Biometrics 63, 1079–1088
December 2007
DOI: 10.1111/j.1541-0420.2007.00799.x

Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models

Dawei Liu,[1,*] Xihong Lin,[2,] and Debashis Ghosh[3,***]**

Center for Statistical Sciences, Brown University, Providence, Rhode Island 02912, U.S.A
Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A
[*] *email:* daweiliu@stat.brown.edu
[**] *email:* xlin@hsph.harvard.edu
[***] *email:* ghoshd@umich.edu

Summary. We consider a semiparametric regression model that relates a normal outcome to covariates and a genetic pathway, where the covariate effects are modeled parametrically and the pathway effect of multiple gene expressions is modeled parametrically or nonparametrically using least-squares kernel machines (LSKMs). This unified framework allows a flexible function for the joint effect of multiple genes within a pathway by specifying a kernel function and allows for the possibility that each gene expression effect might be nonlinear and the genes within the same pathway are likely to interact with each other in a complicated way. This semiparametric model also makes it possible to test for the overall genetic pathway effect. We show that the LSKM semiparametric regression can be formulated using a linear mixed model. Estimation and inference hence can proceed within the linear mixed model framework using standard mixed model software. Both the regression coefficients of the covariate effects and the LSKM estimator of the genetic pathway effect can be obtained using the best linear unbiased predictor in the corresponding linear mixed model formulation. The smoothing parameter and the kernel parameter can be estimated as variance components using restricted maximum likelihood. A score test is developed to test for the genetic pathway effect. Model/variable selection within the LSKM framework is discussed. The methods are illustrated using a prostate cancer data set and evaluated using simulations.

Key words: BLUPs; Kernel function; Model/variable selection; Nonparametric regression; Penalized likelihood; REML; Score test; Smoothing parameter; Support vector machines.

1. Introduction

Analysis of microarray data has been mainly focused on detection of individually significantly expressed genes (Efron et al., 2001; Tusher, Tibshirani, and Chu, 2001). This approach has some major limitations: (1) long list of individually significant genes without a single encompassing theme is difficult to interpret; (2) cellular processes often affect sets of genes and individually highly ranked genes are often downstream genes, so moderate changes in many genes may give more insight into biological mechanisms than dramatic change in a single gene (Mootha et al., 2003); (3) individually highly ranked genes can be poorly annotated and are often not reproducible across studies (Fortunel et al., 2003). Researchers have now become more interested in knowledge-based studies on gene sets, for example, genetic pathways that are more biologically interpretable and reproducible (Goeman et al., 2005; Subramanian et al., 2005).

A data example motivating the proposed research is the data from the Michigan prostate cancer study (Dhanasekaran et al., 2001). Prostate-specific antigen (PSA) has been routinely used as a biomarker for screening prostate cancer. Recently there have been significant breakthroughs in the effort of finding candidate genes related to prostate cancer. The early results of Dhanasekaran et al. (2001) indicate that certain functional genetic pathways seemed dysregulated in prostate cancer relative to noncancerous tissues. One is interested in studying the genetic pathway effects on PSA after adjusting for effects of clinical and demographic covariates. Due to the complicated unknown relationships between genes and PSA, we propose a flexible framework to model the genetic pathway effect parametrically or nonparametrically.

There is a vast literature on multidimensional nonparametric modeling. Methods such as multivariate kernel smoothing (Wand and Jones, 1995), projection pursuit regression (Friedman and Stuetzle, 1981), and multivariate adaptive regression splines (MARS) (Friedman, 1991), are usually computationally expensive. Popular spline-based methods include generalized additive models (GAMs) (Hastie and Tibshirani, 1990), thin-plate splines (Wahba, 1990; Green and Silverman, 1994), penalized regression splines (Ruppert, Wand, and

Carroll, 2004), and smoothing spline ANOVA (Gu, 2002). These methods require the specification of the smoothness condition of an unknown function using differentiability conditions, which is much more involved and awkward in multidimensional settings.

In the past decade, the kernel machine method has been developed in machine learning as a powerful learning technique for multidimensional data (Vapnik, 1998; Schölkopf and Smola, 2002; Suykens et al., 2002; Rasmussen and Williams, 2006). Popular examples of kernel machine methods include support vector machine (SVM) (Vapnik, 1998) and Bayesian Gaussian process (Rasmussen and Williams, 2006). In the context of function approximation, kernel machine methods and spline-based methods share a similar theoretical foundation, but their model-fitting philosophies are different. Kernel machine methods start with a kernel function that implicitly determines the smoothness property of the unknown function. By contrast, spline-based methods start with the smoothness conditions of the unknown function and a corresponding kernel function can usually be derived from these conditions (Wahba, 1990). Kernel machine methods hence greatly simplify specification of a nonparametric model, especially for multidimensional data.

In this article, we propose a semiparametric model for covariate and genetic pathway effects on a continuous outcome (e.g., PSA), where covariates effects are modeled parametrically and genetic pathway effect is modeled parametrically or nonparametrically using least-squares kernel machine (LSKM). We establish a connection between LSKM and linear mixed models, and show that the LSKM estimator of the regression coefficients and the pathway effect can be obtained by fitting a linear mixed model. This connection provides a unified framework for inference of parameters in models with multidimensional covariates, including the regression coefficients, the nonparametric function, and smoothing parameters. Our work extends the connection between univariate smoothing splines and linear mixed models (Speed, 1991; Wang, 1998; Zhang et al., 1998) to multivariate smoothing with an arbitrary kernel function. We also propose a score test to test for the nonparametric genetic pathway effect, and a model/variable selection method within the LSKM framework.

The rest of the article is organized as follows. In Section 2, we present the semiparametric model for Gaussian outcomes. In Section 3, we describe the LSKM method. In Section 4, we establish a connection between LSKMs and linear mixed models and propose a score test for testing for the genetic pathway effect. We discuss the variable selection problem in LSKM in Section 5. The performance of the proposed method is evaluated by simulations in Section 7, and is illustrated using the prostate cancer microarray data in Section 6. The article ends with a discussion in Section 8.

2. Semiparametric Model for Multidimensional Data

2.1 *The Model*

Suppose the data consist of n subjects. For subject i ($i = 1, \ldots, n$), y_i is a normally distributed continuous outcome, $\boldsymbol{x}_i$ is a $q \times 1$ vector of clinical covariates and $\boldsymbol{z}_i$ is a $p \times 1$ vector of gene expressions within a pathway. We assume an intercept is included in $\boldsymbol{x}_i$. The outcome y_i depends on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ through the following partial linear model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{z}_i) + e_i, \tag{1}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, $h(\boldsymbol{z}_i)$ is an unknown centered smooth function, and the errors e_i are assumed to be independent and follow $N(0, \sigma^2)$.

