function. Thus the resulting estimators of $m_\ell(\cdot)$ are functions of $\beta$ and $\alpha$. Estimation of the parameters $\beta$ and $\alpha$ is achieved by replacing the true functions $m_\ell(\cdot)$ with their spline estimators in the objective function. Asymptotic normalities and consistency for estimators of both the parameters and nonparametric function are therefore established.

The rest of this paper is organized as follows. Section 2 introduces the profile quasi-log-likelihood estimation and presents asymptotic properties of the proposed estimators. Section 3 proposes score tests for regression parameters and nonparametric coefficient functions. In Section 4, we evaluate finite sample performance of the proposed estimation and inference procedures via simulation studies. Section 5 illustrates the proposed model and method through the analysis of
two data applications. A concluding discussion is given in Section 6. All technical proofs are provided in the Appendix A which can be found from the on-line Supplemental Materials.

2. Estimation method

Let \((Z_i^T, X_i^T, T_i, Y_i : 1 \leq i \leq n)\) be a random sample from the same distribution as \((Z^T, X^T, T^T, Y)\), where \(Z_i = (Z_{i1}, \ldots, Z_{ip})^T, X_i = (X_{i1}, \ldots, X_{id})^T\) and \(T_i = (T_{i1}, \ldots, T_{id})^T\). In practice, the full likelihood is often unavailable. Therefore, we estimate the conditional mean function \(\mu(Z, X, T)\) by replacing the conditional log-likelihood with a quasi-log-likelihood function \(Q(\mu(Z, X, T), y)\) (McCullagh, 1983). By assuming \(\varphi(Y|Z, X, T) = V[\mu(Z, X, T)]\) for some known positive variance function \(V(\cdot)\), the corresponding \(Q(\mu, y)\) satisfies

\[
\partial Q(\xi, y)/\partial \xi = (y - \xi)/|V(\xi)|.
\]

More explicitly, \(Q(\mu, y) = \int_{\mathbb{R}^d} (y - \xi)/|V(\xi)| d\xi\). Estimation of the mean function can be achieved by maximizing the quasi-log-likelihood of the observed data

\[
\sum_{i=1}^{n} Q\{g^{-1}(\eta(Z_i, X_i, T_i)), Y_i\}.
\]

We propose a quasi-log-likelihood profile estimation procedure. The unknown functions \(m_{\ell}(\cdot)\) are estimated by B-spline functions. Denote \(U(\beta) = X^T \beta\). We assume that \(U(\beta)\) is distributed on a compact interval \([a, b]\), for \(\beta\) in the neighborhood of its true parameter value, so that the range of the B-splines can be well defined. Let \(a = \xi_0, \xi_1 < \xi_2, \ldots, \xi_{N_1} < b = \xi_{N_2 + 1}\) be a partition of \([a, b]\) into \((N + 1)\) subintervals \(I_{\ell,i} = [\xi_{\ell,i}, \xi_{\ell,i+1}]\), for \(j = 0, \ldots, N_1 - 1\) and \(N_1,\xi_{\ell,1} = [\xi_{N_1}, 1]ESM, satisfying

\[
\max_{0 \leq \ell \leq n_{\ell,n}} \left| \xi_{\ell,i+1} - \xi_{\ell,i} \right| / \min_{0 \leq \ell \leq n_{\ell,n}} \left| \xi_{\ell,i+1} - \xi_{\ell,i} \right| \leq M
\]

uniformly in \(n\) for some constant \(0 < M < \infty\), where \(n_{\ell,n}\) increases with the number of subjects \(n\). Furthermore, define the \(qth\) order normalized B-spline basis as \(B_{\ell}(u) = \{B_{\ell,j}(u): 1 \leq j \leq n_{\ell,n}\}\) (de Boor, 2001), where \(j_{n,\ell} = N_\ell + q\). Let \(H_{\ell,n} = H_{n,\ell}^{(q-2)}\) be the space spanned by \(B_{\ell}(u)\). Thus, there exists \(\gamma_{\ell} = (\gamma_{\ell,1}, \ldots, \gamma_{\ell,n_{\ell,n}})^T\) such that \(m_{\ell}(u) = \sum_{j=1}^{n_{\ell,n}} B_{\ell,j}(u) \gamma_{\ell,j}(u)\in H_{\ell,n}\), and under suitable smoothness assumptions, \(m_{\ell}(u)\) can be well approximated by \(m_{\ell}(u) = \sum_{j=1}^{n_{\ell,n}} B_{\ell,j}(u) \gamma_{\ell,j}(u)\in H_{\ell,n}\). Let \(\gamma(Z, X, T)\) be approximated by \(\hat{\gamma}(Z, X, T, \beta, \alpha) = \sum_{\ell=1}^{p} B_{\ell}(X^T \beta) \gamma_{\ell}T_{\ell}\alpha\). Denote \(q_0(\eta, y) = \partial Q(g^{-1}(\eta), y)/\partial \eta^2\), so that

\[
q_1(\eta, y) = \frac{\partial}{\partial \eta} Q(g^{-1}(\eta), y) = (y - g^{-1}(\eta)) \rho_1(\eta),
\]

\[
q_2(\eta, y) = \frac{\partial^2}{\partial \eta^2} Q(g^{-1}(\eta), y) = (y - g^{-1}(\eta)) \rho_2(\eta) - \rho_2(\eta),
\]

where \(\rho_2(\eta) = \left\{dg^{-1}(\eta)\right\}^T / V\left\{g^{-1}(\eta)\right\}\).

Profile estimation of parameters \(\beta and \alpha\) and coefficient functions \(m_{\ell}(\cdot)\). Let \(\gamma = (\gamma_1^T, \ldots, \gamma_p^T)^T\). For given \(\beta\) and \(\alpha\), we consider the quasi-log-likelihood:

\[
L_\alpha(\gamma, \beta, \alpha) = \sum_{i=1}^{n} Q\left\{g^{-1}\left\{\sum_{\ell=1}^{p} B_{\ell}(X_i^T \beta) \gamma_{\ell}T_{\ell}\alpha\right\}, Y_i\right\}.
\]

The estimator \(\hat{\gamma}(\beta, \alpha) = \{\hat{\gamma}_1(\beta, \alpha)^T, \ldots, \hat{\gamma}_p(\beta, \alpha)^T\}^T\) is obtained by maximizing (5). To proceed estimation of \(\beta\) and \(\alpha\), by the assumption that \(\|\beta\|_2 = 1\) and \(\beta_1 > 0\) for identifiability of model (3), we exclude the first component \(\beta_1\) of \(\beta\) by letting \(\beta_1 = \sqrt{1 - \|\beta_{-1}\|_2^2}\), where \(\beta_{-1} = (\beta_2, \ldots, \beta_p)^T\), and reformulate the parameter space of \(\beta\) as follows:

\[
\left\{\left(\sqrt{1 - \|\beta_{-1}\|_2^2}, \beta_2, \ldots, \beta_p\right)^T : \|\beta_{-1}\|_2^2 < 1\right\}.
\]

By replacing \(\gamma_{\ell}\) with \(\hat{\gamma}_{\ell}(\beta, \alpha)\), for estimating \(\beta_{-1}\) and \(\alpha\), we next define

\[
L_\alpha(\beta, \alpha) = \sum_{i=1}^{n} Q\left\{g^{-1}\left\{\sum_{\ell=1}^{p} B_{\ell}(X_i^T \beta) \hat{\gamma}_{\ell}(\beta, \alpha)T_{\ell}\alpha\right\}, Y_i\right\}.
\]

