Semiparametric Mixed Model for Evaluating Pathway-Environment Interaction

Zaili Fang\textsuperscript{1}, Inyoung Kim\textsuperscript{1*}, and Jeesun Jung\textsuperscript{2}

May 1, 2014

1 Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, U.S.A.
2 Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, U.S.A.

*To whom correspondence should be addressed:
Inyoung Kim, Ph.D.
Department of Statistics, Virginia Polytechnic Institute and State University, 410A Hutcheson Hall, Blacksburg, VA 24061-0439, U.S.A.
Tel: (540) 231-5366
Fax: (540) 231-3863
Email: inyoungk@vt.edu
Abstract

A biological pathway represents a set of genes that serves a particular cellular or a physiological function. The genes within the same pathway are expected to function together and hence may interact with each other. It is also known that many genes, and so pathways, interact with other environmental variables. However, no formal procedure has yet been developed to evaluate the pathway-environment interaction. In this article, we propose a semiparametric method to model the pathway-environment interaction. The method connects a least square kernel machine and a semiparametric mixed effects model. We model nonparametrically the environmental effect via a natural cubic spline. Both a pathway effect and an interaction between a pathway and an environmental effect are modeled nonparametrically via a kernel machine, and we estimate variance component representing an interaction effect under a semiparametric mixed effects model. We then employ a restricted likelihood ratio test and a score test to evaluate the main pathway effect and the pathway-environment interaction. The approach was applied to a genetic pathway data of Type II diabetes, and pathways with either a significant main pathway effect, an interaction effect or both were identified. Other methods previously developed determined many as having a significant main pathway effect only. Furthermore, among those significant pathways, we discovered some pathways having a significant pathway-environment interaction effect, a result that other methods would not be able to detect.

Keywords: environmental variable; Gaussian random process; Kernel machine; Pathway analysis; Semiparametric model; Smoothing splines

Running Title: Semiparametric Method for Evaluating Pathway-Environment Interaction
1 Introduction

Gene-related diseases are complex processes associated not only with specific gene or gene sets but also with gene-gene and gene-environment interaction. For decades, statistical methods have focused on analyzing microarray data based on single genes or single-nucleotide polymorphisms (SNPs) analysis (Chatterjee et al., 2006; Hahn et al., 2003; Maity et al., 2009; Moore et al., 2010; Ritchie et al., 2001). However, single-gene based methods have many limitations. For instance, the effect of one gene on a disease is difficult to interpret and current methods are unable to model gene dependencies so that they may not detect genes with moderate changes that give more insight into biological processes but pick up single gene with dramatic changes (Mootha et al., 2003). For these reasons, gene-set or pathway-based approaches have attracted increasing attention in recent years (Goeman et al., 2004, 2005; Liu et al., 2007; Wang, et al., 2007; Pang et al., 2006, 2011; Kim et al., 2011). It is recognized that a joint study of the association between the outcome and a group of genes within the same pathway could complement genes/SNPs analysis for providing insight in understanding complex diseases (Wang, et al., 2007).

A genetic pathway is the interactions of genes that depend on each other’s individual functions and act accordingly to create the aggregate function related to a cellular process (Goeman et al., 2004). There are several special characteristics of pathways, such as various dimensionality (a pathway can contain several genes or over a thousand ones), and interaction network (genes within the a pathway are expected to function together and hence interact with each other). Thus traditional statistical analyses face difficulties in handling these situations. For instance, linear parametric models usually either fail due to the “curse of dimensionality”, or end up with computational explosion in the number of possible interactions among genes within a pathway. To deal with these difficulties, many innovative statistical methods have merged in recent years. Goeman et al. (2004) proposed a global
test derived from a random effects model to determine the significance of the global expression pattern of a group of genes. A random forests approach was proposed by Pang et al. (2006). Liu et al. (2007) proposed a semiparametric model for covariate and genetic pathway effects on continuous outcomes, where the covariate effects and the pathway effect are modeled parametrically and nonparametrically, respectively. They established the connection between the least squares kernel machine (LSKM) and linear mixed models, which simplifies specification of a nonparametric model with multi-dimensional data. Pang et al. (2011) considered more complicated situations with two or more pathway effects presented in the linear mixed model, which allows the researcher to study how multiple pathways relate to the phenotype of interest. A semiparametric Bayesian approach has also been proposed for evaluating pathway effects on clinical outcomes Kim et al. (2011). However, despite the success of analyzing pathways instead of a single gene, all existing methods ignore the environment exposure covariates, and still fewer focus on the interaction between environmental variables and the genetic pathways.

It has been recognized that genetic factors alone cannot account for many cases of gene related disease (Adami, et al., 2008; Chakravarti and Little, 2003). The gene-environment (G-E) or pathway-environment (P-E) interactions are critical in understanding the dynamic process of disease since ignoring them may mask the detection of a genetic effect and may lead to inconsistent association results (Manolio et al., 2006). Furthermore, understanding the G-E interactions can be important for risk prediction and evaluating the benefit of changes in modifiable environmental exposures or environmental regulations. For these reasons, the number of studies utilizing gene-environment interactions has increased dramatically. These range from semiparametric linear or logistic regression models with linear combinations of genes/SNPs as the predictor (Chatterjee et al., 2006; Maity et al., 2009; Park and Hastie, 2008) to the multifactor dimensionality reduction (MDR) as a data mining
technique for identifying genetic and environmental effects associated with either dichoto-
mous or continuous phenotypes (Ritchie et al., 2001; Hahn et al., 2003; Moore et al., 2010).
Unfortunately, these studies are all genes/SNPs based methods, and they possess problems
in dealing with the pathway analysis. For example, representing the pathway effects with
linear combinations of genes has limitations in detecting non-linear patterns of interacting
genes. Furthermore, the number of genes in a pathway can be in the hundreds or thousands,
which makes modeling the gene-gene or gene-environment interaction very consuming.

To capture high order interactions within the high dimensional genes regressor space as
well as the G-E interactions, Zou et al. (2010) employed a nonparametric regression model
with a Gaussian process. With their model the gene and environmental variables are modeled
non-parametrically, and all of the possible interactions effects are considered simultaneously.
However, using one Gaussian process to describe both gene and environmental variable
function spaces results in all the interaction effect being indistinguishable. Thus it is almost
impossible to apply a suitable test for interesting effects such as G-E interaction.

In this paper, we propose a semiparametric mixed effects model to include environmen-
tal variables, genetic pathway effect, and their interaction. By extending Liu et al. (2007)’s
linear mixed model to our model, we evaluate the interaction between an environmental vari-
able and pathway as well as allow nonlinear relationships between the environmental variable
and a continuous outcome. Assuming that both the pathway and interaction effects have
multivariate normal distributions with a zero mean and covariance structure with specific
kernels, we model them within the framework of Gaussian processes. Thus in our model both
pathway and interaction effects are indeed modeled as random effects. Instead of modeling
the environmental variable as a parametric fixed effect, we model it non-parametrically via
natural cubic spline. By modeling environmental variables and pathways in this way, we can
construct the kernel for the P-E interaction based on the analysis-of-variance-like (ANOVA-
like) decompositions of functions (Wahba, 1990; Gu and Wahba, 1993) for a multivariate function. The feature of our method is to model the interaction between environmental and pathway covariates separately from the interactions among genes within the pathway, which are automatically modeled by the Gaussian process for pathway effect. Our model also extends the additive and interaction smoothing splines for univariate functions to multivariate functions with arbitrary kernel.

In a mixed model, the smoothing parameters of the spline and the Gaussian kernels can be considered as the variance components of the random effects, and thus are simultaneously estimated by maximizing the restricted maximum likelihood (REML). By additively modeling the multivariate functions, this model is suitable for analyzing genetic pathway data in which the P-E interaction attracts particular interests. Furthermore, the covariance structure of our model makes the test of the “overall” pathway effect or P-E interaction effect possible. By “overall” we mean either the main effect of a pathway, the interaction effect associated with the pathway, or both. The restricted likelihood ratio test (RLRT) of two zero variance components under non-standard conditions is employed to test the overall pathway effect, while the RLRT of one zero-variance component and score test are applied to test the P-E interaction.

We first define our model in Section 2 and discuss two REML methods to estimate the model parameters in Section 3. Then in Section 4 we introduce PLRT statistics for testing two or one zero-variance components and the score test for testing the P-E interaction effect. In Section 5 we present a set of simulation studies concerning nonparametric function estimates and variance component tests for various settings. In Section 6 we apply our method to the genetic pathway data for Type II diabetes. Finally, in the last Section, we conclude our work and discuss potential extensions of our model.
2 Construction of Semiparametric Linear Mixed Effects Models

2.1 Model Description and the Kernel of the Interaction Function Space

Let us consider that we have a total of \( n \) subjects and the \( i \)th subject has a continuous disease-related outcome \( y_i \), \( i = 1, 2, \ldots n \). We are interested in relating this response \( y = (y_1, y_2, \ldots, y_n)^T \) with one particular pathway gene expression data \( Z = (z_1, z_2, \ldots, z_n)^T \) and \( k \) environmental variables. In a general form, we can write this nonlinear relationship as

\[
y = f + \epsilon,
\]

where \( \epsilon \) and \( f \) are \( n \times 1 \) dimensional vectors with a specific relationship with \( y \) for the \( i \)th entry as \( y_i = f(x_i^T, z_i^T) + \epsilon_i \), in which \( x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ik}) \) is \( 1 \times k \) vector of environmental variables and \( z_i^T = (z_{i1}, z_{i2}, \ldots, z_{ip}) \) is the \( 1 \times p \) vector of gene expression within a pathway and \( p \) is the gene number. In this paper, we only consider the case with one environmental variable, i.e., \( k = 1 \) so that the input \( x^T \) is reduced to univariate \( x \). We assume that the errors \( \epsilon \sim N(0, \sigma^2 I) \) are \( n \times 1 \) iid random variables vector. \( f(\cdot) \) denotes the unknown nonlinear smooth functions for \( x_i, z_i^T \), and their interaction. In this paper, we assume function \( f \) has the following form:

\[
f(x, z^T) = \beta_0 + f_x(x) + f_z(z^T) + f_{xz}(x, z^T),
\]

where \( \beta_0 \) is the intercept term, and \( f_\alpha, \alpha \in \{x, z, xz\} \), represents the nonlinear effect of the environmental variable, the pathway or the interaction respectively. The above equation is similar to the additive model with two univariate variables and their interaction, except \( z^T \) is a multivariate variable. By writing the general model \([1]\) in this way, we can estimate \( f_x, f_z \) and their interaction \( f_{xz} \) separately according to the characteristics of the pathway.
and the environmental variable. We model $f_x(x)$ using the nonparametric function such as a cubic smoothing spline (Wahba, 1990; Lin and Zhang, 1999; Zhang and Lin, 2003). To handle the high dimensional pathway covariates, $z^T$, we may consider a Gaussian process to express $f_z(z^T)$ since the least squares kernel machine method with the Gaussian kernel has achieved success in a genetic pathway data analysis (Liu et al., 2007).

Before we derive the specific representation for the interaction function, we need examine the function space of $f_x$ and $f_z$ respectively. For the smoothing spline $x \in \mathcal{T} = [0, 1]$, $f_x$ is spanned on the function space $\mathcal{H}_x = \mathcal{H}^0_x \oplus \mathcal{H}^1_x$, where “$\oplus$”, $\mathcal{H}^0_x$ and $\mathcal{H}^1_x$ represent the direct sum operator of two subspaces, the null function space and the penalized function space respectively (Wahba, 1990). Assuming $n$ distinct values of $x_i$ such that $0 < x_i^0 < \cdots < x_n^0 < 1$, the $m$th order smoothing spline estimator $f_x(x)$ can be expressed as (Wahba, 1990; Zhang and Lin, 2003),

$$f_x(x) = \sum_{j=1}^m b_j \phi_j(x) + \sum_{i=1}^n c_i k_x(x, x_i^0),$$

where $\phi_j(x)$ is the polynomial basis that span the null space $\mathcal{H}^0_x$ with $\phi_j(x) = x^{j-1}/(j-1)!$, $j = 1, 2, \ldots, m$, and $k_x(x_i, x_i^0) = [(m - 1)!]^{-2} \int (x - u)^{m-1}(x_i^0 - u)^{m-1}du$ is the kernel which uniquely determines the space $\mathcal{H}^1_x$. For $m = 2$, the natural cubic spline that we shall apply in our model, the kernel of $\mathcal{H}^1_x$ can be calculated as (Hastie et al., 2009; Rasmussen and Williams, 2006)

$$k_x(x, x') = \int_0^1 (x - u)_+(x' - u)_+ du = \frac{\min(x, x')^3}{3} + \frac{\min(x, x')^2|x - x'|}{2},$$

where subscript “$+$” indicates the positive part of the expression. For the null space $\mathcal{H}^0_x$, the kernel is calculated as $k^0_x(x, x') = \sum_{j=1}^2 \phi_j(x)\phi_j(x') = 1 + xx'$.