Model (1) models covariate effects parametrically and the pathway effect parametrically or nonparametrically. When $h(\cdot) = 0$, (1) reduces to the standard linear regression model. When $\boldsymbol{x}_i = 1$, it reduces to LSKM regression (Suykens et al., 2002).

2.2 *Specifications of a Function Space of $h(\boldsymbol{z})$ Using a Kernel*

We assume the nonparametric function $h(\boldsymbol{z})$ lies in a function space $\mathcal{H}_K$ generated by a positive definite kernel function $K(\cdot, \cdot)$. From Mercer's theorem (Cristianini and Shawe-Taylor, 2000), under some regularity conditions, a kernel function $K(\cdot, \cdot)$ implicitly specifies a unique function space spanned by a particular set of orthogonal basis functions (features) $\{\phi_j(\boldsymbol{z})\}_{j=1}^J$. In other words, any $h(\boldsymbol{z}) \in \mathcal{H}_K$ can be represented using a set of bases as $h(\boldsymbol{z}) = \sum_{j=1}^J \omega_j \phi_j(\boldsymbol{z}) = \boldsymbol{\phi}(\boldsymbol{z})^T \boldsymbol{\omega}$ (the primal representation), where $\boldsymbol{\omega}$ is a vector of coefficients. Equivalently, $h(\boldsymbol{z})$ can also be represented using a kernel function $K(\cdot, \cdot)$ as $h(\boldsymbol{z}) = \sum_{l=1}^L \alpha_l K(\boldsymbol{z}_l^*, \boldsymbol{z}; \rho)$ (the dual representation), for some integer L, some constants α_l and some $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_L^*\} \in R^p$. For a multidimensional $\boldsymbol{z}$, it is more convenient to specify $h(\boldsymbol{z})$ using the dual representation, because explicit basis functions or features might be complicated to specify, and the number of features might be high or even infinite.

Two popular kernel functions and the corresponding function spaces are as follows: (1) *The dth Polynomial Kernel*: $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = (\boldsymbol{z}_1^T \boldsymbol{z}_2 + \rho)^d$, where ρ and d are tuning parameters. The dth polynomial kernel generates the function space $\mathcal{H}_\mathcal{K}$ spanned by all possible dth-order monomials of the components of $\boldsymbol{z}$. For example, if $d = 1$, the first polynomial kernel generates the linear function space with basis functions $\{\phi_j(\boldsymbol{z})\} = \{z_1, \ldots, z_p\}$. If $d = 2$, the second polynomial kernel corresponds to the quadratic function space with basis functions $\{\phi_j(\boldsymbol{z})\} = \{z_k, z_k z_{k'}\}$ ($k, k' = 1, \ldots, p$), that is, the main effects, all two way interactions and quadratic main effects of the $\boldsymbol{z}_k$'s. (2) *The Gaussian Kernel*: $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \exp\{-||\boldsymbol{z}_1 - \boldsymbol{z}_2||^2/\rho\}$, where $||\boldsymbol{z}_1 - \boldsymbol{z}_2||^2 = \sum_{k=1}^p (z_{1k} - z_{2k})^2$. The Gaussian kernel generates the function space spanned by radial basis functions. See Buhmann (2003) for their mathematical properties and desirable features. Examples of other choices of kernel functions include the sigmoid and neural network kernels, and the B-spline kernel (Schölkopf and Smola, 2002). The choice of a kernel function determines which function space one would like to use to approximate $h(\boldsymbol{z})$.

3. LSKM Estimation in the Semiparametric Model

Assume $h(\cdot) \in \mathcal{H}_K$, the function space generated by a kernel function $K(\cdot, \cdot)$. Estimation of $\boldsymbol{\beta}$ and $h(\cdot)$ in (1) proceeds by maximizing the scaled penalized likelihood function

$$J(h, \boldsymbol{\beta}) = -\frac{1}{2} \sum_{i=1}^n \left\{ y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - h(\boldsymbol{z}_i) \right\}^2 - \frac{1}{2} \lambda \|h\|_{\mathcal{H}_K}^2, \tag{2}$$

where λ is a tuning parameter which controls the tradeoff between goodness of fit and complexity of the model. When $\lambda = 0$, the model interpolates the gene expression data,

whereas when $\lambda = \infty$, the model reduces to a simple linear model without $h(\cdot)$.

By the Representer theorem (Kimeldorf and Wahba, 1970), the general solution for the nonparametric function $h(\cdot)$ in (2) can be expressed as

$$h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot, \boldsymbol{z}_i), \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ are unknown parameters. Substituting (3) back into (2) we have

$$J(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{ y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \sum_{j=1}^{n}\alpha_j K(\boldsymbol{z}_i, \boldsymbol{z}_j)\right\}^2 - \frac{1}{2}\lambda\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}, \tag{4}$$

where $\boldsymbol{K}$ is an $n \times n$ matrix whose (i, j)th element is $K(\boldsymbol{z}_i, \boldsymbol{z}_j)$. Differentiating $J(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, some calculations give

$$\hat{\boldsymbol{\beta}} = \{\mathbf{X}^T(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{X}\}^{-1}\boldsymbol{X}^T(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{y} \tag{5}$$

$$\hat{\boldsymbol{\alpha}} = \lambda^{-1}(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{6}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Plugging (6) into (3), we have that the function $h(\cdot)$ evaluated at the design points $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$ is estimated as

$$\hat{\boldsymbol{h}} = \boldsymbol{K}\hat{\boldsymbol{\alpha}} = \lambda^{-1}\boldsymbol{K}(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{7}$$

Using (3) and (6), $\hat{h}(\cdot)$ at an arbitrary $\boldsymbol{z}$ is

$$\hat{h}(\boldsymbol{z}) = \lambda^{-1}\{K(\boldsymbol{z}, \boldsymbol{z}_1), \ldots, K(\boldsymbol{z}, \boldsymbol{z}_n)\}(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{8}$$

Equivalently, if $h(\boldsymbol{z}) = \boldsymbol{\phi}(\boldsymbol{z})^T\boldsymbol{\omega}$, where $\{\phi_j(\boldsymbol{z})\}$ are orthogonal basis functions, the corresponding LSKM regression coefficients $\hat{\boldsymbol{\omega}}$ are

$$\hat{\boldsymbol{\omega}}(\boldsymbol{z}) = \lambda^{-1}\{\boldsymbol{\phi}(\boldsymbol{z}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{z}_n)\}(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{9}$$

The kernel function $K(\cdot, \cdot)$ usually depends on an unknown parameter ρ, such as the scale parameter in Gaussian kernel. Inference on $\hat{\boldsymbol{\beta}}, \hat{h}(\boldsymbol{z})$ depends on λ, ρ and the residual variance σ^2, which need to be estimated. Cross-validation can be used to estimate λ; however, its computation is often intensive. Little literature is available on the systematic estimation of ρ and σ^2. In the machine learning literature, ρ is often preset at some fixed values. Further, estimation of σ^2 needs to properly account for the loss of degrees of freedom from estimating $\boldsymbol{\beta}$ and $h(\cdot)$. Hence it is desirable to develop a systematic method to estimate these parameters simultaneously. We accomplish this by establishing a connection between LSKM and linear mixed models.

