Published in final edited form as:

J Stat Plan Inference. 2013 April ; 143(4): 745–763.

Statistical properties on semiparametric regression for evaluating pathway effects

Inyoung Kim^{a,*}, Herbert Pang^b, and Hongyu Zhao^{c,d}

^aDepartment of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

^bDepartment of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705, USA

^cDivision of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA

^dDepartment of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

Abstract

Most statistical methods for microarray data analysis consider one gene at a time, and they may miss subtle changes at the single gene level. This limitation may be overcome by considering a set of genes simultaneously where the gene sets are derived from prior biological knowledge. We call a pathway as a predefined set of genes that serve a particular cellular or physiological function. Limited work has been done in the regression settings to study the effects of clinical covariates and expression levels of genes in a pathway on a continuous clinical outcome. A semiparametric regression approach for identifying pathways related to a continuous outcome was proposed by Liu et al. (2007), who demonstrated the connection between a least squares kernel machine for nonparametric pathway effect and a restricted maximum likelihood (REML) for variance components. However, the asymptotic properties on a semiparametric regression for identifying pathway have never been studied. In this paper, we study the asymptotic properties of the parameter estimates on semiparametric regression and compare Liu et al.'s REML with our REML obtained from a profile likelihood. We prove that both approaches provide consistent estimators, have $\sqrt{n}$ convergence rate under regularity conditions, and have either an asymptotically normal distribution or a mixture of normal distributions. However, the estimators based on our REML obtained from a profile likelihood have a theoretically smaller mean squared error than those of Liu et al.'s REML. Simulation study supports this theoretical result. A profile restricted likelihood ratio test is also provided for the non-standard testing problem. We apply our approach to a type II diabetes data set (Mootha et al., 2003).

Keywords

Gaussian random process; Kernel machine; Mixed model; Pathway analysis; Profile likelihood; Restricted maximum likelihood

1. Introduction

Numerous statistical methods have been developed for analyzing microarray data based on single genes (Efron et al., 2004; Fan and Li, 2001; Tibshirani, 1996; Zou and Hastie, 2005).

However, they may not detect coordinated yet subtle changes among a set of genes where each gene only shows modest changes. One way to address the limitation of single gene based analysis is to analyze gene sets derived from prior biological knowledge to uncover patterns among the genes within a set. A number of methods and programs have been developed to consider gene groupings based on gene ontology (GO) (Harris et al., 2004). These methods have been successful to detect subtle changes in expression levels which could be missed using single gene-based analysis (Mootha et al., 2003; Hosack et al., 2003; Rajagopalan and Agarwal, 2005). In our following discussion, we consider a pathway as a predefined set of genes that serve a particular cellular or physiological function.

Limited work has been done in the regression settings to study the effects of clinical covariates and expression levels of genes in a pathway on a continuous clinical outcome. Goeman et al. (2004) proposed a global test derived from a random effects model. Liu et al. (2007) proposed a semiparametric regression model for high dimensional covariate data. Liu et al.'s approach is developed by connecting between a least squares kernel machine in a nonparametric model with REML in a linear mixed model. However, statistical properties of their approach have not been studied. In the rest of this paper, we refer Liu et al.'s approach to LLG's approach.

The goal of our study is to provide the asymptotic properties of these estimators on consistency, convergence rate, and limiting distributions under regularity conditions. We compare LLG's estimators with ours which are estimated using a REML obtained from a profile likelihood. We note that although our likelihood is a penalized likelihood, we simply use it as a profile likelihood. Note that there is a penalized likelihood called penalized Henderson's likelihood (Wang, 1998; Gu and Ma, 2005) which are obtained using smoothing splines.

We show that our REML estimators give more accurate score equations and information matrix, and have smaller mean squared errors (MSEs) than those of LLG's REML. For non-standard testing problems, a profile restricted likelihood ratio test is also provided.

This paper is organized as follows. In Section 2, we give the definition of the least square kernel machine estimation in the semiparametric model. We then explain LLG's likelihood-based approach on semiparametric regression and compare their method with our REML obtained from a profile likelihood approach. Section 3 describes the asymptotic properties for both estimators. In Section 4, we provide a testing procedure based on a profile restricted likelihood ratio test for nonstandard testing problem and also derive its theoretical distribution. In Section 5 we report the simulation results comparing MSEs, type I errors, and power of LLG's estimators with those. In Section 6, we apply our approach to the type II diabetes data set analyzed by Mootha et al. (2003). Section 7 contains concluding remarks.

2. Semiparametric regression

A semiparametric regression model can be written as

$$Y = X\beta + \mathbf{r}(Z) + \epsilon, \quad (1)$$

where Y is an $n \times 1$ vector denoting the continuous outcome measured on n subjects, X is an $n \times q$ matrix representing q clinical covariates of these subjects, β is a $q \times 1$ vector of regression coefficients for the covariate effects, Z is a $p \times n$ matrix denoting the gene expression matrix for p genes ($p \gg n$), $Z = [\mathbf{z}_1, \dots, \mathbf{z}_n]$, and $\mathbf{z}_i$ is a $p \times 1$ vector for the gene expression levels of the i th subject, $\mathbf{r}(\cdot)$ is an unknown nonlinear smooth function, and $\epsilon \sim$

$\mathcal{N}(\mathbf{0}, \sigma^2 I)$, where $\sigma^2 > 0$. Because of the high dimensional space of Z , we make statistical inference for the model (1) by connecting a least squares kernel machine with a Gaussian random process, where $\mathbf{r}(Z) = \{r(\mathbf{z}_1), \dots, r(\mathbf{z}_n)\} \sim \mathcal{N}(\mathbf{0}, \tau \mathbf{K})$ with $r(\cdot)$ following a Gaussian process (GP) with mean 0 and covariance $\text{cov}\{r(\mathbf{z}), r(\mathbf{z}')\} = \tau K(\mathbf{z}, \mathbf{z}')$, $\tau > 0$, $\mathbf{z}$ and $\mathbf{z}'$ represent two different arbitrary $p \times 1$ vectors, $K(\cdot, \cdot)$ is a kernel function which implicitly specifies a unique function space spanned by a particular set of orthogonal basis functions, and $\mathbf{K}$ is an $n \times n$ matrix with the ij th component $K(\mathbf{z}_i, \mathbf{z}_j)$, $i, j = 1, \dots, n$.

The linear effects of the clinical variables are adjusted in this model. This model reduces to the standard linear regression model when $\tau = 0$. If $\tau > 0$, two samples i and j with similar gene expression patterns have correlated random effects $r(\mathbf{z}_i)$ and $r(\mathbf{z}_j)$, and therefore they have a greater probability of having similar outcomes y_i and y_j than samples with less similar expression patterns. This “similarity” is measured using a kernel function.

We consider the following three kernels $K(\cdot, \cdot)$ to model the covariance matrix of pathway effects:

- The d th order polynomial kernel, $K(\mathbf{z}, \mathbf{z}') = (\mathbf{z}\mathbf{z}')^d$, $d = 1, 2$, which quantifies similarity through the inner-product.
- Gaussian kernel, $K(\mathbf{z}, \mathbf{z}') = \exp(-\|\mathbf{z} - \mathbf{z}'\|^2 / \rho)$, where $\|\mathbf{z} - \mathbf{z}'\|^2 = \sum_{k=1}^p (z_k - z'_k)^2$ and ρ is an unknown scale parameter ($\rho > 0$). The “similarity” in this kernel is measured using the Euclidean distance between $\mathbf{z}$ and $\mathbf{z}'$.
- Neural network kernel, $K(\mathbf{z}, \mathbf{z}') = \tanh(\mathbf{z} \cdot \mathbf{z}')$, which uses the hyperbolic tangent (tanh) function to quantify similarity.

By connecting a least squares kernel machine (Cristianini and Shawe-Taylor, 2006) with a restricted maximum likelihood (REML) (Zhang et al., 1998; Wang, 1998), Liu et al. (2007) estimated the nonparametric pathway effects of multiple gene expressions, $\mathbf{r}(Z)$, using the least squares kernel machine, estimated the covariate effects, $\boldsymbol{\beta}$, using the weighted least squares estimator, and estimated the parameters of the variance components using REML. They derived the score equations and information matrix of variance component parameters by treating the estimated $\hat{\boldsymbol{\beta}}$ as if they were the true parameter without uncertainty. However, because $\hat{\boldsymbol{\beta}}$ depends on the variance components parameters, we consider incorporating the relationship between $\hat{\boldsymbol{\beta}}$ and the variance components parameters into the REML to obtain more accurate score equations and information matrix of these parameters.

We adopt the same general approach to estimate the nonparametric function $\mathbf{r}(\cdot)$. $\boldsymbol{\beta}$ and $\mathbf{r}(\cdot)$ are estimated by maximizing the scaled penalized likelihood function with smoothing parameter λ

$$L\{\boldsymbol{\beta}, r(\cdot)\} = -\frac{1}{2} \sum_{i=1}^n \{y_i - \mathbf{x}_i \boldsymbol{\beta} - r(\mathbf{z}_i)\}^2 - \frac{1}{2} \lambda \|\mathbf{r}\|_{H_K}^2,$$

where $\|\cdot\|_{H_K}$ represents the norm induced by Kernel K of H_K , the function space generated by a kernel function K . This estimation is called least square kernel machine estimation (Liu et al., 2007).

By using the dual formulation (Cristianini and Shawe-Taylor, 2006; Liu et al., 2007) to reduce the high dimensional problem into a low dimensional one, the parameters $\boldsymbol{\beta}$ and $\mathbf{r}(\cdot)$ can be estimated as follows:

$$\begin{aligned}\widehat{\beta} &= \{X^T(I + \lambda^{-1}\mathbf{K})^{-1}X\}^{-1}X^T(I + \lambda^{-1}\mathbf{K})^{-1}\mathbf{y}, \\ \widehat{\mathbf{r}} &= \lambda^{-1}\mathbf{K}(I + \lambda^{-1}\mathbf{K})^{-1}(\mathbf{y} - X\widehat{\beta}).\end{aligned}$$

For estimating the variance components parameters, we start from a mixed model formulation with the assumption that the nonparametric function $\iota(\cdot)$ follows a Gaussian process (GP) with mean 0 and covariance $\text{cov}\{\iota(\mathbf{z}), \iota(\mathbf{z}')\} = \tau\mathbf{K}(\mathbf{z}, \mathbf{z}') = \tau \exp(-\|\mathbf{z} - \mathbf{z}'\|^2/\rho)$, where $\rho > 0$ is an unknown scale parameter. Using the marginal distribution of $\mathbf{y} \sim \mathcal{N}(X\beta, \Sigma)$, the regression coefficient estimator $\widehat{\beta}$ can be obtained from the weighted least squares estimator $\widehat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$, where $\Sigma = \sigma^2I + \tau\mathbf{K} = \Sigma(\theta)$, where $\theta = (\tau, \rho, \sigma^2)$ and $\Sigma(\theta)$ are an $n \times n$ positive definite variance matrix. The covariance of $\widehat{\beta}$ and the nonparametric function $\widehat{\mathbf{r}}(\cdot)$ can be estimated as $\text{cov}(\widehat{\beta}) = (X^T\Sigma^{-1}X)^{-1}$, $\text{cov}(\widehat{\mathbf{r}}) = \tau\mathbf{K} - (\tau\mathbf{K})\mathbf{P}(\tau\mathbf{K})$, and $\text{cov}\{\widehat{\mathbf{r}}(\mathbf{z})\} = \tau\mathbf{K}(\mathbf{z}, \mathbf{z}) - (\tau\mathbf{K}_{\mathbf{z}})\mathbf{P}(\tau\mathbf{K}_{\mathbf{z}})$, where $\mathbf{P} = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$ and $\mathbf{K}_{\mathbf{z}} = \{K(\mathbf{z}, \mathbf{z}_1), \dots, K(\mathbf{z}, \mathbf{z}_n)\}^T$. Based on the above estimates, the variance components parameters, $\theta = (\tau, \rho, \sigma^2)$, can be estimated through the REML, a commonly used approach for estimating variance components in the mixed effects model, and the smoothing parameter can be obtained from $\lambda = \tau^{-1}\sigma^2$. The REML is

$$l(\theta) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}(\mathbf{y} - X\beta)^T\Sigma^{-1}(\mathbf{y} - X\beta).$$

LLG's REML and our REML are

$$l_1(\theta) = -\frac{1}{2}(\mathbf{y} - X\widehat{\beta})^T\Sigma^{-1}(\mathbf{y} - X\widehat{\beta}) - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}\log|\Sigma|, \quad (2)$$

$$l_2(\theta) = -\frac{1}{2}\{\mathbf{y} - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}\}^T\Sigma^{-1}\{\mathbf{y} - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}\} - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}\log|\Sigma|. \quad (3)$$

LLG calculate the score equations and information matrix of $\theta = (\tau, \rho, \sigma^2)$ at given $\widehat{\beta}$. Unlike LLG's approach (2007), we consider a profile likelihood and replace $\widehat{\beta}$ by $(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}$ because $\widehat{\beta}$ depends on the parameters θ . We then obtain the score equations and information matrix of θ (see Appendix A.1). The parameters are estimated using the Newton-Raphson method to estimate all parameters.

3. Asymptotic properties

In this section, we will provide the asymptotic properties on consistency, convergence rate, and limiting distribution for the estimators. We will also show that the estimators from our REML have asymptotically smaller MSE than those of LLG's REML.

