

## Varying coefficient model for gene-environment interaction: a non-linear look

Shujie Ma<sup>1</sup>, Lijian Yang<sup>1</sup>, Roberto Romero<sup>2</sup>, and Yuehua Cui<sup>1,\*</sup>

<sup>1</sup>Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824

<sup>2</sup>The Perinatology Research Branch, NICHD, NIH, DHHS, Bethesda, MD, and Detroit, 48201

\**cui@stt.msu.edu*

**SUMMARY:** The genetic basis of a complex trait often involves the function of multiple genetic factors, their interactions and the interaction between the genetic and environmental factors. Gene-environment (G×E) interaction is considered pivotal in determining trait variations and susceptibility of many genetic disorders such as neurodegenerative diseases or mental disorders. Regression-based method assuming a linear relationship between a disease response and the genetic and environmental factors as well as their interaction is the commonly used approach in detecting G×E interaction. The linearity assumption, however, could be easily violated due to the nonlinear interaction between the genetic and environment factors. In this work, we propose to relax the linearity assumption and allow for non-linear G×E interaction under a varying coefficient model framework. We propose to estimate the varying coefficients with regression spline technique. The model allows one to assess the non-linear penetrance of a genetic variant under different environmental stimuli, therefore could help us to gain novel insights into the etiology of a complex disease. Various hypothesis tests are proposed toward a complete dissection of G×E interaction. A wild bootstrap method is adopted to assess the statistical significance. Both simulation and real data analysis demonstrate the power and utility of the proposed method. Our method provides a powerful alternative and a testable framework for assessing non-linear G×E interaction.

## 1 Introduction

The genetic basis of a complex trait often involves multiple genetic factors functioning in a coordinated manner. The extent on how our genetic blueprint expresses also depends on the interactions between genetic and environmental factors. Increasing evidences have shown that gene-environment ( $G \times E$ ) interactions play pivotal roles in determining the risk of diseases, for instance, the psychiatric diseases (reviewed in Caspi and Moffitt, 2006), the neurodegenerative and cardiovascular diseases (Costa and Eaton, 2006), and cancer (Ulrich et al. 1999). Due to the complex nature of the form and mechanism of  $G \times E$  interaction in different living organisms, hunting down the molecular machinery of  $G \times E$  interaction has been a daunting task in the post genomic era. There is a pressing need in developing efficient and powerful statistical methods for a rigorous investigation of  $G \times E$  interaction.

$G \times E$  interaction refers to how genotypes influence phenotypes differently in different environments (Falconer, 1952). In a typical  $G \times E$  interaction study design, environment is often defined as different conditions coded as a discrete variable in a statistical model. For example, in a study of  $G \times E$  interaction related to lung cancer, smoking status can be defined as an environment condition coded as 1 (smoking) or 0 (no smoking). In many other studies, the environment condition is defined as a continuous measure. For one example, studies show that about 80% of type II diabetes and 70% of cardiovascular disease are related to obesity (defined by body mass index (BMI)). To track down genetic factors responsible for diabetes or cardiovascular disease, obesity can be defined as an environment factor which may induce or reduce the expression of particular genes to affect the disease status. The contribution of the same gene to a disease status may be largely different under different BMI levels. As another example, the peak bone mineral density (BMD) in adulthood varies a lot across different age groups. The amount of nutrition intake (e.g., Vitamin D) is also an important environment factor influencing the variation of BMD (Peacock et al. 2002). Individuals carrying the same

gene may respond differently to the rate of density decrease as they get older. Also the peak BMD measure may vary a lot across groups with different nutrition intake, potentially due to the interaction of specific genes with the amount of nutrition intake (e.g., Vitamin D).

In the above mentioned examples, one is interested in understanding how genes respond differently across different environment conditions in determining the variation of a trait or the risk of a disease. We focus our attention to environment conditions measured on a continuous scale. From a statistical point of view, “interaction” is typically modeled as a product term. A simple model to detect interaction would be a simple linear regression model with the form

$$Y = \alpha_0 + \alpha_1 X + \beta_1 G + \beta_2 XG + \varepsilon, \quad (1)$$

where  $Y$  is the phenotypic response;  $\alpha_0$  is the overall mean;  $\alpha_1$  and  $\beta_1$  are the effects of the environment ( $X$ ) and genetic ( $G$ ) variables, respectively;  $\beta_2$  is the effect for  $G \times E$  interaction; and  $\varepsilon$  is the error term with mean 0 and variance  $\sigma^2$ . A simple rearrangement of model (1) leads to

$$Y = \alpha_0 + \alpha_1 X + (\beta_1 + \beta_2 X)G + \varepsilon. \quad (2)$$

With this representation, it is clear that the contribution of a gene to the variation of a phenotype  $Y$  is restricted to a linear function in  $X$ . From a biological point of view,  $G \times E$  interaction can be better viewed as the genetic responses to environment changes or stresses (McClintock, 1984; Hoffmann and Parsons, 1991). The form and pattern of the responses are typically unknown and may not follow a linear relationship as described in model (1).

Statistical methods in testing  $G \times E$  interaction can be broadly categorized into two areas: the model-based method, parametrically, non-parametrically or semi-parametrically (e.g., Guo 2000; Kraft et al. 2007; Maity et al. 2009; Chatterjee and Carroll 2005), and the model-free method such as the multifactor dimensionality reduction method (Hahn et al. 2003). In a

model-based regression framework, methods typically assume linear interactions as given in model (1). This assumption, however, could be easily violated due to the underlying nonlinear machinery between the genetic and environment factors. In addressing the limitation of the linear model assumption in dissecting the role of a gene under different environment conditions, one can relax the linearity assumption of  $G \times E$  interaction and allow for a nonlinear interaction by replacing the linear  $G \times E$  interaction coefficient  $\beta_1 + \beta_2 X$  in model (2) by a smooth non-linear function  $\beta(X)$  and apply a varying coefficient (VC) model to detect non-linear  $G \times E$  interaction. A VC model has the form

$$Y = \alpha(X) + \beta(X)G + \sigma(X)\varepsilon, \quad (3)$$

for given covariates  $(X, G)^T$  and the response  $Y$  with  $E(\varepsilon|X, G) = 0$  and  $\text{Var}(\varepsilon|X, G) = 1$ .  $\beta(X)$  is a smoothing function in  $X$  and  $\sigma^2(X) = \text{Var}(Y|X, G)$  is the conditional variance function. VC models have gained considerable attention in recent years, see for example, the work of Hastie and Tibshirani (1993), Hoover et al. (1998), Fan and Zhang (1999), Cai, Fan and Li (2000), Fan and Zhang (2000), Huang, Wu and Zhou (2004) among others. Under the VC modeling framework, the effect of a gene is allowed to vary as a function of environmental factors, either linearly or non-linearly, captured by the model itself. Thus, the VC model has the potential to dissect the non-linear penetrance of genetic variants.

Methods for the estimation of VC models have flourished in the literature, see Fan and Zhang (1999), Xia and Li (1999) and Cai, Fan and Li (2000) for kernel type estimators; Hoover, Rice, Wu and Yang (1998), Chiang, Rice and Wu (2001) and Huang, Wu and Zhou (2004) for spline estimators; and Zhou and You (2004) for wavelet estimators. In this work, we propose to use the B-spline function to estimate the coefficient function  $\beta(\cdot)$  for two major reasons. Firstly, it is computationally expedient, which is much necessary for analyzing high-dimensional genetic data with hundreds of thousands of markers. Secondly, it is theoretically reliable guarded by the asymptotic consistency and normality property of the spline estimator

$\widehat{\beta}(\cdot)$ , see Huang, Wu and Zhou (2004). Our another goal is to draw inferences about the coefficient function of  $\beta(X)$  in Model (3), to test whether it is significantly different from zero or a constant. Because of the distribution free nature of nonparametric models, we adopt the wild bootstrapping approach as in Härdle and Mammen (1993) for its simplicity and reliability. See the examples in Härdle and Mammen (1993) for the great performances of wild bootstrap over other bootstrap approaches.

The paper is organized as follows. In Section 2, we introduce the methodology of applying varying coefficient models to genetic data to detect G×E interaction. We introduce the B-spline fitting technique and its necessary notations. We introduce the test statistics for the hypothesis testing evaluated by the wild bootstrap strategy. In Section 3, we study the finite sample properties of the proposed procedure using the simulated example. Furthermore, the utility of the method is illustrated through the analysis of a real data set detailed in Section 4, followed by the discussion in Section 5.