4. LSKMs and Linear Mixed Models

4.1 *Connection Between LSKMs and Linear Mixed Models*

Linear mixed models have commonly been used for analyzing longitudinal and hierarchical data (Harville, 1977; Laird and Ware, 1982). A connection between smoothing splines and linear mixed models has been established (Speed, 1991; Wang, 1998; Zhang et al., 1998). We show here that the LSKM estimator in model (1) corresponds to the best linear unbiased predictor (BLUP) estimator from a linear mixed model, and the regularization parameters (τ, ρ) and the residual variance σ^2 can be treated as variance components and estimated simultaneously using restricted maximum likelihood (REML).

To see this connection, simple calculations show that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{h}}$ from equations (5) and (7) can be equivalently obtained from the equations

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1} \\ \boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{R}^{-1} + (\tau\boldsymbol{K})^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{h} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}, \tag{10}$$

where $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ and $\tau = \lambda^{-1}\sigma^2$. Equation (10) corresponds exactly to the normal equation of the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{h} + \boldsymbol{e}, \tag{11}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, $\boldsymbol{h}$ is an $n \times 1$ vector of random effects with distribution $N(\mathbf{0}, \tau\boldsymbol{K})$, and $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$. A comparison of (11) with model (1) indicates that they have exactly the same form except that $\boldsymbol{h}$ is now treated as random effects. It follows that the BLUPs of the regression coefficients $\hat{\boldsymbol{\beta}}$ and the random effects $\hat{\boldsymbol{h}}$ under the linear mixed model (11) correspond to the LSKM estimator given in Section 3. In fact, one can easily see that the regression coefficient estimator $\hat{\boldsymbol{\beta}}$ in (5) is the weighted least-squares estimator under the linear mixed model representation (11) using the marginal covariance of $\boldsymbol{y}$ under (11) as $\boldsymbol{V} = \sigma^2\boldsymbol{I} + \tau\boldsymbol{K}$, i.e., $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$.

The linear mixed model representation of the LSKM in the semiparametric model (1) can also be considered as a Bayesian Gaussian process regression (Schölkopf and Smola, 2002). Note that this Bayesian correspondence is finite-dimensional (Wahba, 1990; Green and Silverman, 1994). It is not strictly equivalent to a continuous Bayesian Gaussian process (Rasmussen and Williams, 2006), because the finite-dimensional representation of $h(\cdot)$ does not lead to a coherent Bayesian model (Green and Silverman, 1994; Tipping, 2001; Sollich, 2002. To see the Bayesian representation, we can treat $\{h(\boldsymbol{z})\}$ as a random vector with a Gaussian process (GP) prior, with mean 0 and covariance $\mathrm{cov}\{h(\boldsymbol{z}_1), h(\boldsymbol{z}_2)\} = \tau K(\boldsymbol{z}_1, \boldsymbol{z}_2)$. Note that the positive definiteness of the kernel function $K(\cdot, \cdot)$ ensures it is a proper covariance function. Now we assume

$$y \,|\, (\boldsymbol{\beta}, h(\boldsymbol{z})) \sim N\{\boldsymbol{x}^T\boldsymbol{\beta} + h(\boldsymbol{z}), \sigma^2\},$$
$$h(\cdot) \sim \mathrm{GP}\{0, \tau K(\cdot, \cdot)\}, \quad \boldsymbol{\beta} \propto \mathbf{1}.$$

One can easily see that under this Bayesian model, the semiparametric model (1) becomes the linear mixed model representation (11). This connection extends the connection between scalar smoothing splines and mixed models and their Bayesian formulations (Wang, 1998; Zhang et al., 1998) to multidimensional regression problems under the kernel machine framework.

The covariances of $\hat{\boldsymbol{\beta}}$ and $\hat{h}(\cdot)$ can be calculated in two ways. The first approach is to treat the true $h(\cdot)$ as a fixed

unknown function and the variance of y_i as σ^2. Using (5) and (7), the covariances of $\hat{\boldsymbol{\beta}}$ and $\hat{h}(\cdot)$ are

$$\text{cov}_F(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}, \tag{12}$$

$$\text{cov}_F(\hat{\boldsymbol{h}}) = \sigma^2(\tau\boldsymbol{K})\boldsymbol{P}^2(\tau\boldsymbol{K}), \tag{13}$$

$$\text{cov}_F\{\hat{h}(\boldsymbol{z})\} = \sigma^2\big(\tau\boldsymbol{K}_z^T\big)\boldsymbol{P}^2(\tau\boldsymbol{K}_z) \quad \text{for arbitrary } \boldsymbol{z},$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$ and $\boldsymbol{K}_z = \{K(\boldsymbol{z}, \boldsymbol{z}_1), \ldots, K(\boldsymbol{z}, \boldsymbol{z}_n)\}^T$ for an arbitrary $\boldsymbol{z}$. We term these covariances as frequentist covariances.

The second approach is to use the linear mixed model representation (11) and treat the true $h(\cdot)$ as a random function following the mean zero Gaussian process with covariance $\tau K(\cdot, \cdot)$. The covariances of $\hat{\boldsymbol{\beta}}$ and $\hat{h}(\cdot)$ can then be calculated as a byproduct of the covariance of the fixed and random effects of the linear mixed model (11) and are

$$\text{cov}_B(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}, \tag{14}$$

$$\text{cov}_B(\hat{\boldsymbol{h}}) = \text{cov}(\hat{\boldsymbol{h}} - \boldsymbol{h}) = \tau\boldsymbol{K} - (\tau\boldsymbol{K})\boldsymbol{P}(\tau\boldsymbol{K}), \tag{15}$$

$$\begin{aligned}\text{cov}_B\{\hat{\boldsymbol{h}}(\boldsymbol{z})\} &= \text{cov}\{\hat{h}(\boldsymbol{z}) - h(\boldsymbol{z})\} \\ &= \tau K(\boldsymbol{z}, \boldsymbol{z}) - (\tau\boldsymbol{K}_z)\boldsymbol{P}(\tau\boldsymbol{K}_z).\end{aligned}$$

We term these covariances as Bayesian covariances.

