Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, 2016, *17*, 388-395

Statistical Identification of Gene-gene Interactions Triggered By Nonlinear Environmental Modulation

Xu Liu[1], Honglang Wang[1] and Yuehua Cui[1,2,*]

[1]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; [2]Division of Health Statistics, School of Public Health, Shanxi Medical University, Shanxi, 030001, P.R. China

Abstract: Complex diseases are often caused by the function of multiple genes, gene-gene (G×G) interactions as well as gene-environment (G×E) interactions. G×G and G×E interactions are ubiquitous in nature. Empirical evidences have shown that the effect of G×G interaction on disease risk could be largely modified by environmental changes. Such a G×G×E triple interaction could be a potential contributing factor to phenotypic plasticity. Although the role of environmental factors moderating genetic influences on disease risk has been broadly recognized, no statistical method has been developed to rigorously assess how environmental changes modify G×G interactions to affect disease risk. To address this issue, we developed a G×G×E triple interaction model in this work. We modeled the environmental modification effect via a varying-coefficient model where the structure of the varying effect is determined by data. Thus the model has the flexibility to assess nonlinear environmental moderation effect on G×G interaction. Simulation and real data analysis were conducted to show the utility of the method. Our approach provides a quantitative framework to assess triple interactions hypothesized in literature.

Keywords: Associate study, Epistasis, Gene-gene-environment triple interaction, Nonlinear interaction, Varying coefficient model.

Received: February 16, 2015 **Revised: September 26, 2015** **Accepted: October 30, 2015**

1. INTRODUCTION

Gene-gene (G×G) interactions are ubiquitous in nature. It is less likely that a disease status is resulted from the function of single genes working separately, but rather due to the complex interactions of genes functioning together [1]. Statistical methods for the identification of gene-gene interactions have been flourished in literature [2]. Some are focused on parametric models based on simple linear or generalized linear regression models, while others are based on nonparametric models such as the Multidimensional Reduction (MDR) method [3]. As for the unit of analysis, some are focused on single nucleotide polymorphism (SNP) interaction analysis. Others are focused on gene level interaction analysis (e.g., [4, 5]), and such knowledge-driven gene-based interaction analyses are attractive given that genes are the functional units in living organisms.

Although the topic of G×G interaction has been studied for several decades, the field has not been advanced much until the recent wave of genome-wide association studies (GWAS). Recent GWAS have identified thousands of disease variants. However, such successes are undermined due to limited heritability explained by these variants in many complex diseases [6]. Some investigators suggest that G×G interaction may potentially contribute to the missing heritability of many complex diseases [7]. Although many statistical methods have been developed for the identification of G×G interaction [2, 8], due to the limitation of model assumptions as well as the underlying mechanism of gene action modes in different diseases, there are still large needs for the development of advanced methods with biological relevance and statistical flexibility.

In addition to G×G interaction, gene-environment (G×E) interaction could also be a contributing factor for the missing heritability [9, 10]. G×E interaction, defined as how genotypes influence phenotypes differently under different environments [11], is also interpreted as genetic sensitivity to environmental stimulus. Vast amount of studies have reported the role of G×E interaction in many diseases, such as mental illness [12], Parkinson disease [13, 14], and type 2 diabetes [15]. The development of statistical methods has also been evolving, from the identification of linear interaction with traditional simple linear or logistic regression models to more flexible nonparametric methods for the identification of nonlinear moderation of environmental influences on genetic risk [16, 17].

Similar as the function of single genes, the influence of G×G interaction on disease risk could also be modified by

[*]Address correspondence to this author at the Division of Health Statistics, School of Public Health, Shanxi Medical University, Shanxi, 030001, P.R. China; Tel: (517) 432-7098; Fax: (517) 432-1405;
E-mails: cui@stt.msu.edu; cuiy@msu.edu

1875-5488/16 $58.00+.00 ©2016 Bentham Science Publishers

environmental changes. In a study of rheumatoid arthritis (RA), Padyukov *et al.* [18] reported the role of smoking in affecting G×G interaction on developing rheumatoid factor (RF)-seropositive disease. The authors found that the relative risk of developing RF-seropositive disease is higher in smokers who carrying double shared epitope (SE) HLA-DR genes than that in those who carrying either SE gene. Zouachel et at. [19] recently showed that a three-way gene-gene-environment (G×G×E) interaction analysis explains the differences in chikungunya virus transmission by *Ae. albopictus* populations at different temperatures. These empirical evidences underline the importance of evaluating three-way G×G×E interactions on disease risk, and further dissect the mechanism in which how G×G interactions on disease risk are moderated by environmental changes.

Recently, Hu *et al.* [20] proposed a three-way gene-gene-gene interaction model for pure gene epistasis analysis based on information theory. Such model can be applied to study three-way G×G×E interactions. However, the model can only be applied for discrete environment factors and it cannot estimate the interaction size (only a detection test). As discussed in [16], when continuous environmental factors are considered, a varying-coefficient nonparametric regression model shows its flexibility in capturing potential nonlinear G×E interactions. In this work, we extend our previous model on nonlinear G×E detection to study how environmental changes modify G×G interactions on disease risk. Simulation and real data analysis are conducted to show the utility of the model.

2. STATISTICAL METHODS

2.1. The Model

From a G×E perspective, we propose the following partial linear varying-coefficient model (PLVCM):

$$Y = \boldsymbol{\alpha}^T \boldsymbol{X} + \beta_0(U) + \beta_1(U)G_1 + \beta_2(U)G_2 + \beta_3(U)G_1G_2 + \varepsilon$$

where Y is the disease trait response; $\boldsymbol{X}$ represents a p-dimensional vector of covariates containing clinical covariates such as age, smoking and gender; G_1 and G_2 are two SNP variables; U is the environmental variable of interest; ε is an *i.i.d.* error term with mean 0 and finite variance; $\boldsymbol{\alpha}$ is a p-dimensional unknown parameter vector; and $\beta_1(u)$, $\beta_2(u)$ and $\beta_3(u)$ are parameters of interest which are varying functions of variable U. We are interested in evaluating how the effects of each SNP variable as well as the interaction of the two are modified by the environmental variable U to affect the trait distribution of Y, in particular the effect of $\beta_3(U)$ which can be interpreted as the triple G×G×E interaction effect. In the model, we also allow the nonlinear marginal effect of U on Y, denoted as $\beta_0(U)$ which can be determined by the data.