With the orthonormal polynomial basis, $\mathcal{H}^0_x = \{1\} \oplus \{x\}$, where $\{1\}$ and $\{x\}$ stand for the linear function spaces spanned by the constant 1 and the linear basis $x$ which is centered (Guo, 2002). Since the kernel of the function space of the direct sum of two subspaces is
expressed by the direct sum of the kernel of the subspaces (Aronszajn, 1950; Wahba, 1990), we can derive the kernel of the function space without the constant term for the cubic smoothing spline, \( \{ x \} \oplus \mathcal{H}_x^1 \), as \([x x' + k_x(x, x')].\)

For the function space of \( f_z \), we consider a similar argument by MacKay (1998) that starting from a parametric model, we can span the function of \( f_z \) by a radial basis

\[
f_z(z^T) = \sum_{h=1}^{H} c_h \phi_h(z^T),
\]

where \( \phi_h(z^T) = \exp \left[ -\frac{\|z - z_h\|^2}{2\rho} \right] \) is the radial basis functions centered at fixed points \( \{z_h\}_{h=1}^H \).

Assuming \( c = (c_1, ..., c_h)^T \sim N(0, \tau_z I) \), the entry of the covariance matrix of \( f_z \) is expressed as

\[
R = \tau_z \sum_{h} \phi_h(z) \phi_h(z').
\]

Taking as an example a one-dimensional case, MacKay (1998) shows that in the above expression the sum over \( h \) becomes an integral when taking the limit \( H \to \infty \) such that \( R = \tau_z \exp \left[ -(z - z')^2 / \rho \right] \). Generalizing from this particular case, we can define the Gaussian kernel of the function space \( \mathcal{H}_z^1 \) on \( z \)

\[
k_z\left( z^T, z'^T \right) = \exp \left( -\|z - z'\|^2 / \rho \right), \quad (5)
\]

and we assume that \( f_z \) is generated from a zero mean Gaussian process with the kernel matrix produced by \( k_z \).

Since the tensor product of the kernels of two function spaces determines a new function space (Aronszajn, 1950), we use the tensor product of the kernels of \( \{ x \} \oplus \mathcal{H}_x^1 \) and \( \mathcal{H}_z^1 \) to construct a new function space, \( \mathcal{H}_{xz}^1 \), which contains any order interaction \( f_{xz} \) between \( x \) and \( z^T \). Now we can express the kernel of the interaction function space as

\[
k_{xz} \left( x, z^T; x', z'^T \right) = [x x' + k_x(x, x')] \cdot k_z \left( z^T, z'^T \right). \quad (6)
\]
Therefore, we are able to represent the nonparametric interaction function using a zero mean Gaussian process with the kernel matrix produced by this kernel function.

In the rest of this paper, we use $K_x$, $K_z$ and $K_{xz}$ to stand for the Gram or kernel matrices produced by $k_x$, $k_z$ and $k_{xz}$ respectively. In a specific problem, the environmental variable $x$ must be scaled into $T = [0, 1]$ to construct the interaction kernel. Notice the model expression (2) is not the analysis of variance (ANOVA) decomposition of the smoothing function $f$ since $H^1_z$ and $H^1_{xz}$ are not orthogonal to each other. This may cause the identifiability problem between $f_z$ and $f_{xz}$. However, in practice, this problem only happens to our model in extreme situations such as when the entries of matrix $xx' + k_x(x, x')$ are close to each other. In general, $f_z$ and $f_{xz}$ can be identified well as shown in the simulation and application study.

2.2 Linear Mixed model Representation

Now we are prepared to pose the optimization problem. Based on the above argument, the corresponding function spaces that are penalized are $H^1_x$, $H^1_z$ and $H^1_{xz}$. Analogous to the additive models (Hastie and Tibshirani [1990]), the estimation problem for model (1) becomes: for a given set of predictors $(x_i, z_i^T)$, $i = 1, 2, ..., n$, find $f$ to maximize

$$-\frac{1}{2}(y - f)^T(y - f) - \frac{1}{2} \sum_{\alpha} \lambda_{\alpha} \|f_{\alpha}\|^2_{H_{\alpha}},$$

where $\|f_{\alpha}\|_{H_{\alpha}}$’s are the norms induced by $K_{\alpha}$ of $H^1_{\alpha}$, $\alpha \in \{x, z, xz\}$, and $\lambda_{\alpha}$’s are the penalty parameters that balance the tradeoff between goodness-of-fit and smoothing of the curve or high dimensional surface. The solutions to expression (7) are called the least square kernel machine estimation, and Liu et al. (2007) showed the equivalence of the least square kernel machine to the linear mixed model without interaction effects.

The model (2) can be represented in terms of a mixed model as follows. According to the Representer Theorem (Kimeldorf and Wahba [1971]), the nonparametric function can be expressed by the kernel, $f_z(\cdot) = \sum_{i=1}^n a_i k_z(\cdot, z_i^T)$ and $f_{xz}(\cdot) = \sum_{i=1}^n b_i k_{xz}(\cdot; x_i, z_i^T)$. So the
vectors of these functions are
\[
\begin{align*}
\mathbf{f}_z &= K_z \mathbf{a}, \\
\mathbf{f}_{xz} &= K_{xz} \mathbf{b},
\end{align*}
\]
where \(a_i \in \mathbb{R}, b_i \in \mathbb{R}\). Based on the properties of reproducing kernels, the squared norms of \(\mathcal{H}_z^1\) and \(\mathcal{H}_{xz}^1\) can be expressed as
\[
\begin{align*}
\|\mathbf{f}_z\|_{\mathcal{H}_z^1}^2 &= \mathbf{a}^T K_z \mathbf{a} = \mathbf{f}_z^T K_z^{-1} \mathbf{f}_z, \\
\|\mathbf{f}_{xz}\|_{\mathcal{H}_{xz}^1}^2 &= \mathbf{b}^T K_{xz} \mathbf{b} = \mathbf{f}_{xz}^T K_{xz}^{-1} \mathbf{f}_{xz}.
\end{align*}
\]

To represent the remaining part of model (2), \(\beta_0 + f_x(\cdot)\), we follow Lin and Zhang (1999); Zhang et al. (1998); Green (1987); Green and Silverman (1994)’s procedure. The vector of \(f_x\), \(\mathbf{f}_x\) (note here the constant \(\beta_0\) is absorbed into \(f_x\)), can be expressed in terms of \(\mathbf{f} = (\beta_0, \beta_1)^T\) and \((n - 2) \times 1\) random vector \(\mathbf{r}_x\) as
\[
\mathbf{f}_x = X \mathbf{f} + B \mathbf{r}_x  
\]
for \(n\) distinct input \(x\) values, where \(\mathbf{r}_x \sim N(0, \tau_x I)\) and \(X\) is the design matrix of the null space \(\mathcal{H}_x^0\) spanned by the orthogonal polynomial basis, i.e., \(X = (1, \mathbf{x})\) and \(\mathbf{x}\) is the \(n \times 1\) vector of centered \(x\). \(B\) is a matrix defined as \(B = L(L^T L)^{-1}\), where \(L\) is \(n \times (n - 2)\) full rank matrix with \(M = LL^T\). \(M\) is a penalty matrix defined by Green and Silverman (1994) such that the squared norm of \(\mathcal{H}_{xz}^1\),
\[
\|\mathbf{f}_x\|_{\mathcal{H}_x^1}^2 = \int_0^1 (f''_x(t))^2 \, dt = f_x^T M f_x = \mathbf{r}_x^T \mathbf{r}_x.
\]
More details to define \(B\) and \(M\) can be found in Green and Silverman (1994), Zhang et al. (1998) and Appendix A.

Plugging those representations of square norms and \(f_x\)’s back into (7), we have
\[
-\frac{1}{2} (\mathbf{y} - \mathbf{f})^T (\mathbf{y} - \mathbf{f}) - \frac{1}{2} \left( \lambda_x \mathbf{r}_x^T \mathbf{r}_x + \lambda_z \mathbf{f}_z^T K_z^{-1} \mathbf{f}_z + \lambda_{xz} \mathbf{f}_{xz}^T K_{xz}^{-1} \mathbf{f}_{xz} \right).
\]
If we define $\lambda_x = \sigma^2/\tau_x$, $\lambda_z = \sigma^2/\tau_z$ and $\lambda_{xz} = \sigma^2/\tau_{xz}$, and have random vectors $r_z = f_z, r_z \sim N(0, \tau_zK_z)$ and $r_{xz} = f_{xz}, r_{xz} \sim N(0, \tau_{xz}K_{xz})$, then the above equation is equivalent to

$$-\frac{1}{2\sigma^2}(y - f)^T(y - f) - \frac{1}{2\tau_x}r_x^T r_x - \frac{1}{2\tau_z}K_z^{-1}r_z - \frac{1}{2\tau_{xz}}r_{xz}^T K_{xz}^{-1}r_{xz},$$

which is the triple penalized log likelihood function of the linear mixed model

$$y = f + \epsilon = X\beta + Br_x + r_z + r_{xz} + \epsilon.$$  \hspace{1cm} (10)

From the Bayesian point-of-view, $f$ is interpreted as the sum of four zero-mean stationary Gaussian processes, each with a prior covariance function $\tau_\alpha K_\alpha$ ($\beta$ can be viewed with infinite variance). The vectors $r_z$ and $r_{xz}$ have more specific meanings as the pathway main effect and the P-E interaction effect. Although $r_x$ does not have such a meaning, it can be interpreted as the nonlinear contribution of the relationship of the response and the environmental variable.

Differentiating expression (10) with respect to $\beta$ and $\tau_\alpha$’s, it is easy to show that the best linear unbiased prediction (BLUP) estimate of the random effects, given $\sigma^2$ and $\tau_\alpha$’s as fixed, is obtained from solving

$$\begin{bmatrix}
X^TX & X^TB & X^T & X^T \\
B^TX & B^TB + \lambda_x I & B^T & B^T \\
X & B & I + \lambda_z[K_z]^{-1} & I \\
X & B & I & I + \lambda_{xz}[K_{xz}]^{-1}
\end{bmatrix}
\begin{bmatrix}
\beta \\
r_x \\
r_z \\
r_{xz}
\end{bmatrix}
= 
\begin{bmatrix}
X^T\mathbf{y} \\
B^T\mathbf{y} \\
\mathbf{y} \\
\mathbf{y}
\end{bmatrix}.$$  \hspace{1cm} (11)

Equation (11) shows that the BLUP estimate of $\beta$ and $\tau_\alpha$’s are unique if $X^TX$ is full rank which is usually satisfied.

### 2.3 Estimate Pathway and Interaction Effects

Given the fixed parameters $\sigma^2$ and $\tau_\alpha$’s, the covariance of $y$ is obtained as follows using model (10),

$$\Sigma = \text{Cov}(\mathbf{y}) = \sigma^2 I + \tau_x BB^T + \tau_z K_z + \tau_{xz} K_{xz},$$

\hspace{1cm} (12)
Instead of solving expression (11) directly, we perform recursive steps to simultaneously achieve the approximate expressions of $\beta$ and $r_\alpha$'s, $\alpha \in \{x, z, xz\}$,

$$\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y,$$

$$\hat{r}_x = (B^T\Delta^{-1}B + \tau^{-1}_xI)^{-1}B^T\Delta^{-1}(y - X\hat{\beta}),$$

$$\hat{r}_z = (\Delta^{-1}_2 + \tau^{-1}_zK^{-1}_z)^{-1}\Delta^{-1}_2(y - X\hat{\beta} - B\hat{r}_z),$$

$$\hat{r}_{xz} = (\Delta^{-1}_3 + \tau^{-1}_{xz}K^{-1}_{xz})^{-1}\Delta^{-1}_3(y - X\hat{\beta} - B\hat{r}_x - \hat{r}_z),$$

where $I$ is the $(n - 2) \times (n - 2)$ identity matrix, and $\Delta_j, j = 1, 2, 3$, are covariances for the following distributions,

$$y = X\beta + \epsilon_0, \quad \epsilon_0 \sim N(0, \Delta_0 = \Sigma),$$

$$y - X\hat{\beta} = Br_x + \epsilon_1, \quad \epsilon_1 \sim N(0, \Delta_1 = \sigma^2I + \tau_zK_z + \tau_{xz}K_{xz}),$$

$$y - X\hat{\beta} - B\hat{r}_x = r_z + \epsilon_2, \quad \epsilon_2 \sim N(0, \Delta_2 = \sigma^2I + \tau_{xz}K_{xz}),$$

$$y - X\hat{\beta} - B\hat{r}_x - \hat{r}_z = r_{xz} + \epsilon, \quad \epsilon \sim N(0, \Delta_3 = \sigma^2I).$$

The above expressions for $\hat{\beta}$ and $\hat{r}_\alpha$'s are all linear transformations of $y$; thus, their covariances are easily determined using identity $\text{Cov}(Ay) = A\text{Cov}(y)A^T = A\Sigma A^T$, where $A$ is the transformation matrix in expressions (13).

## 3 REML Estimation of the Variance Components

### 3.1 REML Approach for Estimating Variance Components

In the previous Section, when solving the equation (11) we assume that the regularization parameters, $\tau_x, \tau_z$ and $\tau_{xz}$, the scale parameter $\rho$ for Gaussian processes, and the error variance $\sigma^2$ are already known. In this linear mixed model framework, we can estimate the parameter $\theta = (\sigma^2, \tau_x, \tau_z, \tau_{xz}, \rho)^T$ simultaneously using restricted maximum likelihood (REML) estimation. REML is superior to the maximum likelihood (ML) method in terms of adjusting the small sample bias (Zhang and Lin, 2003). The REML of our model is derived
routinely (Harville, 1977) up to the usual additive constant

\[ l_R = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} X^T \Sigma^{-1} X - \frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} (y - X\hat{\beta}) + c, \tag{15} \]

where \( c \) is constant. Another advantage of using REML is that it accounts for the degrees-of-freedom adjustment of replacing \( \beta \) with \( \hat{\beta} \) in expression (15) (Breslow and Clayton, 1993).

