Biometrics 59, 000–000 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2010

Kernel Machine Approach to Testing the Significance of Multiple Genetic Markers for Risk Prediction

Tianxi Cai

Department of Biostatistics, Harvard University, 655 Huntington Ave, Boston, MA 02115, USA

**email: tcai@hsph.harvard.edu*

and

Giulia Tonini

Department of Statistics, University of Florence, Italy

and

Xihong Lin

Department of Biostatistics, Harvard University, 655 Huntington Ave, Boston, MA 02115, USA

Summary: There is growing evidence that genomic and proteomic research holds great potentials for changing irrevocably the practice of medicine. The ability to identify important genomic and biological markers for risk assessment can have a great impact in public health from disease prevention, to detection, to treatment selection. However, the potentially large number of markers and the complexity in the relationship between the markers and the outcome of interest impose a grand challenge in developing accurate risk prediction models. The standard approach to identifying important markers often assesses the marginal effects of individual markers on a phenotype of interest. When multiple markers relate to the phenotype simultaneously via a complex structure, such a type of marginal analysis may not be effective. To overcome such difficulties, we employ a kernel machine Cox regression framework and propose an efficient score test to assess the overall effect of a set of markers, such as genes within a pathway or a network, on survival outcomes. The proposed test has the advantage of capturing the potentially non-linear effects without explicitly specifying a particular non-linear functional form. To approximate the null distribution of the score statistic, we propose a simple resampling procedure that can be easily implemented in practice. Numerical studies suggest that the test performs well with respect to both empirical size and power even when the number of variables in a gene set is not small compared to the sample size.

Key words: Genetic Association, Gene-set analysis, Genetic Pathways, Kernel Machine, Kernel PCA, Risk Prediction, Score Test, Survival Analysis.

1. Introduction

Genomic technologies permit systematic approaches to discovery that have begun to have a profound impact on biological research, pharmacology, and medicine. The ability to obtain quantitative information about the complete transcription profile of cells promises to be powerful means to explore basic biology, diagnose disease, facilitate drug development, and tailor therapeutics to specific pathologies (Young, 2000). The standard approach to analyzing genetic data is to identify important genes by assessing the marginal effects of individual genes on the phenotype of interest. However, when multiple genes are related to the phenotype simultaneously via a complex structure, such a single gene analysis may not be effective and may lead to a large number of false positives and false negatives especially when the signals are weak and results that are not reproducible (Vo et al., 2007). To overcome such difficulties, biological knowledge based learning methods have been advocated to integrate biological knowledge into statistical learning (Brown et al., 2000). One important approach is through the use of genetic pathways or networks. Results from pathway/network analysis are often more reliable and reproducible (Goeman et al., 2004, 2005).

To identify pathways that are associated with disease progression, one may test for the overall effect of a pathway on the risk of developing a clinical event. Goeman et al. (2005) proposed a score test under the standard proportional hazards model framework, where linear covariate effects are assumed. However, the functionality of the genes within a pathway are often complicated, and is likely to yield non-linear and non-additive effects on disease progression. We propose to incorporate such complex joint pathway effects with kernel machine learning. The kernel machine framework has been employed in various settings as a powerful machine learning tool to incorporate complex feature spaces (Vapnik, 1998; Scholkopf and Smola, 2002). For example, support vector machine has been successfully used for classifying and validating cancer subtypes (Furey et al., 2000; Ramaswamy et al., 2001; Lee and Lee, 2003). Linear and logistic kernel machine methods have been proposed to model pathway effects on continuous and binary outcomes (Liu et al., 2007, 2008). However, limited work using kernel methods has been done for survival data. Li and Luan (2003) considered a kernel Cox regression model for relating gene expression profiles to survival outcomes. However, no inference procedures were provided for the resulting estimates of the gene expression level effects.

We propose in this paper the use of the kernel machine Cox regression to derive a score test for assessing the pathway effect on survival outcomes. The proposed test has the advantage of capturing both linear and non-linear effects. For nonlinear effects, it does not require the specification of any particular parametric non-linear functional form. To approximate the null distribution of the score test, we propose a simple resampling procedure that can be easily implemented in practice. Numerical studies suggest that the test performs well with respect to both empirical size and power even when the number of variables in a gene set is not small compared to the sample size.