Please cite this article in press as: Ma, S., Xu, S., Semiparametric nonlinear regression for detecting gene and environment interactions. J. Statist. Plann. Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.08.005
The estimators of $\theta_{-1} = (\beta_{-1}^T, \alpha_1^T)^T$ are given as $\hat{\theta}_{-1} = (\hat{\beta}_{-1}, \hat{\alpha})^T = \arg \max_{(\beta_{-1}, \alpha)} \{ L_n^*(\beta, \alpha) \}$. Moreover, one can obtain $\hat{\theta}_{-1}$ as the solution of the estimation equations:

$$
\frac{\partial L_n^*(\beta, \alpha)}{\partial \theta_{-1}} = \sum_{i=1}^n q_i \left[ g^{-1} \left\{ \sum_{t=1}^p B_{t, i} (X_i^T \beta)^T \tilde{Y}_t (\beta, \alpha) Z_{t, i} + T_i^T \alpha \right\}, Y_i \right] 
\times \left[ \sum_{t=1}^p \tilde{m}_t^*(X_i^T \beta, \beta, \alpha) Z_{t, i} (\beta)^T X_i, T_i \right] + \{ \partial \tilde{Y}(\beta, \alpha)^T / \partial \theta_{-1} \} R_i(\beta) 
= 0,
$$

(7)

where $\tilde{m}_t^*(u, \beta, \alpha) = B_t^* (u)^T \tilde{Y}_t (\beta, \alpha)$, in which $B_t^* (u) = \left\{ B_{t, j, i}^*(u) : 1 \leq j \leq j_{n,t} \right\}^T$ and $B_{t, j, i}^*(u)$ is the first order derivative of the B-spline function $B_{t, j, i}^*(u)$.

Then $\beta_1$ is estimated by $\hat{\beta}_1 = \sqrt{1 - \| \hat{\beta}_{-1} \|^2}$. Let $\hat{\theta} = (\hat{\beta}^T, \hat{\alpha})^T$ and $\theta = (\beta^T, \alpha^T)^T$. Let $J(\beta) = \partial \beta / \partial \beta_{-1}$ be the Jacobian matrix of size $p \times (p - 1)$, which is $J(\beta) = \left( -\beta_{-1}/\sqrt{1 - \| \beta_{-1} \|^2} \right)_{i=1}^p$. Denote $J = J(\beta^0)$. Now define the space $\mathcal{M}$ as a collection of functions with finite $L_2$ norm on $[0, 1] \times R^p$ by

$$
\mathcal{M} = \left\{ \omega (u, \mathbf{z}) = \sum_{t=1}^p \omega_t (u) z_t, E \omega_t (U)^2 < \infty \right\},
$$

where $\mathbf{z} = (z_1, \ldots, z_p)^T$. For any vector $\xi = (\xi_1, \ldots, \xi_p)^T$, let $\text{Proj}_\mathcal{M} (\xi_k)$ be the projection of $\xi_k$ onto the space $\mathcal{M}$ with respect to the theoretical weighted $L_2$ norm such that

$$
\text{Proj}_\mathcal{M} (\xi_k) = \arg \min_{\omega \in \mathcal{M}} \mathbb{E} \left[ \left\{ \xi_k - \omega (U (\beta^0), \mathbf{Z}) \right\}^2 \right] \left\{ dg^{-1}(t) \right\}^2 / \sigma (X, \mathbf{Z}, T)^2 .
$$

Let $\text{Proj}_\mathcal{M} (\xi) = \left\{ \text{Proj}_\mathcal{M} (\xi_k) : 1 \leq k \leq d \right\}^T$. Denote $\mathbf{\tilde{X}}_t = \mathbf{X}_t - \text{Proj}_\mathcal{M} (\mathbf{X}_t)$, $\mathbf{\tilde{T}}_t = \mathbf{T}_t - \text{Proj}_\mathcal{M} (\mathbf{T}_t)$, and $\text{var}(Y_t | \mathbf{Z}_t, \mathbf{X}_t, \mathbf{T}_t) = \alpha_2^2 (\mathbf{Z}_t, \mathbf{X}_t, \mathbf{T}_t) = \sigma_2^2$. The asymptotic covariance of the parameter estimators given in (8) of the following theorem contains $\mathbf{\tilde{X}}_t$ and $\mathbf{\tilde{T}}_t$. In practice, we approximate $\text{Proj}_\mathcal{M} (\mathbf{X}_t)$ and $\text{Proj}_\mathcal{M} (\mathbf{T}_t)$ by their spline estimates $\text{Proj}_{\mathcal{M}, a} (\mathbf{X}_t)$ and $\text{Proj}_{\mathcal{M}, a} (\mathbf{T}_t)$ with their explicit forms given in (A.9) of the Appendix A.

For any positive numbers $a_n$ and $b_n$, let $a_n \ll b_n$ denote that $a_n/b_n = o (1)$. Let $r$ with $r > 3/2$ be the smoothness order of the coefficient functions $m_t (u)$ as given in Condition (C2) in the Appendix A. Let $m_t^* (u)$ be the first order derivative of $m_t (u)$. Let $J_{\max} = \max \left\{ J_{n,t}, J_{\min} \right\}$ and $J_{\min} = \min \left\{ J_{n,t} \right\}$.

**Theorem 1.** Under Conditions (C1)–(C4) in the Appendix A, and $J_{\max}/J_{\min} \asymp 1$, max\{ $n^{1/(2r+2)}$, $n^{1/(4r-2)}$ \} $\ll J_{\max} \ll n^{1/4}$, we have (i) (consistency) $\sqrt{n} (\hat{\theta}_{-1} - \theta_{-1}) \rightarrow O_p (n^{-1/2})$; and (ii) (asymptotic normality)

$$
\sqrt{n} (\hat{\theta}_{-1} - \theta_{-1}) = \left[ n^{-1} \sum_{i=1}^n \left\{ \sum_{t=1}^p m_t^*(X_i^T \beta^0) Z_{t, i} \mathbf{\tilde{X}}_t^T \mathbf{\tilde{X}}_t^T \right\} \right]^{-1} \left\{ dg^{-1}(t) \right\}^2 / \sigma_2^2 
\times n^{-1/2} \sum_{i=1}^n (Y_i - g^{-1}(t_i) \left\{ \sum_{t=1}^p m_t^*(X_i^T \beta^0) Z_{t, i} \mathbf{\tilde{X}}_t^T \mathbf{\tilde{X}}_t^T \right\} \right. 
\left. \left\{ dg^{-1}(t_i) \right\}^2 / \sigma_2^2 + o_p (n^{-1/2}) \right).
$$

Moreover $\sqrt{n} (\hat{\theta}_{-1} - \theta^0) \darrow \mathcal{N} (0, \Sigma^{-1})$, as $n \rightarrow \infty$, where

$$
\Sigma = \mathbb{E} \left[ \left\{ \sum_{t=1}^p m_t^*(X_i^T \beta^0) Z_{t, i} \mathbf{\tilde{X}}_t^T \mathbf{\tilde{X}}_t^T \right\} \right] \left\{ \sigma_2^2 \right\}^2 / \sigma_1^2 .
$$

(8)

After we obtain the estimator $\hat{\theta}$, the final estimators of the spline coefficients are given as

$$
\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T = \arg \max_{\beta} \left\{ L_n (\mathbf{y}, \hat{\beta}, \alpha) \right\} .
$$