Our model is motivated by a recent genome-wide association study to identify genetic factors interacting with maternal uterine environments for birth weight [21]. As a fetus resides inside its mother's womb, intensive signalling and chemical exchanges between the two are expected. As a result, the effect of fetal genes on birth weight can be modified by maternal conditions such as mother's glucose level, BMI level and blood pressure. In addition to identify major G×E interactions in the context of the maternal-fetal unit, we are also interested in identifying how G×G interactions in fetal genome are modified by maternal conditions to control birth weight. The varying coefficient function $\beta_3(\cdot)$ has much flexibility to capture the underlying functional mechanism of triple interactions which can be linear or nonlinear that must be determined by the data.

2.2. The Estimation

Denote $\mathcal{F}_{\mathrm{n}}$ as the space of B-spline basis function of order r $(r \geq 2)$ [22] with the B-spline basis $\mathbf{B}_r(u) = (B_{s,r}(u) : 1 \leq s \leq J_n)^T$, $u \in [a,b]$ where $J_n = N + r$ and $N = N_n$ is the number of interior knots for a knot sequence $\xi_1 = \cdots = 0 = \xi_r < \xi_{r+1} < \cdots < \xi_{r+N_n} < 1 = \xi_{r+N_n+1} = \cdots = \xi_{2r+N_n}$ in which N_n increases along with the sample size n. Then function $\beta_l(u), l = 0, \cdots, 3$, can be approximated by the following spline expansion.

$$\widetilde{\beta}_l(u) \approx \sum_{s=1}^{J_n} B_{s,r}(u)\lambda_{s,l} = \mathbf{B}_r^T(u)\lambda_l ,$$

where $\lambda_l = (\lambda_{s,l}, 1 \leq s \leq J_n)^T$. Let $\lambda = (\lambda_0^T, \lambda_1^T, \lambda_2^T, \lambda_3^T)^T$. The B-spline coefficients λ can be estimated by

$$(\hat{\alpha}^T, \hat{\lambda}^T)^T = \underset{\alpha,\lambda}{\arg\min}\, R(\alpha, \lambda) ,$$

Where $R(\alpha,\lambda) = \sum_{i=1}^{n}[Y_i - \alpha^T \mathbf{X}_i - \widetilde{\beta}_0(u_i) - \widetilde{\beta}_1(u_i)G_{1i} - \widetilde{\beta}_2(u_i)G_{2i} - \widetilde{\beta}_3(u_i)G_{1i}G_{2i}]^2$. Let $D_i = [\mathbf{X}_i^T, (D_{i,sl}, 1 \leq s \leq J_n, l = 0,1,2,3)^T]^T$, where $D_{i,s0}(\beta_1) = B_{s,r}$ and $D_{i,s1}(\beta_2) = B_{s,r}G_{1i}$, $D_{i,s2}(\beta_2) = B_{s,r}G_{2i}$ and $D_{i,s3}(\beta_2) = B_{s,r}G_{1i}G_{2i}$. Denote $\mathbf{D} = (D_1, \cdots, D_n)^T$ which is an $n \times (p + 4J_n)$ matrix, and $Y = (Y_1, \cdots, Y_n)^T$, where $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^T$ is an $n \times p$ matrix. Then the least squares estimators of α and λ can be obtained as

$$(\hat{\alpha}^T, \hat{\lambda}^T)^T = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T Y \qquad (2.2)$$

It is easy to obtain the estimator of the nonparametric function $\beta_l(u)$ by

$$\hat{\beta}_l(u) = \mathbf{B}_r^T(u)\hat{\lambda}_l, l\text{=0, 1, 2, 3.} \qquad (2.3)$$

We use the Bayesian Information Criterion (BIC) to select the number of interior knots and the order of basis function based on $\beta_0(U)$ in model $E[Y\,|\mathbf{X}, U] = \alpha^{\mathrm{T}}\mathbf{X} + \beta_0(U)$. More specific, we minimize the following criterion

$$(N, r) = \underset{N \in \{2,3,4,5\}, r \in \{3,4,5\}}{\arg\min} \log(n^{-1}RSS(\breve{\alpha}, \breve{\lambda}_0)) + \log(n)(N + r)/n,$$

where $RSS(\breve{\alpha}, \breve{\lambda}_0) = \sum_{i=1}^{n}\{Y_i - \breve{\alpha}^T\mathbf{X}_i + \breve{\beta}_0(U_i)\}^2$, and $\breve{\alpha}$ and $\breve{\beta}(u)$ are the estimates based on model $E[Y\,|\mathbf{X}, U] = \alpha^{\mathrm{T}}\mathbf{X} + \beta_0(U)$, with some preset interior knots and spline orders. The selected knots and orders are then fixed when estimating functions $\beta_l(\cdot)$, $l = 1, 2, 3$ to save computational time.