Taking the derivatives of (15) with respect to \( \theta \), the estimates of \( \theta \) are obtained by solving

\[ \frac{\partial l_R}{\partial \sigma^2} = -\frac{1}{2} \text{Tr}(P) + \frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} \Sigma^{-1} (y - X\hat{\beta}) = 0, \]

\[ \frac{\partial l_R}{\partial \tau_\alpha} = -\frac{1}{2} \text{Tr} \left( \frac{\partial \Sigma}{\partial \tau_\alpha} P \right) + \frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \tau_\alpha} \Sigma^{-1} (y - X\hat{\beta}) = 0, \quad \alpha \in \{x, z, xz\}, \tag{16} \]

\[ \frac{\partial l_R}{\partial \rho} = -\frac{1}{2} \text{Tr} \left( \frac{\partial \Sigma}{\partial \rho} P \right) + \frac{1}{2} (y - X\hat{\beta})^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \rho} \Sigma^{-1} (y - X\hat{\beta}) = 0, \]

where \( P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \), and \( \frac{\partial \Sigma}{\partial \rho} = \tau_z \frac{\partial K_z}{\partial \rho} + \tau_{xz} \frac{\partial K_{xz}}{\partial \rho} \). The 5 \times 5 information matrix \( I(\theta) \) has the \( i, j \)th entry as

\[ I(\theta)_{ij} = \frac{1}{2} \text{Tr} \left( P \frac{\partial \Sigma}{\partial \theta_i} P \frac{\partial \Sigma}{\partial \theta_j} \right), \tag{17} \]

and the variance of \( \hat{\theta} \) can be estimated through the expression of the information matrix. Equation (16) can be solved using an iteration method such as Fisher’s scoring method. In practice, the sample size \( n \) may be small, for instance the Type II diabetes data contains only 35 observations, while the model (10) includes two fixed-effect parameters and three smoothing parameters. We may have problems with overparameterization, and it may cause a negative estimate of the variance components based on REML. In such case, the step-halving method can be adopted (Jennrich and Schluchter, 1986), but still the corresponding variance component can be estimated as very close to zero.

### 3.2 Profile REML Approach for Estimating Variance Components

In this Section, we suggest a modification to the REML estimation of the variance components so that the estimate of the error components always remains in the parameter
space. This new approach makes the use of the profile restricted maximum likelihood (p-REML). The covariance of $y$ in expression (12) can be written as $\Sigma = \sigma^2 \Sigma_\lambda$, where $\Sigma_\lambda = (I + \lambda^{-1}_x B B^T + \lambda^{-1}_z K_z + \lambda^{-1}_{xz} K_{xz})$. Defining the matrix $P_\lambda = \Sigma_\lambda^{-1} - \Sigma_\lambda^{-1} X (X^T \Sigma_\lambda^{-1} X)^{-1} X \Sigma_\lambda^{-1}$, and $P = P_\lambda / \sigma^2$, the restricted log likelihood function (15) can be rewritten as

$$l_R = -\frac{1}{2} (n - q) \log(\sigma^2) - \frac{1}{2} |\Sigma_\lambda| - \frac{1}{2} \log |X^T \Sigma_\lambda^{-1} X| - \frac{1}{2} \frac{y^T P_\lambda y}{\sigma^2} + c, \quad (18)$$

where $q = 2$ is the rank of $X$. Assuming that $\lambda_\alpha, \alpha \in \{x, z, xz\}$ are known, by solving the derivative of (18) with respect to $\sigma^2$ set equal to zero, the p-REML estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{y^T P_\lambda y}{n - q}. \quad (19)$$

Since $P_\lambda \Sigma_\lambda$ is idempotent, $\frac{y^T P_\lambda y}{\sigma^2} \sim \chi^2_{r(P_\lambda)}$, where $r(P_\lambda) = \text{Tr}(P_\lambda)$ is the rank of $P_\lambda$, the variance of $\text{Var}(\hat{\sigma}^2) \approx 2\hat{\sigma}^4 \text{Tr}(P_\lambda) / (n - q)^2$. Plug $\hat{\sigma}^2$ back into expression (18) and we have the log profile restricted likelihood (PRL) function

$$l_{PR} = -\frac{1}{2} \log |\Sigma_\lambda| - \frac{1}{2} |X^T \Sigma_\lambda^{-1} X| - \frac{n - q}{2} \log(y^T P_\lambda y) + c. \quad (20)$$

Now we can use the similar scoring algorithm to estimate $\theta^* = (\lambda^{-1}_x, \lambda^{-1}_z, \lambda^{-1}_{xz}, \rho)$. By simple algebra the score of the p-REML likelihood is

$$\frac{\partial l_{PR}}{\partial \theta_j} = -\frac{1}{2} \text{Tr} \left( \frac{\partial \Sigma_\lambda}{\partial \theta_j^*} P_\lambda \right) + \frac{1}{2\hat{\sigma}^2} y^T P_\lambda \frac{\partial \Sigma_\lambda}{\partial \theta_j^*} P_\lambda y, \quad j = 1, 2, 3, 4, \quad (21)$$

and the $i, j$th entry of the information matrix $I^*(\theta^*)$ for the PRL can be approximated as

$$I^*(\theta^*)_{ij} = \frac{1}{2(n - q)} \left\{ (n - q - 2) \text{Tr} \left( \frac{\partial \Sigma_\lambda}{\partial \theta_i^*} P_\lambda \frac{\partial \Sigma_\lambda}{\partial \theta_j^*} P_\lambda \right) - \text{Tr} \left( \frac{\partial \Sigma_\lambda}{\partial \theta_i^*} P_\lambda \right) \text{Tr} \left( \frac{\partial \Sigma_\lambda}{\partial \theta_j^*} P_\lambda \right) \right\}. \quad (22)$$

Note that $I^*(\theta^*)$ is positive definite when $n$ is large enough. Claeskens (2004) also showed the convergence of $I^*(\theta^*)$ under regular conditions so that we can apply the restricted likelihood ratio test (RLRT, see Section 4). Since PRL is not a true likelihood, we only use PRL for statistical test purposes, and use p-REML to obtain a better estimate of the variance components. The variances of $\theta$ is found by plugging the p-REML estimates into (17).
4 Test for Pathway Effects

4.1 Test for Two Zero Variance Components

One of the primary problems in the study of pathway based analysis is testing the “overall” pathway effects. Recall that the meaning of “overall” refers to either the main effect of a pathway, the interaction effect associated with the pathway, or both. In model (10), two random effects are involved with the overall pathway effects. Thus, the hypothesis for testing the overall pathway effect is

\[ H_0 : \tau_z = \tau_{xz} = 0 \text{ vs. } H_a : \tau_z > 0 \text{ or } \tau_{xz} > 0, \] (23)

which is equivalent to the following test

\[ H_0 : \lambda_z^{-1} = \lambda_{xz}^{-1} = 0 \text{ vs. } H_a : \lambda_z^{-1} > 0 \text{ or } \lambda_{xz}^{-1} > 0. \] (24)

For this type of test problem, a likelihood ratio test (LRT) is most commonly used. Note that parameter space for \( \theta = (\lambda_z^{-1}, \lambda_{z}^{-1}, \lambda_{xz}^{-1}, \rho)^T \) equals \([0, \infty)^3 \times (0, \infty)\) (to avoid abuse of notation, in this Section, \( \theta \) and \( \mathcal{I} \) stand for counterparts of PRL). The true parameters \( \theta_0 \) are either in the interior or on the boundary of the parameter space, so the LRT is nonstandard. \[ \text{Vu and Zhou (1997)} \] generalized the hypothesis test for both interior and boundary problems within a setting of mixed regression fitting, so it allows the nonidentically distributed response variable \( y_i \)'s to depend on the covariates and allows the random effects to induce dependence between the response values. \[ \text{(Claeskens, 2004)} \] further extended the non-standard LRT test to the profile restricted likelihood ratio test (RLRT), focusing on nonparametric mixed models with spline fitting.

Following \[ \text{(Claeskens, 2004)} \], we apply RLRT to test hypothesis (24). Under this hypothesis, the RLRT test statistics, \( D \), is the deviance of two times the log PRL, \(-2l_{PR}(\theta)\), i.e. \( D = 2l_{PR}(\omega) - 2l_{PR}(\omega_0) \). Note that \( D \) is the same using either \( l_R \) or \( l_{PR} \). Assuming that the
corresponding regular conditions in [Vu and Zhou (1997)] are satisfied for the PRL function model, $D$ converges to

$$
D \rightarrow \inf_{\theta \in \tilde{C}_0} \|U - \theta\|^2 - \inf_{\theta \in \tilde{C}} \|U - \theta\|^2,
$$

where $\tilde{C} = \{\tilde{\theta} : \tilde{\theta} = \mathcal{I}(\theta_0)^{T/2}(\theta - \theta_0), \theta \in \mathcal{C}_\Omega\}$ is the orthonormal transformation of the cone approximation, $\mathcal{C}_\Omega$, of the parameter space $\Omega$ with $\theta_0$ as the vertex, and $\tilde{C}_0 = \{\tilde{\theta} : \tilde{\theta} = \mathcal{I}(\theta_0)^{T/2}(\theta - \theta_0), \theta \in \mathcal{C}_{\Omega_0}\}$ is the orthonormal transformed cone approximation of the parameter space $\Omega_0$ under the null hypothesis. $U$ is a random vector from $N(0,I)$, and $\mathcal{I}(\theta_0)^{T/2}$ is the right Cholesky square root of p-REML information matrix, i.e. $\mathcal{I}(\theta_0) = [\mathcal{I}(\theta_0)]^{1/2} [\mathcal{I}(\theta_0)]^{T/2}$.

Note that under the null hypothesis, $\theta_0 = (\lambda^{-1}_x, 0, 0, \rho)^T$, $\rho$ is inestimable. We suggest estimating the parameters with $\rho$ fixed at the average of $\|z - z'\|^2$ (average on pairwise observations) to not only reduce the parameter space dimensions but also achieve a better fit. Let $\theta = (\lambda^{-1}_x, \lambda^{-1}_z, \lambda^{-1}_{xz})^T = (\theta_1, \theta_2, \theta_3)^T$. Now the cone parameter spaces are reduced to $\mathcal{C}_\Omega = [0, \infty)^3$ and $\mathcal{C}_{\Omega_0} = [0, \infty) \times \{0\} \times \{0\}$. However, in this problem, all three parameters can be on the boundaries and the orthonormal transformation for the nuisance parameter $\theta_1$ is not invariant, which leads to a transformation for 3 dimensional space. The calculation of (25) in a 3 dimensional space becomes considerably more difficult when the information matrix is not diagonal. To simplify the calculation, we consider the special case that $\theta_1 \approx 0$, which is a reasonable consideration for the Type II diabetes data in a later Section, where the p-REML estimates of $\theta_1$'s are very close to zero for most pathways.

Now the parameter space is reduced to 2 dimensions. Under the orthonormal transformation, the cone spaces become to $\tilde{C} = \{\theta : \gamma \theta_3 - \theta_2 \geq 0, \theta_3 \geq 0\}$, and $\tilde{C}_0 = \{\theta : \theta_3 = \theta_2 = 0\}$, where $\gamma = \tilde{I}_{23} \cdot |\tilde{I}(\theta_0)|^{-1/2}$ is the slope of the axis $\theta_3$ after transformation as shown in Figure 1(a). To account for the fact that $\theta_1$ is estimated, $\tilde{I}(\theta_0)$ is defined from the $3 \times 3$ information
matrix \( \mathcal{I}(\theta_0) \) as

\[
\tilde{\mathcal{I}}(\theta_0) = \begin{bmatrix}
\tilde{I}_{22} & \tilde{I}_{23} \\
\tilde{I}_{32} & \tilde{I}_{33}
\end{bmatrix} = \begin{bmatrix}
I_{22} & I_{23} \\
I_{32} & I_{33}
\end{bmatrix} - \begin{bmatrix}
I_{21} \\
I_{31}
\end{bmatrix} I_{11}^{-1} [I_{12}, I_{13}].
\]

From the graphic point of view, the representation of the test statistics (25) is determined by the minimum distance of the independent normal vector \( U = (U_2, U_3)^T \) to \( \theta \). Under the alternative hypothesis, the minimum distance, \( \inf_{\theta \in \tilde{C}} \| U - \theta \|_2 \), can be understood as the projection of \( U \) on the cone space \( \tilde{C} \) when \( U \) is outside of the cone. As shown in Figure I(a), the representations of \( \inf_{\theta \in \tilde{C}} \| U - \theta \|_2 \) are different in the four regions of the plane with coordinates \((\theta_2, \theta_3)\)

\[
\inf_{\theta \in \tilde{C}} \| U - \theta \|_2^2 = \begin{cases}
0 & \theta_3 \geq 0, \gamma \theta_3 - \theta_2 \geq 0, \quad I \\
U_2^2 + U_3^2 - (\gamma U_2 + U_3)^2/(1 + \gamma^2) & \theta_3 + \gamma \theta_2 \geq 0, \gamma \theta_3 - \theta_2 < 0, \quad II \\
U_3^2 & \theta_3 < 0, \theta_2 \geq 0, \quad III \\
U_2^2 + U_3^2 & \theta_3 + \gamma \theta_2 < 0, \theta_2 < 0, \quad IV.
\end{cases}
\]

The area proportions, \((\phi, 1/4, 1/4, 1/2 - \phi)\) as in the aforementioned order, of these four regions determine the probabilities that the vector \( U \) lies in which region, where \( \phi = \cos^{-1} \left( \gamma \cdot (1 + \gamma^2)^{-1/2} \right) = \tilde{I}_{23} \cdot (\tilde{I}_{22} \tilde{I}_{33})^{-1/2} \).