We first give a consistency result of our estimators when the true parameter θ_0 is either in the interior region or on the boundary of parameter space Ω . For this result we need that the parameter space (Ω) near θ_0 behaves like a closed set (Self and Liang, 1987; Vu and Zhou, 1997) which imposes the condition that the intersection of Ω and the closure of neighborhoods centered about θ_0 compose closed subsets. This requirement is satisfied because our Ω is a rectangle which is a cross product of intervals.

Theorem 1

If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are estimators of Θ using REML functions (2) and (3), respectively, then $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are both consistent estimators under regularity conditions.

Theorem 1 proves that the estimators from both approaches are consistent under regularity conditions. These regularity conditions are described in Appendix A.2. The proof of Theorem 1 is summarized in Appendix A.3.

Next, we will show that both approaches have $\sqrt{n}$ convergence rate. Before proving this convergence rate, we first explain the following two lemmas (Lemmas 1 and 2).

Lemma 1

If Σ and Σ_k are simultaneously diagonalizable, where Σ_k is the first derivative of Σ with respect to the k th component of Θ , then $(1/n) \text{tr}(P\Sigma_k)$ uniformly converges. That is, $(1/n) \text{tr}(P\Sigma_\tau)$ and $(1/n) \text{tr}(P\Sigma_\sigma)$ uniformly converge.

Its proof is summarized in Appendix A.4. Lemma 1 implies that $\text{tr}(P\Sigma_\tau P\Sigma_\sigma)^{1/2}$ uniformly converges because $0 \leq \text{tr}(P\Sigma_\tau P\Sigma_\sigma)^{1/2} \leq \text{tr}(P\Sigma_\tau)^{1/2} \text{tr}(P\Sigma_\sigma)^{1/2}$.

Lemma 2

If $\Sigma\Sigma_k = \Sigma_k\Sigma$, then Σ and Σ_k are simultaneously diagonalizable.

It is well known that if $AB = BA$ for two matrices A and B , then A and B are simultaneously diagonalizable. We need to use this fact in our proof.

The $\sqrt{n}$ convergence rate can be proved using (1) these two lemmas, (2) four conditions C1–C4 (see Appendix A.5), which are described in Cressie and Lahiri (1993, 1996), and (3) the condition that the intersection of Ω and the closure of neighborhoods centered about Θ_0 constitute closed subsets (Vu and Zhou, 1997).

Theorem 2

Under conditions C1–C4 in Appendix A.5, $\sqrt{n}(\hat{\Theta}_1 - \Theta) = O_p(1)$ and $\sqrt{n}(\hat{\Theta}_2 - \Theta) = O_p(1)$.

In contrast to Cressie and Lahiri (1993, 1996) who showed the asymptotic property of REML estimator in a linear model with Gaussian error, we consider a semiparametric model. The proof of Theorem 2 is described in Appendix A.6. Because the intersection of Ω and the closure of neighborhoods centered about Θ_0 constitute closed subsets, this result also holds when the true parameter is on the boundary of the parameter space Ω (Vu and Zhou, 1997). In our study we use the Gaussian, polynomial, and neural network kernels. We show that Σ and Σ_τ are simultaneously diagonalizable as well as Σ and Σ_σ . For the Gaussian kernel, using Lemma 2, we show that Σ and Σ_ρ are simultaneously diagonalizable when n goes to ∞ .

Theorem 3

(3.1) Suppose that the true parameters are the interior points of the parameter space. Then under regularity conditions, $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are asymptotically normally distributed.

(3.2) Suppose that one component of the true parameters is a left endpoint of the parameter space. Then under regularity conditions, $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are asymptotically distributed with a mixture of normal distributions.

The proof of (3.1) is explained in Appendix A.7. The proof of (3.2) can be shown by Chant's (1974) case (i).

As for the MSEs, we compare our estimator based on our REML (2) with LLG's estimator based on LLG's REML (1). We prove that the estimators from our REML have asymptotically smaller MSEs than those of LLG's REML in the following theorem.

Theorem 4

$$E\|\hat{\theta}_2 - \theta\|^2 < E\|\hat{\theta}_1 - \theta\|^2 \text{ asymptotically.}$$

Its proof is given in Appendix A.8. This result implies that our score equations and information matrix are more accurate than those of Liu et al. (2007) because we take into account the relationship between $\hat{\beta}$ and the variance components in REML.

Although estimators of parameters in the semiparametric regression approach developed by connecting a least squares kernel machine with REML have useful large sample properties, there are some situations that ρ and τ are not identifiable. This result is summarized in Theorem 5.

Theorem 5

ρ and τ are not identifiable under one of the following situations: (i) $\tau \rightarrow 0$; (ii) $\rho \rightarrow 0$ and $\tau \sim O(1/\rho^m)$ for any positive m ; or (iii) $1/\rho \rightarrow 0$.

This is because the marginal distribution of Y , $f(Y|\tau) = f(Y|\rho) \sim N(0, \sigma^2 I)$ under conditions (i) and (ii) and $f(Y|\tau) = f(Y|\rho) \sim N(0, \sigma^2 I + \tau I)$ under condition (iii). Its proof is given in Appendix A.9.

4. Test for the nonparametric function

Our main interest is to test $H_0: \mathbf{r}(Z) = \text{constant}$, where $\mathbf{r}(\cdot)$ is an unknown nonlinear smooth function under model (1), which means that a null hypothesis under the Gaussian random process using Gaussian kernel, $H_0: \{\mathbf{r}(Z) \text{ is a point mass at zero}\} \cup \{\mathbf{r}(Z) \text{ has a constant covariance matrix as a function of } Z\}$ is equivalent to $H_0: \tau/\rho = 0$ or $\rho = 0$. For the polynomial and neural network kernels, our null hypothesis is $H_0: \tau = 0$. The proof of this equivalence is given in Appendix A.10. We perform this testing based on two approaches: one is a profile restricted likelihood ratio test, which is described in Section 4.1, and the other is a permutation test, which is explained in Section 4.2.

4.1. Profile restricted likelihood ratio test

We derive a profile restricted likelihood ratio test (PRLRT) by taking into account that the parameter of interest under the null hypothesis is on the boundary of the parameter space and the kernel matrix K is not a block diagonal matrix. Theoretical properties of our profile REML estimators, which we have shown in Section 3, are needed for PRLRT. We can perform PRLRT test using both empirical distribution obtained by permutation as well as theoretical distribution derived in the following subsections for each null hypothesis.

4.1.1. Test for $H_0: \tau/\rho = 0$ vs $H_a: \tau/\rho > 0$ —Test for H_0 means that either $H_{01}: \tau = 0$ and ρ is positive and bounded or $H_{02}: \rho = \infty$ and τ is positive and bounded. The PRLRT can be derived as follows.

Test for $H_{01}: \tau = 0$ vs $\tau > 0$, where ρ is positive and bounded. Let Ω be the parameter space for $\theta = (\tau, \rho, \sigma^2)^T$. Denote Ω_0 and $\Omega_1 = \Omega \setminus \Omega_0$ as the parameter spaces under H_0 and H_a . The true parameters θ_0 are either in the interior or on the boundary of the parameter space Ω .

Assume that the parameter spaces Ω_0 and Ω_1 can be approximated at Θ_0 by cones C_{Ω_0} and C_{Ω_1} , respectively, with vertex Θ_0 .

The PRLRT statistic, D , is the deviance of two times the log PRLRT, $D = 2l_2(\Theta) - 2l_2(\Theta_0)$. By Claeskens (2004) who extended the non-standard LRT test to PRLRT, D converges to

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

where $\tilde{C} = \{\tilde{\Theta}: \tilde{\Theta} = I(\Theta_0)^{T/2}(\Theta - \Theta_0), \Theta \in C_{\Omega}\}$ is the orthonormal transformation of the cone approximation, C_{Ω} of the parameter space Ω with Θ_0 as the vertex, and $\tilde{C}_0 = \{\tilde{\Theta}: \tilde{\Theta} = I(\Theta_0)^{T/2}(\Theta - \Theta_0), \Theta \in C_{\Omega_0}\}$ is the orthonormal transformed cone approximation of the parameter space Ω_0 under the null hypothesis. U is a random vector from $N(0, I)$, and $I(\Theta_0)^{T/2}$ is the right Cholesky square root of profile REML information matrix, i.e., $I(\Theta_0) = [I(\Theta_0)]^{1/2}[I(\Theta_0)]^{T/2}$.

Under the null hypothesis, $\Theta_0 = (0, \rho, \sigma^2)^T$, ρ is inestimable. Let $\Theta = (\tau, \sigma^2)$. Now the cone parameter spaces are reduced to $C_{\Omega} = [0, \infty) \times (0, \infty)$. $C_{\Omega_0} = \{0\} \times (0, \infty)$ and $C_{\Omega_1} = (0, \infty)^2$.

Decompose normal vector U and $I(\Theta_0)$ as $U = (U_1, U_2)^T$ and $I(\Theta_0) = \{I_{jk}\}$ corresponding to τ and σ^2 . After some algebra, we can easily show that

$$\inf_{\Theta \in \tilde{C}_0} \|U - \Theta\| = (I_{11} - I_{12}I_{22}^{-1}I_{21})U_1^2$$

and

$$\inf_{\Theta \in \tilde{C}} \|U - \Theta\| = (I_{11} - I_{12}I_{22}^{-1}I_{21})U_1^2(\tilde{U}_1 \leq 0),$$

where $\tilde{U}_1 = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{1/2}U_1$. Therefore, the difference between these two becomes $(I_{11} - I_{12}I_{22}^{-1}I_{21})U_1^2(\tilde{U}_1 > 0)$. Since $(I_{11} - I_{12}I_{22}^{-1}I_{21})^{1/2}U_1 \sim N(0, 1)$, the asymptotic distribution of D under H_{01} is a 50:50 mixture between a point mass at zero and a chi-square distribution with 1 degree of freedom, i.e., $0.5\chi_0^2 + 0.5\chi_1^2$.

Test for H_{02} : $\rho = \infty$, where τ is positive and bounded. This test H_{02} : $\rho = \infty$ means that $1/\rho = 0$. Let us denote that $\rho^* = 1/\rho$. In this case, $\Theta = (\theta_1, \theta_2^T)$, where $\theta_1 = \rho^*$ and $\theta_2 = (\tau, \sigma^2)$. $C_{\Omega_0} = \{0\} \times (0, \infty)^2$ and $C_{\Omega_1} = (0, \infty)^3$. Similar to the previous case, decompose normal vector U and $I(\Theta_0)$ as $U = (U_1, U_2^T)^T$ and $I(\Theta_0) = \{I_{jk}\}$ corresponding to θ_1 and θ_2 . We can then show that

$$\inf_{\Theta \in \tilde{C}_0} \|U - \Theta\| = (I_{11} - I_{12}I_{22}^{-1}I_{21})U_1^2, \quad (4)$$

$$\inf_{\Theta \in \tilde{C}} \|U - \Theta\| = (I_{11} - I_{12} I_{22}^{-1} I_{21}) U_1^2 \quad (\tilde{U}_1 \leq 0). \quad (5)$$

In a similar way, the asymptotic distribution of D under H_{02} : $\rho = \infty$ is a 50:50 mixture between a point mass at zero and a chi-square distribution with 1 degree of freedom, i.e., $0.5\chi_0^2 + 0.5\chi_1^2$.

4.1.2. Test for H_0 : $\rho = 0$ —Test for H_0 means that either H_{03} : $\rho = 0$ and τ is positive and bounded or H_{04} : $\rho = 0$ and $\tau = 0$.

Test for H_{03} : $\rho = 0$, where τ is positive and bounded. In this case, $\Theta = (\theta_1, \theta_2^T)$, where $\theta_1 = \rho$ and $\theta_2 = (\tau, \sigma^2)$. $C_{\mathcal{Q}_0} = \{0\} \times (0, \infty)^2$ and $C_{\mathcal{Q}_1} = (0, \infty)^3$. Similar to the previous case, decompose U and $\mathcal{I}(\Theta_0)$ as $U = (U_1, U_2^T)^T$ and $\mathcal{I}(\Theta_0) = \{I_{jk}\}$ corresponding to θ_1 and θ_2 . In a similar way, the asymptotic distribution of D under H_{03} is a 50:50 mixture between a point mass at zero and a chi-square distribution with 1 degree of freedom, i.e., $0.5\chi_0^2 + 0.5\chi_1^2$.