## 2 Statistical Methods

### 2.1 A two-parameter varying coefficient model

In model (3), we only consider the additive effect of a genetic variant. In real life, we do not know the true gene action mode, hence a more flexible model is to consider both additive and dominance penetrance effects. We assume a continuous response variable  $Y$  which is a function of an environment variable  $X$  and the additive and dominance scales  $G_1$  and  $G_2$  of a genetic factor. Each genetic factor has three possible genotype categories represented by  $AA$ ,  $Aa$  and  $aa$ . The three genotype categories can be coded as 1, 0, and  $-1$  for the additive scale  $G_1$ , and as  $-1/2$ ,  $1/2$ , and  $-1/2$  for the dominance scale  $G_2$ , corresponding to genotypes  $AA$ ,  $Aa$  and  $aa$ , respectively. We assume allele  $A$  is the minor allele with its frequency represented by  $p_A$ . We model the coefficients of  $G_1 \in (1, 0, -1)$  and  $G_2 \in (-1/2, 1/2, -1/2)$

for each genetic factor as smooth functions of the environment variable  $X$ . Since our major interests are the estimation and inference about the coefficient functions for  $G_1$  and  $G_2$ , for simplicity we impose a linear structure on the intercept function  $\alpha(X)$  defined in model (3) by letting  $\alpha(X) = \alpha_0 + \alpha_1 X$ , although a non-parametric smooth function can also be fitted. Thus, the redefined VC model is given as

$$Y = \alpha_0 + \alpha_1 X + \beta_1(X) G_1 + \beta_2(X) G_2 + \sigma(X) \varepsilon, \quad (4)$$

for given covariates  $(X, G_1, G_2)$ , with  $E(\varepsilon | X, G_1, G_2) = 0$ ,  $\text{Var}(\varepsilon | X, G_1, G_2) = 1$  and the conditional mean function of  $Y$  given on  $X, G_1$ , and  $G_2$  is  $E(Y | X, G_1, G_2) = m(X, G_1, G_2) = \alpha_0 + \alpha_1(X) + \beta_1(X) G_1 + \beta_2(X) G_2$ . The same model is fitted separately for each marker, followed by multiple testing corrections. The two-parameter model given in (4) is not only biologically more meaningful than the one-parameter model given in (3), but also is statistically attractive since it is invariant to allele coding (i.e., whether code  $AA$  as 1 or code  $aa$  as 1 for variable  $G_1$ ).

**Remark:** By assuming specific expressions for  $\beta_1(\cdot)$  and  $\beta_2(\cdot)$ , model (4) would become a parametric model. For example, by letting  $\beta_1(X) = \beta_1 + \beta_3 X$ , and  $\beta_2(X) = \beta_2 + \beta_4 X$ , where  $\beta_1, \beta_2, \beta_3$ , and  $\beta_4$  are constants, model (4) can be written as

$$Y = \alpha_0 + \alpha_1 X + \beta_1 G_1 + \beta_2 G_2 + \beta_3 X G_1 + \beta_4 X G_2 + \sigma(X) \varepsilon, \quad (5)$$

which is a linear regression model with both main effects of  $X$  and  $(G_1, G_2)$  as well as their interaction effects (denoted hereafter as LM-I). If we assume a homogeneous residual variance, this is the commonly applied linear regression model for testing  $G \times E$  interaction which reduces to model (1) if only additive effect is considered. If we impose a constant structure on  $\beta_1(X)$  and  $\beta_2(X)$ , i.e.,  $\beta_1(X) = \beta_1$  and  $\beta_2(X) = \beta_2$ , then model (4) is reduced to

$$Y = \alpha_0 + \alpha_1 X + \beta_1 G_1 + \beta_2 G_2 + \sigma(X) \varepsilon, \quad (6)$$

which is a linear regression model without the interaction terms (denoted hereafter as LM). Therefore, the traditional linear regression model for testing  $G \times E$  interaction is a special case of model (4).

Although, their properties are very well established, the conventional parametric approaches are infeasible in this case, since the functional forms of  $\beta_1(\cdot)$  and  $\beta_2(\cdot)$  are unknown to us due to the complexity of the underlying interaction mechanism. Any mis-specification of the model would lead to uncertainty estimates and low power (see simulation). By relaxing the linear assumption for the coefficients  $\beta_1(X)$  and  $\beta_2(X)$ , model (4) has much flexibility to capture the non-linear penetrance of a genetic variant under different environmental stimuli, thus ensures the power of the proposed VC model in detecting non-linear  $G \times E$  interactions. In this paper, we apply the B-spline smoothing technique to estimate  $\beta_1(\cdot)$  and  $\beta_2(\cdot)$ , which solves only one least squares problem to get the estimators. The great advantages of B-spline estimation are simple implementation and fast computation.

As in most works on nonparametric smoothing, estimation of the functional coefficients  $\beta_1(\cdot)$  and  $\beta_2(\cdot)$  is conducted on a compact interval  $[a, b]$ . In this paper, we denote the space of  $p$ -th order smooth function on  $[a, b]$  as  $C^{(p)}[a, b] = \{g \mid g^{(p)} \in C[a, b]\}$ , and  $C[a, b]$  is the space of continuous functions on  $[a, b]$ .

We make the following assumptions on the functional coefficient model.

- (A1) The marginal density  $f(\cdot)$  of  $X_i$  is bounded away from zero and twice continuously differentiable on  $[a, b]$ .
- (A2)  $\sigma^2(\cdot) = \text{Var}(Y_i \mid X_i = x, G)$  is bounded away from 0 and  $\infty$ .
- (A3)  $E[\exp(t\varepsilon)]$  is bounded for  $|t|$  small enough.
- (A4)  $\beta_k(x) \in C^{(p)}[a, b]$ , for  $p \geq 1$ ,  $k = 1, 2$ .

Assumptions (A1)-(A3) are identical with (A1), (A4) and (A5) in Härdle and Mammen

(1993), while Assumption (A4) is the same as (A1) in Wang and Yang (2009). All are typical assumptions for nonparametric regression.

## 2.2 Parameter estimation

Given a random sample  $\{(X_i, G_i, Y_i)\}_{i=1}^n$  from model (4), the polynomial spline modeling will be adopted to estimate  $\beta(\cdot)$ . Let  $\mathcal{F}_n$  be the space of polynomial splines of order  $p \geq 1$ .

We introduce a knot sequence with  $N_n$  interior knots

$$k_{-(p-1)} = \dots = k_{-1} = k_0 = a < k_1 < \dots < k_N < b = k_{N+1} = \dots = k_{N+p},$$

where  $N \equiv N_n$  increases when sample size  $n$  increases, and the precise order is given in Assumption (A5). Then  $\mathcal{F}_n$  consists of functions  $\varpi$  satisfying (i)  $\varpi$  is a polynomial of degree  $p - 1$  on each of the subintervals  $I_s = [k_s, k_{s+1})$ ,  $s = 0, \dots, N_n - 1$ ,  $I_{N_n} = [k_{N_n}, b]$ ; and (ii) for  $p \geq 2$ ,  $\varpi$  is  $p - 2$  time continuously differentiable on  $[a, b]$ . Let  $J_n = N_n + p$ , where  $N_n$  is the number of interior knots and  $p$  is the spline order. We define the normalized B-spline basis as  $\{B_s : 1 \leq s \leq J_n\}^T$  as given in Wang and Yang (2009). Equally-spaced knots are used in this article for simplicity. The distance between neighboring interior or boundary knots is  $h = h_n = (b - a)(N_n + 1)^{-1}$ . For positive numbers  $a_n$  and  $b_n$  and for  $n \geq 1$ , let  $a_n \sim b_n$  mean that  $\lim_{n \rightarrow \infty} a_n/b_n = c$ , where  $c$  is some nonzero constant. The number of interior knots satisfy Assumption (A5) below.

(A5) The number of interior knots  $N = N_n \sim n^{1/(2p+1)}$ , i.e.,  $c_N n^{1/(2p+1)} \leq N \leq C_N n^{1/(2p+1)}$

for some positive constants  $c_N$  and  $C_N$ .