Under the null hypothesis, the parameters space is reduced to the origin of the plane, thus according to Vu and Zhou (1997)

\[
\inf_{\theta \in \tilde{C}_0} \| U - \theta \|_2^2 = U_2^2 + U_3^2.
\]

Then the asymptotic distribution of \( D \) is the difference of the above two representations

\[
D \rightarrow \begin{cases}
U_2^2 + U_3^2 & \text{with probability } \phi, \quad I \\
(\gamma U_2 + U_3)^2/(1 + \gamma^2) & \text{with probability } 1/4, \quad II \\
U_2^2 & \text{with probability } 1/4, \quad III \\
0 & \text{with probability } 1/2 - \phi, \quad IV.
\end{cases}
\]
Note that because $U_2$ and $U_3$ are independent, thus $(\gamma U_2 + U_3)/\sqrt{1 + \gamma^2} \sim N(0, 1)$, and the final approximate asymptotic distribution of $D$ is

$$D \sim \phi \chi_2^2 + 0.5 \chi_1^2 + (0.5 - \phi) \chi_0^2.$$ (28)

In this paper, we suppose $\lim_{n \to \infty} |\gamma| < \infty$. If $\lim_{n \to \infty} |\gamma| \to \infty$, the representation of $\inf_{\theta \in \tilde{C}} \|U - \theta\|^2$ is in different form [Vu and Zhou 1997] and the asymptotic distribution of $D$ may be different. An additional approximation is that we obtain $\gamma$ with a finite sample size under the null hypothesis, so we assume that $n$ is large enough that the finite $\gamma$ is close to the converged value.

### 4.2 Test for the P-E Interaction Effect

The RLRT for two variance components introduced above allows us to test the overall pathway effect. Furthermore, we may be attracted to testing single variance components, such as testing the P-E effect, given that the overall the pathway effect test is significant.

The hypothesis of this problem is

$$H_0 : \lambda_{xz}^{-1} = 0 \text{ vs. } H_a : \lambda_{xz}^{-1} > 0,$$ (29)

which is equivalent to testing $H_0 : \tau_{xz} = 0$ vs. $H_a : \tau_{xz} > 0$. The RLRT test statistics $d = 2l_{PR}(\theta) - 2l_{PR}(\theta_0)$ for one variance component in semiparametric model with PRL was also suggested by [Claeskens 2004], and an exact RLRT algorithm was proposed by [Crainiceanu et al. 2005]. Unfortunately, this exact RLRT method cannot apply to test (29) for model (10). In their work, there are no random effects in the model under the null hypothesis, thus $d$ can be represented exactly as the form of a mixture of chi-square distribution. On the contrary, our model (10) under the null hypothesis of (29) contains two random effects $r_x$ and $r_z$, which makes it impossible to represent $d$ exactly.
The second choice is to use the method described in the previous section using an asymptotic distribution. However, we need the same approximations; that is, we fix $\rho$ and assume that the relationship between the response and the environmental variable is almost linear, i.e. $\tau_x \approx 0$. Then similarly, the parameters cone space is reduced to 2 dimensions. One interesting parameter $\theta_3 = \lambda_{xz}^{-1}$, and one nuisance parameter $\theta_2 = \lambda_z^{-1}$, both have the true values on the boundary. Thus, $C_\Omega = [0, \infty) \times [0, \infty)$ and $C_{\Omega_0} = [0, \infty) \times \{0\}$.

Under the approximations described above, the asymptotic representation of 2 times the log PRL function under the null hypothesis is

$$\inf_{\theta \in \tilde{C}_0} \|U - \theta\|^2 = 0 \cdot I(U_2 > 0) + U_2^2 I(U_2 \leq 0) + U_3^2,$$

where $I(\cdot)$ is the indicator function. The representation under the alternative hypothesis is the same as in (26), but because the cone under the null hypothesis is no longer the origin of the $(\theta_2, \theta_3)$ plane, $\inf_{\theta \in \tilde{C}_0} \|U - \theta\|^2$ has two regions as shown by (30). Now we must divide the plane with coordinates $(\theta_2, \theta_3)$ into five regions and set the approximated asymptotic representation of $d$ as (see Figure 1(b))

$$d \rightarrow \begin{cases} 
U_3^2 & \text{with probability } 1/4, 
U_2^2 + U_3^2 & \text{with probability } \phi - 1/4, 
(\gamma U_2 + U_3)^2/(1 + \gamma^2) & \text{with probability } 1/4, 
0 & \text{with probability } 1/4, 
0 & \text{with probability } 1/2 - \phi,
\end{cases}$$

(31)

Thus, we have the asymptotic distribution of $d$ for testing $\theta_3 = \lambda_{xz}^{-1} = 0$ or $\tau_{xz} = 0$

$$d \sim (\phi - 0.25)\lambda_x^2 + 0.5\chi_1^2 + (0.75 - \phi)\chi_0^2,$$

(32)

where $\phi$ is calculated through $\gamma$ under hypothesis (29).

In many cases, the relationship between the response and the environmental variable is not linear, i.e. $\tau_x$ is significant and not equal to 0, then we are in the 3 dimension space to
derive the asymptotic distribution of the $d$, which becomes arduous. In this situation, we adopt a score test approach based on the REML function (15) which was proposed by Lin (1997) in a mixed model. The asymptotic distribution of the REML score may not converge to a standard normal distribution, Zhang and Lin (2003) suggested using the scaled chi-square approximation of the test statistics. More generally, the REML score for covariance component $\tau_\alpha, \alpha \in \{x, z, xz\}$ of (16) can also be written as

$$\frac{\partial l_R}{\partial \tau_\alpha} = \frac{1}{2} (P\mathbf{y})^T \frac{\partial \Sigma}{\partial \tau_\alpha} P\mathbf{y} - \frac{1}{2} \text{Tr} \left( P \frac{\partial \Sigma}{\partial \tau_\alpha} \right),$$

where we used identity $(\mathbf{y} - X\hat{\beta})^T \Sigma^{-1} = (P\mathbf{y})^T$. $P$ can be expressed as $P = \Gamma(\Gamma^T \Sigma \Gamma)^{-1} \Gamma^T$ (Searle et al., 1992), where $\Gamma^T$ is $(n-q) \times n$ matrix with full row rank $n-q$ ($q=2$ is the rank of $X$). The matrix $\Gamma^T$ satisfies $\Gamma^T \mathbf{X} = 0$ and $\Gamma^T \mathbf{y} \sim N(0, \Gamma^T \Sigma \Gamma)$. Thus the REML version score test statistics can be written as

$$U_{\tau_\alpha} = \frac{1}{2} (P\mathbf{y})^T \frac{\partial \Sigma}{\partial \tau_\alpha} P\mathbf{y} = \tilde{\mathbf{y}}^T M \tilde{\mathbf{y}},$$

where $\tilde{\mathbf{y}} = (\Gamma^T \Sigma)^{-\frac{1}{2}} \Gamma^T \mathbf{y}$ with $\tilde{\mathbf{y}} \sim N(0, I_{n-q})$, and $M = \frac{1}{2} (\Gamma^T \Sigma)^{-\frac{1}{2}} \Gamma^T \frac{\partial \Sigma}{\partial \tau_\alpha} \Gamma (\Gamma^T \Sigma)^{-\frac{1}{2}}$. $U_{\tau_\alpha}$ is the quadratic form of $\mathbf{y}$ with mean $E(U_{\tau_\alpha}) = \frac{1}{2} \text{Tr} \left( P \frac{\partial \Sigma}{\partial \tau_\alpha} \right)$ and variance $\text{Var}(U_{\tau_\alpha}) = \mathcal{I}_{jj}$, where $\mathcal{I}_{jj}$ is the corresponding entry of the information matrix (17) for the interesting variance component of $\tau_\alpha \in \{\tau_x, \tau_z, \tau_{xz}\}$.

Let $r$ denote the number of non-zero eigenvalues of $M$, then $M$ can be further decomposed using the spectral decomposition as $M = H \Xi H^T = \sum_{i=1}^r \xi_i h_i h_i^T$, where $H = (h_1, ..., h_r)$ is $n \times r$ orthogonal normal matrix, i.e. $h_i^T h_j = \delta_{ij}$, and $\Xi = \langle \xi_i \rangle$ is $r \times r$ diagonal matrix. It follows that

$$U_{\tau_\alpha} = \tilde{\mathbf{y}}^T H \Xi H^T \tilde{\mathbf{y}} = \sum_{i}^r \xi_i \tilde{\mathbf{y}}^T h_i h_i^T \tilde{\mathbf{y}} \sim \sum_{i} \xi_i \chi^2_1.$$ 

Therefore, under $H_0$, the distribution of $U_{\tau_\alpha}$ can be represented as a weighted mixture of chi-square distribution. This is because $\tilde{\mathbf{y}}^T h_i h_i^T \tilde{\mathbf{y}} \sim \chi^2_1$ since $h_i h_i^T$ is an idempotent matrix with rank 1. Because the calculation for $\xi_i$’s is intensive, we follow Zhang and Lin (2003) for the scaled chi-square approximation of the test statistics.
in using the Satterthwaite method to approximate the distribution of $U_{\tau_\alpha}$ by a scaled chi-square distribution $\kappa \chi^2_\nu$, where $\kappa = I_{jj}/2E(U_{\tau_\alpha})$, and $\nu = 2E(U_{\tau_\alpha})^2/I_{jj}$. Zhang and Lin (2003) also suggested to further account for the fact that $\theta = (\sigma^2, \tau_x, \tau_z, \tau_{xz}, \rho)^T$ is estimated, so that $\kappa$ and $\nu$ are calculated by replacing $I_{jj}$ with the efficient information

$$\tilde{I}_{jj} = I_{jj} - I_{j\theta}I_{\theta\theta}^{-1}I_{j\theta}^T,$$

where $I_{j\theta}$ and $I_{\theta\theta}$ are the corresponding vector and matrix if we rearrange the $5 \times 5$ information matrix $I(\theta)$ as

$$I(\theta) = \begin{bmatrix} I_{jj} & I_{j\theta} \\ I_{j\theta}^T & I_{\theta\theta} \end{bmatrix}.$$

In this paper, we are particularly interested in testing the P-E interaction effect, i.e., $\tau_{xz}$.

5 Simulation Study

5.1 Parameters Estimation

We carried out the simulation study to evaluate the accuracies of the estimators; 200 runs were performed for each of the simulation scenarios. Let $p$ denote the number of genes in the pathway and $n$ denote the number of observations. We considered a setup that mimics the real diabetes pathway data with a total of 50 genes within a pathway. The true model of the $i$th observations is

$$y_i = f_x(x_i) + f_z(z_i^T) + f_{xz}(x_i, z_i^T) + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2)$$

with nonparametric functions

$$f_x(x_i) = 5.6 + 0.1x_i + \cos(x_i\pi/18),$$

$$f_z(z_i^T) = a \cdot z_i^{(30)} \exp\left(-0.2|z_i^{(30)}|\right)/5,$$

$$f_{xz}(x_i, z_i^T) = b \cdot e^{x_i/10} \sin\left(z_i^{(30)}\right) \cos\left(z_i^{(30)}\right)/8,$$

where $z_i^{(30)}$, $\tilde{z}_i^{(30)}$ and $\bar{z}_i^{(30)}$ stand for $\sum_{j=1}^{30} z_{ij}$, $\sum_{j=1}^{30} |z_{ij}|/30$ and $\sum_{j=1}^{30} z_{ij}/30$. We sample $x_i$ and $z_{ij}$ ($j = 1, ..., 50$) from Uniform[18, 36] and $N(0, 1)$, respectively. Furthermore, $a$ and
are parameters to control the magnitude of the nonparametric functions respectively. In this Section they are fixed at \(a = 1.5\) and \(b = 2\). In the true model (34), a total of 30 genes, \(z_{i1}, \ldots, z_{i30}\), are involved. However in a real situation, we may fit the model with extra genes that are not involved in the true model. Thus we consider the following settings for model (34):

- Setting 1: \(n = 100/150\), true \(p = 30\), fitted \(p = 30\), \(\sigma^2 = 0.2^2\),
- Setting 2: \(n = 100/150\), true \(p = 30\), fitted \(p = 40\), \(\sigma^2 = 0.2^2\),
- Setting 3: \(n = 100/150\), true \(p = 30\), fitted \(p = 50\), \(\sigma^2 = 0.2^2\).

For each setting, two sample sizes \(n = 100\) and 150 were considered.