The rest of the paper is organized as follows. In section 2, we introduce the Cox proportional hazards kernel machine model. In section 3, we present the score test and procedures for approximating the null distribution of the proposed test. Simulation results are presented in Section 4.1 and the proposed procedures are illustrated by assessing the effects of various canonical pathways on breast cancer survival using a breast cancer gene expression study conducted by van de Vijver et al. (2002). In the example section, we also discuss simultaneous testing procedures to adjust for multiple comparisons when there are more than one pathway of interest.

2. The Cox Proportional Hazard Kernel Machine Model

Suppose we are interested in assessing the association between genetic measurements, $\mathbf{Z}_{ptimes1}$, and the survival outcome T adjusting for covariates $\mathbf{U}$. For example, $\mathbf{Z}$ could represent the gene expression levels within a pathway and $\mathbf{U}$ may represent additional covariates such as age and gender. Due to censoring, the survival time T is not always observable. Instead, we observe a bivariate vector (X, Δ), where $X = \min(T, C)$, $\Delta = I(T \leqslant C)$ and C is the censoring time. We require the standard assumption that C is independent of T conditional on $\mathbf{Z}$ and $\mathbf{U}$. The data for analysis consist of n i.i.d copies of random vectors $\{(X_i, \Delta_i, \mathbf{Z}_i, \mathbf{U}_i), i = 1, ..., n\}$. We assume that the survival time T relate to $\mathbf{Z}$ and $\mathbf{U}$ through the proportional hazards model (Cox, 1972):

$$\lambda_{\mathbf{Z},\mathbf{U}}(t) = \lambda_0(t) e^{\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{U} + h(\mathbf{Z})}$$

where $\lambda_{\mathbf{Z},\mathbf{U}}(t)$ is the conditional hazard function given $\mathbf{Z}$ and $\mathbf{U}$, $\lambda_0(t)$ is the baseline hazard function, $\boldsymbol{\gamma}$ is an unknown covariate effect for $\mathbf{U}$, and $h(\mathbf{Z})$ is an unknown centered smooth function of $\mathbf{Z}$.

Here, we are particularly interested in testing the null hypothesis $H_0 : h(\mathbf{Z}) = 0$. If $\mathbf{Z}$ represents a genetic pathway/network, the null suggests that none of the genes in the pathway/network is associated with survival time conditional on $\mathbf{U}$, i.e, no pathway/network effect. To test for such a hypothesis, one may consider a parametric or non-parametric specification for $h(\cdot)$. For example, if $h(\mathbf{z}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{z}$, the model becomes the standard Cox proportional hazard model (Cox, 1972). A score test for $h(\cdot) = 0$ under such a framework has been considered by Goeman et al. (2005). However, such a test for linear effects may have limited power when the covariate effect is non-linear.

We consider a general parametric/non-parametric setting where we allow the functional form of $h(\cdot)$ to belong to a function space $\mathcal{H}_K$, generated by a given positive definite kernel function $K(\cdot, \cdot; \rho)$, where ρ is a possibly unknown scale parameter of the kernel function. By Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), under regularity conditions, a kernel function $K(\cdot, \cdot; \rho)$ implicitly specifies a unique function space spanned by a particular set of orthogonal basis functions $\{\phi_l(\mathbf{z}; \rho))\}_{l=1}^{L}$, where $\phi_l(\mathbf{z}; \rho)) = \lambda_l(\rho)^{\frac{1}{2}}\psi_l(\mathbf{z}; \rho)$, $\{\lambda_l(\rho)\}_{l=1}^{L}$ and $\{\psi_l(\mathbf{z}; \rho))\}_{l=1}^{L}$ are the eigenvalue and eigenfunctions of $K(\cdot, \cdot; \rho)$ such that $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = \sum_{l=1}^{L} \lambda_l(\rho)\psi_l(\mathbf{z}_1; \rho)\psi_l(\mathbf{z}_2; \rho)$ and $\lambda_1(\rho) \geqslant \lambda_2(\rho) \geqslant \cdots \geqslant \lambda_L(\rho) \geqslant 0$. The functional space $\mathcal{H}_K$ is essentially $\mathcal{H}_K = \overline{\{\boldsymbol{\phi}(\mathbf{z}; \rho)\}}$, where $\overline{\{\boldsymbol{a}\}}$ denotes the function space spanned by basis functions $\boldsymbol{a}$. This gives a *primal representation* of $h \in \mathcal{H}_K$: $h(\mathbf{z}) = \sum_{l=1}^{L} \beta_l \phi_l(\mathbf{z}, \rho)$. One popular kernel function is the rth polynomial kernel $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = (\rho + \mathbf{z}_1^{\mathsf{T}}\mathbf{z}_2)^r$, where ρ is the intercept. For our present case, $h(\mathbf{z})$ is centered and thus ρ may be set to 0. The first order polynomial kernel with $r = 1$ corresponds to the linear effect with $h(\mathbf{z}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{z}$ and thus $\mathcal{H}_K = \overline{\{\boldsymbol{\phi}(\mathbf{z})\}} = \overline{\{z_1, \cdots, z_p\}}$. If $r = 2$, $\mathcal{H}_K = \overline{\{\boldsymbol{\phi}(\mathbf{z})\}} = \overline{\{z_j, z_j z_{j'}, j, j' = 1, \cdots, p\}}$, i.e., a model with main effects, quadratic effects and two-way interactions. The kernel function $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = \prod_{j=1}^{p}(1 + z_{1j} z_{2j})$ corresponds to the model with linear effects and all multi-way interactions, i.e. $\overline{\{\phi_j(\mathbf{z})\}} = \overline{\{z_j, \prod_{l=1}^{2} z_{j_l}, \cdots, \prod_{l=1}^{p} z_{j_l}, j_l = 1, \cdots, p, j_l \neq j_{l'} \text{ if } l \neq l'\}}$. Another popular kernel is the Gaussian Kernel $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$, where $\|\mathbf{z}\|^2 = \mathbf{z}^{\mathsf{T}}\mathbf{z}$ and ρ is an unknown parameter. The Gaussian kernel generates the function space spanned by the radial basis functions. See Buhmann (2003) for the mathematical properties associated with this kernel function. Other kernel functions include the sigmoid, spline, Fourier and B-Spline kernels (Vapnik, 1998; Burges, 1998; Scholkopf and Smola, 2002). The kernel function can be viewed as a measure of similarity of gene profiles within the same pathway between two subjects. A choice of the kernels specifies a particular parametric/nonparametric model one is interested in fitting.