For each marker, and  $k = 1, 2$ , the coefficients  $\beta_k(x)$  is estimated by  $\hat{\beta}_k(x) \equiv \sum_{s=1}^{J_n} \hat{\lambda}_{s,k} B_s(x)$  where the coefficients  $\left\{ \left( \hat{\alpha}_0, \hat{\alpha}_1, \hat{\lambda}_{s,1}, \hat{\lambda}_{s,2} \right)_{1 \leq s \leq J_n} \right\}^T$  are solutions of the following least squares problem

$$\underset{\{(\alpha_0, \alpha_1, \lambda_{s,1}, \lambda_{s,2})_{1 \leq s \leq J_n}\} \in R^{2J_n+2}}{\operatorname{argmin}} \sum_{i=1}^n \left\{ Y_i - \alpha_0 - \alpha_1 X_i - \sum_{k=1}^2 \sum_{s=1}^{J_n} \lambda_{s,k} B_s(X_i) G_{ki} \right\}^2. \quad (7)$$



### 2.3 Number of knots $N$ and spline order $p$ selection

One of the challenging issues in the proposed semi-parametric modeling is to select appropriate knots and spline order to avoid over- and under-smoothing. For simplicity, we assume the same spline basis  $\{B_s : 1 \leq s \leq J_n\}^T$  to approximate the coefficient functions  $\beta_1(x)$  and  $\beta_2(x)$ , even though the spline order and knots can be different for the two functions. We use the BIC criteria to select the ‘optimal’  $N$ , denoted by  $\hat{N}^{opt}$ , from  $[\max([0.5n^{1/(2p+1)}], 1), [1.5n^{1/(2p+1)}]]$ , where  $[a]$  is an integer part of  $a$ , and the ‘optimal’ order  $p$  for the spline basis, denoted by  $\hat{p}^{opt}$ , from (3, 4), which minimize the BIC value  $\text{BIC}(N, p) = \log(\hat{\sigma}^2) + (N + p) \log(n)/n$ , where  $\hat{\sigma}^2 = \sum_{i=1}^n \{Y_i - \hat{m}_F(X_i, G_{1i}, G_{2i})\}^2/n$ .  $p = 3$  and 4 are the orders for quadratic and cubic splines, respectively. A search for the combination of hypothesized values for  $N$  and  $p$  can be done and the values of  $N$  and  $p$  corresponding to the minimum of the BIC values are the ‘optimal’ results.

### 2.4 Hypothesis testing

Before we test possible  $G \times E$  interaction, the first step is to assess whether a genetic marker is associated with a phenotype. This can be done by formulating the hypotheses

$$\begin{cases} \mathcal{H}_0 : \beta_1(\cdot) = \beta_2(\cdot) = 0 \\ \mathcal{H}_1 : \text{at least one functional coefficient is not zero} \end{cases} \quad (8)$$

If the null is rejected, then we test significance of the additive effect ( $G_1$ ) and the dominance effect ( $G_2$ ), by formulating the hypotheses

$$\begin{cases} \mathcal{H}_0^{11} : \beta_1(\cdot) = 0 \\ \mathcal{H}_1^{11} : \beta_1(\cdot) \neq 0 \end{cases}, \text{ and } \begin{cases} \mathcal{H}_0^{12} : \beta_2(\cdot) = 0 \\ \mathcal{H}_1^{12} : \beta_2(\cdot) \neq 0 \end{cases} \quad (9)$$

When the null in (8) is rejected, we then test if the coefficient functions  $\beta_1(X)$  and

$\beta_2(X)$  in model (4) are varying or not. The hypotheses for this test are formulated by

$$\begin{cases} \mathcal{H}_0^2 : \beta_k(\cdot) = \beta_k, \text{ for } k = 1, 2 \\ \mathcal{H}_1^2 : \text{at least one functional coefficient is not a constant} \end{cases} \quad (10)$$

where  $\beta_k$ ,  $k = 1, 2$ , are unknown constants, for the selected genetic markers from the first step. Under  $\mathcal{H}_0^2$ , the reduced model can be written as  $Y = \alpha_0 + \alpha_1 X + \beta_1 G_1 + \beta_2 G_2 + \sigma(X) \varepsilon$ , which implies that there is no  $G \times E$  interaction. Thus, hypothesis (10) is essentially a test for  $G \times E$  interaction. Upon rejecting the null, one can also proceed to test  $\mathcal{H}_0^3: \beta_1(X) = \beta_1 + \beta_3 X$  and  $\beta_2(X) = \beta_2 + \beta_4 X$ . Under  $\mathcal{H}_0^3$ , the reduced models can be written as  $Y = \alpha_0 + \alpha_1 X + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_1 X + \beta_4 G_2 X + \sigma(X) \varepsilon$ , a model commonly applied for assessing linear  $G \times E$  interaction. Rejecting the null implies non-linear  $G \times E$  interaction.

Note that the current model does not assume any specific distribution for the error term  $\varepsilon$ , thus there is no likelihood function for the data. Borrowing the idea from Härdle and Mammen (1993), we use the integrated squared deviation between the estimators denoted by  $\hat{m}_F(\cdot)$  and  $\hat{m}_R(\cdot)$  of  $m(X, G_1, G_2)$  for the full and reduced models as the test statistic, which would be  $\mathcal{T}_n = \sum_{i=1}^n \{\hat{m}_F(X_i, G_{1i}, G_{2i}) - \hat{m}_R(X_i, G_{1i}, G_{2i})\}^2 / n$ , where  $\{(X_i, G_{1i}, G_{2i}, Y_i), i = 1, \dots, n\}$  is a random sample of  $(X, G_1, G_2, Y)$ . For the superiority of  $\mathcal{T}_n$  over other goodness-of-fit tests, see the discussion in Härdle and Mammen (1993). The critical values are computed by the wild bootstrap method proposed in Härdle and Mammen (1993).

### 2.5 Wild bootstrap to assess statistical significance

Härdle and Mammen (1993) pointed out that the standard ways of bootstrapping, including the naive resampling method and the adjusted residual bootstrap, fail to compute critical values of  $\mathcal{T}_n$ , due to the reason that the bootstrapped statistic does not have the same limiting behavior, which would lead to very conservative tests. As a result they proposed the wild bootstrapping method, which is adopted in this paper. The coefficient functions are estimated by B-spline estimators. For the  $i$ -th observation, recall that  $\hat{m}_R(X_i, G_{1i}, G_{2i})$  and

$\widehat{m}_F(X_i, G_{1i}, G_{2i})$  are the estimators of  $m(X_i, G_{1i}, G_{2i})$  for the reduced and full model respectively. As discussed in Härdle and Mammen (1993), in order to mimic the *i.i.d.* structure of  $(X_i, G_{1i}, G_{2i}, Y_i)$ , we need to construct the bootstrap procedure so that  $E^*(Y_i^* | X_i^*, G_{1i}^*, G_{2i}^*) = \widehat{m}_R(X_i^*, G_{1i}^*, G_{2i}^*)$ , where  $\{(X_i^*, G_{1i}^*, G_{2i}^*, Y_i^*)\}_{i=1}^n$  is the bootstrap sample drawn from the set  $\{(X_i, G_{1i}, G_{2i}, Y_i)\}_{i=1}^n$ . For this purpose we define  $\widetilde{\varepsilon}_i = Y_i - \widehat{m}_F(X_i, G_{1i}, G_{2i})$  and construct  $\varepsilon_i^* = U_i \widetilde{\varepsilon}_i$ , where  $U_i$  is a two-point distributed random variable independent of  $(X_i, G_{1i}, G_{2i}, Y_i)$  satisfying  $U_i = 1/2 - \sqrt{5}/2$  with probability  $(1 + \sqrt{5}) / (2\sqrt{5})$ ,  $U_i = 1/2 + \sqrt{5}/2$  with probability  $1 - (1 + \sqrt{5}) / (2\sqrt{5})$ . By simple calculation, we obtain that  $E(\varepsilon_i^* | X_i, G_{1i}, G_{2i}) = 0$ ,  $E(\varepsilon_i^{*2} | X_i, G_{1i}, G_{2i}) = \widetilde{\varepsilon}_i^2$  and  $E(\varepsilon_i^{*3} | X_i, G_{1i}, G_{2i}) = \widetilde{\varepsilon}_i^3$ . Then we use  $(X_i, G_{1i}, G_{2i}, Y_i^* = \widehat{m}_R(X_i, G_{1i}, G_{2i}) + \varepsilon_i^*)$  as bootstrap observations and create  $\mathcal{T}^{*,W}$  like  $\mathcal{T}_n$  by the squared deviation between the coefficient estimators under  $\mathcal{H}_0$  and  $\mathcal{H}_1$ . From the Monte Carlo approximation of  $\mathcal{L}^*(\mathcal{T}_l^{*,W}) = \mathcal{L}(\mathcal{T}^{*,W} | (X_i, G_{1i}, G_{2i})_{i=1}^n)$ , then the p-value  $p_v$  is obtained by finding the  $(1 - p_v)$ th quantile  $\widehat{t}_v^W$  which satisfies  $\widehat{t}_v^W = \mathcal{T}_n$ . Multiple testing is then adjusted among the tests for all genetic markers.