4.2 *Estimation of the Regularization Parameters and the Residual Variance*

We discuss in this section estimation of the regularization parameter τ, the residual variance σ^2 and the scale parameter ρ in $K(\cdot, \cdot)$. Using the mixed model representation of LSKM, we propose to estimate (τ, ρ, σ^2) simultaneously by treating them as variance components in the linear mixed model (11) and estimating them using REML.

Specifically, the REML under the linear mixed model (11) can be written as

$$\begin{aligned}\ell_R(\sigma^2, \tau, \rho) = &-\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X}| \\ &- \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),\end{aligned} \tag{16}$$

where $\boldsymbol{\theta} = (\tau, \rho, \sigma^2)^T$. The score equations of (τ, ρ, σ^2) are

$$\begin{aligned}&-\frac{1}{2}\text{tr}(\boldsymbol{K}\boldsymbol{P}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}\boldsymbol{K}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0, \\ &-\frac{1}{2}\text{tr}\left\{\tau\frac{\partial\boldsymbol{K}}{\partial\rho}\boldsymbol{P}\right\} + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1} \\ &\quad\times\left(\tau\frac{\partial\boldsymbol{K}}{\partial\rho}\right)\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0, \\ &-\frac{1}{2}\text{tr}(\boldsymbol{P}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0,\end{aligned} \tag{17}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$. Let $\mathbf{A}$ denote the hat matrix so that $\boldsymbol{X}^T\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{h}} = \boldsymbol{A}\boldsymbol{y}$. Using the identities $\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \{\sigma^2\}^{-1}(\boldsymbol{y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{h}})$ and $\boldsymbol{P} = \{\sigma^2\}^{-1}(\boldsymbol{I} - \boldsymbol{A})$ (Harville, 1977), one can show using equation (17) that $\hat{\sigma^2} = \{n - \text{tr}(\boldsymbol{A})\}^{-1}\sum_{i=1}^n\{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \hat{h}(\boldsymbol{z}_i)\}^2$. Hence $\text{tr}(\boldsymbol{A})$ represents the loss of degrees of freedom from estimating $\boldsymbol{\beta}$ and $h(\cdot)$ when estimating σ^2. The covariance of $\hat{\boldsymbol{\theta}} = (\hat{\tau}, \hat{\rho}, \hat{\sigma}^2)$ can be estimated using the information matrix of the REML likelihood $\mathcal{I}_{\theta_l\theta_{l'}} = \frac{1}{2}\text{tr}\{\boldsymbol{P}\frac{\partial\boldsymbol{V}(\boldsymbol{\theta})}{\partial\theta_l}\boldsymbol{P}\frac{\partial\boldsymbol{V}(\boldsymbol{\theta})}{\partial\theta_{l'}}\}$.

4.3 *Test for the Nonparametric Function*

Because we are interested in the effect of a whole genetic pathway rather than individual genes, it is of significant practical interest to test $\text{H}_0 : h(\boldsymbol{z}) = 0$. In the PSA microarray example, this tests for a genetic pathway effect on PSA controlling for the effects of covariates. Assuming $h(\boldsymbol{z}) \in \mathcal{H}_k$, one can easily see from the linear mixed model representation (11) that $\text{H}_0 : h(\boldsymbol{z}) = 0$ is equivalent to testing the variance component τ as $\text{H}_0 : \tau = 0$ versus $\text{H}_1 : \tau > 0$. Note the null hypothesis places τ on the boundary of the parameter space. Because the kernel matrix $\boldsymbol{K}$ is not block diagonal, unlike the standard case considered by Self and Liang (1987), the likelihood ratio for $\text{H}_0 : \tau = 0$ does not following a mixture χ_0^2 and χ_1^2. We consider a score test in this article.

Zhang and Lin (2002) proposed a score test for $\text{H}_0 : \tau = 0$ to compare a polynomial model with a smoothing spline. Unlike the smoothing spline case, a general kernel function $K(\cdot, \cdot)$ in LSKM might depend on an unknown scale parameter ρ. However, for smoothing splines, $K(\cdot, \cdot)$ does not depend on any unknown parameter. One can easily see from the linear mixed model (11) that under $\text{H}_0 : \tau = 0$, the kernel matrix $\boldsymbol{K}$ disappears, and hence the scale parameter ρ disappears and becomes inestimable.

Davies (1987) studied the problem of a parameter disappearing under H_0 and proposed a score test by treating the score statistic as a Gaussian process indexed by the nuisance parameter and then obtaining an upper bound to approximate the p-value of the score test. This approach, however, does not work for our setting due to the unboundedness of the parameter space.

We here propose to test for $\text{H}_0 : \tau = 0$ using the score test by fixing ρ and varying its value and examining sensitivity of the score test for $\text{H}_0 : \tau = 0$ with respect to ρ. The REML version of the score statistic of τ under $\text{H}_0 : \tau = 0$ can be written as $Q_\tau(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}, \rho) - \text{tr}\{\boldsymbol{P}_0\boldsymbol{K}(\rho)\}$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma^2}$ are the MLEs of $\boldsymbol{\beta}$ and σ^2 under the linear model $\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + e_i$, the model under H_0, $\boldsymbol{P}_0 = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$, and

$$Q_\tau(\boldsymbol{\beta}, \sigma^2, \rho) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{K}(\rho)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

which is a quadratic function of $\boldsymbol{y}$ and follows a mixture of chi-squares under H_0.

Following Zhang and Lin (2002), for each fixed ρ, we use the Satterthwaite method to approximate the distribution of $Q_\tau(\cdot;\ \rho)$ by a scaled chi-square distribution $\kappa\chi_\nu^2$, where the scale parameter κ and the degrees of freedom ν are calculated by equating the mean and variance of $Q_\tau(\cdot;\ \rho)$ and those of $\kappa\chi_\nu^2$. Specifically, one can show that $\kappa = \tilde{I}_{\tau\tau}/2\tilde{e}$ and $\tilde{\nu} = 2\tilde{e}^2/\tilde{I}_{\tau\tau}$, where $\tilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2}I_{\sigma^2\sigma^2}^{-1}I_{\tau\sigma^2}^T$, $I_{\tau\tau} = \text{tr}(\boldsymbol{P}_0\boldsymbol{K}(\rho))^2/2$, $I_{\tau\sigma^2} = \text{tr}(\boldsymbol{P}_0\boldsymbol{K}(\rho)\boldsymbol{P}_0)/2$, and $I_{\sigma^2\sigma^2} = \text{tr}(\boldsymbol{P}_0^2)/2$. $\tilde{e} = \text{tr}(\boldsymbol{P}_0\boldsymbol{K})/2$. Computation of the proposed score test is quite simple, because one only needs to fit the simple linear model $\boldsymbol{y}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. We evaluate the performance of the score test using simulations.