2.3. Testing the Overall Genetic Effect

One merit of our model is that it can assess the interaction effect of environmental exposures with genes. This can be achieved by testing the nonparametric component β_l (·), l = 1, 2, 3, which allows one to discover the dynamic changes of the interaction effects. We first consider the following hypothesis test to detect if there is any genetic effect of two SNP variables on Y, i.e.,

$H_0 : \beta_1(\cdot) = \beta_2(\cdot) = \beta_3(\cdot) = 0$ v.s. H_1 : at least one is not equal zero (2.4)

via a log-likelihood ratio test (LRT). Under H_0, we can estimate $(\alpha_0^T, \lambda_0^T)^T$ by

$$(\hat{\alpha}_0^T, \hat{\lambda}_0^T)^T = \arg\min_{\alpha,\lambda_0} R(\alpha, \lambda_0),$$

where $R(\alpha, \lambda_0) = \sum_{i=1}^n [Y_i - \alpha^T \mathbf{X}_i + \widetilde{\beta}_0(u_i)]^2$. The log-likelihood function under H_0 and H_1 are, respectively,

$$\ell_0(\hat{\alpha}_0, \hat{\lambda}_0) = -n/2 \log(2\pi\hat{\sigma}_0^2) - 1/2\hat{\sigma}_0^{-2} R(\hat{\alpha}_0, \hat{\lambda}_0)$$
$$\ell_1(\hat{\alpha}, \hat{\lambda}) = -n/2 \log(2\pi\hat{\sigma}_1^2) - 1/2\hat{\sigma}_1^{-2} R(\hat{\alpha}, \hat{\lambda})$$

where $\hat{\sigma}_0^2$and $\hat{\sigma}_1^2$are estimators of variance of response Y under H_0 and H_1, respectively. The LRT is defined as $\mathcal{L} = 2(\ell_1(\hat{\alpha}, \hat{\lambda}) - \ell_0(\hat{\alpha}_0, \hat{\lambda}_0))$, which asymptotically follows a χ^2-distribution with $3J_n$ degrees of freedom. Fail to reject H_0 indicates that the effects of G_1 and G_2 on Y are not significant. Otherwise, we pursue further tests to dissect the interaction mechanism.

2.4. Testing the Interaction Effect

If the null hypothesis in (2.4) is rejected, we can further test which component is significant by formulating the following hypotheses,

$H_{0l}^I : \beta_l(\cdot) = 0$ v.s. $H_{1l}^I : \beta_l \neq 0$, $l \in \{1, 2, 3\}$, (2.5)

Of particular interest is the test of $H_{03}^I : \beta_3(\cdot) = 0$ where the effect $\beta_3(\cdot)$ reflects the triple G×G×E interaction. In the following, we show the derivation of the test by focusing on this component. Similar procedure applies to the test of the other two components, i.e., $H_{0l}^I : \beta_l(\cdot)$, l = 1, 2.

Denote $\lambda_{-3} = (\lambda_0^T, \lambda_1^T, \lambda_2^T)^T$. Under H_{03}^I, we can estimate $(\alpha^T, \lambda_{-3}^T)^T$ by

$$(\hat{\alpha}^T, \hat{\lambda}_{-3}^T)^T = \arg\min_{\alpha,\lambda_{-3}} R(\alpha, \lambda_{-3}),$$

where $R(\alpha, \lambda_{-3}) = \sum_{i=1}^n [Y_i - \alpha^T \mathbf{X}_i - \widetilde{\beta}_0(u_i) - \widetilde{\beta}_1(u_i) G_{1i} - \widetilde{\beta}_2(u_i) G_{2i}]^2$. The estimates under H_{13}^I is the same as (2.2). The log-likelihood function under$H_{03}^I$is given by$\ell_0^I(\hat{\alpha}_0, \hat{\lambda}_{-3}) = -n/2 \log(2\pi\breve{\sigma}_0^2) - 1/2\breve{\sigma}_0^{-2} R(\hat{\alpha}_0, \hat{\lambda}_{-3})$, where $\breve{\sigma}_0^2$ is the estimator of variance of response Y under H_{03}^I. We can construct the LRT as $\mathcal{L}^{\mathrm{I}} = 2(\ell_1(\hat{\alpha}, \hat{\lambda}) - \ell_0^I(\hat{\alpha}_0, \hat{\lambda}_{-3}))$, which has asymptotic χ^2-distribution with J_{n} degrees of freedom. Rejecting H_{03}^I indicates significant interaction effect of the two SNPs. However, whether the interaction effect is moderated by environmental variable U needs to be further assessed by statistical tests.

2.5. Assessing the Environmental Moderation Effect on G×G Interaction

If H_{03}^I in test (2.5) is rejected, we can further test if the G× G interaction is modified by environmental variable U by testing H_0^c: $\beta_3(\cdot) = c$ vs H_1^c: $\beta_3(\cdot) \neq c$, where c is a constant. The null hypothesis implies that the G× G interaction effect is not sensitive to the change of U, hence is not modified by U. To do this, we define a transformation matrix Γ such that $\Gamma \mathbf{B}_r(u) = (1, \widetilde{\mathbf{B}}_r(u)^T)^T$ ([23] chapter 4). In practice, we can simply define $\widetilde{\mathbf{B}}_r(u) = (\mathrm{B}_{2,r}(u), \mathrm{B}_{3,r}(u), \cdots, \mathrm{B}_{J_{\mathrm{n}},r}(u))^T$. Thus, $\widetilde{\beta}_3(u)$ can be rewritten as

$$\widetilde{\beta}_3(u) \approx \xi_1 + \sum_{s=2}^{J_n} B_{s,r}(u) \boldsymbol{\xi}_{\mathbf{s}} = \xi_1 + \widetilde{\mathbf{B}}_r^T(u) \boldsymbol{\xi}_{\mathbf{-1}},$$

where $\boldsymbol{\xi}_{-1} = (\xi_2, \cdots, \xi_{Jn})^T$ and $\xi = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_{-1}^T)^T$. Denote $\lambda_C = (\lambda_C, \lambda_1, \lambda_2, \xi_1)^T$. Under H_0^C, we can estimate $(\alpha^T, \lambda_C^T)^T$ by

$$(\hat{\alpha}_0^T, \hat{\lambda}_C^T)^T = \arg\min_{\alpha,\lambda_C} R(\alpha, \lambda_C),$$