### 3 Monte Carlo simulation

#### 3.1 Simulation design

Extensive simulations were conducted to compare the performance of the varying-coefficient and the parametric regression fits. Since a continuous environment measure (e.g., age, diet and body mass) in a genetic study generally follows a normal distribution, to mimic the real situation, we generated  $X_i$  from a normal distribution. Then we transformed  $X$  by  $Z = \Phi\{(X - \mu_X) / \sigma_X\}$  in order to make  $X$  distributed more evenly on each subinterval  $I_s$ , where  $\mu_X$  and  $\sigma_X$  are the mean and standard deviation of  $X$ , estimated by the sample mean and standard deviation, and  $\Phi(\cdot)$  is the cumulative distribution function for the standard normal. We then used the transformed  $Z$  to generate the B-spline basis. For  $k = 1, 2$ ,  $\beta_k(x)$

was estimated by  $\widehat{\beta}_k(x) \equiv \sum_{s=1}^{J_n} \widehat{\lambda}_{k,s} B_s[\Phi\{(x - \mu_X)/\sigma_X\}] = \sum_{s=1}^{J_n} \widehat{\lambda}_{k,s} B_s^*(x)$  where the coefficients  $\left\{ \left( \widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\lambda}_{1,s}, \widehat{\lambda}_{2,s} \right)_{1 \leq s \leq J_n} \right\}^T$  are solutions of the following least squares problem

$$\underset{\{(\alpha_0, \alpha_1, \lambda_{1,s}, \lambda_{2,s})_{1 \leq s \leq J_n}\} \in R^{2J_n+2}}{\operatorname{argmin}} \sum_{i=1}^n \left\{ Y_i - \alpha_0 - \alpha_1 X_i - \sum_{k=1}^2 \sum_{s=1}^{J_n} \lambda_{k,s} B_s^*(X_i) G_{ki} \right\}^2.$$

For a given a minor allele frequency (MAF)  $p_A$  and assuming Hardy-Weinberg equilibrium, SNP genotypes ( $AA$ ,  $Aa$ , and  $aa$ ) were simulated from a multinomial distribution with frequency  $p_A^2$  for  $AA$ ,  $2p_A(1 - p_A)$  for  $Aa$  and  $(1 - p_A)^2$  for  $aa$ . Then  $G_{1i}$  and  $G_{2i}$  were coded separately for genotype  $AA$  ( $G_{1i} = 1, G_{2i} = -1/2$ ),  $Aa$  ( $G_{1i} = 0, G_{2i} = 1/2$ ), and  $aa$  ( $G_{1i} = -1, G_{2i} = -1/2$ ). The random error term  $\varepsilon_i$  was simulated from  $N(0, 1)$ . Different sample sizes (i.e.,  $n = 200, 500, 1000$ ), and different heritability levels (i.e.,  $H^2 = 0.01, 0.03, 0.05$ ) were assumed. For a given genetic effect and a heritability level,  $\sigma(X_i)$  varies for different  $X_i$ , and detailed calculation can be found in the following sections. Data were simulated assuming different gene action modes and were subsequently analyzed by three models, i.e., the proposed VC model (denoted as VC), the linear regression model without interaction (denoted as LM), and the linear regression model with interaction (denoted as LM-I). The hypothesis testing (8) is conducted to assess the overall power of the effect of  $G_1$  and  $G_2$  on  $Y$ .

### 3.2 Responses generated from the null model

We first evaluated how well the false positives can be controlled for the models. Data were generated from the null model, i.e.,  $Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i$ , with  $\alpha_0 = 3$ ,  $\alpha_1 = 0.1$  and  $\varepsilon_i \sim N(0, 1)$ . Figure 1 shows that the type I error rates for the three models. The type I error rate for the VC model is around 0.68 for  $n = 200$ , which is a little inflated under the 0.05 nominal level. However with the sample size  $n$  increasing, it is getting close to 0.05, and end up at around 0.54 for  $n = 1000$ . The type I error rates for the other two models are

also reasonably close to the nominal significant level 0.05 under different sample sizes. These results indicate good false positive control of the VC model under large sample sizes.

[Figure 1 about here.]

### 3.3 Responses generated from the VC model

In this simulation, we generated the phenotype data assuming the following model

$$Y_i = \alpha_0 + \alpha_1 X_i + \beta_1(X_i) G_{1i} + \beta_2(X_i) G_{2i} + \sigma(X_i) \varepsilon_i$$

where  $\alpha_0 = 3.0$ ,  $\alpha_1 = 0.1$  and  $\beta_1(x)$  and  $\beta_2(x)$  were generated from the B-spline basis functions such that  $\beta_1(x) = \sum_{s=1}^4 \lambda_{1s} B_s(x)$  and  $\beta_2(x) = \sum_{s=1}^4 \lambda_{2s} B_s(x)$ , in which  $\lambda_{11} = -0.53$ ,  $\lambda_{12} = 0.31$ ,  $\lambda_{13} = -0.44$ ,  $\lambda_{14} = 0.50$ ,  $\lambda_{21} = -0.87$ ,  $\lambda_{22} = 0.71$ ,  $\lambda_{23} = -1.27$ , and  $\lambda_{24} = 1.15$ . These spline coefficients were calculated from (7) based on SNP 22265753 from a real data set (see Table 1). The variance function  $\sigma^2(x)$  was obtained by solving  $H^2 = V_G/(V_G + V_E)$ , where  $H^2$  is the heritability level;  $V_G(x) = \beta_1^2(x) \text{var}(G_1) + \beta_2^2(x) \text{var}(G_2) + 2\beta_1(x)\beta_2(x) \text{cov}(G_1, G_2)$  is the genetic variance in which  $\text{var}(G_1) = 2p_A(1 - p_A)$ ,  $\text{var}(G_2) = 1/4 \{1 - (2p_A - 1)^4\}$ , and  $\text{cov}(G_1, G_2) = 2p_A(1 - p_A)(2p_A - 1)$ ; and  $V_E = \sigma^2(x)$ . Simple algebra shows that  $H^2 = [1 + \sigma^2(x)/V_G(x)]^{-1}$ , which gives  $\sigma^2(x) = (1/H^2 - 1)V_G(x)$ . Assuming different heritability levels, i.e.,  $H^2 = 0.01, 0.03, 0.05$ , the phenotype  $Y_i$  can be generated assuming  $\varepsilon_i \sim N(0, 1)$ . As can be seen that the genetic variance is a function of the minor allele frequency, so does for the residual variance  $\sigma(X)$ . For a fixed minor allele frequency, the residual variance decreases as the heritability increases. Thus we expect high power under high  $H^2$  value. However, due to the way we defined the calculation of  $V_G$ , it is no longer true that  $\sigma(X)$  decreases as the MAF increases for a fixed  $H^2$  level. So the power no longer monotonically increases with the increase of the MAF as usually assumed in human genetic association studies. Based on the estimated frequency

( $p_A = 0.08$ ) of the SNP from the real data, we fixed the allele frequency and evaluated the power performance of the three methods under different heritability levels.

We checked the power for an association test (hypothesis (8)) with the test statistic defined as  $\mathcal{T}_{n,V} = \sum_{i=1}^n \{\hat{m}_F(X_i, G_{1i}, G_{2i}) - \hat{m}_R(X_i, G_{1i}, G_{2i})\}^2 / n$ , where  $\hat{m}_F$  and  $\hat{m}_R$  are the estimators of the conditional mean function  $E(Y | X, G_1, G_2)$  for the full and reduced model, respectively. Let  $P_V$  be the p-value obtained from  $B(= 10,000)$  wild bootstrap samples. We carried out  $M(= 1,000)$  repetitions and obtained the p-values, denoted by  $\{P_V^{(m)}, m = 1, \dots, M\}$ , then the power of a hypothesis test at a given  $\alpha (= 0.05)$  level was estimated by  $\hat{P}_{V,\alpha} = \hat{P}(P_V < \alpha) = \sum_{i=1}^M \mathbf{I}(P_V^{(m)} < \alpha) / M$ . The same simulated data sets were also analyzed by the LM and LM-I model, and the significance was assessed by the likelihood ratio test.

[Figure 2 about here.]

From Figure 2, one can see that the testing power increases as the sample size  $n$  and heritability level  $H^2$  increase under the three models. For a fixed genetic effect, large heritability level leads to small residual variance, and consequently leads to increased power. In a power comparison of the three models, it is clear that the VC model outperforms the other two models in all the cases. Since the linear model with interaction (LM-I) is closer to the VC model in structure, it achieves higher power than the linear model without interaction (LM). The simulation results clearly indicate the power of the VC model over the other two, given that the nature of the  $G \times E$  interaction is non-linear. When there is a strong non-linear penetrance effect of a variant, a miss-specification of an analytical model assuming a linear structure suffers tremendously from power loss.