Test for H_{04} : $\rho = 0$ and $\tau = 0$, where $\rho < o(\tau)$. In this case, $\Theta = (\theta_1, \theta_2, \theta_3)$, where $\theta_1 = \rho$, $\theta_2 = \tau$, and $\theta_3 = \sigma^2$. $C_{\mathcal{Q}_0} = \{0\}^2 \times (0, \infty)$ and $C_{\mathcal{Q}_1} = (0, \infty)^3$. Decompose normal vector U and $\mathcal{I}(\Theta_0)$ as $U = (U_1, U_2, U_3)^T$ and $\mathcal{I}(\Theta_0) = \{I_{jk}\}$ corresponding to θ_1 , θ_2 , and θ_3 . Under the orthonormal transformation, the cone spaces become to $\tilde{C} = \{\Theta; \eta\theta_2 - \theta_1 = 0, \theta_2 = 0\}$ and $\tilde{C}_0 = \{\Theta; \theta_2 = \theta_1 = 0\}$, where $\eta = \tilde{I}_{12} |\tilde{\mathcal{I}}(\Theta_0)|^{-1/2}$ is the slope of the axis θ_2 after transformation and

$$\tilde{\mathcal{I}}(\Theta_0) = \begin{pmatrix} \tilde{I}_{11} & \tilde{I}_{12} \\ \tilde{I}_{21} & \tilde{I}_{22} \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} - \begin{pmatrix} I_{13} \\ I_{23} \end{pmatrix} I_{33}^{-1} (I_{31} \ I_{32}).$$

We can then show that

$$\inf_{\Theta \in \tilde{C}_0} \|U - \Theta\|^2 = U_1^2 + U_2^2.$$

The representations of $\inf_{\Theta \in \tilde{C}} \|U - \Theta\|$ are different in the four regions of the plane with coordinates $(\theta_1, \theta_2)^T$

$$\inf_{\Theta \in \tilde{C}} \|U - \Theta\|^2 = \begin{cases} 0, & \theta_2 \geq 0, \ \eta\theta_2 - \theta_1 \geq 0, \\ U_1^2 + U_2^2 - (\eta U_1 + U_2)^2 / (1 + \eta^2), & \theta_2 + \eta\theta_1 \geq 0, \ \eta\theta_2 - \theta_1 < 0, \\ U_2^2, & \theta_2 < 0, \ \theta_1 \geq 0, \\ U_1^2 + U_2^2, & \theta_2 + \eta\theta_1 < 0, \ \theta_1 < 0. \end{cases}$$

The area proportions $(\pi, 1/4, 1/4, 1/2 - \pi)$ as in the aforementioned order, of these four regions determine the probabilities that the vector U lies in which region, where $\pi = \cos^{-1}\{\eta \cdot (1 + \eta^2)^{-1/2}\} = \tilde{I}_{12} \cdot (\tilde{I}_{11} \tilde{I}_{22})^{-1/2}$. Then the asymptotic distribution of D is the difference of the above two representations

$$D \rightarrow \begin{cases} U_1^2 + U_2^2 & \text{with probability } \pi, \\ (\eta U_1 + U_2)^2 / (1 + \eta^2) & \text{with probability } \frac{1}{4}, \\ U_1^2 & \text{with probability } \frac{1}{4}, \\ 0 & \text{with probability } \frac{1}{2} - \pi. \end{cases}$$

Because U_1 and U_2 are independent, $(\eta U_1 + U_2) / \sqrt{1 + \eta^2} \sim N(0, 1)$, and the final approximate asymptotic distribution of D is $\eta\chi_2^2 + 0.5\chi_1^2 + (0.5 - \eta)\chi_0^2$.

4.2. Permutation test

For identifying significant pathways, we also consider using the following permutation procedures.

- Step 1: estimate τ/ρ , ρ , τ using the observed data by fitting the semiparametric model and calculate the residuals $\hat{\varepsilon}_{0i}$ from $y_i = \mathbf{x}_i\beta + \varepsilon_{0i}$.
- Step 2: permute the residual $\hat{\varepsilon}_{0i}$ and simulate outcomes as $y_i^* = \mathbf{x}_i\hat{\beta} + \hat{\varepsilon}_{0i}^*$.
- Step 3: based on $\mathbf{y}^*$, $\mathbf{x}$, and $\mathbf{z}$, fit the semiparametric model using the likelihood-based approach and then estimate $\hat{\tau}^*/\hat{\rho}^*$, $\hat{\rho}^*$, and $\hat{\tau}^*$.
- Step 4: repeat Step 1–Step 3 a large number of times, e.g., 10,000 times.
- Step 5: estimate the statistical significance by the percentage of times either $\hat{\tau}^*/\hat{\rho}^* > \hat{\tau}/\hat{\rho}$ or $\hat{\rho}^* > \hat{\rho}$, where $\hat{\tau}/\hat{\rho}$, $\hat{\rho}$, $\hat{\tau}$ are the estimated values from the observed data.

Significant pathways can be selected based on this percentage, which is equivalent to a statistical significance.

To rank the importance of genes within a significant pathway related to clinical outcomes, we perform the following steps:

- Step 6: bootstrap the samples B times. For each bootstrapped sample, we estimate $\hat{\tau}_{(-g)}^b, \hat{\rho}_{(-g)}^b$, and $(\hat{\tau}/\hat{\rho})_{(-g)}^b$ by fitting the following model, where $b = 1, \dots, B$,

$$Y = X\beta + \mathbf{r}\{Z_{(-g)}\} + \varepsilon.$$

- Step 7: calculate the absolute value of difference, $|\hat{\tau}_{(-g)}^b - \hat{\tau}|$, $|\hat{\rho}_{(-g)}^b - \hat{\rho}|$, and $|(\hat{\tau}/\hat{\rho})_{(-g)}^b - (\hat{\tau}/\hat{\rho})|$, under the Gaussian kernel. For other kernels, only calculate the absolute value of difference $|\hat{\tau}_{(-g)}^b - \hat{\tau}|$.
- Step 8: rank the importance of the genes by the mean absolute difference $(1/B) \sum_{b=1}^B |\hat{\tau}_{(-g)}^b - \hat{\tau}|$. If a gene plays an importance role in a pathway, this difference will be large.

We can also rank the importance of gene pairs by performing the same procedure except for fitting the following model

$$Y = X\beta + \mathbf{r}\{z_{-(g,g')}\} + \varepsilon.$$

5. Simulation studies

5.1. Mean squared error and coverage probability

We conducted simulations to compare the MSEs and coverage probabilities of LLG's estimators with those of our estimators.

We considered the following cases and simulated 1000 data sets with two sets of n and p values. One set is $(n, p) = (60, 5)$ with a relatively large sample size compared to the number of genes and the other set is $(n, p) = (60, 200)$ with a relatively small sample size compared to the number of genes.

- Case 1: $y_i = \beta x_i + r(z_{i1}, z_{i2}, \dots, z_{ip}) + \varepsilon_i$ with $\beta = 1$, $x_i = 3 \cos(z_{i1}) + 2u_i$ with u_i being independent of z_{i1} and following $N(0, 1)$, $z_{ij} \sim \text{Uniform}(0, 1)$, and the true $r(\mathbf{z}) \sim GP\{0, \tau K_{g(\rho)}\}$, where $K_{g(\rho)}(\mathbf{z}, \mathbf{z}') = \exp(-\|\mathbf{z} - \mathbf{z}'\|^2/\rho)$ and $\|\cdot\|$ was used as Euclidean norm.
- Case 2: the same setting as that of case 1 except that $r(\mathbf{z}) \sim GP\{0, \tau K_{p2}\}$, where $K_{p2}(\mathbf{z}, \mathbf{z}') = (\mathbf{z}\mathbf{z}')^2$.
- Case 3: the same setting as that of case 1 except that $r(\mathbf{z}) \sim GP\{0, \tau K_{p1}\}$, where $K_{p1}(\mathbf{z}, \mathbf{z}') = (\mathbf{z}\mathbf{z}')^1$.

For each case, we estimated parameters in the semiparametric regression model using the LLG's REML and our REML. We calculated the average MSEs and coverage probability of the parameter estimates. The results are summarized in Table 1. For both $n > p$ and $n < p$ cases, our REML estimators have smaller MSEs than that of LLG's REML estimators. Both approaches have comparable coverage probability.

5.2. Consistency

We also performed additional simulation to study the increased accuracy of our estimates as n increases when p is fixed. We considered $n = 15, 30, 60$, and 600 , for one small, two medium, and one relatively large sample sizes. For each case, we simulated 1000 data sets and average MSEs are summarized in Table 2.

5.3. Type I error and power

For the assessment of type I error and power for both approaches, we considered the following Case 4 for type I error and Case 5 for power, respectively. We consider Case 5 because the nonparametric function $r(z_{i1}, z_{i2}, \dots, z_{ip})$ is allowed to have a complex form with nonlinear functions of the $\mathbf{z}$'s and interactions among the $\mathbf{z}$'s. It allows for x_i and $(z_{i1}, \dots, z_{ip})$ to be correlated. We obtained type I error and power using a profile restricted likelihood ratio test (PRLRT) described in Section 4.1. We performed PRLRT using both empirical distribution as well as theoretical distribution. We denote them as "PRLRT(e)" and "PRLRT(t)", respectively. We also compared them with the permutation test described in Section 4.2. This permutation is denoted as "PERM".

- Case 4: $y_i = \beta x_i + \varepsilon_i$ with $\beta = 1$, $x_i = 3 \cos(\pi/6) + 2u_i$ with $u_i \sim N(0, 1)$ and $\varepsilon_i \sim N(0, \sigma^2)$.
- Case 5: $y_i = \beta x_i + r(z_{i1}, z_{i2}, \dots, z_{ip}) + \varepsilon_i$, $x_i = 3 \cos(z_{i1}) + 2u_i$ with u_i being independent of z_{i1} and following $N(0, 1)$, $z_{ij} \sim \text{Uniform}(0, 1)$, and $r(\mathbf{z}) = 10 \cos(z_1) - 15 z_2^2 + 10 \exp(-z_3) z_4 - 8 \sin(z_5) \cos(z_3) + 20 z_1 z_2$.

The estimated type I error rates from both approaches are summarized in Tables 3–5 and they are all close to the nominal level when $n > p$. However, when $n < p$, both methods differ from the nominal level. The power for both approaches are comparable. The performance

between “PRLRT (e)” and “PRLRT (t)” was similar to each other, so was the performance between “PRLRT” and “PERM”.

6. Real data analysis

6.1. Pathway based analysis for type II diabetes

We applied our profile likelihood approach to a microarray expression data set on type II diabetes (Mootha et al., 2003), where a microarray expression data from 22,283 genes were measured in 17 male patients with normal glucose tolerance and 18 male samples with type II diabetes mellitus. Because their approach is based on a normalized Kolmogorov–Smirnov statistics, their method is unable to model the continuous outcome, such as glucose level, and clinical covariates such as age. Incorporating such information into analysis might be helpful to more efficiently detect subtle differences in gene expression profiles. We studied a total of 277 pathways consisting of 128 KEGG pathways (<http://www.genome.jp/kegg/pathway.html>) and 149 curated pathways. The 149 curated pathways were constructed from known biological experiments by Mootha and colleagues.

In our analysis, let Y be the log transformed glucose level, X be the age, and Z be the $p \times n$ gene expression levels within each pathway, where n is 35, i.e. the number of subjects, p is the number of genes in a specific pathway which varied from 4 to 200 across these pathways. Our goal is to identify pathways that affect the glucose level related to diabetes after adjusting for the age effect and also to rank genes within a significant pathway. To identify significant pathways, we fitted the semiparametric model.

6.2. Identifying significant pathways

We chose the 0.05 cut off for the statistical significance using both theoretical and empirical distributions of PRLRT and the percentage which was described in Section 4.2. We also applied existing multiple comparison methods (Storey, 2002, 2003) to our pathway data, although our pathways are not independent of each other because of shared genes and interactions among pathways. The FDR q -values were between 0.081 and 0.303. Our approach using both theoretical and empirical distributions took about 60 min to run, using a Mac Pro with two 3.0 GHz Quad-Core Intel Xeon processors, and 10 GB of memory. Our code is written in Matlab and is available upon request.

For overlapping pathways across four kernels, we selected top 50 pathways for each kernel and then examined the common pathways across four kernels. We found that a total of seven pathways were common among the four kernels, including *Alanine and aspartate metabolism*, *Oxidative phosphorylation*, and *RNA polymerase* pathways. Since one of them is a subset of another pathway, six pathways are summarized in Table 6.

Pathway 4 is related to the *Alanine and aspartate metabolism* pathway which has been studied for its association with abnormal hepatocellular functions and abnormal fasting glucose level alanine and aspartate metabolism in type II diabetes (Jiamjarasrangsi et al., 2009). Two other pathways out of six pathways, pathways 133 and 229, where all except one gene in pathway 133 belong to pathway 229, are related to the *Oxidative phosphorylation* which is known to be associated to diabetes (Misu et al., 2007; Mootha et al., 2003, 2004). This is a process of cellular respiration in humans (or in general eukaryotes). These pathways contain coregulated genes across different tissues and are related to insulin/glucose disposal. They consist of ATP synthesis, a pathway involved in energy transfer. The other two pathways, pathways 209 and 254, are related to *RNA polymerase*. All genes except two genes in pathway 209 are part of pathway 254. Among all the pathways, pathway 36, *c17_U133_probes*, is the most significant one under the Gaussian kernel. This pathway plays a role in cellular behavioral changes (Saxena, 2001) and is also

one of the seven pathways that are common among the four kernels. It contains several genes, e.g. CAP1, MAPP2K6, ARF6, and SGK, related to human insulin signaling (Dahlquist et al., 2002). Only one gene in pathway 36 is included in pathway 254. These genes were not significant using single gene based analysis. We also note that all genes in the *Oxidative phosphorylation* pathway were not significant using single gene based analysis. Only one gene, GAD2, in pathway 4 is significant using single gene based analysis.

We compared the top 50 pathways identified by global test (Goeman et al., 2004) and GSEA (Subramanian et al., 2005). Global test is based on a random effects model which does not include covariates and uses a linear kernel. GSEA calculates enrichment score which is weighed function of correlation among genes in a pathway. This GSEA method cannot include covariate information into this enrichment score value. The GSEA result and global test results are summarized in supplementary materials.