where $R(\alpha, \lambda_C) = \sum_{i=1}^n [Y_i - \alpha^T \mathbf{X}_i - \widetilde{\beta}_0(u_i) - \widetilde{\beta}_1(u_i) G_{1i} - \widetilde{\beta}_2(u_i) G_{2i} - \xi_1 G_{1i} G_{2i}]^2$. The estimates under H_1^c is the same as (2.2). The log-likelihood function under H_0^c is given by $\ell_0^c(\hat{\alpha}_0, \hat{\lambda}_C) = -n/2 \log(2\pi\breve{\sigma}_C^2) - 1/2\breve{\sigma}_C^{-2} R(\hat{\alpha}_0, \hat{\lambda}_C)$, where $\breve{\sigma}_C^2$ is the estimator of variance of response Y under H_0^c. We can construct the LRT as $\mathcal{L}^C = 2(\ell_1(\hat{\alpha}, \hat{\lambda}) - \ell_0^C(\hat{\alpha}_0, \hat{\lambda}_C))$, which has asymptotic χ^2-distribution with $J_n - 1$ degrees of freedom. Rejecting H_0^c indicates there exists G× G× E triple interaction.

3. MONTE CARLO SIMULATION

The finite sample performance of the proposed method was evaluated by simulation studies. Considering model (2.1), we generated the environmental variable U from a Uniform distribution $U(0, 1)$ and two covariates X_1, X_2 from an independent Normal distribution $N(0, 1)$. The SNP variable G, coded as (2, 1, 0) corresponding to genotypes (AA, Aa, aa), was simulated from a multinomial distribution with corresponding frequencies $(p_A^2, 2p_A(1-p_A), (1-p_A)^2)$ where the frequency of the minor allele is specified as p_A = (0.1, 0.3, 0.5). The error term ε was simulated from a normal distribution N (0, σ^2), where σ = 0.1, 0.5, 1.0. The testing performance was compared under different MAFs and different error distributions.

For the varying coefficient functions, we set $\beta_0(u)$ = cos(πu); $\beta_1(u)$ = sin(π(u − A)/(B−A)) with $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$; $\beta_2(u)$ = sin(πu); $\beta_3(u)$ = cos(πu)/3 + 2 sin(πu)/3; and $\boldsymbol{\alpha}$ = (0.3, 0.5)^T. We run 1000 simulation replicates each with sample size n = 500, 1000, 2000. The number of interior knots N and spline order r were selected by the BIC criterion.

We first evaluated the performance for testing the functional coefficients in hypothesis (2.4), i.e., $H_0 : \beta_1(\cdot) = \beta_2(\cdot) = \beta_3(\cdot) = 0$. The testing power was evaluated under a sequence of alternative models indexed by τ, i.e., H_1^{τ}: $\beta_l^{\tau}(\cdot) = \tau\beta_l(\cdot)$, $l = 1, 2, 3$. When $\tau = 0$, the test results gave the false positive rates. (Fig. **1**) depicts the empirical sizes ($\tau = 0$) and power functions ($\tau > 0$) to detect the overall genetic effect under the different scenarios with triplets (n, p_A, σ), where $n = 500$, 1000, 2000, $p_A = 0.1$, 0.3, 0.5 and $\sigma = 0.1$, 0.5, 1.0. As shown in (Fig. **1**), the empirical type I errors under all scenarios are very close to the nominal level 0.05. As we expected, the power increases as the sample size and MAF increase under a fixed error variance. Fox fixed n and MAF, the power increases as the error variance σ^2 decreases. The results indicate that our method can reasonably control the false positives and has appropriate power to detect genetic effect.

For the performance of testing H_{03}^{I}: $\beta_3(\cdot) = 0$, (Fig. **2**) shows the empirical sizes (τ = 0) and power functions (τ > 0) under different scenarios. Similar as the previous simulation setup, the alternative hypothesis was index by τ. We observed similar trends as described for the overall genetic effect test's results shown in (Fig. **1**). The testing sizes are reasonably controlled. The results indicate relatively good performance of the method.

4. A CASE STUDY

We applied the proposed PLVCM model to a data set from the Gene Environment Association Studies initiative (GENEVA, http://www.genevastudy.org) funded by the trans-NIH Genes, Environment, and Health Initiative (GEI). Low and high birth weights are major causes of neonatal morbidity and mortality. They are also associated with increased risk of metabolic diseases in adult life. New born baby's weight is determined by fetal genes, and also controlled by maternal uterine environmental conditions, leading to complicated gene-environment interactions. For this dataset, we aimed at identifying potential genes as well as gene-gene interactions and further explore the mechanisms in which their effects are modified by maternal environmental conditions.

We focused on the Thai population with 1126 subjects genotyped with the Omni1_Quad_v1_0_B platform after removing potential outliers. We picked mother's one hour OGTT glucose level (denoted as U) as the environmental moderator in our analysis and baby's gender as the covariate (denoted as X). There are total 590,913 SNPs after filtering out SNPs with MAF < 0.05, missing rate < 0.05 and deviating from Hardy-Weinberg equilibrium (p-value< 0.001). We first fitted the SNP data with a simple liner model, i.e.,

$$Y = \alpha X + \beta_0 U + \beta_1 G + \varepsilon \qquad (4.6)$$

and test $H_0 : \beta_1 = 0$ using the Plink software with centered birth weight. There are total 61 SNPs with p-value < 10^{-4}.