Please cite this article in press as: Ma, S., Xu, S., Semiparametric nonlinear regression for detecting gene and environment interactions. J. Statist. Plann. Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.08.005
and the $\ell$th coefficient function $m_\ell (u)$ is estimated by $\hat{m}_\ell (u) = \hat{m}_\ell \left( u, \hat{\beta}, \hat{\alpha} \right) = \textbf{B}_\ell (u)^T \hat{\gamma}_\ell = \sum_{j=1}^{j_{n,\ell}} B_{\ell,j}(u) \hat{\gamma}_{j\ell}$. Denote 
\[
R_{j\ell,i} (\beta_0) = B_{\ell,j\ell} (X_i^T \beta_0) Z_{ij}, 
\text{ and } \ 
R_{i} (\beta_0) = \left\{ R_{j\ell,i} (\beta_0)^T : 1 \leq \ell \leq p \right\}^T,
\]
where $R_{j\ell,i} (\beta_0) = \left\{ R_{j\ell,i} (\beta_0)^T : 1 \leq j \leq j_{n,\ell} \right\}$ and 
\[
\mathbb{I} (u) = \begin{bmatrix} B_{1,1}(u) & \cdots & B_{1,j_{n,1}}(u) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{p,1}(u) & \cdots & B_{p,j_{n,p}}(u) & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}_{p \times j_{n,1}}.
\]

**Theorem 2.** Under Conditions (C1)–(C4) in the Appendix A, and $j_{n,\ell}/j_{n,1} \rightarrow 1$, 
$\max \left\{ n^{1/(2r+2)}, n^{1/(4r-2)} \right\} \equiv j_{n,\ell} \equiv n^{1/4}$, we have (i) (consistency) 
$|\hat{m}_\ell (u) - m_\ell (u)| = O_p \left( \sqrt{j_{n,\ell}/n} + j_{n,\ell}^{-1} \right)$ uniformly in $u \in [0, 1]$; 
and (ii) under $n^{1/(2r+1)} \equiv j_{n,\ell} \equiv n^{1/4}$, as $n \rightarrow \infty$, (asymptotic normality) for $1 \leq \ell \leq p$,
\[
\sigma_{n,\ell}^{-1} (u) (\hat{m}_\ell (u) - m_\ell (u)) \rightarrow N (0, 1),
\]
where 
\[
\sigma_{n,\ell}^2 (u) = e_i^T \mathbb{I} (u) \sum_{i=1}^{n} R_i (\beta_0) R_i (\beta_0)^T \left\{ \frac{dg^{-1} (\eta_i)}{\gamma_2} \right\}^2 / \sigma_2^2 \mathbb{I} (u)^T e_i,
\]
and $e_i$ is the $p$-dimensional vector with “1” as its $i$th element and “0” as other elements.

**Remark.** The same convergence rate in (i) of Theorem 2 is established in Zhou et al. (1998) for spline estimation in univariate nonparametric regression. Under the order assumptions for $j_{n,\ell}$ in Theorem 1 and (i) of Theorem 2, if we let $N_\ell = j_{n,\ell} - q \asymp n^{1/(2r+1)}$, which is the optimal order for the number of interior knots, then the convergence rate for $|\hat{m}_\ell (u) - m_\ell (u)| = O_p \left( n^{-r/(2r+1)} \right)$.

3. Hypothesis tests

Applying the profile estimation method described in Section 2, we propose Rao-score-type tests for nonparametric coefficient functions and regression parameters.

1. Testing the nonparametric component. To test whether there are interaction effects between $X$ and some $Z_i$s, without loss of generality, we set up the null and alternative hypotheses as $H_0$: $m_\ell (\cdot) = c_\ell$ for $1 \leq \ell \leq p$ versus $H_1$: 
$m_\ell (\cdot) \neq c_\ell$ for some $\ell \in \{ p_1 + 1, \ldots, p \}$. We perform a transformation of the B-spline basis by letting be $B_{\ell,1}(u) = \{ 1, B_{\ell,\ell}(u) : 2 \leq j \leq j_{n,\ell} \}$. Since by the B-spline property that $\sum_{j=1}^{j_{n,\ell}} B_{\ell,j}(u) = 1$, $B_{\ell,1}(u)$ generates the same function space as $B_{\ell,1}(u) = \{ 1 : 1 \leq j \leq j_{n,\ell} \}$. Thus $m_\ell (u)$ can be approximated by $m_\ell (u) = \gamma_1 \ell + \sum_{j=1}^{j_{n,\ell}} B_{\ell,j}(u) \gamma_j \ell = B_{\ell,1}(u)^T \gamma_1 \ell$, where $\gamma_1 \ell = (\gamma_{p_1,1}, \ldots, \gamma_{p_1,p_1})^T$ and $\gamma_{p,1-1} \ell = (\gamma_{p,1-1}, \ldots, \gamma_{p,p,1-1})^T$. When $\gamma_{p,1-1} \ell = 0_{(j_{n,\ell}-1),p_1}$, $m_\ell (u) = \gamma_1 \ell$, which is a constant. Then the null and alternative hypotheses can be written as $H_0 : \gamma_{p,1-1} \ell = 0_{(j_{n,\ell}-1),p_1}$ for $\ell = p_1 + 1, \ldots, p$ versus $H_1 : \gamma_{p,1-1} \ell \neq 0_{(j_{n,\ell}-1),p_1}$ for some $\ell$. Let $\hat{\gamma}_N = \left\{ \hat{\gamma}_{p,1-1}, \ldots, \hat{\gamma}_{p,1-1} \right\}^T$. Thus $\hat{\gamma}_{N,1-1} = 0_{(j_{n,\ell}-1),p_1}$, for which $\hat{\gamma}_{p,1-1} = \hat{\gamma}_{p,1-1} \ell$. Let $\hat{\gamma}_{(1)} = \left\{ \hat{\gamma}_{1}, \ldots, \hat{\gamma}_{p_1} \right\}^T$, $\hat{\gamma}_{(2)} = \left\{ \hat{\gamma}_{p_1+1}, \ldots, \hat{\gamma}_{p} \right\}^T$, $\hat{\gamma}_{(3)} = \left\{ \hat{\gamma}_{p_1+1}, \ldots, \hat{\gamma}_{p} \right\}^T$, and $\hat{\gamma}_{(4)} = \hat{\gamma}_{(1)}, \ldots, \hat{\gamma}_{(p_1)}, \hat{\gamma}_{(p_1+1)}, \ldots, \hat{\gamma}_{(p)}$. Let 
$j_{n,1} = \sum_{\ell=1}^{p_1} j_{n,\ell}$. Define the score function as
\[
s_{2n} (\hat{\gamma}_{(2)}, \hat{\hat{\beta}}, \hat{\hat{\alpha}}) = \partial L_n (\hat{\gamma}_{(2)}, \hat{\hat{\beta}}, \hat{\hat{\alpha}})/\partial \gamma_{(2)}
\]
\[
= \sum_{i=1}^{p_1} \left\{ g^{-1} \left\{ \sum_{\ell=1}^{p_1} B_{\ell,i} (X_i^T \hat{\hat{\beta}})^T \hat{\gamma}_{p,1-1} \ell Z_{i\ell}, Y_{i1} \right\} \times \left\{ B_{\ell,i} (X_i^T \hat{\hat{\beta}})^T \hat{\gamma}_{p,1-1} \ell Z_{i\ell} : 2 \leq j \leq j_{n,\ell} , p_1 + 1 \leq \ell \leq p \right\} \right\}^T.
\]
Let
\[
R_{11} (\beta_0) = \left[ \begin{bmatrix} B_{\ell,1}(X_i^T \beta_0)^T Z_{i1}, \ldots, B_{\ell,1}(X_i^T \beta_0)^T Z_{ip_1}, (Z_{i\ell} : p_1 + 1 \leq \ell \leq p)^T \end{bmatrix}_{(j_{n,1}+p-p_1) \times 1},
\]
\[
R_{21} (\beta_0) = \left[ \begin{bmatrix} B_{\ell}(X_i^T \beta_0)^T Z_{i\ell} : 2 \leq j \leq j_{n,\ell} , p_1 + 1 \leq \ell \leq p \end{bmatrix}_{(j_{n,1}+p-p_1) \times 1},
\]
Define $\Omega_n = \left( \Omega_{n,11} \quad \Omega_{n,12} \mid \Omega_{n,21} \quad \Omega_{n,22} \right)^T$, where
\[
\Omega_{n,j'} = n^{-1} \sum_{i=1}^n R_n \left( \beta^0 \right) R_{i,j'} \left( \beta^0 \right)^T \left\{ dg^{-1}(\eta_i) \right\}^2 / \sigma_i^2, \quad j, j' = 1, 2.
\]

Define
\[
T_n = n^{-1} S_n \left( \bar{\gamma}^N, \bar{\beta}, \bar{\alpha} \right)^T \Omega_n \left( \bar{\gamma}^N, \bar{\beta}, \bar{\alpha} \right),
\]
where $\Omega_n = \left( \Omega_{n,22} - \Omega_{n,21} \Omega_{n,12} \Omega_{n,11} \right)^{-1}$.