### 3.4 Responses generated from the linear model without interaction

In the second simulation, phenotype data were generated assuming a linear model without interaction, i.e.,

$$Y_i = \alpha_0 + \alpha_1 X_i + \beta_1 G_{1i} + \beta_2 G_{2i} + \sigma_0 \varepsilon_i \quad (11)$$

where  $\alpha_0 = 3$ ,  $\alpha_1 = 0.1$ ,  $\beta_1 = 0.5$ , and  $\beta_2 = 0.3$ . This model assumes constant coefficients, corresponding to the null  $\mathcal{H}_0^2$ :  $\beta_k(\cdot) = \beta_k$ ,  $k = 1, 2$ , in hypothesis (10). The purpose of this simulation was to assess how robust the VC model is if the true model coefficient is not varying with  $X$  under finite sample size. Similarly as above,  $\sigma_0$  is obtained by solving  $H^2 = (1 + \sigma_0^2 / V_G)^{-1}$ , where  $V_G = \beta_1^2 \text{var}(G_1) + \beta_2^2 \text{var}(G_2) + 2\beta_1\beta_2 \text{cov}(G_1, G_2)$ , thus  $\sigma_0^2 = (H^{-2} - 1) V_G$ .

[Figure 3 about here.]

We used the likelihood ratio test to test:  $\mathcal{H}_0 : \beta_1 = \beta_2 = 0$ . The test statistic follows a 2 degrees of freedom (df) chi-square distribution. The simulation results were summarized in Figure 3. Results with  $p_A = 0.3$  were reported. As we expected, the power increases as the sample size and heritability increase for the three models. When the underlying true model is linear without interaction, the results obtained with the LM model has the highest power among the three fitted models. This is not surprising because optimal power is obtained when the data are fitted with the true model. However, the power differences between the VC and LM models are not as large as the ones observed above. As the heritability increases, the difference in power between the VC and LM models vanishes, especially under a large sample size. For example, the power for VC and LM is indistinguishable when the sample size increases from 200 to 500 under high heritability levels ( $H^2 = 0.03, 0.05$ ).

### 3.5 Responses generated from the linear model with interaction

In this simulation, phenotype data were generated assuming a linear model with interaction, i.e.,

$$Y_i = \alpha_0 + \alpha_1 X_i + \beta_1 G_{1i} + \beta_2 G_{2i} + \beta_3 X_i G_{1i} + \beta_4 X_i G_{2i} + \sigma(X) \varepsilon_i$$

where  $\alpha_0 = 3$ ,  $\alpha_1 = 0.1$ ,  $\beta_1 = 0.3$ ,  $\beta_2 = 0.3$ ,  $\beta_3 = 0.5$ , and  $\beta_4 = 0.5$ . The above model can be rearranged as

$$Y_i = \alpha_0 + \alpha_1 X_i + (\beta_1 + \beta_3 X_i) G_{1i} + (\beta_2 + \beta_4 X_i) G_{2i} + \sigma(X) \varepsilon_i$$

Thus,  $V_G(x) = (\beta_1 + \beta_3 x)^2 \text{var}(G_1) + (\beta_2 + \beta_4 x)^2 \text{var}(G_2) + 2(\beta_1 + \beta_3 x)(\beta_2 + \beta_4 x) \text{cov}(G_1, G_2)$ .

As a function of heritability described before,  $\sigma(x)$  can be calculated from  $V_G(x)$ , i.e.,  $\sigma^2(x) = (1/H^2 - 1) V_G$ . Thus, the heteroscedasticity of the residual variance is taken into account in the simulation.

[Figure 4 about here.]

We used the likelihood ratio test to test:  $\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ . The test statistic follows a 4 df chi-square distribution. The simulation results were summarized in Figure 4. Again, results for  $p_A = 0.3$  were reported. As we expected, the highest power was observed when data were fitted with the true model assuming interaction. Low power was observed when the heritability level was low. As the sample size and heritability increase, the power for the three models increases and the difference between the VC model and the other two diminishes. We also observed that the power of the VC model is closer to the LM-I model than the LM model does. This is what we expected as the VC model is closer to the LM-I in structure.

In summary, we conclude that: (1) When the underlying true interaction model is non-linear, the proposed VC model has the highest power among the three. The other two parametric linear models suffer tremendously from power loss; (2) When the underlying true



model is linear without or with interaction, the linear model assuming interaction or no interaction has the best power. However, as the sample size and heritability level increase, the power difference between the VC model and the other two decreases significantly. ; and (3) In real data analysis, the VC model can not substitute the other two models before we know the true functional effect. We can first do a hypothesis testing to check if the coefficient functions  $\beta_k(X)$ ,  $k = 1, 2$ , are constant or linear in  $X$ , then apply the optimal model in the analysis. The non-linear VC model would be the choice if the constant or linear function is rejected. Otherwise, a linear model is suggested, especially when sample size is small. The final results (with p-values) should be combined.

#### 4 Real data analysis

To show the utility of the VC model in detecting  $G \times E$  interaction, we apply the method to a real data set. The data contain 1536 new born babies, recruited through the Department of Obstetrics and Gynecology at Sotero del Rio Hospital in Puente Alto, Chile. Total 648 single nucleotide polymorphisms (SNPs) covering 189 unique genes were left after eliminating SNPs with minor allele frequency less than 0.05 and those departure from Hardy-Weinberg equilibrium. When fitting to the VC model, we found that the spline design matrix could be exactly singular when there are extremely unbalanced genotype distributions, especially when only two genotypes categories were present for a SNP. Thus, we eliminated additional 143 SNPs and only 505 SNPs were included in our analysis. Phenotypes were dichotomized as small for gestational age (SGA) or large for gestational age (LGA) depending on the babies' birth weight and the mother's gestational age. The initial study were designed to identify genetic risk factors associated with SGA or LGA. We took the original birth weight (kg) measure as the response and merged the two data sets together to form one data set for an analysis.

It is postulated that baby's birth weight might be related to mother's body mass index

(MBMI). When a baby resides inside of its mother's womb, the environmental conditions are defined through its mother, for instance, mother's age and obesity condition (measured by MBMI). Under different environmental stimuli (e.g., MBMI), fetus carrying the same genes might trigger different responses, which consequently leads to different birth weights. This is due to the complex interaction between a mother's obesity condition and fetus' genes. With the combined data, we were interested in identifying genetic factors that can explain the normal variation of birth weight, and if any, influenced by MBMI.

To explore the non-linear relationship between birth weight (BW) and genetic factors, we let the coefficients of  $G_1$  and  $G_2$  evolve with MBMI. We applied model (3) to the data set by fitting each SNP as  $(G_1, G_2)$  and letting MBMI and BW be  $X$  and  $Y$  respectively, and then estimate  $\beta_k(\cdot)$  by B-spline estimators  $\hat{\beta}_k(\cdot)$ ,  $k = 1, 2$ . Extreme observations in both  $X$  and  $Y$  (with the  $3 \times$  inter-quartile range criterion) were removed. Using the BIC criterion described in Section 3, we found  $\hat{N}^{opt} = 1$  and  $\hat{p}^{opt} = 3$  (order for quadratic splines) and they were used to fit all SNPs. Defining the vector  $\mathbf{R} = \{R_j\}_{1 \leq j \leq n}^T = \{(Y_i - \hat{m}(X_i, G_i))^2\}_{1 \leq i \leq n}^T$ , the estimator of  $\sigma^2(x)$  can be obtained by  $\hat{\sigma}^2(x) = \sum_{s=1}^{J_n} \hat{\lambda}_s B_s^*(x)$ , where the coefficients  $\{\hat{\lambda}_1, \dots, \hat{\lambda}_{J_n}\}^T$  are solutions of the least squares problem:

$$\{\hat{\lambda}_1, \dots, \hat{\lambda}_{J_n}\}^T = \underset{\{\hat{\lambda}_1, \dots, \hat{\lambda}_{J_n}\} \in R^{J_n}}{\operatorname{argmin}} \sum_{i=1}^n \left\{ R_j - \sum_{s=1}^{J_n} \lambda_s B_s^*(X_{ij}) \right\}^2.$$

Following the procedure described in Section 2.4, we first tested the null hypothesis  $\mathcal{H}_0 : \beta_1(\cdot) = \beta_2(\cdot) = 0$ , and obtained p-values for the 648 SNP markers. To obtain the p-values, the Monte Carlo approximation to  $\mathcal{L}^*(\mathcal{T}^{*,W})$  was performed by  $B = 10,000$  repetitions using the wild bootstrap algorithm. Data were also analyzed by fitting a linear model without interaction (LM) and with interaction (LM-I). The results were tabulated in Table 1.

[Table 1 about here.]

The first three columns list the SNP ID, the gene and location each SNP belongs to. When we applied the false discovery rate (FDR) control method (Benjamini and Hochberg,

1995), only two SNP showed statistical significance (indicated by \* in Table 1). To illustrate the method, we also listed SNPs with p-values that are less than 0.005. The p-values for the overall genetic effect tests, i.e.,  $\mathcal{H}_0: \beta_1(\cdot) = \beta_2(\cdot) = 0$ , are given in the column denoted by P\_VC, P\_LM and P\_LMi when fitting the data with the VC, LM and LM-I models, respectively. The upper panel shows the results with the VC model. Testing constant coefficients ( $\mathcal{H}_0^2$ ) indicates that the function of these SNPs does vary across MBMI (P\_const<0.05). Further tests ( $\mathcal{H}_0^3$ ) for linear relationship show that the function of these SNPs do not follow a linear structure either. Therefore, it is not surprising that the p-values obtained with the VC model are all smaller than the ones obtained by fitting the LM and LM-I models.