5. Model Selection within the Kernel Machine Framework

The kernel machine method requires a kernel function to be explicitly specified. Section 2.2 provides wide choices of kernel functions. A question of substantial interest is which kernel function to choose. This kernel selection problem has much broader implications. We consider two types of kernel selection problems. The first is to choose between different parametric and nonparametric models with different smoothness properties. The second problem involves variable selection.

As stated in Section 2.2, a kernel function fully specifies a function space $\mathcal{H}_K$ where the unknown function $h(\cdot)$ resides. Hence this function space determines the type of models used to fit $h(\cdot)$. For example, a dth-degree polynomial kernel specifies a parametric model with dth order monomials; the kernel $K(s,u)=\int_0^1 (s-t)_+(t-u)_+\,dt$ specifies a cubic smoothing spline model (Wahba, 1990); and the Gaussian kernel assumes an infinitely smooth function. It is therefore clear that model selection within the kernel machine framework is in fact a special case of kernel selection.

Variable selection can also be treated as a kernel selection problem within the kernel machine framework. For example, let $\boldsymbol{z}_p$ be a p-dimensional vector and $\boldsymbol{z}_{p'}$ a p' dimensional subvector of $\boldsymbol{z}_p$ with $p'<p$. Then two kinds of kernel functions can be specified: one based on $\boldsymbol{z}_p$ and another one based on $\boldsymbol{z}_{p'}$. The unknown function can then be fitted separately based on each kernel. If the fitted curves are not "far away" from each other, then the model using $\boldsymbol{z}_{p'}$ provides an equally good but more parsimonious fit than that using $\boldsymbol{z}_p$. This demonstrates that variable selection is also a special case of kernel selection.

These discussions show that model selection is a very interesting and important topic within the kernel machine framework. However, little work has been done in this area. We propose AIC and BIC as kernel selection criteria within the kernel machine framework. Equations (5) and (7) show that the estimated response $\hat{\boldsymbol{y}}$ can be expressed as $\hat{\boldsymbol{y}}=\boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}=(\boldsymbol{I}+\lambda^{-1}\boldsymbol{K})^{-1}[\lambda^{-1}\boldsymbol{K}+\boldsymbol{X}\{\boldsymbol{X}^T(\boldsymbol{I}+\lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{X}\}^{-1}\boldsymbol{X}^T(\boldsymbol{I}+\lambda^{-1}\boldsymbol{K})^{-1}]$ is the LSKM smoothing matrix. Let $r=\mathrm{trace}(\boldsymbol{A})$ be the degree-of-freedom of the kernel machine smoother $\boldsymbol{A}$. We define the least squares kernel machine (KM) AIC and BIC as

$$\mathrm{KM\_AIC}=n\log(\mathrm{RSS})+2r,$$

$$\mathrm{KM\_BIC}=n\log(\mathrm{RSS})+r\log(n),$$

where $\mathrm{RSS}=(\boldsymbol{y}-\hat{\boldsymbol{y}})^T(\boldsymbol{y}-\hat{\boldsymbol{y}})$. Models with smaller KM_AIC/KM_BIC values are preferred.

6. Application to the Prostate Cancer Genetic Pathway Data

We applied the proposed semiparametric model to the analysis of prostate cancer genetic pathway data described in Section 1. The data set contained 59 patients who were clinically diagnosed with local or advanced prostate cancer. The objective of the study was to evaluate whether a genetic pathway has an overall effect on PSA after adjusting for covariates. We focus in this article on the cell growth pathway, which contains five genes. The outcome was pre-surgery PSA level. A log transformation was performed to make the normality assumption plausible. Two covariates included age and Gleason score, a well-established histological grading system for prostate cancer.

Table 1
Parameter estimates of the semiparametric model and the score test for the genetic pathway effect for the PSA data using the LSKM via the linear mixed model representation

Covariate	Estimate	SE	p-value
Intercept	−1.7722	1.1915	0.1425
Age	0.0177	0.0114	0.1259
Gleason	0.4461	0.1055	0.0001
τ	2.8182	3.7720	·
ρ	6.3635	13.5708	·
σ^2	0.3712	0.0816	0.001
ρ	$\mathcal{S}$	ν	p-value
Score test for the genetic pathway effect $\mathrm{H}_0: h(\boldsymbol{z})=0$			
3	31.010	14.924	0.0085
5	28.750	11.223	0.0028
10	26.598	8.295	0.0010
30	23.264	5.970	0.0007

The semiparametric model (1) provides a convenient framework to evaluate the effect of the cell growth pathway on PSA by allowing for complicated interactions among the genes within the pathway. Specifically, we consider the model

$$\log(\mathrm{PSA})=\beta_0+\beta_1\mathrm{age}+\beta_2\mathrm{gleason}+h(\mathrm{gene}_1,\ldots,\mathrm{gene}_5)+e, \tag{18}$$

where $h(\cdot)$ is a nonparametric function and $e\sim N(0,\sigma^2)$. We fit this model using the LSKM method via the linear mixed model representation (11) and using the Gaussian kernel in estimating $h(\cdot)$. Under the linear mixed model representation, we estimated $(\beta_0,\ \beta_1,\ \beta_2)$ and $h(\cdot)$ using BLUPs, and estimated the smoothing parameter τ, the kernel parameter ρ and the residual variance σ^2 simultaneously using REML. The results are presented in Table 1, indicating Gleason score was highly significant, while age was not.

We tested for the cell growth pathway effect on PSA, $\mathrm{H}_0: h(\boldsymbol{z})=0$ versus $\mathrm{H}_1: h(\boldsymbol{z})\in H_K$ using the score test described in Section 4.3. Table 1 gives the score test statistics and p-values for a range of ρ values. The p-values are not sensitive to the choice of ρ and range from 0.0007 to 0.0085, suggesting a strong cell growth pathway effect on PSA.

Even though the five genes are believed to function together biologically, it is of interest to investigate whether there are a small number of relatively important genes in the cell growth pathway that most affect PSA. We investigated this problem using the proposed variable selection method. An all-possible-subset selection procedure of genes was performed using the Gaussian kernel. The kernel machine AIC and BIC proposed in Section 5 were used as the model selection criteria. The result shows that the model with the lowest AIC and BIC values is the one containing genes FGF2 and IGFBP1. The detailed results are given in Web Table 1 in the Supplementary Materials. These two genes can be studied further in laboratory settings to explore their detailed relationship with PSA.