In Section 3 we introduced two methods to estimate the variance components using REML and p-REML. We are particularly interested in comparing the performance of these two methods. One of the difficulties of solving equation (16) or (21) using a scoring method is finding the initial values for \(\theta\) or \(\theta^*\), since there are no analytic expressions to roughly obtain those initial values. Breslow and Clayton (1993) suggested starting the variance parameters from small positive values within a complex situation. We started the variance components with \((\sigma^2, \tau_x, \tau_z, \tau_{xz})^T = (0.001, 0.001, 0.001, 0.001)^T\), which is equivalent to starting with \((\sigma^2, \lambda_x^{-1}, \lambda_z^{-1}, \lambda_{xz}^{-1})^T = (0.001, 1, 1, 1)^T\) for p-REML. For scale parameter \(\rho\), we can either fix or estimate it. In this simulation study, we choose the initial value \(\rho = 2\) which is the average of \(\|z - z'\|^2\) on all pairwise observations if it is estimated. We also compare the results with \(\rho\) fixed at 2. Note that if \(\rho\) is estimated, we consider two possible ways. One way is to perform a two-step procedure where we first fix \(\rho\) at 2 and evaluate \((\sigma^2, \tau_x, \tau_z, \tau_{xz})\) until convergence and then use the results with \(\rho = 2\) as the initial values to evaluate \((\sigma^2, \tau_x, \tau_z, \tau_{xz}, \rho)\) until convergence. The other way is to evaluate \((\sigma^2, \tau_x, \tau_z, \tau_{xz}, \rho)\) together from an initial value \((0.001, 0.001, 0.001, 0.001, 2)^T\). The simulation results show that the former method is more stable, so only these results are shown. Similarly, a two-step procedure was used for p-REML.
when $\rho$ is estimated.

To demonstrate the fitting results, Figure 2 shows one selected example of setting 1 comparing estimated $f, f_x, r_z$ and $r_{xz}$ with the true ones. The overall response $f$ is fitting very well as shown in Figure 2(d). As shown in Figure 2(b) and (c), there is not much identifiability issue since both the fitted pathway effect $\hat{r}_z$ and fitted the interaction effect $\hat{r}_{xz}$ capture the patterns of the true ones very well.

To have an overall evaluation of the goodness-of-fit of the nonparametric function $f_x, f_z$ and $f_{xz}$, we followed the techniques used by Liu et al. (2007), who suggested regressing the true nonparametric functions on the fitted ones. By reporting the average intercepts, slopes and $R^2$'s from these regressions, the goodness-of-fit of the fitted nonparametric functions can be assessed empirically. The closer to 0 and 1 of the intercepts and slopes are and the closer to 1 of $R^2$ is, the better the performance of the estimation is.

In Table 1 we summarized the goodness-of-fit of $f_\alpha, \alpha \in \{x, z, xz\}$ for 200 hundred runs. The scenarios of three settings were used in four procedures: I) REML with $\rho$ estimated, II) REML with $\rho$ fixed at 2, III) p-REML with $\rho$ estimated, and IV) p-REML with $\rho$ fixed at 2. It can be seen that the performance of using procedure I) is not so good; $\rho$ goes to an extremely large value and $f_\alpha$'s deviate from $\hat{f}_\alpha$'s. This may be because the REML likelihood function does not have a maximum and the likelihood increases or becomes flat with $\rho$. In such a case, the entries of $K_z$ becomes a matrix of ones. One solution when the REML function becomes flat with $\rho$ is to fix $\rho$ at the turning point of the REML function. In procedure II) we fixed $\rho$ at 2. The average of $\|z - z'\|^2$ on all pairwise observations is very close to 2 and using this $\rho$ allows us to avoid having extreme values for the entries of $K_z$. The performance of this procedure is improved significantly; all the $R^2$ values are over 90% and close to 1, and the intercepts and slopes of the regressions are close to 0 and 1. However, $\hat{\sigma}^2$ values are all close to zero. The zero error component happens in REML estimation (Searle...
et al. [1992], especially with high dimensional parameter spaces.

Table 1 shows that the performance is much better for the two p-REML procedures. Not only is the fitting of nonparametric functions very good, but the estimate of error variance component $\hat{\sigma}^2$ is close to the true value. As expected, fitting with extra genes introduces more error, which results in the increase of $\hat{\sigma}^2$. This is because fitting irrelevant genes is equivalent to introducing more noise into the model. However, the results show little difference in fitting $f_\alpha$'s for differently used gene numbers. Increasing the number of observations is expected to improve the fitting performance. Although overall there is no much difference between $n = 100$ and 150, there is slight improvement in fitting the P-E interaction effect. This can be seen from the fact that $R^2$ increases and the slope of regressing $f_{xz}$ on $\hat{f}_{xz}$ is closer to 1 for $n = 150$.

The overall goodness-of-fit using p-RMEL is very good, except there are small biases: the regression slope of $f_z$ on $\hat{f}_z$ is slightly smaller than 1 and the the regression slope of $f_{xz}$ on $\hat{f}_{xz}$ is slightly larger than one. This means that $f_z$ is overestimated and $f_{xz}$ is slightly underestimated. However, for each $f_z$ and $f_{xz}$, the fitted results can explain most of the variations as all the $R^2$ values are very close to 1. We also realized that the fitting of $f_z + f_{xz}$ is much better than individual ones (the regression parameters of $f_z + f_{xz}$ on $\hat{f}_z + \hat{f}_{xz}$ are not shown), which is easy to be understood if we can treat $r = r_z + r_{xz}$ as one random effect with covariance $\tau_z K_z + \tau_{xz} K_{xz}$. This indicates that there is no bias in fitting $f_z + f_{xz}$, but the weight between $f_z$ and $f_{xz}$ might be biased. The reason for this can be understood from the interaction kernel expression (6). It can be seen that if the entries of matrix $xx' + k_z(x, x')$ are close to each other, then $\tau_z K_z + \tau_{xz} K_{xz}$ is nothing more than a scalar times $K_z$, and we will have overestimation of $f_z$. However, this bias is not too significant, because the good fit of $f_z + f_{xz}$ and the high $R^2$ values of fitting $f_{xz}$ indicate that it has little influence on testing either the overall pathway or the P-E interaction effect.
5.2 Test Study

To obtain better convergence, for the rest of this paper we adopt the Marquardt procedure as a scoring method. With the Marquardt method we have flexible iteration steps, this is

\[ \theta^{(k+1)} = \theta^{(k)} + \left[ I(\theta^{(k)}) + \delta^{(k)} I \right]^{-1} \frac{\partial l_R}{\partial \theta} \bigg|_{\theta^{(k)}} , \]

where \( l_R, I, \) and \( \theta \) are replaced by the counterparts of the p-REML procedure when it is required. The scalar \( \delta^{(k)} \) partially determines the step size and \( I \) is the identity matrix. If \( \delta^{(k)} \) is small, the procedure approximates a scoring method. If \( \delta^{(k)} \) is large, a small step is taken in approximately the direction of the scoring method. We modify \( \delta^{(k)} \) accordingly to achieve increasing likelihood. In this paper, we start from \( \delta^{(k)} = (1.0 \times 10^{-5}) \times \frac{\text{Tr}(I(\theta))}{\text{number of } \sigma^2_i} \) to make the initial step size as large as possible.

We first studied the performance of RLRT of testing two zero variance components under hypothesis (24). In this simulation study we are particularly interested in two issues: how RLRT performs at different fixed \( \rho \) values since we prefer to estimate the parameters with \( \rho \) fixed and how the performance degrades with irrelevant genes included in the model. The true model used and the data generating method are the same as described for (34) in Section 5.1. For both issues, we first set \( a = 0 \) and vary \( b \), and then set \( b = 0 \) and vary \( a \). It turns out the test is very powerful when both \( a \) and \( b \) are not equal to zero, so no simulation on this situation is shown here. For all cases, the total number of simulation runs is 1000 times. In addition, the function \( f_x(\cdot) \) in (34) has a trivial nonlinear component, so we can apply RLRT in this simulation.

For the first issue, we consider the case where the sample size is \( n = 100 \), and both the true and used gene numbers are \( p = 30 \). Table 2 presents the Type II errors and powers of testing hypothesis (24) for 2 sets of \( \{a, b\} \) values at 4 different \( \rho \) values (one is estimated). In general, the power curve of RLRT does not depend on \( \rho \) very much. Liu et al. (2007)
revealed the same phenomena for the score test of a single variance component within a model with only one random effect. This is because moderate differences of $\rho$ do not change the structure of the covariance matrix very much, except for extreme values such as $\rho \to 0$ or $\rho \to \infty$, with which the covariance matrix turns to an identity matrix or a matrix of ones. Note that the empirical Type II errors of all situations are around 0.03, smaller than the nominal one. The reason could be the approximation of (28) due to the assumption, $\theta_1 = \lambda_x^{-1} \approx 0$.

To test two zero variance components with extra genes, we consider simulations with the sample sizes $n = 60$ and $n = 35$. The latter mimics the Type II diabetes data where the total subjects under study are $n = 35$. Fitting with the equal true and used gene numbers is compared to fitting with an extra 20 irrelevant genes. The results in Table 3 show that, when fitting with extra genes, the power decreases as expected but not dramatically, which means that the model we proposed can be applied to pathway data for which only some of the genes are related to the responses. In addition, comparing Table 2 and 3 shows that the power does decrease with the sample size $n$.

The simulation study for testing P-E interaction using RLRT and the score test is carried out using a new setup for the data generation. We continue using the same nonparametric expression (34) except with true gene number $p = 5$; that is, simply replacing $f_z(\cdot)$ and $f_{xz}(\cdot)$ as $f_z(z_i^T) = a \cdot z_i^{(5)} \exp\left(-0.2 |\bar{z}_i|^{(5)}\right)/5$ and $f_{xz}(x_i, z_i^T) = b \cdot e^{x_i/10} \sin\left(\bar{z}_i^{(5)}\right) \cos\left(\bar{z}_i^{(5)}\right)/8$, where $z_i^{(5)} = \sum_{j=1}^5 z_{ij}$, $|\bar{z}_i|^{(5)} = \sum_{j=1}^5 |z_{ij}|/5$ and $\bar{z}_i^{(5)} = \sum_{j=1}^5 z_{ij}/5$. $x_i$, $z_{ij}$ and $\epsilon_i$ are generated the same way as before. Note the function form changes when the gene number is different in (34). We use this setup to compare two test procedures for testing (29). For the score test, we first estimate the parameters using p-REML and then calculate the statistics using expressions (17) and (33). The results are listed in Table 4. Again, we see that the test’s power does not depend on $\rho$. The results indicate that the RLRT are slightly lower in power.
and that the type I errors of the two test methods are all closer to the nominal 5% from different directions. These results indicate we can apply both test methods under suitable conditions.

6 Application to Type II Diabetes Data

We applied our mixed model (10) to a set of diabetes data from Mootha et al. (2003). They utilized the HGC-133a Affymetrix genechip with 22,283 genes to study 17 normal glucose tolerance individuals vs. 18 Type II diabetes mellitus patients. The 22,283 genes make up a total of 251 pathways. The goal of this study is to identify pathways with the highest significant overall pathway effect when an environmental variable, body mass index, is present in the model, and from them identify pathways with significant P-E interaction effect. Therefore, there are a total of 251 sets of data, each having \( n = 35 \) observations. Corresponding to each individual pathway, the data set contains \((y, X, Z)\), where \(y\) is the outcomes of glucose level, \(X\) has the same meaning as before with the first column of 1’s and the second column as the body mass index data of 35 subjects, and \(Z(n \times p)\) is the gene expression levels of each pathway, which contains the number of genes ranging from \(p = 3\) to \(p = 543\).

The fitting results of the top 20 pathways are listed in Table 5 ranked ascendingly in the \(p\)-value of testing the overall pathway effect using RLRT \(D\). It has almost an identical order of the magnitude as the \(D\). It can be seen that 19 out of the 251 pathways are significant. For each pathway, the variance components are estimated using p-REML methods and the standard error of those parameters including \(\hat{\sigma}^2\) are calculated using information matrix \((17)\) with the p-REML estimates plugged in. Again, the initial values for the variance parameters are \((\sigma^2, \lambda_x^{-1}, \lambda_z^{-1}, \lambda_{xz}^{-1})^T = (0.001, 1, 1, 1)^T\) and \(\rho\) is fixed at the average of \(||z' - z'||^2\) of different pairwise observations, which ranges from 0.1 to 1.8 for different pathways.
To show an overall view of the fitting results for 251 pathways, Figure 3 plots the four estimated variance components in the same order of the $p$-value of RLRT $D$. The straight dashed line divides the significant and insignificant pathways of RLRT. The error components, $\sigma^2$'s, are around the constant 3.0 except for those top significant pathways. This is consistent with the test results indicating that for those pathways with genes relevant to the responses, the error is reduced since part of the variation of the responses is explained by pathway main effect or P-E interaction effect. The variations of $\hat{\tau}_x$ and $\hat{\tau}_z$ seems to compensate for each other. For the top 50 pathways, $\hat{\tau}_z$'s are close to zero and $\hat{\tau}_z$ values are large. On the other side, for those pathways which are ranked as lower than 50, $\hat{\tau}_z$ values are very small and $\hat{\tau}_x$ values increase. This indicates that for those pathways not relevant enough to the response, part of the variation of response is explained by the nonlinear relationship of the responses and the environmental variable. The variation of $\hat{\tau}_{xz}$ seems less dramatic than other random effects. It does not decrease to zero for those non significant pathways, and stabilizes after the top 100 pathways. However, using the test of RLRT $d$, we show that the lower ranked pathways, ranked as [50, ..., 251], are not significant in the interaction effect. These results suggest that the body mass index is important in explaining the relationship between the glucose level and the genetic pathway since many pathways that are significant in the overall pathway effect are either significant in the interaction effect or not.

Because the distribution for $D$ is asymptotic, the $p$-value calculated based on 35 observations may not be as accurate as expected. Hence, we carried out a permutation test process to obtain the exact distribution of $D$ as follows:

- **Step 1:** We fit the observed data with the full model (10) and reduced model under hypothesis (24) using the p-REML approach. In both models, we set $\tau_x = 0$ since we assume that $\tau_x$ is insignificant when deriving (28). Then we obtained test statistics $D$, and calculated the residual $\hat{\epsilon}_0 = \hat{\epsilon}_z + \hat{\epsilon}_{xz} + \hat{\epsilon}$ using the fitted results of the full model.
from \( \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_z + \mathbf{r}_{xz} + \mathbf{e} \).