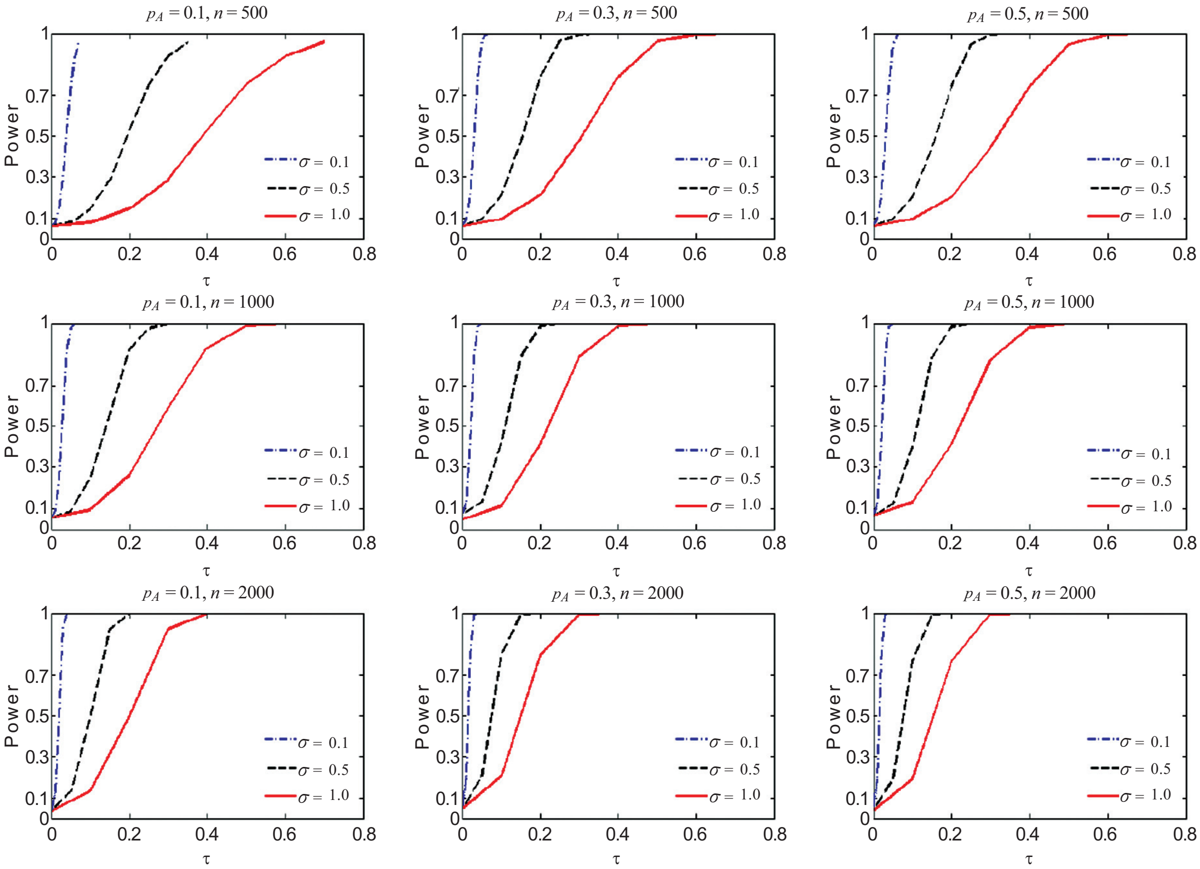

Fig. (1). The power functions for testing H_0: $\beta_1(\cdot) = \beta_2(\cdot) = \beta_3(\cdot) = 0$ under different samples sizes, MAFs and error variances.

The left panel in (Fig. **3**) shows the QQ plot of the -log$_{10}$(p-values). No significant deviation from the expected diagonal line was observed. The right panel in (Fig. **3**) shows the Manhattan plot of the signals.

We then fitted the data with the following varying-coefficient model to allow the effect of G vary over U, and allow nonlinear marginal effect of U, i.e.,

$$Y = \alpha X + \beta_0(U) + \beta_1(U)G + \varepsilon, \tag{4.7}$$

then we test $H_0 : \beta_1(\cdot) = 0$. There are total 59 SNPs remained significant with a p-value threshold 10^{-4}. The left panel in (Fig. **4**) shows the QQ plot of the -log$_{10}$ (p-values). Again, we did not observe significant inflation of the p-values. The right panel in (Fig. **4**) shows the Manhattan plot of the signals.

Since doing a pairwise interaction search with nearly 600K SNPs is technically infeasible, we merged the two SNP sets obtained in previous tests and got 115 SNPs in total (the two sets share 5 SNPs in common). Then we applied the proposed triple interaction PLVCM model (2.1) to the selected 115 SNPs. This strategy is also statistically valid since we only focused on the interaction analysis for SNPs showing marginal significance. The results are summarized in (Table **1**). We assessed the overall genetic effect by testing $H_0 : \beta_1(\cdot) = \beta_2(\cdot) = \beta_3(\cdot) = 0$. The corresponding p-values are denoted by p_{vc}^{O}. The p-values for testing H_{03}^{I}: $\beta_3(\cdot) = 0$ are denoted by p_{vc}^{I}. The last column of the table is the p-value for testing $H_0 : \beta(\cdot) = c$. As a comparison, we also fitted the 115 SNPs with a linear interaction model, i.e.,

$$Y = \alpha X + \beta_0 U + \beta_1 G1 + \beta_2 G_2 + \beta_3 G_1 G_2 + \varepsilon, \tag{4.8}$$

then test H_0^O: $\beta_1 = \beta_2 = \beta_3 = 0$ and H_0^I: $\beta_3 = 0$ with the Plink software. The corresponding p-values are denoted by p_L^O and p_L^I, respectively in the table.

We reported the results in (Table **1**) based on the interaction p-values, i.e., $p_{vc}^{I} < 0.001$ or $p_L^I < 0.001$. The top panel shows the SNP pairs with $p_{vc}^{I} < 0.001$ when fitting the PLVCM model. We observed consistently smaller p-values for testing the overall genetic effect compared to the p-values fitted with the linear model (p_{vc}^{O} vs p_L^O). When fitting the linear interaction model, the interaction effect β_3 does not show significance ($p_L^I > 0.05$) at the 0.05 significance level. Further tests show that the interaction effects are not constant for these SNP pairs ($p_{vc}^{O} < 0.05$) at the 0.05 level, which implies that the interactions are significantly modified by mother's glucose level to affect baby's birth weight. If we fit a linear interaction model, we could miss these interaction effects.