7. Simulation Studies

7.1 *Simulation Study for the Parameter Estimates*

We conducted a simulation study to evaluate the performance of the proposed LSKM estimation method for the semiparametric model (1) by fitting the linear mixed model (11). We considered the following model

$$y_i = x_i + h(z_{i1}, \ldots, z_{ip}) + e_i, \tag{19}$$

where $e_i \sim N(0, 1)$. To allow for x_i and $(z_{i1}, \ldots, z_{ip})$ to be correlated, x_i was generated as $x_i = 3\cos(z_{i1}) + 2u_i$ with u_i being independent of z_{i1} and following $N(0, 1)$, $z_{ij} (j = 1, \ldots, p)$ were generated from Uniform(0, 1). The nonparametric function $h(\cdot)$ was allowed to have a complex form with nonlinear functions of the z's and interactions among the z's. In our simulations, we first fit the model using the same set of z's as that in the true model. In practice, without advanced knowledge, the true set of z's is often unknown and the set of z's that is used might be larger than the true set and contains some noisy z's that are irrelevant to the outcome y. To mimic such a scenario, in the second set of simulations, we added some noisy z's in the set of z's and fit (19).

We considered four configurations by varying n (the sample size) and p (the number of covariates z's). For each setting, only the Gaussian kernel is used and 300 simulations were run.

Setting 1: $n = 60$, $p = 5$, true $h(\boldsymbol{z}) = 10\cos(z_1) - 15z_2^2 + 10\exp(-z_3)z_4 - 8\sin(z_5)\cos(z_3) + 20z_1z_5$. Fit the model with the five true z's. This setting mimics the PSA data.

Setting 2: $n = 100$, $p = 8$, $h(\cdot)$ is the same as setting 1. Fit the model (19) by including 3 additional irrelevant z_6, z_7, z_8 besides the true $z_1, \ldots, z_5$.

Setting 3: $n = 200$, $p = 10$, true $h(z_1, \ldots, z_{10}) = 10\cos(z_1) - 15z_2^2 + 10\exp(-z_3)z_4 - 8\sin(z_5)\cos(z_3) + 20z_1z_5 + 9z_6\sin(z_7) - 8\cos(z_6)z_7 + 20z_8\sin(z_9)\sin(z_{10}) - 15z_8^3 - 10z_8z_9 - \exp(z_{10})\cos(z_{10})$. Fit the model assuming these 10 true z's are used.

Setting 4: $n = 300$, $p = 15$, $h(\cdot)$ is the same as that in setting 3. Fit the model with additional 5 irrelevant noisy predictors $z_{11}, \ldots, z_{15}$ besides the true $z_1, \ldots, z_{10}$.

The point estimate results are presented in Table 2. Because it is difficult to graphically display the fitted value of $h(\cdot)$ as a function of $\boldsymbol{z}$, we summarized the goodness of fit of $h(\cdot)$ in the following way. For each simulation data set, we regressed the true h on the fitted $\hat{h}$, both evaluated at the design points. We then empirically summarized the goodness of fit of $\hat{h}(\cdot)$ by reporting the average intercepts, slopes, and R^2's obtained from these regressions over the 300 simulations. If the intercept from this regression is close to zero and the slope is close to one and R^2 is close to one, it would provide empirical evidence that the estimated multi-dimensional function $h(\cdot)$ is close to the true manifold.

The results in Table 2 show that, when the true set of z's was included in fitting $h(\cdot)$ and all the model parameters $\{\boldsymbol{\beta}, h(\cdot), \tau, \rho, \sigma^2\}$ were estimated simultaneously, the LSKM method via the mixed model framework performed well in estimating $\boldsymbol{\beta}$, $h(\cdot)$ and σ^2. However, if the scale parameter ρ in the Gaussian kernel was fixed, which is often done in traditional machine learning, the model estimators could be subject to considerable bias, especially for the estimate of σ^2. When ρ was fixed at values close to the estimated one, the bias was small. Because in practice, ρ is unknown, our results suggest it is useful to estimate the scale parameter ρ using the data. When extra irrelevant covariates z's besides the true set of z's were used in fitting $h(\cdot)$, the proposed method still performed well if all model parameters were estimated.

Table 3 compares the estimated standard errors of $\hat{\beta}$ using the frequentist method (12) and the Bayesian method (14) with the empirical ones. The results show that both the frequentist and the Bayesian standard error estimates were close to their empirical counterparts. Table 3 also compares the estimated standard errors of $\hat{h}$ (including intercept) using the frequentist method (13) and the Bayesian method (15) with the empirical standard errors. For the ease of presentation, for

Table 2
Simulation results of estimated regression coefficients $\boldsymbol{\beta}$ and the nonparametric function $h(\cdot)$ in model $y = x\beta + h(\boldsymbol{z}) + e$ based on 300 runs. True $\beta = 1$ and true $\sigma^2 = 1$

				Model parameter estimates			Reg of h on $\hat{h}$		
Setting	True # z	Used # z	n	β	σ^2	ρ	Intercept	Slope	R^2
1	5	5	60	1.00	0.96	5.34[a] (estimated)	−0.04	1.00	0.99
			100	1.01	0.96	7.24 (estimated)	−0.01	1.00	0.99
			100	1.00	0.92	1.00 (fixed)	−0.01	1.00	0.99
			100	1.00	1.01	100.00 (fixed)	−0.02	1.00	0.99
2	5	8	100	1.05	0.89	6.74 (estimated)	0.16	1.00	0.98
			100	1.06	0.30	1.00 (fixed)	0.36	0.98	0.97
			100	1.12	2.15	100.00 (fixed)	0.23	1.01	0.96
3	10	10	200	0.98	0.93	12.83 (estimated)	−0.07	1.00	0.99
			200	0.92	0.30	1.00 (fixed)	−0.18	0.99	0.98
			200	0.98	1.15	100.00 (fixed)	−0.04	1.00	0.99
4	10	15	300	1.01	0.82	14.02 (estimated)	0.03	1.00	0.99
			300	1.01	0.75	10.00 (fixed)	0.02	1.00	0.99
			300	1.01	1.17	100.00 (fixed)	0.02	1.00	0.99

[a]Average of the estimated $\hat{\rho}$ from 300 simulations.