**Theorem 3.** Under Conditions (C1)–(C4) in the Appendix A, and $n^{1/2} \ll J_{\max} \ll n^{1/4}$, we have under $H_0$, as $n \to \infty$, 
\[
\left\{ 2 \left( J_n - J_{n-1} \right) - p - p_1 \right\} \to \mathcal{N} (0, 1).
\]

**Remark.** Note that $T_n$ defined in (10) contains popular parameters, so that it cannot be directly used as a test statistic. To carry out the Rao-score-type hypothesis test, we instead use $\hat{T}_n$ as the test statistic. We define $\hat{T}_n$ in the same way as $T_n$, but only replace $\Omega_{n,j'}$ by its consistent estimate $\hat{\Omega}_{n,j'} = n^{-1} \sum_{i=1}^n R_n \left( \hat{\beta} \right) R_{i,j'} \left( \hat{\beta} \right)^T \left\{ dg^{-1}(\hat{\eta}_i) \right\}^2 / \hat{\sigma}_i^2$, where
\[
\hat{\eta}_i = \sum_{\ell=1}^p B^*_\ell \left( X_i^\ell \hat{\beta} \right)^T \hat{\gamma}^N z_{\ell i} + T^\ell \hat{\alpha}.
\]
and $\hat{\sigma}_i^2$ is a consistent estimate of $\sigma_i^2$. For example, when $Y_i$ are binary, we let $\hat{\sigma}_i^2 = g^{-1}(\eta_i) \left( 1 - g^{-1}(\eta_i) \right)$. When $Y_i$ are continuous, we let $\hat{\sigma}_i^2 = \hat{\sigma}_i^2 = \sum_{i=1}^n \left( Y_i - g^{-1}(\hat{\eta}_i) \right)^2 / n$. For implementation, the critical value of the test statistic is calculated from the chi-square distribution with $(J_n - J_{n-1} - p - p_1)$ degrees of freedom.

**Remark.** Before testing interaction effects between $X$ and $Z^\ell$'s, we first can assess whether the genetic factors are associated with the phenotype by formulating the hypotheses as $H_0 : m_\ell(\cdot) = 0$, for $\ell = 2, \ldots, p$ versus $H_1 : m_\ell(\cdot) \neq 0$ for some $\ell \in \{2, \ldots, p\}$. Thus, the null and alternative hypotheses can be further written as $H_0 : \gamma_\ell = 0_{d \times 1}$, for $\ell = 2, \ldots, p$ versus $H_1 : \gamma_\ell \neq 0_{d \times 1}$ for some $\ell$. Then a similar test statistic denoted by $\hat{T}_n$ as $T_n$ given in (10) can be constructed.

2. Testing parametric components. Next, we test whether some components in $\beta$ and $\alpha$ are zero or not by a score test procedure. Without loss of generality, we write the parameter vectors as $\beta = \left( \beta_1^T \mid \beta_2^T \right)^T$ and $\alpha = \left( \alpha_1^T \mid \alpha_2^T \right)^T$, where $\beta_1 = \left( \beta_1(\cdot) - \beta_1(\cdot) \right)^T$, correspondingly, $X_i$ and $T_i$ are written as $X_i = \left( X_i^0(\cdot), X_i^1(\cdot) \right)^T$ and $T_i = \left( T_i^0(\cdot), T_i^1(\cdot) \right)^T$. Then the null and alternative hypotheses are given as: $H_0 : \beta = 0_{(d_i - d_{i-1})}$ and $\alpha = 0_{(d_i - d_{i-1})}$ versus $H_1$; not all components in $\beta$ and $\alpha$ are zero, where $0_{d \times 1}$ is the $d \times 1$ vector with ‘0’ as its components. Let $\hat{\beta}_N^N = \left( \hat{\beta}_N^N \right)^T$, $\hat{\alpha}_N^N = \left( \hat{\alpha}_N^N \right)^T$, where $\hat{\beta}_N^{N-1} = \left( \hat{\beta}_N^{N-1} \right)^T$, $\hat{\alpha}_N^{N-1} = \left( \hat{\alpha}_N^{N-1} \right)^T$, $\hat{\beta}_N^{N+1} = \left( \hat{\beta}_N^{N+1} \right)^T$ and $\hat{\alpha}_N^{N+1} = \left( \hat{\alpha}_N^{N+1} \right)^T$ be the maximizer of (6) under the null hypothesis $H_0$. Let $\hat{\theta}^N = \left( \hat{\theta}^N \right)^T$, $\hat{\theta}^{N+1} = \left( \hat{\theta}^{N+1} \right)^T$ be the maximizer of (6) under the null hypothesis $H_0$. Let $\hat{\theta}^N = \left( \hat{\theta}^N \right)^T$, $\hat{\theta}^{N+1} = \left( \hat{\theta}^{N+1} \right)^T$ be the maximizer of (6) under the null hypothesis $H_0$.

Define the score function as
\[
s_{2n}^N (\hat{\theta}^{-1}) = \partial L_n(\hat{\beta}^N, \hat{\alpha}^N) / \partial \left( \beta_2^N, \alpha_2^N \right)^T
\]
\[
= \sum_{i=1}^n \frac{q_i}{\left( \sum_{k=1}^p B_i \left( X_i^\ell \hat{\beta}^T \right)^T \hat{\gamma}^N \hat{\alpha}_N Z_{2 \ell i} + T^\ell \alpha \right) \gamma_i}, \quad Y_i
\]
\[
= \left[ \sum_{k=1}^p \frac{m_i( X_i^\ell \hat{\beta}^T \hat{\alpha}_N Z_{2 \ell i} + T^\ell \alpha ) \psi_{j i} }{ \partial \left( \beta_2^N, \alpha_2^N \right)^T \partial \left( \beta_2^N, \alpha_2^N \right)^T } \right] R_i (\hat{\beta}^N).
\]
Let $J_1(\beta_1) = \left( -\frac{1}{\left( \beta_1 - 1 \right)^2} \right)$, $J_1 = J_1(\beta_1)$, $\tilde{X}_j = X_j - \text{Proj}_M( X_j )$, and $\tilde{T}_{ji} = T_{ji} - \text{Proj}_M( T_{ji} )$, for $j = 1, 2$. Denote $\tilde{X}_i = \tilde{X}_i^1 \tilde{X}_i^2$, $\tilde{X}_i^2 = \tilde{X}_i^2$, and
\[
\psi_{ji} = \left[ \sum_{k=1}^p \frac{ m_i( X_i^\ell \hat{\beta}^T \hat{\alpha}_N Z_{2 \ell i} + T^\ell \alpha ) \psi_{j i} }{ \partial \left( \beta_2^N, \alpha_2^N \right)^T \partial \left( \beta_2^N, \alpha_2^N \right)^T } \right] R_i (\hat{\beta}^N)^T,
\]
for \(j = 1, 2\). Define \(\Sigma_{n} = \begin{pmatrix} \Sigma_{n,11} & \Sigma_{n,12} \\ \Sigma_{n,21} & \Sigma_{n,22} \end{pmatrix} \), where
\[
\Sigma_{n,j'} = n^{-1} \sum_{i=1}^{n} \left[ \psi_{ji} \psi_{ji}^{T} \frac{dg^{-1}(\eta_i)}{\sigma_i^2} \right],
\]
for \(j, j' = 1, 2\). Let \(\Sigma_{n}^{-1} = \begin{pmatrix} \Sigma_{n,11}^{-1} & \Sigma_{n,12}^{-1} \\ \Sigma_{n,21}^{-1} & \Sigma_{n,22}^{-1} \end{pmatrix} \). We define
\[
T_{n}^{*} = s_{2n}^{*} \left( \frac{\Sigma_{n}}{\Sigma_{n,22} - \Sigma_{n,21} \Sigma_{n,11}^{-1} \Sigma_{n,12}} \right) / n,
\]
where \(\Sigma_{n,22} = \begin{pmatrix} \Sigma_{n,22} & \Sigma_{n,21} \Sigma_{n,11}^{-1} \Sigma_{n,12} \end{pmatrix}^{-1} \).