- **Step 2**: We permuted the residual \( \hat{\mathbf{e}}_0 \) to get new \( \hat{\mathbf{e}}_0^* \) and simulate outcomes as \( \mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}}_0^* \).

- **Step 3**: Based on \( \mathbf{y}^* \), \( \mathbf{X} \) and \( \mathbf{Z} \), we fit the full model (10) and reduced model under hypothesis (23) again using the p-REML approach and then calculated the test statistics \( D^* \).

- **Step 4**: We repeated Steps 2-3 for a large number of times (e.g. 10,000 times).

- **Step 5**: We obtained the empirical p-value of the RLRT by formula

\[
\text{p-value} = \frac{\text{number of } D^*\text{'s greater than } D}{\text{total number of } D^*\text{'s}}.
\]

The p-value of the permutation test of \( D \) as well as the RLRT \( D \) are listed in Table 6 in the same order of Table 5 for the top 20 pathways. Note that for RLRT if the sample size is too small such that the information matrix (22) is non positive definite, \( \phi \) in (28) cannot be calculated, so we are not able to get the asymptotic distribution of \( D \). However the information matrices of the 251 pathways under hypothesis (24) are all positive definite (not true under hypothesis (29)), so we are able to test the overall pathway effect for all using RLRT \( D \). The results of both tests are similar to each other with respect to the general rank of the significance, specifically both tests have the same top 3 pathways, which are pathways 73, 274, and 230. In addition, most of the p-values of the permutation tests are slightly larger than those of RLRT, as expected, since the permutation test is usually more conservative. Table 6 also labels those significant pathways ranked in the top 50 list according to the global score test (Goeman et al., 2004) and the forest tree method (Pang et al., 2006; Pang and Zhao, 2008), which do not take into account the environmental variable in their models. Our approach identified pathways that have either significant main pathway effect,
the interaction effect, or both, while other methods determined many as having a significant main pathway effect only. Through following one zero variance component test, we also discovered that some pathways have a significant P-E interaction effect although they may not have a significant main pathway effect.

Furthermore, the \( p \)-values of RLRT \( d \) are also listed in Table 3. There are pathways for which we are unable to calculate \( d \) because the information matrix is not positive definite. In Figure 4, the \( p \)-values of RLRT \( D \) and RLRT \( d \) of all pathways are plotted for comparison. Among the top 50 that are significant in overall pathway effect, only part of them are significant in the interaction effect, but for the remaining 151 pathways, none are significant in either interaction effect or overall pathway effect. Similar to RLRT \( D \), a permutation test process for the exact distribution of RLRT \( d \) is introduced here:

- **Step 1**: We fit the observed data with the full model (10) and reduced model under hypothesis (29) using the p-REML approach. Again in both models we assume that \( \tau_x \) is negligible. Then we obtained \( d \), and calculated the residual \( \hat{e}_0 = \hat{r}_z + \hat{\epsilon} \) using the fitted results of the full model from \( y = X\beta + r_z + r_{xz} + \epsilon \).

- **Step 2**: We permuted the residual \( \hat{e}_0 \) to get new \( \hat{e}_0^* \) and simulated outcomes as \( y^* = X\hat{\beta} + \hat{r}_z + \hat{\epsilon}_0^* \).

- **Step 3**: Based on \( y^* \), \( X \) and \( Z \), we fit the full model and reduced model under hypothesis (29) again using the p-REML approach and then calculated the test statistics \( d^* \).

- **Step 4**: We repeated Steps 2-3 a large number of times (e.g. 10,000 times).

- **Step 5**: We obtained the empirical \( p \)-value of the RLRT by formula \( p \)-value = \( \frac{\text{number of } d^*\text{'s greater than } d}{\text{total number of } d^*\text{'s}} \).
The permutation test results of RLRT $d$ are close to those of RLRT $d$ in the 20 pathways, but it is difficult to tell which one is more conservative.

We also calculated the $p$-values of testing $H_0$ using the score test approach for the top 20 pathways. Compared with the RLRT $d$ and RLRT $d$ permutation tests, the $p$-values of the score test is similar in sense of determining the significant pathways at the 5% level. Among these top 20 pathways with significant overall pathway effect, the pathways with insignificant interaction effect are $\{229, 152, 16, 236, 144, 151, 103, 271, 101, 158\}$ according to the score test, and $\{229, 152, 16, 236, 144, 151, 14, 103, 271, 150, 158\}$ according to the RLRT $d$ permutation test. Note that the difference of the two sets, $\{14, 101, 150\}$, all have marginal $p$-values for the two tests at the 5% level. If they are removed from the two sets, both tests have identical pathways which have insignificant P-E environment interaction effects.

Based on the three tests procedures, we identified the pathways with a significant P-E environment interaction effect for all tests among the top 20 pathways. They are $\{73, 274, 230, 173, 228, 172\}$ pathways at the 5% level. These pathways are known to be related to Type II diabetes. Pathway 73 is a Cysteine metabolism pathway. It is known that taurine (a semi-essential sulphur amino acid) derived from cysteine metabolism can prevent diabetes mellitus and/or insulin resistance (Franconi et al., 2006). Pathway 274 is involved in the Urea cycle and metabolism of amino groups, which has also been reported to be related to Type II diabetes (Czyzyk et al., 1989). Pathway 230 is OXPHOS_HG-U133A_probes pathway. It has been reported that genes involved in oxidative phosphorylation are coordinately upregulated with fasting hyperglycaemia in the livers of patients with Type II diabetes (Misu et al., 2007). The transcription levels of a class of genes involved in oxidative phosphorylation mechanisms are consistently lower in diabetics than in controls (Mootha et al., 2003; Misu et al., 2007). Pathway 173 is MAP00531_Glycosaminoglycan degradation pathway. It is known that Type II diabetes mellitus also induces an increased urinary
excretion of total glycosaminoglycans (Juretić et al., 2002). Pathway 228 is involved in Oxidative phosphorylation. It is known to be related to diabetes (Misu et al., 2007; Mootha et al., 2003, 2004). This pathway is a process of cellular respiration in humans (or in general eukaryotes) and contains coregulated genes across different tissues and is related to insulin/glucose disposal. It is associated with ATP synthesis, a pathway involved in energy transfer. Pathway 172 is MAP00530_Aminosugars metabolism pathway. Aminosugars (= glucosamine) have no effect on fasting blood glucose levels, glucose metabolism, or insulin sensitivity at any oral dose level in healthy subjects, individuals with diabetes, or those with impaired glucose tolerance (Simon et al., 2011).

7 Discussion

The development of a pathway-based mixed model to relate the response with genetic pathways is motivated by the fact that genes always interact with the environmental variables. Modeling the P-E interaction effect can help in further understanding the biological mechanisms underlying diseases and facilitate the discovery of potential biomarkers. However, no existing approaches are able to jointly analyze pathways with the environmental variables when P-E interaction exists.

In this paper, we have addressed a mixed effects model connecting with kernel machine methods and smoothing spline, so that we can analyze the genetic pathway data with a continuous clinical outcome when the P-E interaction effect is present in the model. We demonstrated the application of our method to a pathway data of Type II diabetes. Our approach allows us to evaluate the pathway effect and its interaction with the environmental variables by estimating the corresponding variance components and testing the significance of those parameters. Because of the high dimensional parameters space, there are usually some difficulties in solving the REML equations, such as non-positive error estimated. We
reduced the parameter space dimension in solving REML equations by introducing the p-REML approach to estimate the variance components so that the error component is always in the parameter space. The p-REML approach not only allows us to solve the REML equations efficiently, but also provides an efficient choice in testing one or two zero variance components besides the global score test, i.e. the profile restricted likelihood ratio test for testing the overall pathway effect or P-E interaction.

Modeling the linear mixed model with a kernel machine has other advantages. It allows us to choose appropriate kernels to construct the variance matrix of the random effect as well as the interaction random effect in accordance with the data structure. In this paper, we focused on the Gaussian kernel, but when the sample size is large so that the computation becomes expensive, some less computational intensive alternatives to Gaussian kernel are available, such as rational quadratic kernel: \( k(z^T, z'^T) = 1 - \|z - z'\|^2 / (\|z - z'\|^2 + c) \).

Other kernels, such as a polynomial kernel, an exponential kernel, an inverse multiquadric kernel, etc., have also been examined and can replace the Gaussian kernel in appropriate situations. Note that these kernels are similar to the Gaussian kernel in terms of reducing the dimension of the covariates through measuring the similarity of \( z \) and \( z' \). To some extent, this may be a disadvantage of the kernel method since there may be some information lost beyond the similarity of the two attributes.

Possible extensions of our method include applying the interaction kernel machine to generalized linear models. Logistic kernel machine regression with a Gaussian kernel has been developed by Liu et al. (2008), but no interaction between the genetic pathway effect and environmental variable has been considered. By adding the interaction kernel machine to a generalized linear model, our method can be applied in more general genomewide association studies, especially in the case-control studies of G/P-E interaction. The second potential extension of our method is to consider a higher dimension of environmental variables \( x_i^T \),
such as bivariate $x_i^T = (x_{i1}, x_{i2})$, longitude and latitude data, and the nonparametric function $f_x(x_i^T)$ can be fitted using thin plate splines (Gu and Wahba, 1993). With the kernel of the thin plate splines, we can construct the interaction function space kernel similarly. This extension may have wider applications such as in spatial data where the interaction between location and other high dimensional covariates are particularly interesting.

We note that we evaluate the interaction between each pathway and environmental variable. It is known that pathways are not independent of each other because of shared genes and interactions among pathways as well as their interaction with environmental variables, making it difficult to adjust the $p$-value due to the complex dependency structure. Because existing multiple comparison methods based on false discovery rates (Benjamini and Hochberg, 1995; Storey, 2002) were developed only for single gene based analysis that did not take into account the interaction between genes and environmental variables, they are not applicable in such a complicated situation as our problem. Developing a multiple comparison method will be an interesting and challenging problem because of the complex dependence structure among pathways and environmental variables.

**Acknowledgements**

This study was supported in part by the National Science Foundation grant number 0964680.
References

Adami, H. O., Hunter, D., and Trichopoulos, D. (2008). *Textbook of Cancer Epidemiology*. New York: Oxford University Press.

Aronszajn, N. (1950). Theory of Reproducing Kernels, *Transactions of the American Mathematical Society*, 68, 337-404.

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.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37, 373-384.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88, 9-25.

Chakravarti, A. and Little, P. (2003). Nature, Nurture, and Human Disease. *Nature*, 421, 412-414.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful Multilocus Tests of Genetic Association in the Presence of Gene-gene and Gene-environment Interactions. *The American Journal of Human Genetics*, 79, 1002-1016.

Claeskens, G. (2004). Restricted Likelihood Ratio Lack-of-fit Tests Using Mixed Spline Models. *Journal of the Royal Statistical Society, Series B*, 66, 909-926.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. P. (2005). Exact Likelihood Ratio Tests for Penalized Splines. *Biometrika*, 92, 91-103.
Czyzyk, A., Lao, B., Orowska, K., Szczepanik, Z., and Bartosiewicz, W. (1989). Effect of Anti-diabetics on Post-exercise Alaninemia in Patients with Non-insulin-dependent Diabetes Mellitus (Type 2). *Polskie Archiwum Medycyny Wewntrznej, 81*, 193-206.

Franconi, F., Loizzo, A., Ghirlanda, G., and Seghieri, G. (2006). Taurine Supplementation and Diabetes Mellitus. *Current Opinion in Clinical Nutrition & Metabolic Care, 9*, 32-36.

Goeman, J. J., van de Geer, S. A., de Kort F., and van Houwelingen, H. C. (2004). A Global Test for Groups of Genes: Testing Association with a Clinical Outcome. *Bioinformatics, 20*, 93-99.

Goeman, J. J., Oosting, J., Cleton-Jansen, A., Anninga, J. K., and van Houwelingen, H. C. (2005). Testing Association of a Pathway with Survival Using Gene Expression data. *Bioinformatics, 21*, 1950-1957.

Green, P. J. (1987). Penalized Likelihood for General Semi-parametric Regression Models. *International Statistical Review, 55*, 245-259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Gu, C. and Wahba, G. (1993). Semiparametric Analysis of Variance with Tensor Product Thin Plate. *Journal of the Royal Statistical Society, Series B, 55*, 353-368.

Guo, W. (2002). Inference in Smoothing Spline Analysis of Variance. *Journal of the Royal Statistical Society, Series B, 64*, 887-898.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association, 72*, 320-338.
Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London; New York: Chapman and Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hahn, L. W., Ritchie, M. D., and Moore, J. H. (2003). Multifactor Dimensionality Reduction Software for Detecting Gene-gene and Gene-environment Interaction. *Bioinformatics*, **19**, 376-382.

Jennrich, R. J. and Schluchter, M. D. (1986). Unbalanced Repeated Measures Models with Structured Covariance Matrices. *Biometrics*, **42**, 805-820.

Juretić, D., Krajnović, V., and Lukac-Bajalo, J. (2002). Altered Distribution of Urinary Glycosaminoglycans in Diabetic Subjects. *Acta Diabetologica*, **39**, 123-8.