The lower panel in (Table **1**) shows the SNP pairs with $p_L^I < 0.001$ when fitting the linear interaction model. There are total 7 SNP pairs showing significant interaction effects based on the test $H_0 : \beta_3 = 0$. The p-values for testing the overall genetic effect when fitting both models are quite similar to each other. However, the p-values (p_{vc}^{I}) for testing interaction effect with the PLVCM model are larger than the

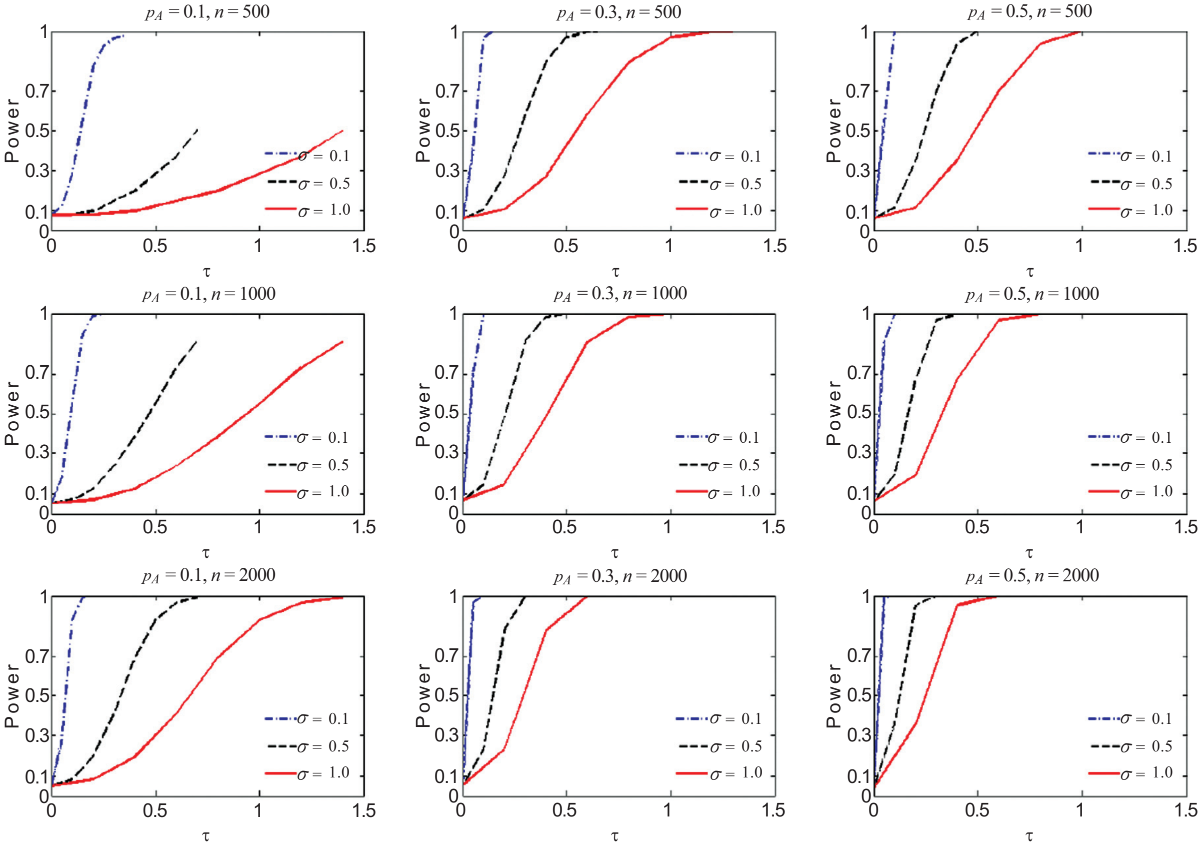

Fig. (2). The power functions for testing H_{03}^{I} : $\beta_1(\cdot) = \beta_2(\cdot) = \beta_3(\cdot) = 0$ under different samples sizes, MAFs and error variances.

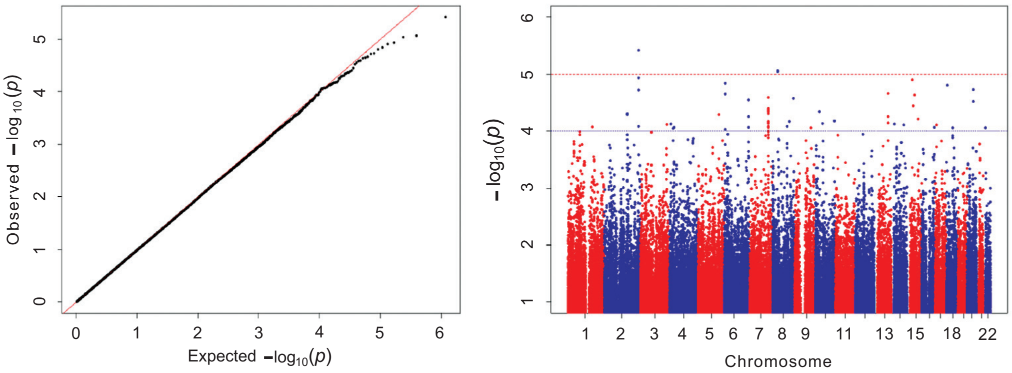

Fig. (3). QQ plot (left) and Manhattan plot (right) of the testing signals by fitting Model (4.6). The dotted blue and red lines in the Manhattan plot correspond to the level of 10^{-4} and 10^{-5}, respectively.