The proportions of overlap among global, GSEA, and our approach were calculated. The proportions of overlap between the global and GSEA was 0.36. The largest proportion among the GSEA and our approaches was 0.41 using the quadratic kernels. These results suggest that the proportions of overlap among different methods were small, meaning that they detected different pathways. When our approach was used with four different kernels, the largest overlap was 0.92 between linear and quadratic kernels. Global test and GSEA approaches found that pathways 4, 140, and 229 are significant. But they could not detect pathways 36, 133, and 254, while our approach detects them all.

For kernel selection, we used Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978), where $AIC = n \log \{ (Y - \hat{Y})^T (Y - \hat{Y}) \} + 2r$ and $BIC = n \log \{ (Y - \hat{Y})^T (Y - \hat{Y}) \} + r \log(n)$, $\hat{Y} = LY$, $L = (I + \lambda^{-1}K)^{-1}[\lambda^{-1}K + X^T(I + \lambda^{-1}K)^{-1}X]^{-1}X^T(I + \lambda^{-1}K)^{-1}$, and $r = \text{rank}(L)$.

We found that all seven pathways are in the top 15 pathway list having the smallest AIC and BIC values for all kernels. Pathway 36, *c17_U133_probes*, has the smallest AIC and BIC values with the Gaussian kernel among four kernels, whereas pathway 229, *Oxidative phosphorylation*, has the smallest values with the quadratic kernel among four kernels. Pathway 254, *RNA polymerase*, has the smallest values with the Neural network kernel among four kernels. Pathways 4, 36, and 229 have small values among all pathways with quadratic kernel.

7. Discussion

In this paper, we have derived the asymptotic properties of semiparametric regression for identifying pathway effects on clinical outcomes after controlling for covariates. We compared LLG's REML with our REML obtained from a profile $\sqrt{n}$ likelihood. We showed that the estimators obtained from our REML and LLG's REML are both consistent, have n convergence, and asymptotically follow normal distribution when the true parameters are interior points of the parameter space and a mixture of normal distribution when the one component of true parameters is a left endpoint of the parameter space. However, our REML gives more accurate score equations and information matrix, and have smaller MSEs than those of LLG's REML. A profile restricted likelihood ratio test is also provided for the non-standard testing problem in our application.

The choice of an appropriate kernel poses a significant practical issue. We used kernel selection approaches based on AIC and BIC (Liu et al., 2007). Kim et al. (2012) proposed Bayes factor-based approach on kernel selection. However, this Bayes factor-based approach is computational expensive. Kernel selection can be considered as a model

selection problem with the kernel machine framework. More general and flexible methods on model selection for covariance matrix estimation may be explored here and are worth future research.

We note that we analyze each pathway separately in our analysis. It is known that pathways are not independent of each other because of shared genes and interactions among pathways, which makes difficult to adjust p-value. Because existing multiple comparison methods based on false discovery rates (Benjamini and Hochberg, 1995; Storey, 2002, 2003) are developed in single gene based analysis under independence assumption or known positive dependence structure among genes, they are not applicable in a pathway based analysis, where pathways are not independent of each other. Developing such multiple comparison method for pathway based analysis will be a challenging problem because of complex dependence structure among pathways.

It is also important to generalize the semiparametric model (Hastie and Tibshirani, 1990; Tibshirani, 1996) to incorporate multiple pathways using generalized additive model and the multivariate adaptive regression splines (Friedman, 1991).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by NIH Grants GM-59507, N01-HV-28186 and P30-DA-18343, National Science Foundation Grant DMS 1106738, and a pilot grant from the Yale Pepper Center P30AG021342.

References

- Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19:716–723.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*. 1995; 57:289–300.
- Chant D. On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika*. 1974; 61:291–298.
- Chernoff. On the Distribution of the Likelihood Ratio. *Annals of Mathematical Statistics*. 1954; 25:573–578.
- Claeskens G. Restricted likelihood ratio lack of fit tests using mixed spline models. *Journal of the Royal Statistical Society: Series B*. 2004; 66:909–926.
- Cressie N, Lahiri SN. The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*. 1993; 45:217–233.
- Cressie N, Lahiri SN. Asymptotic for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference*. 1996; 50:327–341.
- Cristianini, N.; Shawe-Taylor, J. *Kernel Methods for Pattern Analysis*. Cambridge University Press; Cambridge: 2006.
- Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*. 2002; 31:19–20. [PubMed: 11984561]
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics*. 2004; 32:407–499.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Friedman JH. Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*. 1991; 19:1–141.

- Geyer CJ. On the asymptotics of constrained M-estimation. *The Annals of Statistics*. 1994; 22:1993–2010.
- Goeman J, van de Geer SA, Kort F, Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004; 20:93–99. [PubMed: 14693814]
- Gu C, Ma P. Optimal smoothing in nonparametric mixed effect models. *Annals of Statistics*. 2005; 33:1357–1379.
- Harris MA, et al. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*. 2004; 32:D258–D261. [PubMed: 14681407]
- Hastie, TJ.; Tibshirani, RJ. *Generalized Additive Models*. Chapman & Hall, CRC; 1990.
- Hosack DA, Dennis G Jr, Sherman BT, Clifford H, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biology*. 2003; 4 (10):R70. [PubMed: 14519205]
- Jiamjarasrangsri W, Lertmaharit S, Sangwatanaroj S, Lohsoonthorn V. Type 2 diabetes, impaired fasting glucose, and their association with increased hepatic enzyme levels among the employees in a University Hospital in Thailand. *Journal of the Medical Association of Thailand*. 2009; 92:961–968. [PubMed: 19626817]
- Kim I, Pang H, Zhao H. Bayesian semiparametric regression models for evaluating pathway effects on clinical continuous and binary outcomes. *Statistics in Medicine*. 2012; 31:1633–1651. [PubMed: 22438129]
- Lehmann, EL. *Theory of Point Estimation*. John Wiley; New York: 1983.
- Liu D, Lin X, Ghosh D. Semiparametric regression of multi-dimensional genetic pathway data: least square kernel machines and linear mixed models. *Biometrics*. 2007; 63:1079–1088. [PubMed: 18078480]
- Misu H, Takamura T, Matsuzawa N, Shimizu A, Ota T, Sakurai M, Ando H, Arai K, Yamashita T, Honda M, Yamashita T, Kaneko S. Genes involved in oxidative phosphorylation are coordinately upregulated with fasting hyperglycaemia in livers of patients with type 2 diabetes. *Diabetologia*. 2007; 50:268–277. [PubMed: 17187250]
- Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson P, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop L. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*. 2003; 34:267–273. [PubMed: 12808457]
- Mootha VK, Handschin C, Arlow D, Xie X, Pierre JS, Sihag S, Yang W, Altshuler D, Puigserver P, Patterson N, Willy PJ, Schulman IG, Heyman RA, Lander ES, Spiegelman BM. Err α and Gabpa/b specify PGC-1 α -dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proceedings of the National Academy of Sciences*. 2004; 101:6570–6575.
- Rajagopalan DA, Agarwal P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*. 2005; 21:788–793. [PubMed: 15509611]
- Saxena, V. PhD Dissertation of Mechanical Engineering. Massachusetts Institute of Technology; 2001. Genomic Response, Bioinformatics, and Mechanics of the Effects of Forces on Tissues and Wound Healing.
- Schwarz GE. Estimating the dimensional of a model. *The Annals of Statistics*. 1978; 6:461–464.
- Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of American Statistical Association*. 1987; 82:605–610.
- Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*. 2003; 31:2013–2035.
- Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*. 2002; 64:479–498.
- Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 2005; 102 (43):15545–15550.
- Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society, Series B*. 1996; 58:267–288.

- Vu HTV, Zhou S. Generalization of likelihood ratio tests under nonstandard conditions. *The Annals of Statistics*. 1997; 25:916–987.
- Wang Y. Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B*. 1998; 60:159–174.
- Zhang D, Lin X, Raz J, Sowers M. Semi-parametric mixed models for longitudinal data. *Journal of American Statistical Association*. 1998; 93:710–719.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B*. 2005; 67:301–320.

Appendix A. Technical complements

A.1. Score equations and information matrix

We introduce the following notations to simplify our discussion:

$$K(\mathbf{z}, \mathbf{z}') = \exp \left\{ - \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho} \right\},$$

$$K_{\rho}(\mathbf{z}, \mathbf{z}') = \frac{\partial K(\mathbf{z}, \mathbf{z}')}{\partial \rho} = \exp \left\{ - \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho} \right\} \left\{ \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho^2} \right\},$$

$$K_{\rho\rho}(\mathbf{z}, \mathbf{z}') = \frac{\partial^2 K(\mathbf{z}, \mathbf{z}')}{\partial \rho^2} = \exp \left[- \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho} \left\{ \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho^2} \right\}^2 \right] - 2 \exp \left\{ - \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho} \right\} \left\{ \sum_{k=1}^p \frac{(z_k - z'_k)^2}{\rho^3} \right\},$$

where $K_{\rho}(\cdot, \cdot)$ and $K_{\rho\rho}(\cdot, \cdot)$ are the first and second derivatives of $K(\cdot, \cdot)$ with respect to ρ . Let $\mathbf{K}$, $\mathbf{K}_{\rho}$, and $\mathbf{K}_{\rho\rho}$ be $n \times n$ matrices. Their entries consist of $K(\cdot, \cdot)$, $K_{\rho}(\cdot, \cdot)$, and $K_{\rho\rho}(\cdot, \cdot)$, respectively. We recall the following notations:

$$\Sigma = \sigma^2 I + \tau \mathbf{K} = \Sigma(\Theta),$$

$$P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}, \quad H = I - X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}.$$

Based on these notations and the calculations of

$$\frac{\partial \Sigma^{-1}}{\partial \Theta} = - \Sigma^{-1} \sum_{\Theta} \Sigma^{-1} \quad \text{and} \quad \frac{\partial \log |\Sigma|}{\partial \Theta} = \Sigma^{-1} \sum_{\Theta}$$

we obtain the first and second derivatives of Σ and Σ^{-1} with respect to τ as follows $\Sigma_{\tau} = \mathbf{K}$,

$$\Sigma_{\tau\tau} = 0 \times I, \quad \sum_{\tau}^{-1} = - \sum_{\tau}^{-1} \sum_{\tau} \sum_{\tau}^{-1}, \quad \text{and}$$

$$\sum_{\tau\tau}^{-1} = - \left(\sum_{\tau}^{-1} \sum_{\tau} \sum_{\tau}^{-1} + \sum_{\tau}^{-1} \sum_{\tau\tau} \sum_{\tau}^{-1} + \sum_{\tau}^{-1} \sum_{\tau} \sum_{\tau}^{-1} \right).$$

We also calculate the first and second derivatives of Σ and Σ^{-1} with respect to ρ as follows:

$$\Sigma_{\rho} = \tau \mathbf{K}_{\rho}, \quad \Sigma_{\rho\rho} = \tau \mathbf{K}_{\rho\rho}, \quad \sum_{\rho}^{-1} = - \sum_{\rho}^{-1} \sum_{\rho} \sum_{\rho}^{-1}, \quad \text{and}$$

$$\sum_{\rho\rho}^{-1} = - \left(\sum_{\rho}^{-1} \sum_{\rho} \sum_{\rho}^{-1} + \sum_{\rho}^{-1} \sum_{\rho\rho} \sum_{\rho}^{-1} + \sum_{\rho}^{-1} \sum_{\rho} \sum_{\rho}^{-1} \right).$$

The first and second derivatives of Σ and Σ^{-1} with respect to σ^2 are $\Sigma_{\sigma^2} = I$, $\Sigma_{\sigma^2\sigma^2} = 0 \times I$,

$$\sum_{\sigma^2}^{-1} = - \sum_{\sigma^2}^{-1} \sum_{\sigma^2} \sum_{\sigma^2}^{-1}, \quad \text{and}$$

$$\sum_{\sigma^2\sigma^2}^{-1} = - \left(\sum_{\sigma^2}^{-1} \sum_{\sigma^2} \sum_{\sigma^2}^{-1} + \sum_{\sigma^2}^{-1} \sum_{\sigma^2\sigma^2} \sum_{\sigma^2}^{-1} + \sum_{\sigma^2}^{-1} \sum_{\sigma^2} \sum_{\sigma^2}^{-1} \right).$$

We obtain the second derivatives of Σ and Σ^{-1} with respect to τ and ρ , τ and σ^2 , and ρ and σ^2 as follows:

$$\Sigma_{\tau\rho} = \mathbf{K}_\rho, \quad \Sigma_{\tau\sigma^2} = 0 \times I, \quad \Sigma_{\rho\sigma^2} = 0 \times I, \quad \Sigma_{\tau\rho}^{-1} = -(\Sigma_\rho^{-1} \Sigma_\tau \Sigma^{-1} + \Sigma^{-1} \Sigma_{\tau\rho} \Sigma^{-1} + \Sigma^{-1} \Sigma_\tau \Sigma_\rho^{-1}),$$

$$\Sigma_{\tau\sigma^2}^{-1} = -(\Sigma_{\sigma^2}^{-1} \Sigma_\tau \Sigma^{-1} + \Sigma^{-1} \Sigma_{\tau\sigma^2} \Sigma^{-1} + \Sigma^{-1} \Sigma_\tau \Sigma_{\sigma^2}^{-1}), \quad \text{and} \quad \Sigma_{\rho\sigma^2}^{-1} = -(\Sigma_{\sigma^2}^{-1} \Sigma_\rho \Sigma^{-1} + \Sigma^{-1} \Sigma_{\rho\sigma^2} \Sigma^{-1} + \Sigma^{-1} \Sigma_\rho \Sigma_{\sigma^2}^{-1}).$$