Kim, I., Pang, H., and Zhao, H. (2011). Semiparametric Methods for Evaluating Pathway Effects on Clinical Outcomes Using Gene Expression Data. *Technical Report*.

Kimeldorf, G. and Wahba, G. (1971). Some Results on Tchebychefian Spline Functions. *Journal of Mathematical Analysis and Applications*, **33**, 82-95.

Lin, X. (1997). Variance Component Testing in Generalized Linear Models with Random Effects. *Biometrika*, **84**, 309-326.

Lin, X. and Zhang, D. (1999). Inference in Generalized Additive Mixed Models by Using Smoothing Splines. *Journal of the Royal Statistical Society, Series B*, **61**, 381-400.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and Testing for the Effect of a Genetic Pathway on a Disease Outcome Using Logistic Kernel Machine Regression via Logistic Mixed Models. *BMC Bioinformatics*, **9**, 292.
Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multi-Dimensional Genetic Pathway Data: Least Squares Kernel Machines and Linear Mixed Models. *Biometrics, 63*, 1079-1088.

MacKay, D. J. C. (1998). Introducing to Gaussian Process. In Bishop, C. M., editor, *Neural Networks and Machine Learning*. New York: Springer-Verlag.

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.

Manolio, T. A., Bailey-Wilson, J. E., and Collins, F. S. (2006). Genes, Environment and the Value of Prospective Cohort Studies. *Nature Review Genetics, 7*, 812-820.

Misu, H., Takamura, T., Matsuzawa, N., Shimizu, A., Ota, T., Sakurai, M., Ando, H., Arai, K., Yamashita, T., Honda, M., Yamashita, T., and Kaneko, S. (2007). Genes Involved in Oxidative Phosphorylation are Cooperatively Upregulated with Fasting Hyperglycaemia in Livers of Patients with Type 2 Diabetes. *Diabetologia, 50*, 268-277.

Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics Challenges for Genome-wide Association Studies. *Bioinformatics, 26*, 445-455.

Mootha, V. K., Handschin, C., Arlow, D., Xie, X., Pierre, J. S., Sihag, S., Yang, W., Altshuler, D., Puigserver, P., Patterson, N., Willy, P. J., Schulman, I. G., Heyman, R. A., Lander, E. S., and Spiegelman, B. M. (2004). Errα and Gabpa/b Specify PGC-1α-dependent Oxidative Phosphorylation Gene Expression that is Altered in Diabetic Muscle. *Proceedings of the National Academy of Sciences, 101*, 6570-6575.

Mootha, V. K., Lindgren, C. M., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J.,
Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1 alpha-Responsive Genes Involved in Oxidative Phosphorylation are Coordinately Downregulated in Human Diabetes. *Nature Genetics*, **34**, 267-273.

Pang, H., Lin, A., Holford, M., Enerson, B., Lu, B., Lawton, M. P., Floyd, E., and Zhao, H. (2006). Pathway Analysis Using Random Forests Classification and Regression. *Bioinformatics*, **22**, 2028-2036.

Pang, H. and Zhao, H. (2008). Building Pathway Clusters form Random Forest Classification Using Class Votes. *BMC Bioinformatics*, **9**, 87.

Pang, H., Kim, I., and Zhao, H. (2011). Random Effect Model for Multiple Pathway Analysis with Applications to Type II Diabetes Microarray Data. *Technical Report*.

Park, M. Y. and Hastie, T. (2008). Penalized Logistic Regression for Detecting Gene Interactions. *Biostatistics*, **9**, 30-50.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Process for Machine Learning*. Cambridge: MIT Press.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality Reduction Reveals High-order Interactions Among Estrogen-metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*, **69**, 138-147.

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, **1**, 211-244.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.
Simon, R., Marks, V., Leeds, A., and Anderson, J. (2011). A Comprehensive Review of Oral Glucosamine Use and Effects on Glucose Metabolism in Normal and Diabetic Individuals. *Diabetes Metabolism Research and Reviews*, **27**, 14-27.

Storey, J. D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479-498.

Vu, H. T. V. and Zhou, S. (1997). Generalization of Likelihood Ration Tests under Nonstandard Conditions. *Annals of Statistics*, **25**, 897-916.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wang, K., Li M., and Bucan, M. (2007). Pathway-based Approaches for Analysis of Genomewide Association Studies. *The American Journal of Human Genetics*, **81**, 1278-1283.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric Stochastic Mixed Models for Longitudinal data. *Journal of the American Statistical Association*, **93**, 710-719.

Zhang, D. and Lin, X. (2003). Hypothesis Testing in Semiparametric Additive Mixed Models. *Biostatistics*, **4**, 57-74.

Zou, F., Huang, H., Lee, S., and Hoeschele, I. (2010). Nonparametric Bayesian Variable Selection with Applications to Multiple Quantitative Trait Loci Mapping with Epistasis and Gene-Environment Interaction. *Genetics*, **186**, 385-394.
Table 1: Assessments of estimating $f_x$, $f_z$ and $f_{xz}$ simulated by (34) using REML and p-REML procedures with $\rho$ estimated from initial value 2 or fixed at 2. Total runs number 200 for each scenario, and the average values are reported.

| $\rho$ | fitted $\rho$ | $\hat{\rho}$ | $\hat{f}_x \sim \hat{f}_x$ | $f_z \sim \hat{f}_z$ | $f_{xz} \sim \hat{f}_{xz}$ |
|--------|---------------|----------------|----------------------------|----------------|----------------|
|        | $n$ (true $p$) | $\hat{\sigma}^2$ (initial $\rho$) | Int | Slope | $R^2$ | Int | Slope | $R^2$ | Int | Slope | $R^2$ |
| REML   | 30(30) | 0.34 | 2130(2) | -0.38 | 1.00 | 0.97 | -0.01 | 10.51 | 0.90 | -0.14 | 5.19 | 0.46 |
|        | 100    | 40(30) | 0.29 | 1824(2) | -0.55 | 1.06 | 0.96 | 0.01 | 11.65 | 0.89 | -0.11 | 4.17 | 0.50 |
|        | 50(30) | 0.32 | 1929(2) | -1.53 | 1.26 | 0.96 | -0.02 | 16.07 | 0.87 | -0.13 | 5.28 | 0.48 |
|        | 150    | 30(30) | 0.26 | 1604(2) | -1.15 | 1.17 | 0.98 | 0.09 | 5.87 | 0.93 | -0.17 | 3.70 | 0.54 |
|        |        | 40(30) | 0.29 | 1814(2) | -0.68 | 1.18 | 0.97 | -0.09 | 8.65 | 0.91 | -0.15 | 3.95 | 0.48 |
|        |        | 50(30) | 0.35 | 2054(2) | -1.24 | 1.18 | 0.97 | 0.06 | 12.32 | 0.90 | -0.15 | 4.79 | 0.45 |
|        | 150    | 30(30) | 6.9e-10 | 2 | 0.10 | 0.99 | 0.99 | 0.01 | 0.85 | 0.99 | 0.01 | 1.44 | 0.90 |
|        |        | 40(30) | 8.6e-10 | 2 | 0.13 | 0.98 | 0.98 | 0.02 | 0.86 | 0.98 | 0.00 | 1.40 | 0.90 |
|        |        | 50(30) | 8.5e-10 | 2 | 0.16 | 0.98 | 0.96 | 0.01 | 0.86 | 0.98 | 0.01 | 1.41 | 0.88 |
|        | 150    | 30(30) | 8.5e-10 | 2 | 0.05 | 0.99 | 0.99 | 0.01 | 0.84 | 0.99 | 0.01 | 1.41 | 0.93 |
|        |        | 40(30) | 8.7e-10 | 2 | -0.00 | 1.00 | 0.99 | 0.00 | 0.84 | 0.99 | -0.01 | 1.40 | 0.92 |
|        |        | 50(30) | 7.1e-10 | 2 | 0.10 | 0.99 | 0.99 | 0.02 | 0.85 | 0.99 | 0.00 | 1.38 | 0.91 |
|        | 150    | 30(30) | 0.04 | 3.96(2) | -0.24 | 1.04 | 1.00 | 0.01 | 0.85 | 0.99 | -0.04 | 1.38 | 0.90 |
|        |        | 40(30) | 0.07 | 3.36(2) | -0.19 | 1.03 | 1.00 | -0.01 | 0.87 | 0.99 | -0.05 | 1.46 | 0.89 |
|        |        | 50(30) | 0.09 | 4.72(2) | -0.31 | 1.04 | 1.00 | 0.06 | 0.90 | 0.98 | -0.04 | 1.44 | 0.88 |
|        | 150    | 30(30) | 0.02 | 3.00(2) | -0.28 | 1.04 | 1.00 | 0.01 | 0.85 | 0.99 | -0.05 | 1.29 | 0.92 |
|        |        | 40(30) | 0.02 | 3.63(2) | -0.29 | 1.04 | 1.00 | 0.01 | 0.86 | 0.99 | -0.04 | 1.29 | 0.91 |
|        |        | 50(30) | 0.04 | 3.19(2) | -0.13 | 1.02 | 1.00 | 0.01 | 0.85 | 0.99 | -0.02 | 1.37 | 0.91 |
|        | 150    | 30(30) | 0.04 | 2 | -0.08 | 1.01 | 1.00 | 0.02 | 0.85 | 0.99 | -0.01 | 1.64 | 0.91 |
|        |        | 40(30) | 0.11 | 2 | -0.17 | 1.03 | 0.99 | -0.00 | 0.88 | 0.98 | -0.03 | 1.52 | 0.91 |
|        |        | 50(30) | 0.11 | 2 | -0.12 | 1.02 | 0.99 | -0.00 | 0.90 | 0.98 | -0.01 | 1.38 | 0.91 |
|        | 150    | 30(30) | 0.02 | 2 | -0.08 | 1.01 | 1.00 | 0.02 | 0.86 | 0.99 | -0.01 | 1.34 | 0.93 |
|        |        | 40(30) | 0.03 | 2 | -0.11 | 1.02 | 1.00 | -0.01 | 0.85 | 0.99 | -0.05 | 1.37 | 0.92 |
|        |        | 50(30) | 0.04 | 2 | -0.09 | 1.01 | 1.00 | 0.02 | 0.86 | 0.99 | -0.02 | 1.44 | 0.92 |
Table 2: Simulation study for RLRT of overall pathway effect with $\rho$ fixed at different values and estimated. Simulated samples size $n = 100$, and both used and true gene number equal to $p = 30$.

| $\rho$ | $b = 0$ | 0.2 | 0.35 | 0.5 | 1          |
|--------|---------|-----|------|-----|------------|
| $a = 0$ |         |     |      |     |            |
| 2      | 0.03    | 0.34| 0.91 | 1.00| 1.00       |
| 5      | 0.02    | 0.34| 0.89 | 0.99| 1.00       |
| 10     | 0.02    | 0.30| 0.88 | 0.99| 1.00       |
| estimated | 0.03 | 0.33| 0.87 | 0.99| 1.00       |
| $a = 0$ |         |     |      |     |            |
| 2      | 0.03    | 0.07| 0.37 | 0.96| 1.00       |
| 5      | 0.02    | 0.07| 0.37 | 0.95| 1.00       |
| 10     | 0.02    | 0.06| 0.34 | 0.91| 1.00       |
| estimated | 0.03 | 0.06| 0.34 | 0.93| 1.00       |

Table 3: Simulation study for RLRT of overall pathway effect with fitted genes number $p$ equal or larger than true one $p = 30$. Simulated samples size $n = 60$ and $n = 35$. The parameter $\rho$ is fixed at 2.

|     | $n$ | used $p$ | $b = 0$ | 0.2 | 0.35 | 0.5 | 1 |
|-----|-----|----------|---------|-----|------|-----|---|
| $a = 0$ | 60  | 30       | 0.03    | 0.18| 0.57 | 0.88| 1.00 |
|      |     | 50       | 0.03    | 0.15| 0.48 | 0.76| 0.99 |
|      | 35  | 30       | 0.04    | 0.10| 0.27 | 0.46| 0.85 |
|      |     | 50       | 0.03    | 0.08| 0.23 | 0.38| 0.78 |
| $b = 0$ | 60  | 30       | 0.03    | 0.15| 0.51 | 0.72| 0.72 |
|      |     | 50       | 0.03    | 0.13| 0.41 | 0.72| 0.76 |
|      | 35  | 30       | 0.04    | 0.09| 0.25 | 0.56| 0.63 |
|      |     | 50       | 0.03    | 0.05| 0.18 | 0.43| 0.55 |
Table 4: Simulation study for PLRT and score test of P-E interaction with \( \rho \) fixed at different values. Fitted and used gene numbers are equal to \( p = 5 \), and \( n = 100 \).

| \( \rho \) | \( b = 0 \) | 0.1  | 0.2  | 0.35 | 0.5  | 0.8  | 1  |
|----------|------------|------|------|------|------|------|----|
| 2        | 0.04       | 0.24 | 0.58 | 0.95 | 1.00 | 1.00 | 1.00 |
| RLRT     | 5          | 0.04 | 0.24 | 0.64 | 0.98 | 1.00 | 1.00 | 1.00 |
|          | 10         | 0.03 | 0.24 | 0.67 | 0.97 | 1.00 | 1.00 | 1.00 |
| score    | 2          | 0.08 | 0.31 | 0.68 | 0.98 | 1.00 | 1.00 | 1.00 |
| test     | 5          | 0.06 | 0.30 | 0.72 | 0.97 | 1.00 | 1.00 | 1.00 |
|          | 10         | 0.06 | 0.26 | 0.72 | 0.98 | 1.00 | 1.00 | 1.00 |
Table 5: Estimated parameters of top 20 pathways obtained from p-REML and ranked by $p$-values of testing RLRT $D$. The numbers in the round brackets are the standard errors.