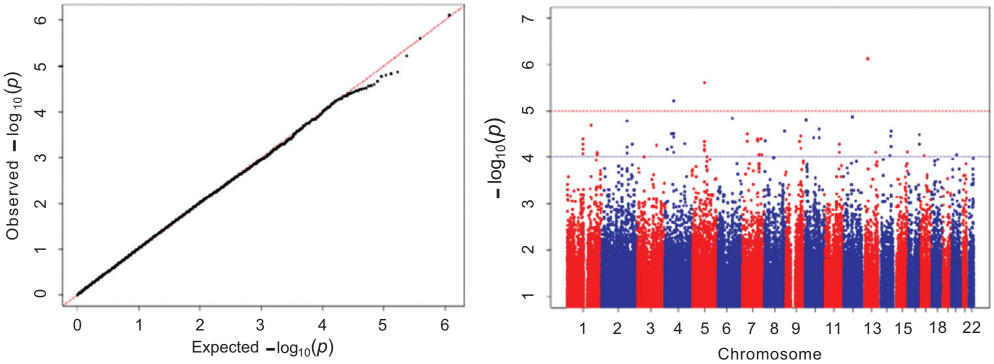

Fig. (4). QQ plot (left) and Manhattan plot (right) of the testing signals by fitting Model (4.7). The dotted blue and red lines in the Manhattan plot correspond to the level of 10^{-4} and 10^{-5}, respectively.

ones (p_L^I) obtained when fitting the data with a linear interaction model. This is not surprise since we failed to reject the null hypothesis $H_0 : \beta(\cdot) = c$. The large p-values ($p_{vc}^C > 0.05$) suggest that the coefficient functions ($\beta(u)$) are all constant. We could miss these interaction effects if we only fit the data with the proposed PLVCM model [22].

To show the estimated varying-coefficient function for $\beta_3(\cdot)$, we picked SNP rs6744005 in gene PLB1 and SNP rs969981 in gene CEP112 and plotted their interaction effect in (Fig. **5**). The constant coefficient fitted with a linear model (4.8) is shown as the dash-dotted line in the figure. Clearly the estimated function is not a constant. The increasing pattern of the function indicates that the baby's birth weight increases as the mother's glucose level increases.

In summary, we identified 5 SNP pairs in which their interactions are significantly modified by mother's glucose level based on the proposed PLVCM model. Such triple G×G×E interaction effects could be missed by fitting a linear interaction model. On the other hand, if a G × G interaction effect is not sensitive to environmental changes, fitting the PLVCM model could lead to potential model misspecification, hence losing power. Thus, a practical guidance when fitting the PLVCM model is to assess the functional form of the varying coefficient first. If the coefficient is a constant, then we fit the data with a constant coefficient model. Otherwise, one can fit the proposed varying coefficient model.

5. DISCUSSION

Identifying gene-gene and gene-environment interactions underlying complex disease traits has been one of the central foci in genetic association studies. Vast amount of empirical studies have supported the role of both types of interaction in understanding the etiology of human diseases. Our previous investigations on nonlinear gene-environment interaction studies [16, 17] have suggested the power of statistical methods in hunting for nonlinear environmental modification effect on genetic risk. Although empirical studies have suggested the role of environmental changes on the effect of gene-gene interactions [18, 19], there has been no rigorous statistical treatment to assess the role of gene-gene interaction on complex diseases triggered by environmental stimulus [23].

In this work, we proposed a triple G×G×E interaction model in which we allow for nonlinear modification effect of environmental changes on gene-gene interactions to affect a disease trait. The proposed PLVCM model has the flexibility to incorporate both parametric (linear part) and nonparametric (nonlinear part) interaction effect. The varying coefficient function is estimated through nonparametric B-spline techniques, hence has the flexibility to capture the underlying functional form, either linear or nonlinear which can be evaluated via statistical tests. Our model is biologically attractive in which it exhibits two important features: (1) It addresses a long-term question on how environmental exposures moderate G×G influences on disease risk; and (2) It has the flexibility to detect nonlinear interaction, thus more

Table 1. List of SNP pairs with p_{vc}^{I} <0.001 or p_{L}^{I} <0.001.

SNP 1					SNP 2					P-value				
SNP ID	MAF	Chr	Gene	Alleles	SNP ID	MAF	Chr	Gene	Alleles	p_{vc}^{O}	p_{vc}^{I}	p_{L}^{O}	p_{L}^{I}	p_{vc}^{C}
rs9824819	0.40	3	-	A/G	rs1332160	0.20	9	-	C/A	1.27E-8	**1.96E-4**	1.84E-8	0.717	7.49E-5
rs1385331	0.44	3	-	G/A	rs1332160	0.20	9	-	C/A	5.22E-9	**2.02E-4**	3.99E-8	0.592	7.85E-5
rs2548754	0.22	5	-	A/G	rs10982000	0.39	9	ZNF618	A/G	3.60E-10	**3.18E-4**	1.01E-4	0.402	2.12E-4
rs6744005	0.24	2	PLB1	A/C	rs969981	0.13	17	CEP112	A/G	1.11E-7	**3.25E-4**	2.79E-7	0.726	1.13E-4
rs11030106	0.37	11	STIM 1	G/A	rs10152064	0.32	14	-	G/A	7.07E-9	**4.22E-4**	4.11E-4	0.656	1.62E-4
rs7903175	0.48	10	-	G/A	rs1144913	0.14	14	EML5	G/A	4.65E-9	1.14E-3	6.55E-9	**9.92E-5**	0.25
rs7903175	0.48	10	-	G/A	rs10144475	0.13	14	ZC3H14	C/G	5.23E-8	5.89E-3	6.65E-8	**1.10E-4**	0.61
rs2584364	0.10	8	-	A/G	rs8027826	0.29	15	-	G/A	4.58E-9	3.12E-3	6.45E-9	**2.24E-4**	0.71
rs3008203	0.34	1	-	G/A	rs903957	0.13	8	-	A/C	1.02E-8	1.69E-2	1.39E-8	**3.11E-4**	0.49
rs1509869	0.34	1	-	G/A	rs903957	0.13	8	-	A/C	8.09E-9	1.56E-2	1.11E-8	**4.06E-4**	0.44
rs9989745	0.18	2	THSD7B	A/G	rs1565190	0.34	12	ITPR2	G/A	1.10E-6	1.59E-2	1.26E-6	**7.36E-4**	0.36
rs2083429	0.43	1	-	C/A	rs11800549	0.15	1	DNM3	G/A	2.43E-9	3.74E-3	3.51E-9	**8.98E-4**	0.14

powerful when G×G effects are nonlinearly modified by environmental stimuli.