The first derivatives of $X^T \Sigma X$ with respect to τ , ρ , σ^2 are

$$(X^T \Sigma^{-1} X)_\tau^{-1} = -(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)^{-1},$$

$$(X^T \Sigma^{-1} X)_\rho^{-1} = -(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\rho^{-1} X) (X^T \Sigma^{-1} X)^{-1},$$

$$(X^T \Sigma^{-1} X)_{\sigma^2}^{-1} = -(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\sigma^2}^{-1} X) (X^T \Sigma^{-1} X)^{-1}.$$

The second derivatives of $(X \Sigma X)^{-1}$ with respect to τ , ρ , σ^2 are

$$(X^T \Sigma^{-1} X)_{\tau\tau}^{-1} = -[\{(X^T \Sigma^{-1} X)_\tau^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\tau\tau}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)_\tau^{-1}\}],$$

$$(X^T \Sigma^{-1} X)_{\rho\rho}^{-1} = -[\{(X^T \Sigma^{-1} X)_\rho^{-1} (X^T \Sigma_\rho^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\rho\rho}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\rho^{-1} X) (X^T \Sigma^{-1} X)_\rho^{-1}\}],$$

$$(X^T \Sigma^{-1} X)_{\sigma^2\sigma^2}^{-1} = -[\{(X^T \Sigma^{-1} X)_{\sigma^2}^{-1} (X^T \Sigma_{\sigma^2}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\sigma^2\sigma^2}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\sigma^2}^{-1} X) (X^T \Sigma^{-1} X)_{\sigma^2}^{-1}\}],$$

$$(X^T \Sigma X)_{\tau\rho}^{-1} = -[\{(X^T \Sigma^{-1} X)_\tau^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\tau\rho}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)_\rho^{-1}\}],$$

$$(X^T \Sigma X)_{\tau\sigma^2}^{-1} = -[\{(X^T \Sigma^{-1} X)_{\sigma^2}^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\tau\sigma^2}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\tau^{-1} X) (X^T \Sigma^{-1} X)_{\sigma^2}^{-1}\}],$$

$$(X^T \Sigma X)_{\rho\sigma^2}^{-1} = -[\{(X^T \Sigma^{-1} X)_{\sigma^2}^{-1} (X^T \Sigma_\rho^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_{\rho\sigma^2}^{-1} X) (X^T \Sigma^{-1} X)^{-1}\} + \{(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma_\rho^{-1} X) (X^T \Sigma^{-1} X)_{\sigma^2}^{-1}\}].$$

The first and second derivatives of H with respect to τ are

$$H_\tau = -\{X(X^T \Sigma^{-1} X)_\tau^{-1} X^T \Sigma^{-1} + X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_\tau^{-1}\},$$

$$H_{\tau\tau} = -[\{X(X^T \Sigma^{-1} X)_{\tau\tau}^{-1} X^T \Sigma^{-1}\} + \{X(X^T \Sigma^{-1} X)_\tau^{-1} X^T \Sigma_\tau^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\tau\tau}^{-1}\} + \{X(X^T \Sigma^{-1} X)_\tau^{-1} X^T \Sigma_\tau^{-1}\}].$$

The first and second derivatives of H with respect to ρ are

$$H_\rho = -\{X(X^T \Sigma^{-1} X)_\rho^{-1} X^T \Sigma^{-1} + X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_\rho^{-1}\},$$

$$H_{\rho\rho} = -[\{X(X^T \Sigma^{-1} X)_{\rho\rho}^{-1} X^T \Sigma^{-1}\} + \{X(X^T \Sigma^{-1} X)_\rho^{-1} X^T \Sigma_\rho^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\rho\rho}^{-1}\} + \{X(X^T \Sigma^{-1} X)_\rho^{-1} X^T \Sigma_\rho^{-1}\}].$$

The first and second derivatives of H with respect to σ^2 are

$$H_{\sigma^2} = -\{X(X^T \Sigma X)_{\sigma^2}^{-1} X^T \Sigma^{-1} + X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\sigma^2}^{-1}\},$$

$$H_{\sigma^2\sigma^2} = -[\{X(X^T \Sigma^{-1} X)_{\sigma^2\sigma^2}^{-1} X^T \Sigma^{-1}\} + \{X(X^T \Sigma^{-1} X)_{\sigma^2}^{-1} X^T \Sigma_{\sigma^2}^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\sigma^2\sigma^2}^{-1}\} + \{X(X^T \Sigma^{-1} X)_{\sigma^2}^{-1} X^T \Sigma_{\sigma^2}^{-1}\}].$$

The second derivatives of H with respect to τ and ρ , τ and σ^2 , ρ and σ^2 are

$$\begin{aligned}
H_{\tau\rho} &= -\{(X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}) + (X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\rho}^{-1}) + (X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\tau}^{-1}) + (X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\tau\rho}^{-1})\}, \\
H_{\tau\sigma^2} &= -[\{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\tau\sigma^2}^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\sigma^2}^{-1}\} + \{X(X^T \Sigma X)^{-1} X^T \Sigma_{\tau}^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\tau\sigma^2}^{-1}\}], \\
H_{\rho\sigma^2} &= -[\{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\rho\sigma^2}^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\sigma^2}^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\rho}^{-1}\} + \{X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\rho\sigma^2}^{-1}\}].
\end{aligned}$$

The first derivatives of P with respect to τ , ρ , and σ^2 are

$$\begin{aligned}
P_{\tau} &= \Sigma_{\tau}^{-1} - [\{\Sigma_{\tau}^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}\} + \{\Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\tau}^{-1}\} + \{\Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}\}], \\
P_{\rho} &= \Sigma_{\rho}^{-1} - [\{\Sigma_{\rho}^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}\} + \{\Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\rho}^{-1}\} + \{\Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}\}], \\
P_{\sigma^2} &= \Sigma_{\sigma^2}^{-1} - [\{\Sigma_{\sigma^2}^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}\} + \{\Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma_{\sigma^2}^{-1}\} + \{\Sigma^{-1} X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}\}].
\end{aligned}$$

The score equations based on our REML $l_2(\Theta)$ are

$$\begin{aligned}
\frac{\partial l_2(\Theta)}{\partial \tau} &= -\frac{1}{2} \text{tr}(\mathbf{K}P) - \frac{1}{2} Y^T \{(H_{\tau}^T \Sigma^{-1} H) + (H^T \Sigma^{-1} H_{\tau}) + (H^T \Sigma_{\tau}^{-1} H)\} Y, \\
\frac{\partial l_2(\Theta)}{\partial \rho} &= -\frac{1}{2} \text{tr}(\tau \mathbf{K}_{\rho} P) - \frac{1}{2} Y^T \{(H_{\rho}^T \Sigma^{-1} H) + (H^T \Sigma^{-1} H_{\rho}) + (H^T \Sigma_{\rho}^{-1} H)\} Y, \\
\frac{\partial l_2(\Theta)}{\partial \sigma^2} &= -\frac{1}{2} \text{tr}(P) - \frac{1}{2} Y^T \{(H_{\sigma^2}^T \Sigma^{-1} H) + (H^T \Sigma^{-1} H_{\sigma^2}) + (H^T \Sigma_{\sigma^2}^{-1} H)\} Y.
\end{aligned}$$

The second derivatives of $l_2(\Theta)$ with respect to τ , ρ , and σ^2 are

$$\begin{aligned}
\frac{\partial^2 l_2(\Theta)}{\partial \tau^2} &= -\frac{1}{2} \text{tr}(\mathbf{K}P_{\tau}) - \frac{1}{2} Y^T \{(H_{\tau\tau}^T \Sigma^{-1} H) + (H^T \Sigma_{\tau\tau}^{-1} H) + (H^T \Sigma^{-1} H_{\tau\tau}) + 2(H_{\tau}^T \Sigma^{-1} H_{\tau} + H_{\tau}^T \Sigma_{\tau}^{-1} H + H^T \Sigma_{\tau}^{-1} H_{\tau})\} Y, \\
\frac{\partial^2 l_2(\Theta)}{\partial \rho^2} &= -\frac{1}{2} \text{tr}(\tau \mathbf{K}_{\rho\rho} P + \mathbf{K}_{\rho} P_{\rho}) - \frac{1}{2} Y^T \{(H_{\rho\rho}^T \Sigma^{-1} H) + (H^T \Sigma_{\rho\rho}^{-1} H) + (H^T \Sigma^{-1} H_{\rho\rho}) + 2(H_{\rho}^T \Sigma^{-1} H_{\rho} + H_{\rho}^T \Sigma_{\rho}^{-1} H + H^T \Sigma_{\rho}^{-1} H_{\rho})\} Y, \\
\frac{\partial^2 l_2(\Theta)}{\partial (\sigma^2)^2} &= -\frac{1}{2} \text{tr}(P_{\sigma^2}) - \frac{1}{2} Y^T \{(H_{\sigma^2\sigma^2}^T \Sigma^{-1} H) + (H^T \Sigma_{\sigma^2\sigma^2}^{-1} H) + (H^T \Sigma^{-1} H_{\sigma^2\sigma^2}) + 2(H_{\sigma^2}^T \Sigma^{-1} H_{\sigma^2} + H_{\sigma^2}^T \Sigma_{\sigma^2}^{-1} H + H^T \Sigma_{\sigma^2}^{-1} H_{\sigma^2})\} Y, \\
\frac{\partial^2 l_2(\Theta)}{\partial \tau \partial \rho} &= -\frac{1}{2} \text{tr}(\mathbf{K}_{\rho} P + \mathbf{K} P_{\rho}) - \frac{1}{2} Y^T \{(H_{\tau\rho}^T \Sigma^{-1} H) + (H^T \Sigma_{\tau\rho}^{-1} H) + (H^T \Sigma^{-1} H_{\tau\rho}) + (H_{\tau}^T \Sigma_{\rho}^{-1} H) + (H_{\tau}^T \Sigma^{-1} H_{\rho}) + (H_{\rho}^T \Sigma_{\tau}^{-1} H) + (H^T \Sigma_{\tau}^{-1} H_{\rho}) + (H_{\rho}^T \Sigma^{-1} H_{\tau})\} Y, \\
\frac{\partial^2 l_2(\Theta)}{\partial \tau \partial \sigma^2} &= -\frac{1}{2} \text{tr}(\mathbf{K} P_{\sigma^2}) - \frac{1}{2} Y^T \{(H_{\tau\sigma^2}^T \Sigma^{-1} H) + (H^T \Sigma_{\tau\sigma^2}^{-1} H) + (H^T \Sigma^{-1} H_{\tau\sigma^2}) + (H_{\tau}^T \Sigma_{\sigma^2}^{-1} H) + (H_{\tau}^T \Sigma^{-1} H_{\sigma^2}) + (H_{\sigma^2}^T \Sigma_{\tau}^{-1} H) + (H^T \Sigma_{\tau}^{-1} H_{\sigma^2}) + (H_{\sigma^2}^T \Sigma^{-1} H_{\tau})\} Y, \\
\frac{\partial^2 l_2(\Theta)}{\partial \rho \partial \sigma^2} &= -\frac{1}{2} \text{tr}(\tau \mathbf{K}_{\rho} P_{\sigma^2}) - \frac{1}{2} Y^T \{(H_{\rho\sigma^2}^T \Sigma^{-1} H) + (H^T \Sigma_{\rho\sigma^2}^{-1} H) + (H^T \Sigma^{-1} H_{\rho\sigma^2}) + (H_{\rho}^T \Sigma_{\sigma^2}^{-1} H) + (H_{\rho}^T \Sigma^{-1} H_{\sigma^2}) + (H_{\sigma^2}^T \Sigma_{\rho}^{-1} H) + (H^T \Sigma_{\rho}^{-1} H_{\sigma^2}) + (H_{\sigma^2}^T \Sigma^{-1} H_{\rho})\} Y.
\end{aligned}$$

The information matrix based on our REML $l_2(\Theta)$ is

$$I_{\theta} = \begin{pmatrix} \frac{\partial^2 l_2(\Theta)}{\partial \tau^2} & \frac{\partial^2 l_2(\Theta)}{\partial \tau \partial \rho} & \frac{\partial^2 l_2(\Theta)}{\partial \tau \partial \sigma^2} \\ \frac{\partial^2 l_2(\Theta)}{\partial \tau \partial \rho} & \frac{\partial^2 l_2(\Theta)}{\partial \rho^2} & \frac{\partial^2 l_2(\Theta)}{\partial \rho \partial \sigma^2} \\ \frac{\partial^2 l_2(\Theta)}{\partial \tau \partial \sigma^2} & \frac{\partial^2 l_2(\Theta)}{\partial \rho \partial \sigma^2} & \frac{\partial^2 l_2(\Theta)}{\partial (\sigma^2)^2} \end{pmatrix}.$$