SNPs with p-values less than 0.005 when fitting the LM model are listed in the middle panel of the table. Testing constant coefficient indicates that the functional coefficients for these three SNPs do not vary across MBMI (P\_const> 0.05). Thus, we observed the smallest p-values for the three SNPs when they were fitted with the LM model. The bottom panel in the table lists nine SNPs with p-values less than 0.005 when fitting the LM-I model (P.LMi< 0.005). Testing constant coefficient indicates that the functional coefficients for these nice SNPs do vary across MBMI (P\_const< 0.05). Further tests ( $\mathcal{H}_0^3$ ) show that the functional coefficients for the nine SNPs follow a linear structure (P\_linear>0.05). Thus the p-values (P.LMi) fitted with the LM-I model are the smallest among the three models. Testing linear interaction indicates that the nine SNPs do have strong interaction effects (P\_i<0.05). In summary, the real data analysis results are consistent with the simulation results in which optimal p-value is always obtained by fitting the data with the “true” model. If we only fit the data with a regular linear model with or without interaction, we could potentially miss the ones detected by the VC model, especially those with large p-values tested with the LM and LM-I models. In checking the function of the SNPs, some of those are growth factors which

are directly related to fetal growth, for example, platelet-derived growth factor B (PDGFB) and fibroblast growth factor 4 (FGF4). FGF4 is essential for mammalian embryogenesis and fetal growth (Lamb and Rizzino, 1998). SNP 634043245 in Exon 3 located in FGF4 was also identified by a different model showing a strong dominance effect on small for gestational age along with maternal body weight when searching for genetic conflict effect (Li et al. 2009).

To further demonstrate the performance of the VC model, we picked SNP 11575857 located in gene PDGFB as an example. Figure 5A plots the fitted baby's birth weight (in kg) against mother's body mass index (MBMI) for individuals carrying different genotypes. The three curves correspond to the fitted BW for three different genotypes. The sample mean is indicated by the dashed straight line. The minor allele for this SNP is T and the estimated MAF is 0.1. From the fitted plot, we can see the non-linear interaction effect between this SNP and MBMI on infant's birth weight. When MBMI is low, infants carrying genotype CC have low birth weight, but not for those carrying the other two types of genotype. As MBMI increases, mother's body size has a positive effect on infant's birth weight, so we saw a slightly increasing trend for infant birth weight. However, infants carrying different genotypes show a clearly different response pattern on birth weight corresponding to the increase of MBMI. For example, infants carrying genotype TT show a sharp increase in their body weight compared to other two genotypes as MBMI passing 25. So mother's obesity condition triggers a stronger effect on TT genotypes than the other two genotypes.

[Figure 5 about here.]

Figure 5B plots the heritability estimation under different mother's BMI conditions. The plot also shows the non-linear penetrance of the variant under different MBMI conditions. Strong penetrance effects (corresponding to large  $H^2$  values) are observed when MBMI is between 25-30. The genetic effect (penetrance) tends to stabilize when MBMI reaches 35. This result fits to our intuition as we do not expect a fetus grow unlimited when mother's body

size increases. If the phenotype of interest is a disease status measurement, prevention efforts should be geared toward those environment conditions corresponding to large heritability estimate.

The spline estimators  $\widehat{\beta}_k(\cdot)$  of the coefficient functions  $\beta_k(\cdot)$ ,  $k = 1, 2$  are plotted in Figure 5C and 5D. It is clearly seen that  $\widehat{\beta}_k(x)$ ,  $k = 1, 2$ , does vary across MBMI. The additive effect  $\beta_1(X)$  shows a quadratic pattern and levels off as MBMI passes 33. This implies that the additive effect of this SNP variant approaches a limit for obese mothers (MBMI > 33), so does for the dominance effect but with a more varying pattern of effect under low MBMI. Due to the non-linear penetrance effect of this SNP under different environment stimuli (measured by mother's obese condition), this SNP could be missed if we fitted the data with the traditional linear interaction model. This example demonstrates the advantage of the VC model in the identification of important genetic variants with non-linear penetrance under different environment stimuli.

## 5 Discussion

The natural variation of a quantitative phenotype is not only determined by the inherited genetic factors, but also can be explained by how sensitive a genetic factor responds to environmental stimuli. Gene environment interaction, the genetic control of sensitivity to environment, plays a pivotal role in determining trait variations. In humans, most diseases results from a complex interaction between an individual's genetic blueprint and the associated environmental condition. For example, type II diabetes and cardiovascular disease are often due to the complex interaction between an individual's genes and obesity condition. The more we learn about how genes interact with environment in determining trait variations and disease risks, the more we can achieve in prevention and treatment of illnesses.

The importance of G×E interaction in human disease has been historically recognized (e.g., Costa and Eaton, 2006). Many statistical methods have been proposed to target

$G \times E$  interaction. In this work, we relaxed the linear  $G \times E$  interaction assumption, and proposed a new method considering non-linear  $G \times E$  interaction. We focused our attention on environment with continuous measurement (e.g., dietary intake, obesity condition and the amount of addictive substances). We adopted the well-known varying coefficient model into a genetic mapping framework and proposed to estimate the functional coefficient by the non-parametric B-spline technique. The asymptotic property of the non-parametric estimator has been established. The superior performance of the VC model in detecting non-linear  $G \times E$  interaction has been demonstrated with extensive Monte Carlo simulations. When the genetic contribution to the variation of a phenotype varies largely across environmental conditions, the proposed VC model achieves the optimal power compared to models assuming constant or linear coefficient.

Although in theory, the B-spline estimator converges to the true underlying function, depending on various factors, the VC model may not achieve the optimal power when the true function is constant or linear. The effect of sample size and heritability level on testing power was shown by simulations. In real data analysis, often the heritability level is unknown before we fit a model. Thus, it is necessary to conduct a hypothesis test to assess the true underlying functional coefficient. If the underlying coefficient is constant or linear, we recommend investigators to fit a constant coefficient or linear interaction model as the VC model suffers from power loss, especially under low heritability level. Results obtained by fitting different models can then be combined followed by multiple testing corrections.

We applied the method to a real data set to identify genetic factors interacting with mother's MBI to explain the normal variation of baby's birth weight. We adopted a two-parameter model which is biologically more attractive than a one-parameter model. We found a few SNPs showing non-linear penetrance across different environmental stimuli (different MBMI levels) (see Table 1). Even though only two SNPs showed statistical significance

after multiple testing adjustments following the FDR procedure (Benjamini and Hochberg, 1995), we still found a few others with relatively strong signals ( $p\text{-value} < 0.005$ ). Based on the results from simulation and real data analysis, we conclude that the VC model cannot completely substitute the linear parametric model in  $G \times E$  analysis. In real data analysis, our recommendation is to do a hypothesis test first to assess the functional form of the coefficient, then fit the appropriate model. The final results should be combined.

In this study, we focused on continuous quantitative phenotype. Extension to other type of phenotype such as a binary disease phenotype or phenotype measured as count is straightforward. A generalized linear model framework can be adopted and a link function can be chosen depending on the underlying phenotypic distribution. In human genetic studies, often a binary disease phenotype is collected such as in a case-control study. When a binary phenotype is considered, the estimation and inference procedure developed in this work can not be directly applied. Such investigation will be considered in our future work. The computational code written in R for implementing the work is available upon request.

#### ACKNOWLEDGEMENTS

This work was supported by NSF awards DMS-0706518 (Yang) and DMS-0707031 (Cui), and by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, NIH, DHHS. The computation of the work is supported by Revolution R (<http://www.revolutionanalytics.com/>).