**Theorem 4.** Under Conditions (C1)-(C4) in the Appendix A, and \(\max \{n^{1/(2r+2)}, n^{1/(4r-2)}\} \ll J_{\max} \ll n^{1/4}\), we have under \(H_0\) as \(n \to \infty\), \(T_{n}^{*}\) has an asymptotic chi-square distribution on \((d_1 - d_{10} + d_2 - d_{20})\) degrees of freedom.

**Remark.** For implementation, we use \(\hat{T}_{n}^{*}\) as the Score test statistic, which is defined in the same way as \(T_{n}^{*}\) given in (11) by only replacing \(\Sigma_{n,j'}\) with its consistent estimate \(\hat{\Sigma}_{n,j'} = n^{-1} \sum_{i=1}^{n} \left[ \hat{\psi}_{ji} \hat{\psi}_{ji}^{T} \frac{dg^{-1}(\hat{\eta}_i)}{\hat{\sigma}_i^2} \right]\), where
\[
\hat{\sigma}_i^2 = \hat{\Sigma}_{i}^{*} + \hat{\Sigma}_{i}^{T},
\]
\(\hat{\Sigma}_{i}^{*}\) is a constant estimate of \(\sigma_i^2\), and
\[
\hat{\psi}_{ji} = \left\{ \sum_{\ell=1}^{p} \hat{m}_{\ell} (X_i^T \hat{\beta}^N, \hat{\beta}^N, \hat{\alpha}^N) Z_{\ell} + T_i \hat{\alpha}^N \right\}^{T},
\]
with \(\hat{X}_{ij} = j (\hat{\beta}^N) \hat{X}_{ij}, \hat{X}_{ij} = \hat{X}_{ij} - \hat{\text{Proj}}_{ij} (X_i)\) and \(\hat{T}_{ij} = T_{ij} - \hat{\text{Proj}}_{ij} (T_{ij})\), for \(j = 1, 2\), in which \(\hat{\text{Proj}}_{ij}\) is the identity matrix. Let \(\hat{\alpha}^N\) and \(\hat{T}_{ij}\) be consistent estimates of \(\hat{\text{Proj}}_{ij} (X_i)\) and \(\hat{\text{Proj}}_{ij} (T_{ij})\) defined similarly as in (A.9) in the Appendix A with the true parameters replaced by their estimates. Moreover, the score function \(s_{2n}^{*} (\hat{\sigma}_i^2)\) is approximated by
\[
\sum_{i=1}^{n} q_i \left[ g^{-1} \left\{ \sum_{\ell=1}^{p} \hat{m}_{\ell} (X_i^T \hat{\beta}^N, \hat{\beta}^N, \hat{\alpha}^N) Z_{\ell} + T_i \hat{\alpha}^N \right\}, Y_i \right] \times \hat{\psi}_{2i}.
\]

4. Simulation

**Example 1 (Continuous Response).** We generate continuous responses \(Y_i\) from the GPLSiCM given as
\[
Y_i = \eta (Z_i, X_i, T_i) + \epsilon_i = m_0 (X_i^T \hat{\beta}^0) + \sum_{\ell=1}^{2} m_i (X_i^T \hat{\beta}^0) Z_{\ell} + T_i \hat{\alpha}^0 + \epsilon_i,
\]
where \(X_i = (X_{ik}, 1 \leq k \leq 3)^T\) are simulated environmental effects, which are generated from independent uniform distributions on \([0, 1]\), \(T_i = (t_k, 1 \leq k \leq 3)^T\) are covariates generated from the multivariate normal distribution with mean 0, marginal variance 1, and an AR-1 correlation matrix with autocorrelation coefficient 0.5, and \(Z_{\ell}, 1 \leq \ell \leq 2\) are simulated genetic effects. We generate SNPs which have three possible genotype categories represented by AA, Aa and aa with frequency \((P_A^2, 2P_A (1 - P_A), (1 - P_A)^2)\), with \(P_A = 0.5\) which is the allele frequency for allele A. The genetic variables \(Z_1\) and \(Z_2\) are coded as \((1, 0, -1)\) and \((-1/2, 1/2, -1/2)\) for genotypes \((aa, Aa, Aa)\) as additive and dominance scales, respectively, following an orthogonal genetic model (Cockerham, 1964). We use the empirical centered value of \(Z_\ell\) to generate \(Y_i\). The random error terms \(\epsilon_i\) are simulated from independent \(N (0, 1)\). We let \(\hat{\beta}^0 = (3, 2, 1) / \sqrt{14}\), \(\hat{\alpha}^0 = (2, 1, -1, 1), m_0 (u) = 1 + 3 \sin (\pi u)\), and \(m_i (u) = 1 + \sin (2\pi u)\) for \(1 \leq \ell \leq 2\). Different sample sizes (i.e., \(n = 250, 500\)) are considered. We generate 2000 simulated samples. The parameters \(\beta^0\) and \(\alpha^0\) in model (12) are estimated by the profile procedure given in Section 2 with initial estimates \(\hat{\beta}^{ini} = (1, 1, 1) / \sqrt{3}\) and \(\hat{\alpha}^{ini}\) obtained by fitting a linear model
\[
Y_i = a_0 + X_i^T \hat{b} + Z_i^T \hat{c} + \sum_{\ell=1}^{3} c_{\ell i} Z_{\ell} X_{i\ell} + T_i \hat{\alpha}.
\]
Then \(\hat{\beta}^{ini} = \text{sgn}(\hat{b}_1) \hat{b} / \|\hat{b}\|_2\) and \(\hat{\alpha}^{ini} = \hat{\alpha}\), where \(\hat{b}\) and \(\hat{\alpha}\) are least squares estimates of \(b\) and \(\alpha\).
Then one selects the optimal number of interior knots $N_ℓ$ by minimizing the BIC criterion given as

$$BIC(N_1, \ldots, N_p) = -2\ln(\hat{\gamma}, \hat{\beta}, \hat{\alpha}) + \sum_{ℓ=1}^p (N_ℓ + q) \log n.$$ 