Also the score equations and information matrix of Liu et al. (2007) are

$$\begin{aligned}
\frac{\partial l_1(\Theta)}{\partial \tau} &= -\frac{1}{2} \text{tr}(\mathbf{K}P) + \frac{1}{2} (Y - X\widehat{\beta})^T \Sigma^{-1} \mathbf{K} \Sigma^{-1} (Y - X\widehat{\beta}), \\
\frac{\partial l_1(\Theta)}{\partial \rho} &= -\frac{1}{2} \text{tr}(\tau \mathbf{K}_{\rho} P) + \frac{1}{2} (Y - X\widehat{\beta})^T \Sigma^{-1} (\tau \mathbf{K}_{\rho}) \Sigma^{-1} (Y - X\widehat{\beta}), \\
\frac{\partial l_1(\Theta)}{\partial \sigma^2} &= -\frac{1}{2} \text{tr}(P) + \frac{1}{2} (Y - X\widehat{\beta})^T \Sigma^{-1} \Sigma^{-1} (Y - X\widehat{\beta}),
\end{aligned}$$

and

$$I_{\Theta} = \left\{ \frac{1}{2} \text{tr} \left(P \sum_{\Theta_i} P \sum_{\Theta_{i'}} \right) \right\}.$$

A.2. Regularity conditions

RC1

The observations $\mathbf{D} = (X, Z, Y)$ have probability density $f(\mathbf{D}, \Theta)$ with respect to some measure μ and parameter $\Theta = (\theta_1, \dots, \theta_k)$. $f(\mathbf{D}, \Theta)$ has a common support and the model is identifiable. Furthermore, the first and second derivatives of $\log(f)$ satisfy

$$E_{\Theta} \left\{ \frac{\partial \log f(\mathbf{D}, \Theta)}{\partial \theta_i} \right\} = 0 \quad \text{for } i=1, \dots, k$$

and

$$I_{ij}(\Theta) = E_{\Theta} \left\{ \frac{\partial}{\partial \theta_i} \log f(\mathbf{D}, \Theta) \frac{\partial}{\partial \theta_j} \log f(\mathbf{D}, \Theta) \right\} = E_{\Theta} \left\{ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\mathbf{D}, \Theta) \right\}.$$

RC2

The Fisher information matrix

$$I(\Theta) = E \left[\left\{ \frac{\partial}{\partial \Theta} \log f(\mathbf{D}, \Theta) \right\} \left\{ \frac{\partial}{\partial \Theta} \log f(\mathbf{D}, \Theta) \right\}^T \right]$$

is finite and positive definite at $\Theta = \Theta_0$, where Θ_0 is the true parameter value.

RC3

There exists an open subset ω of Θ containing the true parameter vector Θ_0 such that for almost all $\mathbf{D}$ the density $f(\mathbf{D}, \Theta)$ admits all third derivatives

$$\frac{\partial^3 \log f(\mathbf{D}, \Theta)}{\partial \theta_i \partial \theta_j \partial \theta_l} \quad \text{for all } \Theta \in \omega.$$

Furthermore there exists functions M_{ijl} such that

$$\left| \frac{\partial^3 \log f(\mathbf{D}, \Theta)}{\partial \theta_i \partial \theta_j \partial \theta_l} \right| \leq M_{ijl}(\mathbf{D}) \quad \text{for all } \Theta \in \omega,$$

where $m_{ijl} = E_{\Theta_0}[M_{ijl}(\mathbf{D})] < \infty$ for i, j, l .

RC4

Near Θ_0 the parameter space Ω behaves like a closed set; this condition was used by Self and Liang (1987) and Vu and Zhou (1997).

A.3. Proof of Theorem 1

Recall that LLG's REML and our REML are denoted as $I_1(\cdot)$ and $I_2(\cdot)$, and $\hat{\Theta}_1$ and $\hat{\Theta}_2$ denote LLG's estimator and our estimator, respectively. In the following derivation, we use $\ell(\cdot)$ and $\hat{\Theta}$ to denote the loglikelihood and estimator of either approach to simplify discussion. The true parameter value is denoted by Θ_0 .

Let $\alpha_n = n^{-1/2} + a_n$, where

$$\alpha_n = \frac{1}{n} \max \left\{ \frac{\partial \log(|X^T \Sigma^{-1} X|)}{\partial \Theta} \right\} = \frac{1}{n} \max \{ (X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} \Sigma_{\Theta} \Sigma^{-1} X) \} = o(1).$$

We want to show that for any $\varepsilon > 0$ there exists a constant C such that

$$P \left\{ w: \sup_{\|\mathbf{u}\|=C} \ell(\Theta_0 + \alpha_n \mathbf{u}) < \ell(\Theta_0) \right\} \geq 1 - \varepsilon. \quad (\text{A.1})$$

This implies with probability at least $1 - \varepsilon$ that there exists a local maximum in the ball $\{\Theta_0 + \alpha_n \mathbf{u}: \|\mathbf{u}\| \leq C\}$. Hence, there exists a local maximizer $\hat{\Theta}$ such that $\|\hat{\Theta} - \Theta_0\| = O(\alpha_n)$. Since the intersection of the parameter space Ω and the closure of a neighborhood about Θ_0 constitute of closed intervals, there also exist a local maximizer on this set although Θ_0 is on the boundary of parameter space Ω . Let $L'(\Theta_0)$ be the gradient vector of loglikelihood function L . By the standard argument on the Taylor expansion of the likelihood function and if $a_n = o(1)$, we have

$$\begin{aligned} \ell(\Theta_0 + \alpha_n \mathbf{u}) - \ell(\Theta_0) &= L(\Theta_0 + \alpha_n \mathbf{u}) - L(\Theta_0) + \frac{1}{2} \log(|X^T \Sigma^{-1}(\Theta_0 + \alpha_n \mathbf{u}) X|) - \frac{1}{2} \log(|X^T \Sigma^{-1}(\Theta_0) X|) \\ &\leq \alpha_n L'(\Theta_0)^T \mathbf{u} - \frac{1}{2} \mathbf{u}^T I(\Theta_0) \mathbf{u} n \alpha_n^2 \{1 + o(1)\} + \frac{1}{2} \log \left(\frac{|X^T \Sigma^{-1}(\Theta_0 + \alpha_n \mathbf{u}) X|}{|X^T \Sigma^{-1}(\Theta_0) X|} \right). \end{aligned} \quad (\text{A.2})$$

Note that $n^{-1/2} L'(\Theta_0) = O(1)$. Thus, the first term on the right-hand side of (A.2) is on the order $O(n^{1/2} \alpha_n) = O(n \alpha_n^2)$. By choosing a sufficiently large C , the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. The third term is also dominated by the second term. Hence, by choosing a sufficiently large C , (A.1) holds when Θ_0 is interior of Ω as well as when Θ_0 is on the boundary of Ω . This completes the proof of (1.1) in Theorem 1.

A.4. Proof of Lemma1

Since $\Sigma(\Theta)$ is symmetric matrix, we can express it as

$$\Sigma(\Theta) = \begin{pmatrix} \lambda_{1n}(\Theta) & 0 & \cdots & 0 \\ 0 & \lambda_{2n}(\Theta) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{nn}(\Theta) \end{pmatrix},$$

where $\lambda_{in}(\Theta)$ denotes the ordered eigenvalues of $\Sigma(\Theta)$ and $\Theta = (\theta_1, \dots, \theta_k)$. For easy notation, we let $\lambda_{in}(\Theta) \equiv \lambda_{in}$ and $\Sigma(\Theta) \equiv \Sigma$. We define

$$\frac{\partial \Sigma}{\partial \theta_k} = \Sigma_k = \begin{pmatrix} \partial_k(\lambda_{1n}) & 0 & \cdots & 0 \\ 0 & \partial_k(\lambda_{2n}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \partial_k(\lambda_{nn}) \end{pmatrix}.$$

Since

$$P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}$$

$$= \begin{pmatrix} \lambda_{1n}^{-1} & 0 & \cdots & 0 \\ 0 & \lambda_{2n}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{nn}^{-1} \end{pmatrix} - \frac{1}{\left(\frac{x_1^2}{\lambda_{1n}} + \cdots + \frac{x_n^2}{\lambda_{nn}}\right)} \begin{pmatrix} \lambda_{1n}^{-1} & 0 & \cdots & 0 \\ 0 & \lambda_{2n}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{nn}^{-1} \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & x_i x_j & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \lambda_{1n}^{-1} & 0 & \cdots & 0 \\ 0 & \lambda_{2n}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{nn}^{-1} \end{pmatrix},$$

we have

$$P \Sigma_k = \left[\begin{pmatrix} \lambda_{1n}^{-1} & 0 & \cdots & 0 \\ 0 & \lambda_{2n}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{nn}^{-1} \end{pmatrix} - \frac{1}{\left(\frac{x_1^2}{\lambda_{1n}} + \cdots + \frac{x_n^2}{\lambda_{nn}}\right)} \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \frac{x_i x_j}{\lambda_{in} \lambda_{jn}} & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \right] \begin{pmatrix} \partial_k(\lambda_{1n}) & 0 & \cdots & 0 \\ 0 & \partial_k(\lambda_{2n}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \partial_k(\lambda_{nn}) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial_k(\lambda_{1n})}{\lambda_{1n}} & 0 & \cdots & 0 \\ 0 & \frac{\partial_k(\lambda_{2n})}{\lambda_{2n}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{\partial_k(\lambda_{nn})}{\lambda_{nn}} \end{pmatrix} - \frac{1}{\left(\frac{x_1^2}{\lambda_{1n}} + \cdots + \frac{x_n^2}{\lambda_{nn}}\right)} \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \frac{x_i x_j \partial_k \lambda_{jn}}{\lambda_{in} \lambda_{jn}} & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}.$$

Hence, the m th diagonal element of $P \Sigma_k$ is

$$(P \Sigma_k)_{mm} = \frac{\partial_k \lambda_{mn}}{\lambda_{mn}} - \frac{1}{\sum_{\lambda_{in}} \frac{x_i^2}{\lambda_{in}^2}} \left(\frac{x_m^2}{\lambda_{mn}^2} \right) \partial_k(\lambda_{mn}) = \left(\frac{\sum_{i \neq m} \frac{x_i^2}{\lambda_{in}}}{\sum_{i=1}^n \frac{x_i^2}{\lambda_{in}}} \right) \frac{\partial_k(\lambda_{mn})}{\lambda_{mn}}.$$

Therefore, we have

$$\frac{1}{n} \text{tr}(P \sum_k) = \frac{1}{n} \sum_{m=1}^n \left(\frac{\frac{1}{n} \sum_{i \neq m} \frac{x_i^2}{\lambda_{in}}}{\frac{1}{n} \sum_{i=1}^n \frac{x_i^2}{\lambda_{in}}} \right) \frac{\partial_k(\lambda_{mn})}{\lambda_{mn}}.$$

Since

$$\frac{1}{n} \sum_{i \neq m} \frac{x_i^2}{\lambda_i} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \frac{x_i^2}{\lambda_i}$$

are the same when n goes ∞ ,

$$\frac{1}{n} \text{tr}(P \sum_k) \rightarrow E \left(\frac{\partial_k(\lambda_{mn})}{\lambda_{mn}} \right)$$

by the law of large numbers.

A.5. Conditions C1–C4

Cressie and Lahiri (1993, 1996) showed that a general result for the asymptotic property of REML estimator $\hat{\theta}_{reml}$ in a parametric linear model with Gaussian error, $Y \sim N_n(X\beta, \Sigma(\theta))$, holds under conditions C1–C4, which we describe later in this section, where Y is an $n \times 1$ data vector $(Y_1, \dots, Y_n)^T$, X is an $n \times q$ matrix of explanatory variables, β is a $q \times 1$ vector of unknown large scale effects, and $\Sigma(\theta)$ is an $n \times n$ positive definite variance matrix which is known up to a $k \times 1$ vector of small scale effects $\theta = (\theta_1, \dots, \theta_k)$. In our studies, $\Sigma(\theta) = \sigma^2 I + \tau K$ is a positive definite matrix. When we use the Gaussian kernel, $\theta = (\tau, \rho, \sigma^2)$ and $k = 3$. For the polynomial and neural network kernels, $\theta = (\tau, \sigma^2)$ and $k = 2$.