#### REFERENCES

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

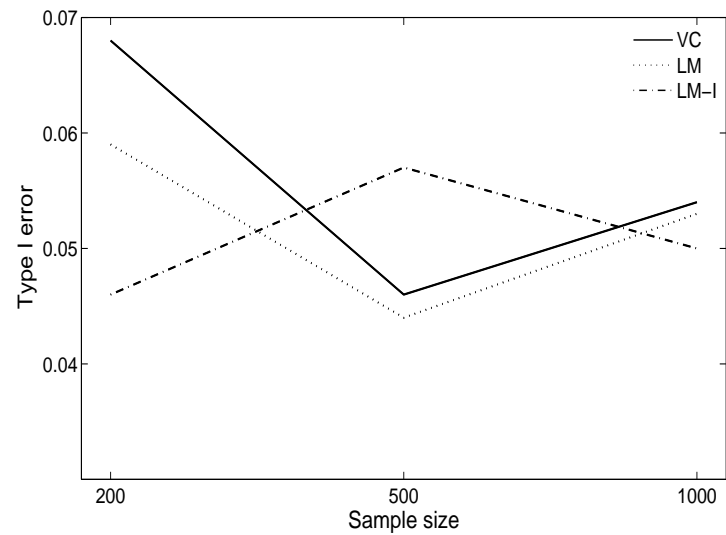
- Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, Springer-Verlag, New York.
- Chatterjee, N. and Carroll, R. J. (2005), “Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies,” *Biometrika*, 92, 399–418.
- Costa, L. G. and Eaton, D. L. (2006), *Gene-Environment Interactions: Fundamentals of Ecogenetics*, Hoboken, NJ: John Wiley & Sons.
- Cai, Z., Fan, J. and Li, R. (2000), “Efficient estimation and inferences for varying-coefficient models,” *Journal of American Statistical Association*, 95, 888–902.
- Caspi, A. and Moffitt, T. E. (2006), “Gene-environment interactions in psychiatry: joining forces with neuroscience,” *Nature Reviews Neuroscience*, 7, 583–590.
- Chiang, C. T., Rice, J. A. and Wu, C. O. (2001), “Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables,” *Journal of American Statistical Association*, 96, 605–619.
- Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A. and Poulton, R. (2003), “Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene,” *Science*, 301, 386–389.
- de Boor, C. (2001), *A Practical Guide to Splines*, Springer-Verlag, New York.
- DeVore, R. A. and Lorentz, G. G. (1993), *Constructive Approximation*, Springer-Verlag, New York.
- Falconer, D. S. (1952). The problem of environment and selection. *The American Naturalist*, 86, 293–298.
- Fan, J. and Zhang, W. (1999), “Statistical estimation in varying coefficient models,” *The Annals of Statistics*, 27, 1491–1518.
- Fan, J. and Zhang, W. (2000), “Simultaneous confidence bands and hypothesis testing in



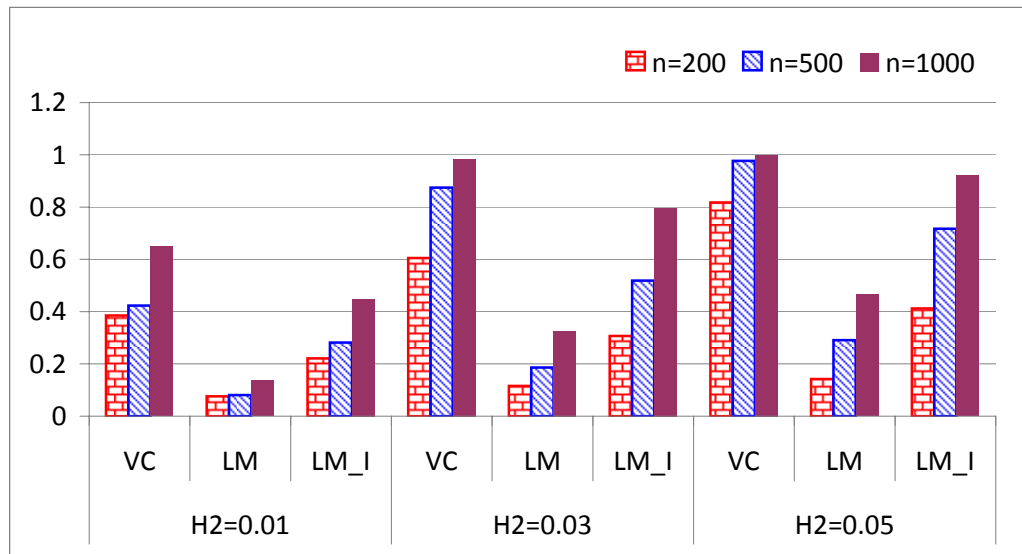
- varying-coefficient models," *Scandinavian Journal of Statistics*, 27, 715–731.
- Guo, S. W. (2000), "Gene-environment interaction and the mapping of complex traits: some statistical models and their implications," *Human Heredity*, 50, 286–303.
- Härdle, W. and Mammen, E. (1993), "Comparing nonparametric versus parametric regression fits," *The Annals of Statistics*, 21, 1926–1947.
- Hahn, L. W, Ritchie, M. D., Moore, J. H. (2003), "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, 19, 376–382.
- Hoover, D., Rice, J., Wu, C. O. and Yang, L. (1998), "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika*, 85, 809–822.
- Huang, J., Wu, C. and Zhou, L. (2004), "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," *Statistica Sinica*, 14, 763–788.
- Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. and Gauderman, W. J. (2007), "Exploiting gene-environment interaction to detect genetic associations," *Human Heredity*, 63, 111–119.
- Lamb, K., and Rizzino, A. (1998), Effects of differentiation on the transcriptional regulation of the FGF-4 gene: Critical roles played by a distal enhancer. *Molecular Reproduction and Development*, 51, 218–224.
- Li, S.Y., Lu, Q., Fu, W., Romero, R., and Cui, Y. (2009), "A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy". *Statistical Applications in Genetics and Molecular Biology* Vol. 8m Iss. 1, Article 45.
- Maity, A., Carroll, R. J., Mammen, E. and Chatterjee, N. (2009), "Testing in semiparametric models with interaction, with applications to gene-environment interactions," *Journal of the Royal Statistical Society, Series B*, 71, 75–96.
- Peacock, M., Turner, C. H., Econs, M. J. and Foroud, T. (2002), "Genetics of osteoporosis,"

*Endocrine Reviews*, 23, 303–326.

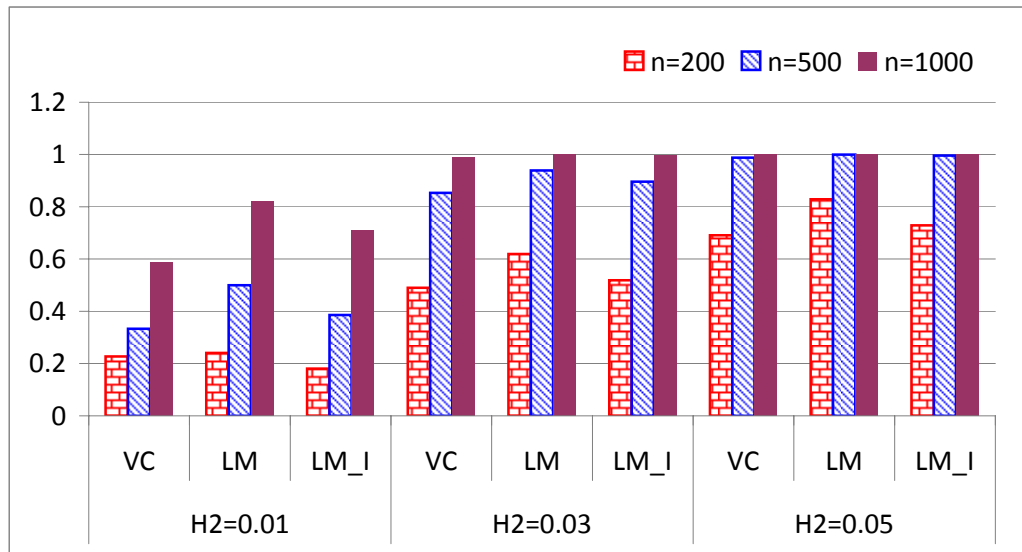
- Ulrich, C. M., Kampman, E., Bigler, J., Schwartz, S. M., Chen C, Bostick, R., Fosdick, L., Beresford, S., Yasui, Y. and Potter, J. (1999), “Colorectal adenomas and the C677T MTHFR polymorphism: evidence for gene-environment interaction,” *Cancer Epidemiology, Biomarkers & Prevention* 8, 659–668.
- Wang, L. and Yang, L. (2007), “Spline-backfitted kernel smoothing of nonlinear additive autoregression model,” *The Annals of Statistics*, 35, 2474–2503.
- Wang, J. and Yang, L. (2009), “Polynomial spline confidence bands for regression curves,” *Statistica Sinica*, 19, 325–342.
- Xia, Y. and Li, W. K. (1999), “On the estimation and testing of functional-coefficient linear models,” *Statistica Sinica*, 3, 735–757.
- Zhou, X. and You, J. (2004), “Wavelet estimation in varying-coefficient partially linear regression models,” *Statistics & Probability Letters*, 68, 91–104.