To evaluate the estimation results by using the proposed profile quasi-log-likelihood method, Table 1 reports the empirical standard errors (ESE) based on 2000 replications and average asymptotic standard errors (ASE) calculated from (8) for the estimates of $\beta^0 = (\beta^0_1, \beta^0_2, \beta^0_3)^T$ and $\alpha^0 = (\alpha^0_1, \alpha^0_2, \alpha^0_3)^T$. Table 2 reports the sample bias (Bias), the empirical coverage probabilities (CP) of the 95% confidence intervals and the mean squared errors (MSE) defined as the sample mean of $\|\hat{\beta} - \beta^0\|^2$ and $\|\hat{\alpha} - \alpha^0\|^2$ among 2000 replications. From Table 1, we observe that the standard errors become smaller as $n$ increases, due to the fact of root-n consistency of the parameter estimators. More importantly, the ASEs are very similar to the corresponding ESEs for all cases, suggesting that the asymptotic covariance matrix is correctly derived. Table 2 shows that the biases are close to 0 for all cases. This result confirms the asymptotic property that the parameter estimators are asymptotically unbiased as given in Theorem 1. It also indicates that estimation consistency is achieved even with a relatively small sample size $n = 250$. Moreover, the MSEs get closer to 0 as $n$ increases. We also observe that the empirical coverage rates are close to the nominal confidence level 95%. This result is confirmatory to the asymptotic normal distribution of the parameter estimators established in Theorem 1.

To evaluate the performance of the spline estimators of the nonparametric functions $m_ℓ(·)$ for $0 ≤ ℓ ≤ 2$, we define the mean integrated squared error (MISE) as the average of

$$\text{ISE}(\hat{m}_ℓ) = n^{-1} \sum_{i=1}^n \left\{ \hat{m}_ℓ(X_i^T \hat{\beta}) - m_ℓ(X_i^T \beta^0) \right\}^2,$$

for $0 ≤ ℓ ≤ 2$ among the 2000 replications. Table 3 shows the MISEs for the spline estimates $\hat{m}_ℓ(·)$ for $0 ≤ ℓ ≤ 2$, $n = 250, 500$. We observe that the MISE values decrease to zero as $n$ increases. This result confirms the asymptotic consistency of the nonparametric functional estimators established in Theorem 2. For visualization of the actual function estimates, Fig. 2 shows the estimated curves $\hat{m}_ℓ(·)$ (solid lines), $ℓ = 0, 1$, the true functions $m_ℓ(·)$ (dashed lines), and the upper and lower 95% pointwise confidence intervals (upper and lower solid lines) for $n = 250$. It is evident that the proposed estimators perform well.

Next we illustrate the performance of the proposed score test statistic $S_n$ for testing importance of the genetic factors by generating data from model (12) with $m_ℓ(u) = \lambda \left[ 1 + \sin (2\pi u) \right]$, where $λ$ ranges from 0 to 0.5 with increment 0.1. We
consider the hypothesis testing given as
\[ H_0 : m_\ell(u) = 0 \text{ for all } \ell = 1, 2, \quad \text{versus} \quad H_1 : m_\ell(u) \neq 0 \text{ for some } \ell. \]

The testing procedure is performed by using the score test statistic \( T_n \). The critical value is calculated from the chi-square distribution with \((N_1 + N_2 + 8)\) degrees of freedom at significance level \( \alpha = 0.005, 0.01, 0.05 \). Table 4 presents the empirical sizes of power for \( n = 250, 500 \) based on 2000 replications. When \( \lambda = 0 \) (\( H_0 \) is true), the empirical sizes of power are close to but slightly larger than the nominal significance levels. We can also observe that the empirical size increases to 1 as the value of \( \lambda \) increases with different type I errors, and it increases faster when the sample size is larger. These results indicate that the proposed score test is a reasonable test.

To illustrate the score test statistic \( T_n \) given in (10) for interaction effects, we generate data from model (12) with \( m_\ell(u) = 1 + \lambda \sin(2\pi u) \), where \( \lambda \) ranges from 0 to 0.5 with increment 0.1. The null and alternative hypothesis tests are set up as
\[ H_0 : m_\ell(u) = c_\ell \quad \text{for all } \ell = 1, 2, \quad \text{where } c_\ell \text{ are unknown constants}, \quad \text{versus} \quad H_1 : m_\ell(u) \neq c_\ell \quad \text{for some } \ell. \]

The testing procedure is performed by using the score test statistic \( T_n \) given in (10). The critical value is calculated from the chi-square distribution with \((N_1 + N_2 + 6)\) degrees of freedom at significance level 0.05.

We compare the score test performance in the proposed GPLSiCM (12) with a parametric linear model by assuming that the main effect of \( X \) as well as the interaction effects with \( Z \) are linear and a semiparametric model by assuming the main effect \( m_0(.) \) of \( X \) is nonlinear but the interaction effects with \( Z \) are linear. Thus the linear model (LM) is given as
\[ E(Y_i|Z_i, X_i, T_i) = a_0 + \sum_{k=1}^{3} b_kX_{ik} + \sum_{\ell=1}^{2} c_{\ell}Z_{i\ell} + \sum_{\ell=1}^{2} \sum_{k=1}^{3} c_{k\ell}Z_{i\ell}X_{ik} + T_i^T \alpha_0. \]
and the semiparametric model is given as

\[ E(Y_i | Z_i, X_i, T_i) = m_0(X_i^T \beta_0^\circ) + \sum_{l=1}^{2} c_l Z_{il} + \sum_{l=1}^{2} \sum_{k=1}^{3} c_{lk} Z_{il} X_{ik} + T_i^T \alpha^0, \tag{16} \]

where \( a_0, b_k, c_l \) and \( c_{lk} \) are unknown constants. The null and alternative hypothesis tests in (14) are equivalent to \( H_0 : c_{lk} = 0 \) for all \( 1 \leq l \leq 2 \) and \( 1 \leq k \leq 3 \) versus \( H_1 : c_{lk} \neq 0 \) for some \( 1 \leq l \leq 2 \) or \( 1 \leq k \leq 3 \). Note that the linear model (15) by assuming a linear main effect is misspecified, but the semiparametric model (16) is correctly specified under the null hypothesis \( H_0 \). This semiparametric model (16) is a partially linear single-index model (PLSiM) (Carroll et al., 1997).

Score tests are performed by assuming (15) and (16) as the alternative models, respectively. The critical value for both tests is calculated from the chi-square distribution with 6 degrees of freedom at significance level 0.05.

Table 5 reports the powers of the score tests for the three models GPLSiCM, LM and PLSiM for \( n = 250, 500 \) and \( \lambda = (0, 0.1, 0.2, 0.3, 0.4, 0.5) \). Clearly we observe that for the GPLSiCM, the power size at \( \lambda = 0 \) (\( H_0 \) is true) is close to the nominal significance level 0.05 for \( n = 250, 500 \), which confirms the asymptotic null distribution of the test statistic established in Theorem 3. The power increases to 1 as the \( \lambda \) value increases. For larger sample size, the power size increases faster. The results illustrate that the proposed score test is a powerful test. For the LM, the power is much larger than 0.05 when \( \lambda = 0 \), since the model is misspecified under \( H_0 \). For the PLSiM, the power is around 0.05 when \( \lambda = 0 \), since the model is correctly specified under \( H_0 \), so that the score test works well under \( H_0 \). For both of these two misspecified models under \( H_1 \), the power increases very slowly as the value of \( \lambda \) increases. This result indicates that when the true model which is GPLSiCM is misspecified, the score test will become less powerful.