Let $U = I'Y$ represent a vector of $n-s$ linearly independent error contrast and; that is, the $(n-s)$ columns of U are linearly independent and $U \sim N(0, I' \Sigma(\theta) I)$. If a set of $(n-s)$ linearly independent contrasts are used to define U , the new negative loglikelihood function is

$$L_U(\theta) = \frac{n-s}{2} \log(2\pi) - \frac{1}{2} \log(|X^T X|) + \frac{1}{2} \log(\sum(\theta)) + \frac{1}{2} \log(|X^T \sum(\theta)^{-1} X|) + \left(\frac{1}{2}\right) Y^T P Y,$$

where $P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}$. A REML estimator of θ can be obtained by minimizing this equation. Let $\phi(\theta) = (-2L_U(\theta)/\theta_i \theta_j)$, the $k \times k$ matrix of the second-order partial derivatives of the negative loglikelihood function $L_U(\cdot)$. Then $E(\phi(\theta)) = \text{tr}(P \Sigma_k(\theta) P \Sigma_j(\theta)/2)$. Under conditions C1–C4, Cressie and Lahiri (1993, 1996) showed that

$$[E_{\theta}(\phi_n(\theta))]^{1/2} (\hat{\theta}_{reml} - \theta) = O(1).$$

In our semiparametric regression setting, we show that their result holds under certain situations which are described in Lemmas 1 and 2. By using Lemmas 1 and 2 of our paper

and the result of Cressie and Lahiri (1993, 1996), there exists a $\bar{c}$ such that $(1/n) \text{tr}(P\Sigma_A(\theta)P\Sigma_A(\theta)/2) \rightarrow_u \bar{c}$, where $\rightarrow_u$ denotes uniform convergence. Therefore, we can show that

$$\sqrt{n}(\widehat{\Theta}_{remt} - \Theta) = O(1).$$

The conditions needed for applying the result of Cressie and Lahiri (1993, 1996) are

- C1. $\Sigma(\theta)$ is twice continuously differentiable on θ .
- C2. For all $c > 0$, $\eta > 0$, let $\|B\| = \{\text{tr}(B'B)\}^{1/2}$,

$$\text{i.} \quad \sup \left\{ \left\| \{E_{\theta}\varphi_n(\Theta)\}^{-1/2} \{E_{\theta}\varphi_n(\Theta^0)\}^{1/2} - I_k \right\| : \left\| [E_{\theta}\varphi(\Theta)]^{-1/2} (\Theta - \Theta^0) \right\| \leq c; \Theta, \Theta^0 \in \Omega \right\} \rightarrow_u 0$$

$$\text{ii.} \quad \text{with } M = (\theta_1^0, \dots, \theta_k^0)$$

$$P_{\theta}(\sup \left\{ \left\| \{E_{\theta}\varphi_n(\Theta)\}^{-1/2} \{\varphi_n(M) - \varphi_n(\Theta)\} \{E_{\theta}\varphi_n(\Theta)\}^{-1/2} \right\|^T : \left\| \{E_{\theta}\varphi(\Theta)\}^{-1/2} (\Theta - \Theta_i^0) \right\| \leq c, 1 \leq i \leq k \right\} > \eta) \rightarrow_u 0.$$

- C3. There exists a positive definite matrix $W(\theta)$, continuous in θ , such that $Q_n(\theta)$ uniformly converge to $W(\theta)$, where

$$Q_n(\Theta)_{ij} = \text{tr}(P \sum_i(\Theta) P \sum_j(\Theta)) / \left(\left\| P \sum_i(\Theta) \right\| \left\| P \sum_j(\Theta) \right\| \right).$$

- Let $|\lambda_{1n}(\theta)| \dots |\lambda_{nn}(\theta)|$ denote the ordered absolute eigenvalues of $\Sigma(\theta)$, $|\lambda_{1n}^i(\theta)| \leq \dots \leq |\lambda_{nn}^i(\theta)|$ denote those of $\Sigma_{ij} = \Sigma(\theta) / \theta_i \theta_j$. Suppose there is a sequence $(r_n)_{n=1}^{\infty}$ with $\limsup_{n \rightarrow \infty} r_n/n = 1 - \delta^*$, for some $\delta^* \in (0, 1)$, such that for any compact subset $\Omega_s \subseteq \Omega$, there exist constants $0 < C_1(\Omega_s) < \infty$ and $\eta_1(\Omega_s) > 0$ such that

$$\limsup_{n \rightarrow \infty} \max \{ |\lambda_n(\Theta)|, |\lambda_n^i(\Theta)|, |\lambda_n^{ij}(\Theta)| : 1 \leq i, j \leq k \} < C_1(\Omega_s) < \infty, \\ \limsup_{n \rightarrow \infty} \min \{ |\lambda_1(\Theta)|, |\lambda_{r_n}^i(\Theta)| : 1 \leq i \leq k \} > \eta_1(\Omega_s) > 0$$

uniformly in $\theta \in \Omega_s$.

A.6. Proof of Theorem 2

We check the validity of assumptions (C1), (C2), (C3), and (C4) as follows:

- Conditions (C1) and (C2,ii) require smoothness of the variance–covariance matrix $\Sigma(\theta)$ as a function of θ . Since our $\Sigma(\theta) = \tau \mathbf{K}(\rho) + \sigma^2 I$ is a smooth function of θ , the conditions (C1) and (C2,ii) hold.
- Since $[E_{\theta}\{\varphi_n(\theta)\}^{1/2}]_{n=1}^{\infty}$ is smooth to guarantee (C2.i), condition (C2.i) holds by the arguments of Cressie and Lahiri (1996).
- C3. We show that condition (C3) holds under certain situation described in Lemmas 1 and 2. When we use the Gaussian, polynomial, and neural network

kernels, Σ and Σ_τ are simultaneously diagonalizable as well as Σ and Σ_ρ are. For the Gaussian kernel, we also need to show that Σ and Σ_ρ are simultaneously diagonalizable when n goes to ∞ . We can show this using Lemma 2 as follows:

The ij th elements of matrix $\mathbf{K}\mathbf{K}_\rho$ and $\mathbf{K}_\rho\mathbf{K}$ are

$$(\mathbf{K}\mathbf{K}_\rho)_{ij} = \sum_k \mathbf{K}_{ik}(\mathbf{K}_\rho)_{kj} = \left(1 + \sum_{k=1}^n |z_k - z_j|^2\right) \exp\left\{-\frac{|z_i - z_k|^2 + |z_k - z_j|^2}{\rho}\right\},$$

$$(\mathbf{K}_\rho\mathbf{K})_{ij} = \left(1 + \sum_{k=1}^n |z_i - z_k|^2\right) \exp\left\{-\frac{|z_i - z_k|^2 + |z_k - z_j|^2}{\rho}\right\}.$$

Since $E_Z(\mathbf{K}\mathbf{K}_\rho - \mathbf{K}_\rho\mathbf{K}) = 0$ when $n \rightarrow \infty$, $\mathbf{K}\mathbf{K}_\rho = \mathbf{K}_\rho\mathbf{K}$ as $n \rightarrow \infty$. Therefore, $\mathbf{K}$ and $\mathbf{K}_\rho$ are almost simultaneously diagonalizable for large n . This means that, Σ and Σ_ρ are almost simultaneously diagonalizable for large n . From Lemma 1, $(1/n) \text{tr}(P\Sigma_\rho)$ converges. Therefore, condition C3 holds if Σ and Σ_k are simultaneously diagonalizable.

- C4. Since $\Sigma(\theta)$ and $\Sigma_k(\theta)$ are nonsingular and uniformly bounded over a compact subset of Ω , this condition holds.

The proof of Theorem 2 is done when θ_0 is interior of parameter space Ω .

Since the intersection of parameter space Ω and the closure of a neighborhood about θ constitute of closed intervals, Theorem 2 also holds when θ_0 is on the boundary of parameter space Ω by arguments in Geyer (1994) and Vu and Zhou (1997) who extended the result of Chernoff (1954) to nonidentically distributed sampling. Geyer's (1994) result is based on a sampling model that is essentially a stationary process, while Vu and Zhou (1997) has no such restriction and allows general nonidentically distributed sampling so that models with covariances can be included.

We can demonstrate that our results hold using Vu and Zhou (1997) by relating their conditions with ours. Vu and Zhou (1997) provided that the existence, consistency, and asymptotic properties of the local maximum estimators covering a large class of estimation problems which allow sampling from nonidentically distributed random variable under certain conditions, A1–A2, B1–B4, in their paper: Regularity condition RC3 of our paper infers A1 in -Vu and Zhou (1997); RC4 of our paper suggest A2; RC1 and 2 of our paper imply B1; RC2 infers B2; Condition C2 in A.5 of our paper infers B3 and B4 in Vu and Zhou (1997). Therefore, under regularity conditions, Theorem 2 also holds when θ_0 is on the boundary of parameter space Ω .

We can also show our results using Geyer (1994) but we need to add his assumption D. Geyer (1994) assumed that the sampling distribution has a stationary process and showed that the existence, consistency, and asymptotic properties of the local maximum estimators under Assumptions A–D: since our density function $f(\mathbf{D}, \theta)$ is in $\mathcal{C}^3$ in θ , Assumption A in Geyer (1994) is satisfied; RC3 infers Assumption B in Geyer (1994); Assumption C in Geyer (1994) is

$$\sqrt{n} \left(\sum_i \frac{\partial f(D_i, \Theta)}{\partial \Theta} - E \left(\frac{\partial f(D_i, \Theta)}{\partial \Theta} \right) \right) \rightarrow N(0, A)$$

for some covariance matrix A . This assumption C is satisfied by the weak law large number theory and central limit theory; Assumption D requires that the estimating sequence $\{\hat{\Theta}_n\} = \Theta_0 + o_p(1)$. Therefore, Theorem 2 also holds under regularity conditions.

A.7. Proof of Theorem 3

For consistent estimators $\hat{\Theta}$ using a first order Taylor expansion of $l(\cdot)$ near Θ_0 yields,

$$0 = \frac{l(\Theta)}{\partial \Theta} \Big|_{\hat{\Theta}} = \frac{l(\Theta)}{\partial \Theta} \Big|_{\Theta_0} + \frac{l(\Theta)}{\partial \Theta} \Big|_{\bar{\Theta}} (\hat{\Theta} - \Theta_0),$$

where $\bar{\Theta}$ is a vector between $\hat{\Theta}$ and Θ_0 . Consequently, we have

$$\sqrt{n}(\hat{\Theta} - \Theta_0) = \left\{ -\frac{1}{n} \frac{\partial^2 l(\Theta)}{\partial \Theta \partial \Theta^T} \Big|_{\bar{\Theta}} \right\}^{-1} \frac{1}{\sqrt{n}} \frac{\partial l(\Theta)}{\partial \Theta} \Big|_{\Theta_0}.$$

Next, we show that, (i) the limit distribution of

$$\frac{1}{\sqrt{n}} \frac{\partial l(\Theta)}{\partial \Theta} \Big|_{\Theta_0} \rightarrow N(0, I(\Theta_0))$$

and (ii)

$$-\frac{1}{n} \frac{\partial^2 l(\Theta)}{\partial \Theta \partial \Theta^T} \Big|_{\bar{\Theta}} \rightarrow I(\Theta_0)$$

as $n \rightarrow \infty$

To prove (i), note that

$$\frac{1}{\sqrt{n}} \frac{\partial l(\Theta)}{\partial \Theta} \Big|_{\Theta_0} = \frac{1}{\sqrt{n}} \frac{\partial L(\Theta)}{\partial \Theta} \Big|_{\Theta_0} + \frac{1}{2\sqrt{n}} (X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} \sum_{\Theta} \Sigma^{-1} X).$$

The first term converges in distribution to $N(0, I(\Theta_0))$; the second term goes to zero with $o(n^{-1/2})$ because $(X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} \sum_{\Theta} \Sigma^{-1} X) = o(1)$. Thus, (i) holds.

For (ii), since

$$-\frac{1}{n} \frac{\partial^2 l(\Theta)}{\partial \Theta \partial \Theta^T} \Big|_{\bar{\Theta}} = -\frac{1}{n} \frac{\partial^2 L(\Theta)}{\partial \Theta \partial \Theta^T} \Big|_{\bar{\Theta}} - \frac{1}{2n} \{ (X^T \Sigma^{-1} X)^{-1} X^T (\sum_{\Theta} \Sigma^{-1} X) + (X^T \Sigma^{-1} X)^{-1} X^T (\sum_{\Theta} \Sigma^{-1} X) \}$$

and again the second term of the right side goes to zero, the first term converges in probability to $I(\theta_0)$ (Lehmann, 1983).

Thus, applying Slutsky's lemma and converting back to the original parameter space via the delta method, the asymptotic normality holds. The proof of (3.1) is done.

The proof of (3.2) can be shown by Chant's (1974) case (i). Suppose that $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$, θ_1 is a left endpoint of Ω_1 , and other components of θ are interior points of Ω_i , $i = 2, 3$. Then $\hat{\theta}$ can be expressed as

$$\begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix} I\{N_1 > 0\} + \begin{pmatrix} 0 \\ N_2 - (I_{21}/I_{11})N_1 \\ N_3 - (I_{31}/I_{11})N_1 \end{pmatrix} I\{N_1 > 0\},$$

where $\mathbf{N}$ is a random variable with a multivariate Gaussian distribution with mean θ and covariance $I^{-1}(\theta_0)$, where θ is restricted to lie in $C_\Omega - \theta_0$ and C_Ω is a cone with vertex at θ_0 . This expression corresponds to the result of Chant's (1974) case (i).

A.8. Proof of Theorem 4

If $\hat{\theta}$ is consistent then by a Taylor series approximation

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left\{ -\frac{1}{n} \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} \Big|_{\theta} \right\}^{-1} \frac{1}{\sqrt{n}} \frac{\partial l(\theta)}{\partial \theta} \Big|_{\theta_0}.$$

Therefore,

$$\begin{aligned} \hat{\theta}_1 - \theta_0 &\approx (-l_{1,\theta\theta}|_{\bar{\theta}})^{-1} l_{1,\theta}|_{\theta_0}, \\ \hat{\theta}_2 - \theta_0 &\approx (-l_{2,\theta\theta}|_{\bar{\theta}})^{-1} l_{2,\theta}|_{\theta_0}, \end{aligned}$$

where $\bar{\theta}$ is a vector between $\hat{\theta}$ and θ_0 , $l_{i,\theta} = I_i(\theta)/\theta^T$, and $l_{i,\theta\theta} = I_{i,\theta}(\theta)/\theta^T$, $i = 1, 2$.