| pathway | ID | gene# | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\sigma}^2$ | $\hat{\tau}_x$ | $\hat{\tau}_z$ | $\hat{\tau}_{xz}$ | $\rho$ | $D$ | $p$-value |
|---------|----|-------|------------------|------------------|------------------|-----------------|-----------------|-----------------|------|-----|-----------|
| 73      | 11 | 5.09(1.51) | -0.01(0.21) | 0.08(0.39) | 1.0e-11(0.02) | 6.09(3.12) | 17.7(11.8) | 0.457 | 12.2 | 0.001 |
| 274     | 16 | 7.25(1.35) | 0.20(0.16) | 0.66(0.71) | 2.1e-09(0.02) | 4.74(3.09) | 9.74(8.90) | 0.581 | 7.68 | 0.006 |
| 230     | 121 | 5.69(1.39) | 0.15(0.14) | 0.10(1.03) | 7.3e-11(0.02) | 5.75(3.42) | 6.17(6.99) | 0.330 | 7.81 | 0.006 |
| 229     | 133 | 5.82(1.13) | 0.15(0.12) | 1.29(1.28) | 1.7e-03(0.02) | 3.25(2.99) | 3.96(6.31) | 0.289 | 6.65 | 0.012 |
| 152     | 11 | 6.13(1.12) | 0.21(0.15) | 2.16(0.91) | 8.6e-09(0.02) | 1.57(2.21) | 7.48(8.69) | 1.266 | 6.20 | 0.014 |
| 16      | 49 | 5.76(1.00) | 0.14(0.13) | 1.98(1.24) | 1.5e-08(0.02) | 1.89(2.55) | 4.57(6.82) | 0.308 | 5.93 | 0.017 |
| 173     | 11 | 6.06(1.07) | 0.19(0.15) | 2.14(0.92) | 2.1e-09(0.01) | 1.57(2.22) | 7.10(7.93) | 0.756 | 5.77 | 0.017 |
| 236     | 22 | 6.27(1.06) | 0.23(0.15) | 2.10(1.06) | 1.4e-08(0.02) | 1.41(2.24) | 7.24(8.08) | 0.862 | 5.63 | 0.019 |
| 144     | 7  | 5.43(1.21) | 0.15(0.20) | 2.35(0.85) | 1.6e-03(0.02) | 1.16(2.26) | 11.5(11.7) | 0.411 | 5.35 | 0.019 |
| 151     | 20 | 6.08(1.04) | 0.22(0.14) | 2.15(1.06) | 7.5e-09(0.02) | 1.52(2.24) | 6.21(7.51) | 0.937 | 5.62 | 0.019 |
| 14      | 49 | 6.09(1.20) | 0.16(0.14) | 1.57(1.27) | 1.3e-09(0.02) | 2.76(2.91) | 5.72(7.42) | 0.706 | 5.30 | 0.024 |
| 228     | 43 | 6.16(0.77) | 0.20(0.14) | 2.88(1.18) | 7.4e-11(0.02) | 0.03(1.73) | 5.91(6.86) | 0.374 | 4.95 | 0.028 |
| 103     | 37 | 6.09(0.90) | 0.20(0.14) | 2.58(1.20) | 9.8e-09(0.02) | 0.74(2.08) | 5.76(7.42) | 0.751 | 4.82 | 0.030 |
| 271     | 37 | 6.20(0.92) | 0.22(0.14) | 2.45(1.23) | 7.5e-12(0.02) | 0.94(2.19) | 5.73(7.19) | 0.702 | 4.83 | 0.030 |
| 150     | 21 | 5.98(0.94) | 0.19(0.14) | 2.54(1.12) | 7.5e-11(0.02) | 0.97(2.10) | 5.75(7.65) | 1.161 | 4.66 | 0.033 |
| 172     | 8  | 5.85(0.92) | 0.15(0.18) | 2.75(0.99) | 2.6e-02(0.02) | 3.5e-10(1.61) | 10.1(9.8) | 0.812 | 4.22 | 0.039 |
| 133     | 58 | 6.01(0.83) | 0.18(0.14) | 2.71(1.29) | 1.8e-03(0.02) | 0.32(2.04) | 6.28(7.18) | 0.339 | 4.15 | 0.044 |
| 8       | 27 | 5.87(0.78) | 0.18(0.15) | 2.92(1.15) | 1.6e-02(0.04) | 3.0e-09(1.72) | 5.96(7.21) | 0.527 | 4.08 | 0.045 |
| 101     | 13 | 6.08(0.90) | 0.19(0.16) | 3.01(1.01) | 5.7e-10(0.02) | 0.23(1.59) | 6.81(8.79) | 0.458 | 3.88 | 0.045 |
| 158     | 8  | 5.79(1.00) | 0.15(0.14) | 2.55(0.98) | 1.3e-09(0.02) | 1.55(2.24) | 5.39(7.72) | 0.621 | 3.53 | 0.056 |
Table 6: P-values of different tests for top 20 pathway significant in the overall pathway effect. Columns 2 and 3 are labels indicating appearance in the top 50 list of other methods or not. Missing values in column 6 is because the information matrix is not positive definite.

| pathway ID | Global Score Test | Forest Score Test | RLRT test for D | permutation test for D | RLRT test for d | permutation test for d | score test for $U_{rzz}$ |
|------------|-------------------|-------------------|-----------------|------------------------|-----------------|------------------------|------------------------|
| 73         | Yes               | Yes               | 0.001           | 0.001                  | 0.002           | 0.001                  | 0.005                  |
| 274        | Yes               | No                | 0.006           | 0.011                  | 0.025           | 0.013                  | 0.016                  |
| 230        | Yes               | Yes               | 0.006           | 0.010                  | -               | 0.025                  | 0.007                  |
| 229        | Yes               | Yes               | 0.012           | 0.020                  | -               | 0.138                  | 0.062                  |
| 152        | No                | No                | 0.014           | 0.015                  | 0.179           | 0.303                  | 0.163                  |
| 16         | Yes               | Yes               | 0.017           | 0.027                  | 0.126           | 0.147                  | 0.058                  |
| 173        | Yes               | Yes               | 0.017           | 0.020                  | 0.017           | 0.018                  | 0.002                  |
| 236        | No                | No                | 0.019           | 0.021                  | 0.133           | 0.119                  | 0.104                  |
| 144        | Yes               | Yes               | 0.019           | 0.020                  | 0.076           | 0.072                  | 0.106                  |
| 151        | No                | No                | 0.019           | 0.023                  | 0.205           | 0.262                  | 0.146                  |
| 14         | Yes               | No                | 0.024           | 0.031                  | 0.113           | 0.054                  | 0.046                  |
| 228        | Yes               | Yes               | 0.028           | 0.035                  | 0.032           | 0.024                  | 0.006                  |
| 103        | No                | Yes               | 0.030           | 0.039                  | 0.121           | 0.106                  | 0.086                  |
| 271        | No                | No                | 0.030           | 0.037                  | 0.148           | 0.142                  | 0.110                  |
| 150        | No                | No                | 0.033           | 0.034                  | 0.080           | 0.062                  | 0.044                  |
| 172        | No                | No                | 0.039           | 0.044                  | 0.016           | 0.015                  | 0.009                  |
| 133        | No                | No                | 0.044           | 0.057                  | 0.053           | 0.043                  | 0.018                  |
| 8          | Yes               | Yes               | 0.045           | 0.052                  | 0.051           | 0.038                  | 0.032                  |
| 101        | No                | No                | 0.045           | 0.044                  | 0.068           | 0.049                  | 0.056                  |
| 158        | Yes               | No                | 0.056           | 0.054                  | -               | 0.343                  | 0.560                  |
Figure 1: Diagram of the parameter space of RLRT for testing two zero variance components (a), and testing the P-E interaction effect (b).
Figure 2: Selected example of fitting results of setting 1. Because of the high dimensionality, $r_z, r_{xz}$ and $f$ are plotted vs. the observation index only.
Figure 3: The estimated variance components of $\hat{\sigma}^2$, $\hat{\tau}_x$, $\hat{\tau}_z$, $\hat{\tau}_{xz}$ for 251 pathways ordered by $p$-values of testing the overall pathway effect. The dash lines separate the significant and insignificant pathways at 5% level.
Figure 4: The $p$-values of testing overall pathway effect (RLRT $D$) and P-E interaction effect (RLRT $d$) for 251 pathways. The vertical dash line divides the significant and insignificant pathways of overall pathway effect test, and the horizontal dash line indicates 5% significant level. Some $p$-values of RLRT $d$ are missing because the information matrix is not positive definite.
Appendix A  The Representation of the Natural Cubic Spline

Following Green and Silverman (1994), the representation of the natural cubic spline (8) in section 2.2 is called the value-second derivative representation. Details for defining matrices $B$ and $M$ are shown as the following.

Suppose $f_x$ is the natural cubic spline with $n$ distinct $x_1^0, ..., < x_n^0$. Define

$$f_{x,i} = f_x(x_i^0) \text{ and } \gamma_i = f''_x(x_i^0) \text{ for } i = 1, ..., n$$

By the definition of natural cubic spline, $\gamma_1 = \gamma_n = 0$. Let $f_x$ stands for the vector $(f_{x,1}, ..., f_{x,n})^T$ and let $\gamma = (\gamma_2, ..., \gamma_{n-1})^T$ where $\gamma$ is a $(n - 2) \times 1$ vector with the element index starting at $i = 2$. Now define two matrices, $Q$ and $R$. Let $h_i = t_{i+1} - t_i$ for $i = 1, ..., n - 1$. Let $Q$ be the $n \times (n - 2)$ matrix with entries $q_{ij}$, for $i = 1, ..., n - 1$ and $j = 2, ..., n - 1$, given by

$$q_{j-1,j} = h_{j-1}^{-1}, \quad q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \quad \text{and} \quad q_{j+1,j} = h_j^{-1}$$

for $j = 2, ..., n - 1$ and $q_{ij} = 0$ for $|i - j| \geq 2$. The columns of $Q$ are indexed in the same way as the elements of $\gamma$ starting at $j = 2$, so that the first element of $Q$ is $q_{12}$.

$R$ is a $(n - 2) \times (n - 2)$ symmetric matrix with elements $r_{ij}$, for $i$ and $j$ running from 2 to $n - 1$, given by

$$r_{ii} = \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, ..., n - 1, \quad r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i \text{ for } i = 2, ..., n - 2,$$

and $r_{ij} = 0$ for $|i - j| \geq 2$.

The matrix $R$ is strictly diagonal dominant and strictly positive definite. Using the Cholesky factorization that avoids taking the square roots (Green and Silverman 1994).
Section 2.6.1, we can factorize $R$ as

$$R = U \Lambda U^T,$$

where $\Lambda$ is a diagonal matrix and $U$ is a lower triangular band matrix with diagonal elements all equal to 1. Since $R$ are strictly positive definite, all diagonal elements of $\Lambda$ are positive, $R^{-1} = (\Lambda^{1/2}U^T)^{-1}(U\Lambda^{1/2})^{-1}$. The penalty matrix $M$ can be expressed as

$$M = QR^{-1}Q^T = Q(\Lambda^{1/2}U^T)^{-1}(U\Lambda^{1/2})^{-1}Q^T = LL^T,$$

(37)

where $L = Q(\Lambda^{1/2}U^T)^{-1}$. The $B$ matrix thus is calculated by

$$B = (LL^T)^{-1} = Q(\Lambda^{1/2}U^T)^{-1}\left\{[(\Lambda^{1/2}U^T)^{-1}]Q^TQ(\Lambda^{1/2}U^T)^{-1}\right\}^{-1}
= Q(\Lambda^{1/2}U^T)^{-1}(\Lambda^{1/2}U^T)(Q^TQ)^{-1}(\Lambda^{1/2}U^T)^T$$
$$= Q(Q^TQ)^{-1}U\Lambda^{1/2}.$$

The Theorem 2.1 in Green and Silverman (1994) states that the vectors $f_x$ and $\gamma$ specify a natural cubic spline $f_x$ if and only if the condition $Q^Tf_x = R\gamma$ is satisfied. If this condition is satisfied then the roughness penalty will satisfy

$$\int_0^1 \{f''_x(x)\}^2 dx = \sum_{j=1}^{n-1} \frac{\gamma_{j+1} - \gamma_j}{h_j} (f_{x,j} - f_{x,j+1}) = \gamma^T Q^T f_x$$
$$= \gamma^T R\gamma = f_x^T QR^{-1}Q^T f_x = f_x M f_x.$$

In the above derivation we assumed that $x^0_i, i = 1, \ldots, n$, were distinct and ordered, so the rank of the penalty matrix $M$ is $n - 2$ and $B$ is a $n \times (n - 2)$ matrix. In our model, we shall have $r$ distinct and ordered $x^0_i, i = 1, \ldots, r$, from the observed data $x_i, i = 1, \ldots, n$, where $r \leq n$ and $x_i$'s may not be ordered. Based the $r$ $x^0_i$'s, $B$ is a $r \times (r - 2)$ matrix. Thus we will use a $n \times r$ incidence matrix $N$ defined in a way similar to that given by Green and Silverman (1994), Section 4.3.1, such that $B = NB$, where the left $B$ is what we shall use in the model, and the right $B$ is calculated based on $r$ distinct $x^0_i$'s.