**Example 2.** In this example, we generate continuous responses \( Y_i \) from the GPLSiCM given in (12), where \( \beta_0^\circ = (3, 2, \lambda, 1, \lambda, \lambda)/\sqrt{14} \) and \( \alpha^0 = (2, 1, \lambda, \lambda, \lambda, -1) \) with \( \lambda \) ranging from 0 to 0.2 with increment 0.02, and \( m_0(u) = 1 + 3 \sin(\pi u) \), and \( m_1(u) = 1 + 0.5 \sin(2\pi u) \) for \( 1 \leq l \leq 2 \). The covariates \( X_i, Z_i \) and \( T_i \) are generated in the same way as in Example 1. We conduct the hypothesis testing:

\[ H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_7 = 0 \quad \text{and} \quad \alpha_3 = \alpha_4 = \alpha_5 = 0, \]

versus \( H_1 : \beta_k \neq 0 \) for some \( k \in (3, 4, 6, 7) \) or \( \alpha_k \neq 0 \) for some \( k \in (3, 4, 5) \). \( \tag{17} \)

The testing procedure is performed by using the score test statistic \( T_n^a \) given in (11). In addition, 2000 realizations were generated with \( n = 250, 500 \) to calculate the power of the test statistic \( T_n^a \) at significance level 0.05. Fig. 3 displays the power function versus the \( \lambda \) values for \( n = 250, 500 \). When \( \lambda = 0 \) such that the null hypothesis \( H_0 \) is true, the empirical sizes of power are 0.040 for \( n = 250 \) and 0.058 for \( n = 500 \), respectively, which are close to the nominal significance level 0.05. This result is confirmative to the asymptotic null distribution of the test statistic \( T_n^a \) established in Theorem 4. We can also observe that the empirical size increases to 1 as the value of \( \lambda \) increases, and it increases faster when the sample size is larger. For \( n = 500 \), the power reaches 1 when \( \lambda = 0.1 \).

**Example 3 (Binary Response).** In this example, we generate the binary responses \( Y_i \) from the logistic regression model given as

\[ \logit \{ \Pr(Y_i = 1 | Z_i, X_i, T_i) \} = m_0(X_i^T \beta_0^\circ) + \sum_{l=1}^{2} m_l(X_i^T \beta_0^\circ) Z_{il} + T_i^T \alpha^0, \tag{18} \]

where \( \beta_0^\circ = (3, 2, 1)/\sqrt{14}, \alpha^0 = (0.5, 0.3, 0.7), m_0(u) = 0.2 + 2 \sin(2\pi u), \) for \( 0 \leq l \leq 2 \). The covariates \( X_i, Z_i \) and \( T_i \) are generated in the same way as in Example 1.

Table 6 reports the empirical standard errors (ESE) based on 500 replications and average asymptotic standard errors (ASE) calculated from (8) for the estimates of \( \beta_0^\circ = (\beta_0^0, \beta_2^0, \beta_5^0)^T \) and \( \alpha^0 = (\alpha_0^0, \alpha_2^0, \alpha_5^0)^T \). Table 7 reports the sample bias (Bias), the empirical coverage probabilities (CP) of the 95% confidence intervals for \( \beta_0^\circ \) and \( \alpha^0 \) and the mean squared errors (MSE) of the estimates for \( \beta_0^\circ \) and \( \alpha^0 \) among 500 replications. In Table 6, same pattern can be observed as in Table 1 in

Table 5

<table>
<thead>
<tr>
<th></th>
<th>GPLSiCM</th>
<th>LM</th>
<th>PLSiM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.054</td>
<td>0.130</td>
<td>0.048</td>
</tr>
<tr>
<td>0.1</td>
<td>0.102</td>
<td>0.144</td>
<td>0.052</td>
</tr>
<tr>
<td>0.2</td>
<td>0.322</td>
<td>0.160</td>
<td>0.068</td>
</tr>
<tr>
<td>0.3</td>
<td>0.622</td>
<td>0.140</td>
<td>0.054</td>
</tr>
<tr>
<td>0.4</td>
<td>0.882</td>
<td>0.152</td>
<td>0.092</td>
</tr>
<tr>
<td>0.5</td>
<td>0.980</td>
<td>0.164</td>
<td>0.104</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th></th>
<th>n = 250</th>
<th>n = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPLSiCM</td>
<td>0.052</td>
<td>0.052</td>
</tr>
<tr>
<td>LM</td>
<td>0.140</td>
<td>0.140</td>
</tr>
<tr>
<td>PLSiM</td>
<td>0.046</td>
<td>0.046</td>
</tr>
</tbody>
</table>
Example 1 that the ASEs and ESEs are similar for all cases, and they get closer as \( n \) increases. Table 7 shows that the sample biases and MSE values decrease to zero as \( n \) increases. The empirical coverage rates get closer to the nominal confidence level 0.95 as \( n \) increases, and they are very close to 95% for all the cases when \( n = 500 \).

For visualization of the actual function estimates, Fig. 4 shows the estimated curves \( \hat{m}_l(\cdot) \) (solid lines), \( \ell = 1, 2 \), the true functions \( m_l(\cdot) \) (dashed lines), and the upper and lower 95% pointwise confidence intervals (upper and lower solid lines) for \( n = 200, 500 \). We can observe that the proposed estimators perform well.

Next, we let \( m_1(u) = 0.2 + \lambda \sin(2\pi u) \), for \( \ell = 1, 2 \), where \( \lambda \) ranges from 0 to 2 with increment 0.4. We consider the hypothesis testing given as \( H_0 : m_1(u) = c_1 \) and \( m_2(u) = c_2 \), where \( c_1 \) and \( c_2 \) are unknown constants, versus, \( H_1 : m_1(u) \neq c_1 \) or \( m_2(u) \neq c_2 \). The testing procedure is performed by using the score test statistic \( T_n \) given in (10). The critical value is calculated from the chi-square distribution with \( 2(N + 3) \) degrees of freedom at significance level 0.05. The left panel of Fig. 5 displays the power function versus the \( \lambda \) values for \( n = 200, 500 \). When \( \lambda = 0 \) (\( H_0 \) is true), the empirical sizes of power are 0.042 for \( n = 200 \) and 0.052 for \( n = 500 \), respectively, which are close to the nominal significance level 0.05. This result is confirmative to the asymptotic distribution of the test statistic \( T_n^* \) established in Theorem 4. We can also observe that the empirical size increases to 1 as the value of \( \lambda \) increases, and it increases faster when the sample size is larger.
Fig. 4. Plots of the estimated curves $\hat{m}_\ell(\cdot)$ (solid lines), $\ell = 1, 2$, the true functions $m_\ell(\cdot)$ (dashed lines), and the upper and lower 95% pointwise confidence intervals (upper and lower solid lines) for $n = 200, 500$.

To evaluate the score test statistic $T_n^\ast$ in (11) for parameters, we let $\beta^0 = (3, 2, \lambda, 1, \lambda, \lambda) / \sqrt{14}$ and $\alpha^0 = (0.5, 0.3, \lambda, \lambda, 0.7)$ with $\lambda$ ranging from 0 to 0.5 with increment 0.1, and $m_\ell(u) = 0.2 + 2 \sin (2\pi u)$, for $0 \leq \ell \leq 2$.

We conduct the hypothesis testing given in (17). When $\lambda = 0$, the powers are 0.058 and 0.054 for $n = 200$ and 500, respectively. From the right panel of Fig. 5, we can observe an increasing pattern for the power function at $n = 200, 500$.