Table 3
Simulation study results of standard error estimates of $\hat{\beta}$ and $\hat{h}(\cdot)$ in model $y = x\beta + h(\boldsymbol{z}) + e$ based on 300 simulations

Setting	True # z	Used # z	n	Empirical SE	Bayesian SE	Frequentist SE	ρ
					SEs of $\hat{\beta}$		
1	5	5	60	0.088	0.088	0.083	5.34 (estimated)
			100	0.054	0.057	0.055	7.24 (estimated)
			100	0.062	0.066	0.058	1.00 (fixed)
			100	0.055	0.056	0.055	100.00 (fixed)
2	5	8	100	0.066	0.065	0.058	6.74 (estimated)
			100	0.070	0.078	0.034	1.00 (fixed)
			100	0.082	0.081	0.078	100.00 (fixed)
3	10	10	200	0.044	0.047	0.042	12.83 (estimated)
			200	0.050	0.077	0.024	1.00 (fixed)
			200	0.041	0.047	0.045	100.00 (fixed)
4	10	15	300	0.039	0.042	0.033	14.02 (estimated)
			300	0.039	0.044	0.032	10.00 (fixed)
			300	0.037	0.041	0.039	100.00 (fixed)
					SEs of $\hat{h}$		
1	5	5	60	0.635	0.662	0.601	5.34 (estimated)
			100	0.482	0.515	0.464	7.24 (estimated)
			100	0.614	0.664	0.576	1.00 (fixed)
			100	0.458	0.470	0.456	100.00 (fixed)
2	5	8	100	0.662	0.683	0.604	6.74 (estimated)
			100	0.933	0.540	0.449	1.00 (fixed)
			100	0.741	0.731	0.645	100.00 (fixed)
3	10	10	200	0.606	0.667	0.583	12.83 (estimated)
			200	0.954	0.541	0.450	1.00 (fixed)
			200	0.559	0.630	0.596	100.00 (fixed)
4	10	15	300	0.712	0.721	0.636	14.02 (estimated)
			300	0.737	0.717	0.634	10.00 (fixed)
			300	0.632	0.732	0.684	100.00 (fixed)

each setting, we averaged the SE estimates across all the grid points and presented these averages. The results show that when the scale parameter ρ was estimated, both the frequentist and the Bayesian standard error estimates were close to their empirical counterparts. When the scale parameter was fixed, the Bayesian and frequentist SEs were still close but could be quite different from the empirical SEs. These results further indicate that it is useful to estimate the scale parameter ρ in practice.

7.2 *The Simulation Study for the Score Test*

We next conducted a simulation study to evaluate the performance of the proposed variance component score test for $\mathrm{H}_0 : h(\cdot) = 0$ versus $\mathrm{H}_1 : h(\cdot) \in \mathcal{H}_k$. The true model is the same as (19), where x and z's were generated in the same way as that in Section 6.1 and $h(\boldsymbol{z}) = ah_1(\boldsymbol{z}), h_1(\boldsymbol{z}) = 2\cos(z_1) - 3z_2^2 + 2e^{-z_3}z_4 - 1.6\sin(z_5)\cos(z_3) + 4z_1z_5$ and $a = 0$, 0.2, 0.4, 0.6, 0.8, 1. We studied the size of the test by generating data under $a = 0$, and studied the power by increasing a. The kernel parameter ρ was fixed at a wide range of values: 0.5, 1, 5, 10, 25, 50, 100, 200. The sample size was 60, mimicking the PSA data example. For the size calculations, the number of simulations was 2000, whereas for the power calculations, the number of runs was 1000.

Table 4 reports the empirical size ($a = 0$) and power ($a > 0$) of the variance component score test for H_0. The results show that the size of the test was very close to the nominal value 0.05 and was not sensitive to the choice of the scale parameter ρ. As a increased, the power quickly approached 1. The power was not much affected by the value of ρ if a moderate ρ was specified, but was more affected if a large value of ρ was specified

Table 4
Simulation results for the score test for H_0: $h(\boldsymbol{z}) = 0$

Scale	Size			Power		
ρ	$\alpha = 0$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$
0.5	0.050	0.158	0.487	0.865	0.989	1.000
1	0.047	0.137	0.509	0.869	0.991	1.000
5	0.050	0.127	0.482	0.865	0.987	1.000
25	0.051	0.139	0.484	0.886	0.990	1.000
50	0.046	0.138	0.508	0.863	0.990	1.000
100	0.048	0.134	0.497	0.867	0.988	1.000
200	0.054	0.148	0.494	0.874	0.991	1.000

7.3 *The Simulation Study for Kernel Selection*

A simulation study was also conducted to assess the performance of kernel selection using the kernel machine AIC and BIC criteria. The true model we considered is

$$y = x + 10\cos(z_1) + 3z_2^2 + \exp(z_3/3)z_4 + 8\cos(z_5) + z_5 z_2 z_1 + e,$$

where $e \sim N(0, 1)$, x was generated as $x = 3\cos(z_1) + 2u$ with u being independent of z_1. All u and $z_j (j = 1, \ldots, 5)$ were generated from $N(0, 1)$. The sample size was 50, and the number of runs was 300. Three types of kernel functions were used in the simulation: the Gaussian kernel $K(\boldsymbol{u}, \boldsymbol{v}) = \exp(-\|\boldsymbol{u} - \boldsymbol{v}\|^2/\rho)$, the second-degree polynomial kernel $K(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^T\boldsymbol{v} + 1)^2$, and the first-degree polynomial kernel that corresponds to ridge regression $K(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$. For each simulated data set, the AIC and the BIC were calculated based on the model with three different kernels.

The mean AIC and BIC across 300 simulations for the Gaussian kernel are 190.79 (51.31) and 284.21 (50.21), respectively (the numbers within parenthesis are standard deviations), those for the second-degree polynomial kernel are 269.07 (10.00) and 308.91 (9.58), respectively, and those for the ridge regression are 363.67 (2.63) and 371.61 (2.51), respectively. The AIC and BIC values from each simulated data set are plotted in Figures 1 and 2. These results show that the kernel machine AIC and BIC of the model with Gaussian kernel are the smallest, whereas those of ridge regression are the largest. Hence the Gaussian kernel is preferred to both the second-degree polynomial kernel and the ridge regression kernel, which is desired in light of the complicated functional forms of the x's.

8. Discussion

In this article, we have developed the LSKM method for semiparametric regression with Gaussian outcomes, where we model the covariate effects parametrically and the genetic pathway effect parametrically or nonparametrically. The kernel machine method does not require an explicit analytical specification of the smoothness conditions on the nonparametric function and unifies the model building procedure in both one- and multiple-dimensional settings. Therefore, it is a more general and flexible method for multidimensional smoothing.

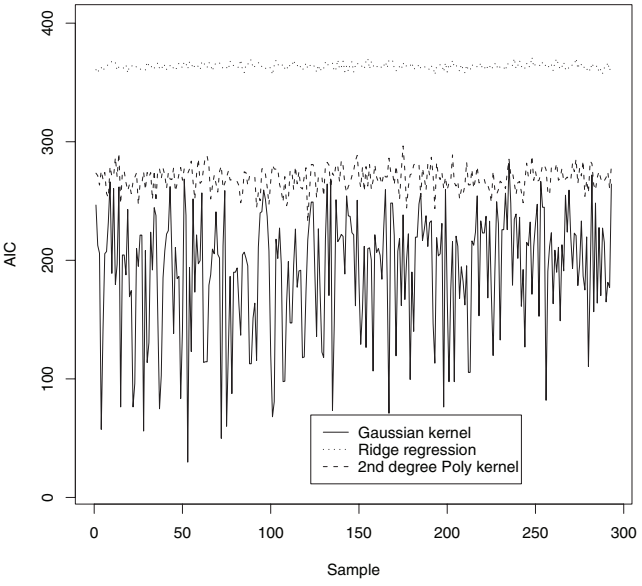

Figure 1. Simulation result of model selection using KM_AIC.