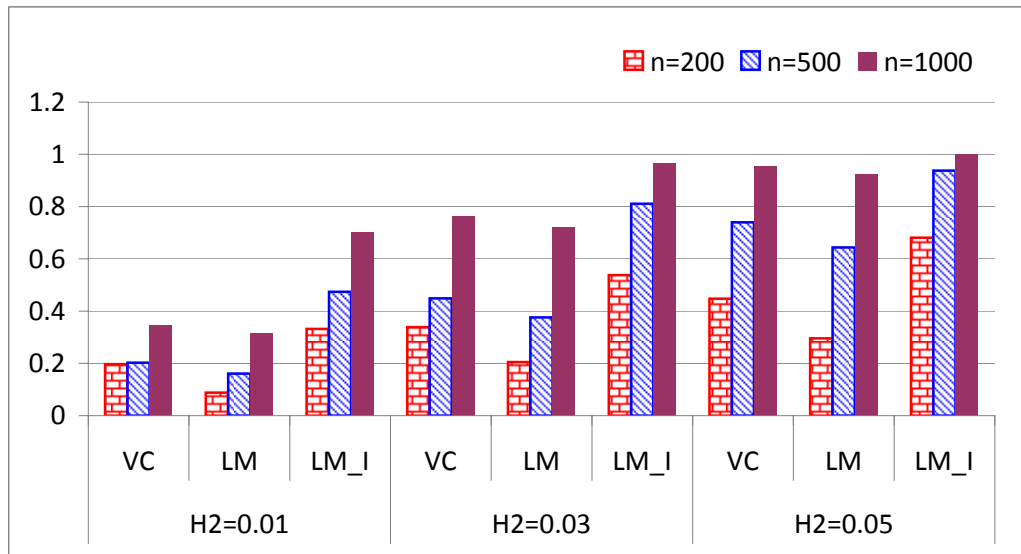
**Figure 1.** The type I error plot under different sample sizes for the three methods. Data were generated under the null model of no genetic effect and were analyzed with the varying-coefficient model (VC), the linear model without interaction (LM) and the linear model with interaction (LM-I).



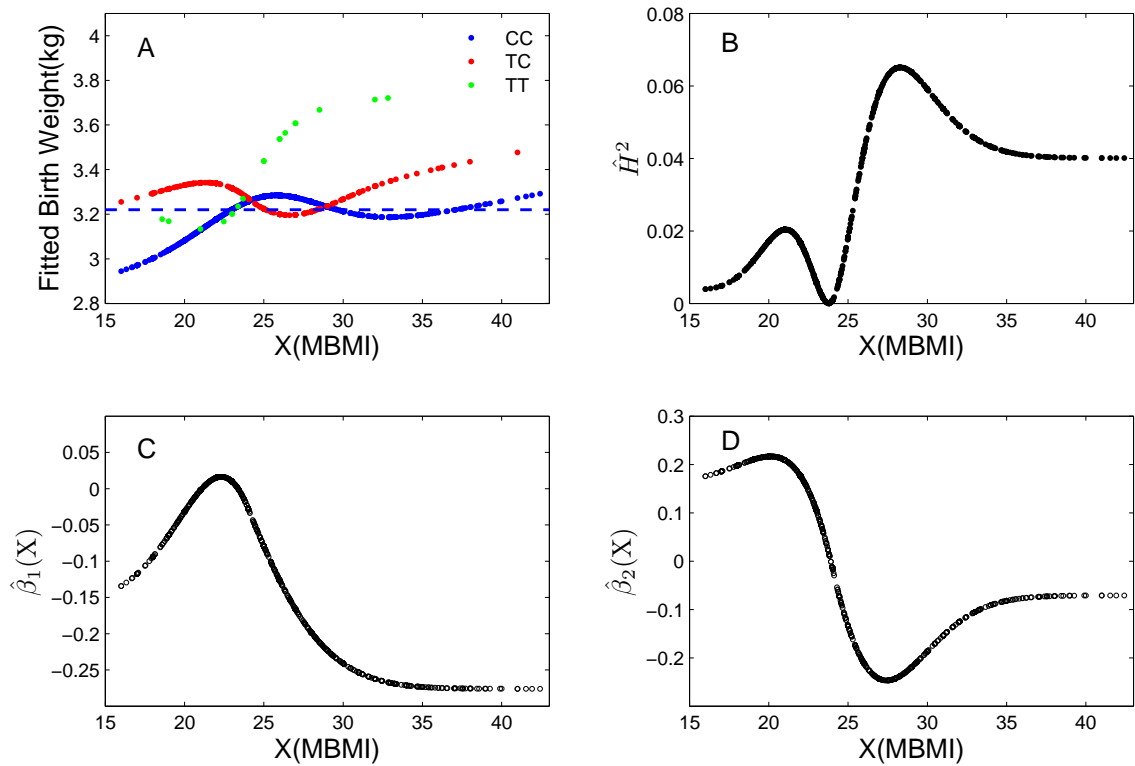
**Figure 2.** The power plot under different sample sizes and heritability levels for the three methods. Data were generated with the VC model and were analyzed with the VC model, the linear model without interaction (LM) and the linear model with interaction (LM-I). MAF is assumed to be 0.1.



**Figure 3.** The power plot under different sample sizes and heritability levels for the three methods (VC, LM and LM-I), where data were generated with the LM model. MAF is assumed to be 0.3.



**Figure 4.** The power plot under different sample sizes and heritability levels for the three methods (VC, LM and LM-I), where data were generated with the LM-I model. MAF is assumed to be 0.3.



**Figure 5.** The plot shows: (A) the fitted birth weight (kg) for the three genotype categories; (B) the estimated heritability value  $\hat{H}^2$  (C) the VC function  $\hat{\beta}_1(X)$ ; and (D) the VC function  $\hat{\beta}_2(X)$ , against mother's body mass index (MBMI) for SNP 11575857 located in gene PDGFB. The horizontal dashed line in (A) denotes the sample mean.

**Table 1**

The lists of SNPs with p-values less than 0.005 when fitting the data with three different models (VC, LM and LM-I).

SNP ID	GeneName	Location	P_VC	P_const	P_linear	P_LM	P_LMi	P_i
fitted with VC model								
22265753*	PLAT	intron 6a	1E-05	<1E-05	3E-05	0.8827	0.0823	0.0182
11575857	PDGFB	intron 2	0.0008	0.0034	0.0056	0.0655	0.0471	0.1237
1465147	ANG	exon 1	0.0013	0.0041	0.0156	0.0930	0.0477	0.0883
632238221	F12	intron 10	0.0018	0.0071	0.0074	0.0234	0.0070	0.0369
634043245	FGF4	exon 3	0.0019	0.0046	0.0016	0.0808	0.1452	0.4070
634085602	HLA-G	exon 3	0.0020	0.0120	0.0239	0.0089	0.0029	0.0360
28138476	COL1A2	intron 28	0.0024	0.0103	0.0038	0.0182	0.0222	0.1811
22185434	FLT4	intron 13	0.0027	0.0017	0.0452	0.4106	0.0090	0.0028
633878892	IL12RB1	intron 15	0.0027	0.0011	9E-05	0.8376	0.5946	0.2968
28139054	COL1A2	intron 19	0.0038	0.0254	0.0381	0.0148	0.0066	0.0544
4476882	TLR9	promoter	0.0048	0.0243	0.1250	0.0061	0.0053	0.1040
fitted with LM model								
619221153	IL1B	exon 5	0.0053	0.1818	-	0.0006	-	-
632272004	IL1A	intron 6	0.0213	0.629	-	0.0007	-	-
8949486	CETP	intron 1	0.0056	0.2073	-	0.0020	-	-
fitted with LM-I model								
3557434*	IL9	intron 4	0.0024	0.0477	0.4773	0.0009	4.9E-05	0.0039
629158954	IL1B	promoter	0.0011	0.0019	0.2477	0.0899	0.0005	0.0005
639694765	MMP8	exon 6	0.0014	0.0009	0.1249	0.4743	0.0009	0.0002
632336450	F5	exon 13	0.0038	0.0237	0.0965	0.0178	0.0032	0.0201
639707157	IL1RN	intron 2	0.0314	0.0136	0.4848	0.9035	0.0041	0.0005
614103201	COL4A2	intron 37	0.0154	0.013	0.1005	0.2072	0.0044	0.0025

P\_VC is the p-value for testing hypothesis (8); P\_const is the p-value for testing hypothesis (10); P\_linear is the p-value for testing linear coefficient ( $\mathcal{H}_0^3$ ); P\_LM is the p-value for testing  $\mathcal{H}_0: \beta_1 = \beta_2 = 0$  for fitting a linear model without interaction (model (11)), a 2 df likelihood ratio test; P\_LMi is the p-value for testing a genetic effect when fitting a linear model with interaction (model (5)), i.e.,  $\mathcal{H}_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ , a 4 df likelihood ratio test; P\_i is the p-value for testing  $\mathcal{H}_0: \beta_3 = \beta_4 = 0$  with model (5), a 2 df likelihood ratio test. SNPs shown significance after the false discovery rate (FDR) control method (Benjamini and Hochberg, 1995) are indicated by \*.