5. Data applications

In this example, we illustrate our method via analysis of the Framingham study to investigate effects of $G \times E$ interactions on body mass index (BMI). We use $X_1 =$ sleeping hours per day, $X_2 =$ hours of light activity per day and $X_3 =$ hours of moderate activity per day as the environmental factors, and use SNPs as the genetic factors. The three possible allele combinations are coded as $Z_1 = (1, 0, -1)$ and $Z_2 = (-1/2, 1/2, -1/2)$ for additive and dominance scales. For details on genotyping, see http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v3.p2. For illustration, we study effects of SNPs located at chromosome 4 on BMI. After eliminating SNPs departure from Hardy–Weinberg equilibrium, there are 24364 SNPs remaining in our study. In addition to genotypes and sleeping and activity hours, we use five additional covariates as the linear part in the GPLSiM (3), which are: $T_1 =$ health condition ($1 =$ EXCELLENT, $2 =$ GOOD, $3 =$ FAIR, $4 =$ POOR), $T_2 =$ whether work ($1 =$ YES, $0 =$ NO), $T_3 =$ systolic blood pressure, $T_4 =$ diastolic blood pressure.
where \( \lambda_1 \) and \( \lambda_2 \) are the centered and standardized values of the additive and dominance scales for the \( \ell \)th SNP.

\[
E(Y_i | Z_i, X_i, T_i) = m_0(X_i^T \beta) + \sum_{\ell=1}^{\ell} m_1(\ell X_i^T \beta) \lambda_{1,\ell} + \sum_{\ell=1}^{2\ell-1} m_2(\ell X_i^T \beta) \lambda_{2,\ell} + T_i^T \alpha, \quad i = 1, \ldots, 299.
\]

(19)
Fig. 6. Plots of the estimated curve \( \hat{m}_0(\cdot) \) against the estimated index and each index component.

\( \beta_1, \ldots, \beta_3 \) are significantly different from zero with \( p \)-values much smaller than 0.05. This result indicates that the BMI is highly related to sleeping and physical activity hours, which confirms the finding in the literature (Knutson, 2012; Wareham et al., 2005). The large \( p \)-values for the parameters \( \alpha_2, \alpha_3, \alpha_4 \) indicate that work status and blood pressures are not important factors to BMI. This result is further corroborated by conducting the hypothesis testing: \( H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0 \), in which we obtain the \( p \)-value 0.999. The testing procedure is performed by using the score test statistic \( T_n^* \) given in (11).

To illustrate the change pattern of the estimated mean curve of BMI with index, Fig. 6 plots the estimated curve \( \hat{m}_0(\cdot) \) against the estimated index. Clearly, \( \hat{m}_0(\cdot) \) shows a nonlinear change pattern with the index as well as its components. Figure (b) shows that the value of \( \hat{m}_0(\cdot) \) increases with sleeping hours in the beginning, and then it shows a decreasing pattern after the sleeping hours exceed 3, which is followed by a sharp decline after 6.5 h. The decreasing pattern after 3 h indicates a negative relationship between BMI and sleeping hours. However, too little sleep with less than 3 h per day may cause significant drop of weight as displayed in Figure (b). Figures (c) shows an increasing trend of \( \hat{m}_0(\cdot) \) from 0 to 1.5 h of light activity. After that, the value of \( \hat{m}_0(\cdot) \) is moving around between 6.2 and 6.8, followed by a steep decline after 4 h of light activity. Similar pattern of \( \hat{m}_0(\cdot) \) can be observe in Figure (d) for moderate activity hours.

Next, to illustrate the effect of a genetic factor interacting with the environmental factors, we plot the estimated coefficient functions \( \hat{m}_{11}(\cdot) \) and \( \hat{m}_{12}(\cdot) \) for the additive and dominance scales of SNP 66155100 in Fig. 7. Again we observe nonlinear patterns of the estimated coefficient functions. The left panel shows that the estimated coefficient for the additive scale drops quickly from 2 to 0 and then it increases slowly to 0.3. The right panel indicates that the estimated coefficient for the dominance scale starts from a negative value and ends up with a positive value, and we can observe two waves in the estimated curve. Moreover, the mean value of the BMI for the three genotypes aa, Aa, AA, where A is the minor
Fig. 7. Plots of the estimated coefficient functions $\tilde{m}_{11}(\cdot)$ and $\tilde{m}_{12}(\cdot)$ for the additive and dominance scales of SNP ss66155100.

Fig. 8. Plots of the estimated mean functions $\tilde{\mu}_{11}(\cdot)$ (thick line), $\tilde{\mu}_{12}(\cdot)$ (thin line) and $\tilde{\mu}_{13}(\cdot)$ (dashed line) against the estimated index for the three genotypes of SNP ss66155100.

allele, can be estimated by $\tilde{\mu}_{11}(X_i^T\beta) = \tilde{m}_0(X_i^T\beta) + 0.5\tilde{m}_{21}(X_i^T\beta)$, $\tilde{\mu}_{12}(X_i^T\beta) = \tilde{m}_0(X_i^T\beta) - 0.5\tilde{m}_{21}(X_i^T\beta)$, and $\tilde{\mu}_{13}(X_i^T\beta) = \tilde{m}_0(X_i^T\beta) - 0.5\tilde{m}_{21}(X_i^T\beta)$, respectively. Fig. 8 shows the plots of the estimated mean functions $\tilde{\mu}_{11}(\cdot)$ (thick line), $\tilde{\mu}_{12}(\cdot)$ (thin line) and $\tilde{\mu}_{13}(\cdot)$ (dashed line) against the estimated index for the three genotypes. Clearly the estimated mean functions show different nonlinear change patterns along with the index for the three different groups. We conduct the hypothesis testing: $H_0 : m_{k\ell}(\cdot) = c_{k\ell}$, where $c_{k\ell}$ is an unknown constant, for all $1 \leq k \leq 2$, $1 \leq \ell \leq 4$, by using the score test statistic $T_n$ given in (10), and obtain the $p$-value which is close to zero. Lastly, we compare the GPLSiCM (20) with the two linear models with and without interaction effects between $X_i$ and $Z_i$, and obtain the coefficients of determination $R^2 = 0.338, 0.239, 0.134$, respectively, for the GPLSiCM and the two linear models. Apparently, the GPLSiCM has the largest $R^2$ value, which improves the model fitting of the linear model without interactions by 152.84%.

6. Discussion

In this paper, we propose a new semiparametric modeling approach to study nonlinear $G \times E$ interactions. We demonstrate that effects of genetic factors on diseases can be altered by environmental factors through the proposed semiparametric GPLSiCM model. The proposed method can help scientists better understand how genetic factors and environmental factors work together to cause human diseases so as to develop new strategies to prevent and treat them. In the GPLSiCM model, we develop a profile quasi-log-likelihood estimation method and Rao-score-type tests for regression
parameters and nonparametric coefficient functions with asymptotic properties studied. Simulation studies and analysis of a body mass index dataset have been conducted to illustrate the proposed method and confirm the asymptotic results. The proposed estimating and testing procedures are conceptually simple, theoretically reliable and computationally efficient.

In practice, complex quantitative traits, such as BMI, may change over time, depending on individual characteristics and environmental exposures. As an extending project, we plan to extend the proposed method to longitudinal study designs. In genomic data analysis, a large number of markers (large $p$) are usually involved. In this paper, we apply the ad hoc single marker scanning approach to the GPLSiCM model by using the proposed score test with a multiple testing procedure. This approach is commonly adopted by the genome-wide association studies (GWAS) community, by which one marker is analyzed at a time and the entire genome is scanned for all markers. As a further work, we plan to apply another popular ad hoc method developed recently in the GWAS community to capture the polygenic effect (genetic background information) via the mixed model approach. We will add a polygenic effect to the proposed GPLSiCM model. The covariance structure of the polygenic effect is obtained by using all marker information prior to the analysis. The theoretical properties and numerical performance of the resulting estimators will be further investigated.

Acknowledgments

The authors thank the Editor, the Associate Editor and the referee for their insightful comments and suggestions that led to substantial improvement of the paper. Ma’s research was partially supported by NSF Award DMS-1306972.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jspi.2014.08.005.

References