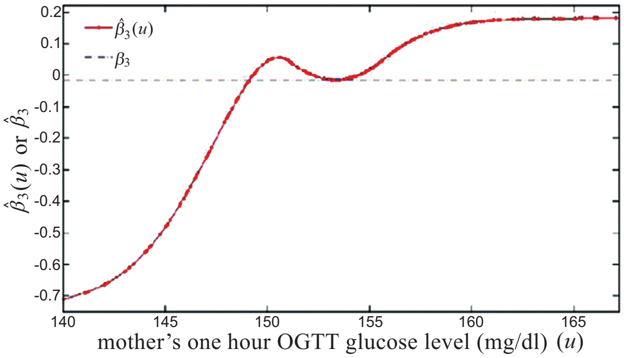

Fig. (5). The estimator $\hat{\beta}_3(u)$ fitted with the PLVCM model (2.1), and the estimator $\hat{\beta}_3$ fitted with the linear interaction model (4.8).

In a typical genetic association study, there are usually large number of genetic variables (e.g., SNPs). When focusing on a gene set based analysis, it is important to fit multiple SNPs within a gene to a single interaction model and select important players moderated by environmental changes. In addition, the proposed model is implemented based on a quantitative trait in current work. The framework can be extended to a binary disease trait with a known link function. The parameter estimation will be a little different due to the nonlinear link function to be adopted. Such a generalized PLVCM has particular power to dissect gene-gene interaction effects triggered by nonlinear environmental moderation. These will be considered in our future investigations. The computational code for implementing the proposed method is available upon request.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was partially supported by grants from NSF (IOS-1237969 and DMS-1209112) and from National Natural Science Foundation of China (31371336). Funding support for the GWA mapping: Maternal Metabolism-Birth Weight Interactions study was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004415). The study is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01 HG04424). The datasets used for the analyses described in this manuscript were obtained from dbGaP at *http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap* through dbGaP accession number phs000096.v4.p1.

REFERENCES

[1] Moore, J.H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.*, **2003**, *56*, 73-82.

[2] Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **2009**, *10*(6), 392-404.

[3] Hahn, L.W.; Ritchie, M.D.; Moore, J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment

interactions. *Bioinformatics*, **2003**, *19*, 376-382.

[4] Li, S.Y.; Cui, Y. Gene-centric gene-gene interaction: a model-based kernel machine method. *Ann. Appl. Stat.*, **2012**, *6*, 1134-1161.

[5] Ma, L,; Clark, A.G.; Keinan, A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet.*, **2013**, *9*(2), e1003321.

[6] Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature*, **2009**, *461*, 747-753.

[7] Zuk, O.; Hechter, E.; Sunyaev, S.R.; Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA,* **2012**, *109*, 1193-1198.

[8] Wei, W.H.; Hemani, G.; Haley, C.S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **2014**, *15*, 722-733.

[9] Kaprio, J. Twins and the mystery of missing heritability: the contribution of gene-environment interactions. *J. Intern. Med.*, **2012**, *272,* 440-448.

[10] Marian, A.J. Elements of missing heritability. *Curr. Opin. Cardiol.*, **2012**, *27,* 197-201.

[11] Falconer, D.S. The Problem of Environment and Selection. *Amer. Natural.*, **1952**, *86*, 293-299.

[12] Caspi, A.; Moffitt, T.E. Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nat. Rev. Neuro.*, **2006**, *7,* 583-590.

[13] Ross, C.A.; Smith, W.W. Gene-environment interactions in Parkinson's disease. *Parkinsonism Relat. Disord.*, **2007**, *13*, S309-S315.

[14] McCulloch, C.C.; Kay, D.M.; Factor, S.A.; Samii, A.; Nutt, J.G.; Higgins, D.S.; Griffith, A.; Roberts, J.W.; Leis, B.C.; Montimurro, J.S.; Zabetian, C.P.; Payami, H. Exploring gene-environment interactions in Parkinson's disease. *Human genet.*, **2008**, *123*, 257-265.

[15] Zimmet, P.; Alberti, K.; Shaw, J. Global and societal implications of the diabetes epidemic. *Nature*, **2001**, *414*, 782-787.

[16] Ma, S.; Yang, L.; Romero, R.; Cui, Y. Varying coefficient model for gene–environment interaction: a non-linear look. *Bioinformatics*, **2011**, *27*, 2119-2126.

[17] Wu, C.; Cui, Y. A novel method for identifying nonlinear gene-environment interactions in case-control association studies. *Human Genet.*, **2013**, *132,* 1413-1425.

[18] Padyukov, L.; Silva, C.; Stolt, P.; Alfredsson, L.; Klareskog, L. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. *Arthritis Rheum.*, **2004**, *50*, 3085-3092.

[19] Zouachel K.; Fontainel, A.; Vega-Rua, A.; Mousson, L.; Thiberge, J.M.; Lourenco-De-Oliveira, R.; Caro, V.; Lambrechts, L.; Failloux, A.B. Three-way interactions between mosquito population, viral strain and temperature underlying chikungunya virus transmission potential. *Proc. Biol. Sci.*, **2014**, *281*, 1-8.

[20] Hu, T.; Chen, Y.; Kiralis, J.W.; Collins, R.L.; Wejse, C.; Sirugo, G.; Williams, S.M.; Moore, J.H. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *J. Am. Med. Inform. Assoc.*, **2013**, *20*(4), 630-636.

[21] HAPO Study Cooperative Research Group. Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study: associations with neonatal anthropometrics. *Diabetes*, **2009**, *58*, 453-459.

[22] de Boor, C. *A practical guide to splines*, Springer: New York, **2001**.

[23] Schumaker, L.L. *Spline Functions: Basic Theory*. Wiley: New York, **1981**.