By applying chain-rule, we obtain the first derivative and second derivatives of REML with respect to θ as follows:

$$l_{2,\theta}(\theta) = l\{\theta, \hat{\beta}(\theta)\}_{\theta} = l_{\theta}(\theta) + l_{\hat{\beta}}(\theta) \hat{\beta}_{\theta}(\theta) = l_{1,\theta}(\theta) + l_{\hat{\beta}}(\theta) \hat{\beta}_{\theta}(\theta) = l_{1,\theta}(\theta) \quad (\text{since } l_{\hat{\beta}} = 0)$$

and

$$\begin{aligned} l_{2,\theta\theta}(\theta) &= l\{\theta, \hat{\beta}(\theta)\}_{\theta\theta} = l_{\theta\theta}(\theta) + l_{\hat{\beta}\theta}(\theta) \hat{\beta}_{\theta}(\theta) + l_{\hat{\beta}}(\theta) \hat{\beta}_{\theta\theta}(\theta) = l_{\theta\theta}(\theta) + \left\{ \frac{\partial l_{\hat{\beta}}(\theta)}{\partial \beta(\theta)} \frac{\partial \hat{\beta}(\theta)}{\partial \theta} \right\} \hat{\beta}_{\theta}(\theta) + l_{\hat{\beta}}(\theta) \hat{\beta}_{\theta\theta}(\theta) \\ &= l_{\theta\theta}(\theta) + l_{\hat{\beta}\theta}(\theta) \{\hat{\beta}_{\theta}(\theta)\}^2 + l_{\hat{\beta}}(\theta) \hat{\beta}_{\theta\theta}(\theta) = l_{1,\theta\theta}(\theta) + l_{\hat{\beta}\theta} \{\hat{\beta}_{\theta}(\theta)\}^2 \quad (\text{since } l_{\hat{\beta}} = 0, l_{\hat{\beta}\theta} \{\hat{\beta}_{\theta}(\theta)\}^2 < 0) \\ &< l_{1,\theta\theta} < 0. \end{aligned}$$

Since $\|l_{2,\theta\theta}\| > \|l_{1,\theta\theta}\|$, we obtain the following inequality:

$$\|\widehat{\Theta}_1 - \Theta\| \approx \|l_{1,\Theta}^{-1} l_{1,\Theta}\| > \|l_{2,\Theta}^{-1} l_{1,\Theta}\| = \|l_{2,\Theta}^{-1} l_{2,\Theta}\| \approx \|\widehat{\Theta}_2 - \Theta\|.$$

Therefore, $E\|\widehat{\Theta}_1 - \Theta\|^2 > E\|\widehat{\Theta}_2 - \Theta\|^2$. This means that the estimators based on our REML have theoretically smaller mean square error than those of LLG's REML. The proof of Theorem 4 is done.

A.9. Non-identifiability between τ and ρ when $\rho \rightarrow 0$

If $\tau \sim O(1/\rho^m)$ for any positive value m and $\rho \sim O\{E(\|\mathbf{z} - \mathbf{z}'\|^2)\}$, $\widehat{\Theta}_2$ is asymptotically normally distributed with mean Θ and covariance matrix $\mathcal{I}(\Theta)$. But if $\rho \rightarrow 0$, $H_\tau = H_{\tau\tau} = P_\tau = P_{\tau\tau} = 0$ and $H_\rho = H_{\rho\rho} = P_\rho = P_{\rho\rho} = 0$, then the asymptotic distributions of $\hat{\tau}_2$ and $\hat{\rho}_2$ are the same and degenerated.

For LLG's estimator $\widehat{\Theta}_1$, as $\rho \rightarrow 0$, the asymptotic distributions of $\hat{\tau}_1$ and $\hat{\rho}_1$ are also the same because the information matrix of Liu et al. (2007) is

$$I_{\theta_i, \theta_{i'}} = \frac{1}{2} \text{tr} \{ P \sum_{\theta_i} P \sum_{\theta_{i'}} \}$$

and $\Sigma_{\theta_i} \rightarrow \mathbf{0}$.

A.10. The proof of equivalence of the two testings

The testing of $H_0: \{\mathbf{r}(\mathbf{Z}) \text{ is a point mass at zero}\} \cup \{\mathbf{r}(\mathbf{Z}) \text{ has a constant covariance matrix as a function of } \mathbf{z}\}$ is equivalent to the testing of $\mathbf{K}(\mathbf{Z})/\mathbf{Z} = \mathbf{0}$.

$$K(\mathbf{z}_i, \mathbf{z}_j) = \tau \exp \left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\rho} \right),$$

$$\frac{\partial K(\mathbf{z}_i, \mathbf{z}_j)}{\partial \mathbf{z}_i} = -2(\|\mathbf{z}_i - \mathbf{z}_j\|) \left(\frac{\tau}{\rho} \right) \exp \left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\rho} \right).$$

If $\rho \rightarrow 0$ and $\tau \rightarrow 0$ with faster rate $O(\rho^m)$, then $\mathbf{K}(\mathbf{Z})/\mathbf{Z} = \mathbf{0}$. That is, if $\tau/\rho \rightarrow 0$, then $\mathbf{K}(\mathbf{Z})/\mathbf{Z} = \mathbf{0}$.

If $\rho \rightarrow \infty$, then $0 \leq \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|^2/\rho) \leq 1$. Therefore,

$$0 \leq \frac{\tau}{\rho} \exp \left(-\frac{\|\mathbf{z} - \mathbf{z}_j\|^2}{\rho} \right) \leq \frac{\tau}{\rho}.$$

Hence, if

$$\frac{\tau}{\rho} \rightarrow 0, \quad \frac{\partial \mathbf{K}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}.$$

If $\rho \rightarrow 0$ and $\tau \sim O(1/\rho^m)$, then $\exp(-\|\mathbf{z}_i - \mathbf{z}_j\|^2/\rho) \rightarrow 0$. Therefore, $\mathbf{K}(\mathbf{Z})/\mathbf{Z} = \mathbf{0}$. Therefore, if $\tau/\rho \rightarrow 0$ or $\rho \rightarrow 0$, then $\mathbf{K}(\mathbf{Z})/\mathbf{Z} = \mathbf{0}$.

Appendix B. Supplementary

Some additional results and information on pathway and genes are available in a separate file for supplementary materials.

Supplementary data associated with this paper can be found in the online version at <http://dx.doi.org/10.1016/j.jspi.2012.09.009>.

Table 1

The coverage probability (cvpr) of 95% confidence intervals and average mean squared error values (mse) of parameter estimates. LLIKE=LLG's likelihood-based approach with kernel K ; PLIKE=our profile likelihood-based approach with kernel K ; GK=Gaussian kernel; P2K=Quadratic kernel; P1K=Linear kernel.

The number of sample size and gene		K	Method	Case	Coverage probability and mean squared error				β	σ	τ	ρ
$n=60$ $p=5$	1	GK	LLIKE	1	cvpr $\times$ 100	mse	97	96	0.0053	0.0247	0.0483	90
			PLIKE		cvpr	mse	97	96	0.0048	0.0151	0.0322	91
	2	P2K	LLIKE	2	cvpr $\times$ 100	mse	98	93	0.0053	0.0485	0.0252	N/A
			PLIKE		cvpr $\times$ 100	mse	98	93	0.0050	0.0474	0.0243	N/A
$n=60$ $p=200$	1	GK	LLIKE	1	cvpr $\times$ 100	mse	97	93	0.0051	0.0104	0.0165	N/A
			PLIKE		cvpr $\times$ 100	mse	97	93	0.0042	0.0095	0.0142	N/A
	2	P2K	LLIKE	2	cvpr $\times$ 100	mse	89	87	0.0321	0.0436	0.0424	80
			PLIKE		cvpr $\times$ 100	mse	89	89	0.0304	0.0295	0.0421	80
$n=60$ $p=200$	1	GK	LLIKE	1	cvpr $\times$ 100	mse	86	63	0.0478	0.4683	0.0388	N/A
			PLIKE		cvpr $\times$ 100	mse	86	63	0.0475	0.4620	0.0379	N/A
	2	P2K	LLIKE	2	cvpr $\times$ 100	mse	86	65	0.0486	0.2984	0.0583	N/A
			PLIKE		cvpr	mse	86	66	0.0480	0.2839	0.0564	N/A

Table 2

The average mean squared error values (mse) of parameter estimates for different sample sizes. LLIKE=LLG's likelihood-based approach with kernel K ; PLIKE=our profile likelihood-based approach with kernel K ; GK=Gaussian kernel.

The number of sample size and gene	K	Method	Case	Coverage probability and mean squared error	β	σ	τ	ρ
$n=15$	GK	LLIKE	1	mse	0.0865	0.0774	0.0799	0.0974
$p=5$		PLIKE		mse	0.0741	0.0643	0.0673	0.0791
$n=30$	GK	LLIKE	1	mse	0.0067	0.0319	0.0529	0.0723
$p=5$		PLIKE		mse	0.0062	0.0234	0.0436	0.0558
$n=60$	GK	LLIKE	1	mse	0.0053	0.0247	0.0483	0.0623
$p=5$		PLIKE		mse	0.0048	0.0151	0.0322	0.0458
$n=600$	GK	LLIKE	1	mse	0.0025	0.0114	0.0235	0.0312
$p=5$		PLIKE		mse	0.0020	0.0107	0.0197	0.0259

Table 3

Estimated type I error rate and power based on profile restricted likelihood ratio test (PRLRT) and permutation test (PERM). LLIKE=LLG's likelihood-based approach with kernel K ; PLIKE=our profile likelihood-based approach with kernel K ; GK=Gaussian kernel; PRLRT(e) and PRLRT(t) are profile restricted likelihood ratio tests which are performed using empirical distribution and theoretical distribution, respectively.

The number of sample size and gene	Type I error and power	Kernel (GK)	LLIKE		PLIKE			
			PERM	PRLRT(e)	PRLRT(t)	PERM	PRLRT(e)	PRLRT(t)
Case								
$n = 60$	Type I	4	0.04	0.05	0.05	0.04	0.05	0.05
$p = 5$	Power	5	0.99	0.99	1	0.99	0.99	1
$n = 60$	Type I	4	0.03	0.04	0.04	0.03	0.04	0.04
$p = 200$	Power	5	0.84	0.84	0.84	0.84	0.85	0.85

Table 4

Estimated type I error rate and power based on profile restricted likelihood ratio test (PRLRT) and permutation test (PERM). LLIKE=LLG's likelihood-based approach with kernel K ; PLIKE=our profile likelihood-based approach with kernel K ; P2K=Quadratic kernel; PRLRT(e) and PRLRT(t) are profile restricted likelihood ratio tests which are performed using empirical distribution and theoretical distribution, respectively.

The number of sample size and gene	Type I error and power	Kernel (P2K)	LLIKE		PLIKE			
			PERM	PRLRT(e)	PRLRT(t)	PERM	PRLRT(e)	PRLRT(t)
Case								
$n = 60$	Type I	4	0.04	0.05	0.05	0.04	0.05	0.05
$p = 5$	Power	5	0.97	1	1	0.98	1	1
$n = 60$	Type I	4	0.03	0.04	0.04	0.03	0.04	0.04
$p = 200$	Power	5	0.83	0.85	0.85	0.84	0.85	0.85

Table 5

Estimated type I error rate and power based on profile restricted likelihood ratio test (PRLRT) and permutation test (PERM). LLIKE=LLG's likelihood-based approach with kernel K ; PLIKE=our profile likelihood-based approach with kernel K ; P1K=Linear kernel; PRLRT(e) and PRLRT(t) are profile restricted likelihood ratio tests which are performed using empirical distribution and theoretical distribution, respectively.

The number of sample size and gene	Type I error and power	Kernel (P1K)	LLIKE		PLIKE			
			PERM	PRLRT(e)	PRLRT(t)	PERM	PRLRT(e)	PRLRT(t)
Case								
$n = 60$	Type I	4	0.04	0.05	0.05	0.04	0.05	0.05
$p = 5$	Power	5	0.97	1	1	0.97	1	1
$n = 60$	Type I	4	0.03	0.04	0.04	0.03	0.04	0.04
$p = 200$	Power	5	0.82	0.83	0.83	0.82	0.83	0.83

Seven pathways which are significant over four kernels (Linear, Quadratic, Gaussian, Neural network kernels) using pathway data of type II diabetes; P1K=linear kernel; P2K=quadratic kernel; GK=Gaussian kernel; NNK=neural network kernel; P-values are obtained using both profile restricted likelihood ratio test (PRLRT) and permutation test (PERM); PRLRT(t) is a profile restricted likelihood ratio test which is performed using theoretical distribution.

Table 6

Pathway ID	Pathway name (# of genes)	P1K		P2K		GK		NNK	
		PERM	PRLRT(t)	PERM	PRLRT(t)	PERM	PRLRT(t)	PERM	PRLRT(t)
4	<i>Alanine and aspartate metabolism</i> (18)	0.009	0.008	0.003	0.002	0.034	0.030	0.016	0.014
36	<i>c17_U133_probes</i> (116)	0.029	0.031	0.002	0.001	0.003	0.002	0.015	0.013
133	<i>MAF00190_Oxidative_phosphorylation</i> (58)	0.025	0.027	0.007	0.007	0.032	0.029	0.022	0.025
229	<i>Oxidative_phosphorylation</i> (113)	0.023	0.021	0.008	0.009	0.035	0.039	0.024	0.022
209	<i>MAF03020_RNA_polymerase</i> (21)	0.034	0.031	0.027	0.026	0.040	0.041	0.025	0.026
254	<i>RNA polymerase</i> (25)	0.037	0.038	0.028	0.023	0.041	0.043	0.022	0.021