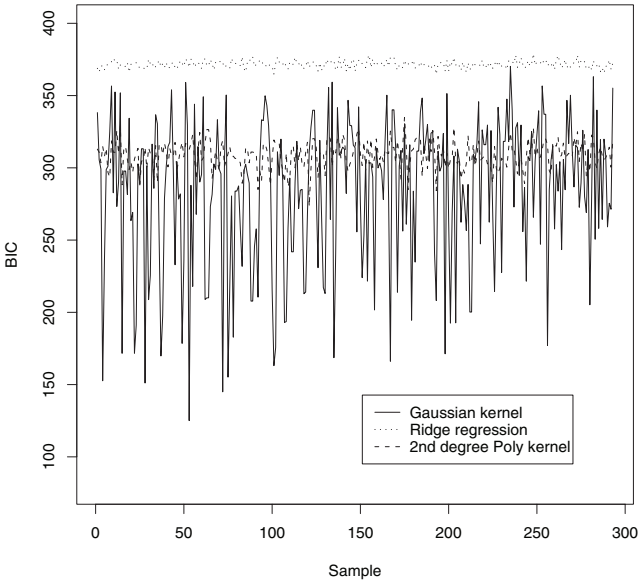

Figure 2. Simulation result of model selection using KM_BIC.

A key contribution of this article is that we have established a close connection between kernel machine methods and linear mixed models and all the model parameters can be estimated within the unified linear mixed model framework. This mixed model connection greatly facilitates the estimation and inference for multidimensional nonparametric regressions and can be easily implemented using familiar statistical software such as `SAS PROC MIXED` or `Splus NLME`.

We proposed a score test for the genetic pathway effect. This can be easily implemented using existing software. Although it requires fixing the scale parameter ρ, our results show that the test is not sensitive to the choice of ρ and has good performance. Alternatively, a Bayesian approach, such as the one proposed by Chen and Dunson (2003), might be used. This method has the advantage that there is no need to fix the scale parameter by proper prior specifications. However, its theoretical properties are unknown. It is of further research interest to study the performance of this Bayesian method and to develop better frequentist methods of testing τ in the kernel machine setting.

Kernel selection within the kernel machine framework is an important and complicated problem. It includes model selection and variable selection as special cases. In this article we propose to use kernel machine AIC/BIC as kernel selection criteria. Our simulation results show AIC/BIC performs well. Further research is still needed to examine their theoretical properties in detail before they can be adopted as a universal criteria.

We have considered in this article a single nonparametric function of multi-dimensional covariates. One could generate the proposed semiparametric model to incorporate multiple

multi-dimensional nonparametric functions. For example, if one is interested in modeling multiple genetic pathway effects, one could consider an semiparametric additive model

$$y = \boldsymbol{X}^T\boldsymbol{\beta} + h_1(\boldsymbol{z}_1) + \cdots + h_m(\boldsymbol{z}_m) + e,$$

where $\boldsymbol{z}_j (j = 1, \ldots, \mathrm{m})$ denotes a $p_j \times 1$ vector of genes in the jth pathway and $h_j(\cdot)$ denotes the nonparametric function associated with the jth genetic pathway.

Machine learning is an emerging area of research in statistics. The field has experienced a rapid development in the past decade mainly by computer scientists dealing with multidimensional data. It has shown increasing promises and wide applications in biomedical research, especially in bioinformatics. These techniques however are somewhat disconnected with well-established biostatistical methods. Our effort of establishing a close connection between LSKMs and linear mixed models is an attempt to build a bridge between kernel machines that are familiar to computer scientists but less familiar to biostatisticians. This connection opens a door for adopting other well-established statistical techniques used in mixed models, such as Bayesian approaches, to handle multidimensional data via the machine learning framework. It also opens a new research direction for model/variable selection methods within the kernel machine framework. Such an interface is still in its infancy and has a lot of room for further developments.

9. Supplementary Materials

The kernel machine AIC and BIC estimates of models containing all the subsets of genes in the cell growth pathway for the analysis of the prostate cancer data are given in Web Table 1 at the *Biometrics* website `http://www.tibs.org/biometrics`.

Acknowledgements

DL and XL's research was supported by a grant from the National Cancer Institute (CA–76404). DG's research was supported by a grant from the National Institute of Health (GM072007). We thank the associate editor and three reviewers for their helpful comments that have improved the article.

References

Buhmann, M. D. (2003). *Radial Basis Functions*. Cambridge, U.K.: Cambridge University Press.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59,** 762–769.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge University Press.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74,** 33–43.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412,** 822–826.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96,** 1151–1160.

Fortunel, N. O., Otu, H. H., Ng, H. H., Chen, J., Mu, X., Chevassut, T., Li, X., Joseph, M., et al. (2003). Comment on "'Stemness': Transcriptional Profiling of Embryonic and Adult Stem Cells" and "A Stem Cell Molecular Signature." *Science* **302,** 393.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19,** 1–141.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76,** 817–823.

Goeman, J. J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J. K., and van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21,** 1950–1957.

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

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Kimeldorf, G. S. and Wahba, G. (1970). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33,** 82–95.

Laird, N. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., et al. (2003). PGC-1alpha responsive genes involved in oxidative phosphorylation are coordinately Downregulated in human diabetes. *Nature Genetics* **34,** 267–273.

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

Ruppert, D., Wand, M. P., and Carroll, R. J. (2004). *Semiparametric Regression*. Cambridge, U.K.: Cambridge University Press.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. Cambridge, Massachusetts: MIT Press.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association* **82,** 605–610.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning* **46,** 21–52.

Speed, T. (1991). Discussion to "BLUP is a good thing: The estimation of random effects" by Robinson, G. K. *Statistical Sciences* **6,** 15–51.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., and Mesirov, J. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, J., and Vandewalle, J. (2002). *Least Squares Support Vector Machines.* Singapore: World Scientific.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1,** 211–244.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98,** 5116–5124.

Vapnik, V. (1998). *Statistical Learning Theory.* New York: Wiley.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM Press.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* **93,** 341–348.

Zhang, D. and Lin, X. (2002). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93,** 710–719.

Received September 2005. *Revised November* 2006.
Accepted January 2007.