The explicit forms of the basis functions corresponding to K are generally difficult if not impossible to specify especially when p is not small and the function h is complex. Thus, it is generally difficult to estimate h based on its *primal representation*. On the other hand, estimating $h(\cdot) \in \mathcal{H}_K$ can be achieved by obtaining a *dual representation* of h with a given dataset. By the representer theorem (Kimeldorf and Wahba, 1970), any regularized estimator of $h(\cdot) \in \mathcal{H}_K$ with $\|h\|_{\mathcal{H}_K}^2$ being the penalty function can be represented as $h(\mathbf{Z}_i) = \boldsymbol{\alpha}^{\mathsf{T}}\mathbb{K}(\rho) = \sum_{j=1}^{n} \alpha_j K(\mathbf{Z}_i, \mathbf{Z}_j; \rho)$, where $\boldsymbol{\alpha}$ is the unknown regression parameter and $\mathbb{K}(\rho)$ is the $n \times n$ matrix with $(i, j)^{th}$ element being $K_{ij}(\rho) = K(\mathbf{Z}_i, \mathbf{Z}_j; \rho)$. The dual representation of h provides us a convenient way of assessing the effects of $\mathbf{z}$ on the outcome of interest without necessitating the specification of the basis functions.

Under the dual representation, the estimation of the unknown parameters $\{\boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ could be facilitated by maximizing the penalized partial likelihood (PPL) function

$$\sum_{i=1}^{n} \int \log \left[\frac{\exp\{\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{K}_i(\rho) + \boldsymbol{\gamma}^{\mathsf{T}}\mathbf{U}_i\}}{\sum_{l=1}^{n} Y_l(s) \exp\{\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{K}_l(\rho) + \boldsymbol{\gamma}^{\mathsf{T}}\mathbf{U}_l\}} \right] dN_i(s) - \frac{\mathfrak{c}}{2}\boldsymbol{\alpha}^{\mathsf{T}}\mathbb{K}(\rho)\boldsymbol{\alpha}. \qquad (1)$$

where $Y_l(t) = I(X_l \geqslant t)$, $\mathbf{K}_i(\rho) = [K_{i1}(\rho), ..., K_{in}(\rho)]^\intercal$, $N_i(t) = I(X_i \leqslant t)\Delta_i$ and $\mathfrak{c}$ is a penalty parameter. The PPL function (1) is closely related to the random effect Cox model:

$$\lambda_{\mathbf{Z}_i,\mathbf{U}_i}(t) = \lambda_0(t)\exp\{\boldsymbol{\alpha}^\intercal \mathbf{K}_i(\rho) + \boldsymbol{\gamma}^\intercal \mathbf{U}_i\} \quad \text{with} \quad \boldsymbol{\alpha} = \tau\boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = 0, \quad \text{var}(\boldsymbol{\epsilon}) = \mathbb{K}(\rho)^- \qquad (2)$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^\intercal$ are the unobserved random effects, τ is the unknown variance component, and $\mathbb{K}(\rho)^-$ is Moore-Penrose generalized inverse of $\mathbb{K}(\rho)$. Specifically, if $\boldsymbol{\epsilon}$ is further assumed to follow $N\{0, \tau\mathbb{K}(\rho)^-\}$, equation (1) corresponds to fitting (2) using the PQL approximation (Breslow and Clayton, 1993) and the penalty parameter $\mathfrak{c}$ corresponds to the reciprocal of τ^2. Similar connections between the kernel machine penalized regression and the mixed effects models have been discussed in (Liu et al., 2007, 2008) for analyzing non-censored data.

3. Score Test for the Parametric/Non-parametric Function

The above connection between the PPL and the mixed effects model motivates us to employ the mixed effects model given in (2) and test the null hypothesis of no pathway effect by testing

$$H_0: \quad \tau = 0$$

since for a centered h, $h(\cdot) = 0$ if and only if $\text{var}\{h(\mathbf{z})\} = 0$ for all $\mathbf{z}$. Under the mixed effects framework, testing the variance being 0 for all $\mathbf{z}$ is equivalent to testing $\tau = 0$.

3.1 *Score Statistic*

Since the null value of $\tau = 0$ is on the boundary of the parameter space and the kernel matrix $\mathbb{K}(\rho)$ is not block diagonal, the distribution of the likelihood ratio statistic for $H_0 : \tau = 0$ is non-standard. Here, we propose a score test based on the working model (2) which can be derived along the lines of Commenges and Andersen (1995). Specifically, let the partial likelihood function conditional on $\boldsymbol{\epsilon}$ be

$$\mathcal{L}_{\boldsymbol{\epsilon}}(\tau; \rho, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \int \log\left[\frac{\exp\{\tau\mathbf{K}_i(\rho)^\intercal\boldsymbol{\epsilon} + \boldsymbol{\gamma}^\intercal\mathbf{U}_i\}}{\sum_{l=1}^{n} Y_l(s)\exp\{\tau\mathbf{K}_l(\rho)^\intercal\boldsymbol{\epsilon} + \boldsymbol{\gamma}^\intercal\mathbf{U}_l\}}\right] dN_i(s).$$

If $\boldsymbol{\gamma}$ is known, then the score statistic is $\widehat{Q}(\rho, \boldsymbol{\gamma}) = E[\{\partial \mathcal{L}_{\boldsymbol{\epsilon}}(\tau; \rho, \boldsymbol{\gamma})/\partial\tau\}^2 \mid \mathcal{O}] + E\{\partial^2 \mathcal{L}_{\boldsymbol{\epsilon}}(\tau; \rho, \boldsymbol{\gamma})/\partial\tau^2 \mid \mathcal{O}\}$, where $\mathcal{O}$ denotes the observed data. When $\boldsymbol{\gamma}$ is unknown as in most practical settings, we obtain the score statistic as $\widehat{Q}(\rho) = \widehat{Q}(\rho, \widehat{\boldsymbol{\gamma}})$, where $\widehat{\boldsymbol{\gamma}}$ is the maximum partial likelihood estimator of $\boldsymbol{\gamma}$ under $H_0 : h(\mathbf{z}) = 0$. It is straightforward to see that the resulting score statistic takes the form

$$\widehat{Q}(\rho) = \widehat{\mathbf{M}}(\infty)^{\mathsf{T}} \mathbb{K}(\rho) \widehat{\mathbf{M}}(\infty) - n\widehat{q}(\rho).$$

where $\widehat{\mathbf{M}}(s) = \{\widehat{M}_1(s), \cdots, \widehat{M}_n(s)\}^{\mathsf{T}}$, $\widehat{M}_i(s) = N_i(s) - \int_0^s Y_i(u) e^{\widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{U}_i} d\widehat{\Lambda}_0(u)$, $\widehat{\mathcal{S}}^{(0)}(s) = \sum_{j=1}^n Y_j(s) e^{\widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{U}_j}$, $\widehat{\mathcal{S}}^{(k)}(s; \rho) = \sum_{j=1}^n Y_j(s) e^{\widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{U}_j} \{\mathbf{K}_j(\rho)^{\mathsf{T}} \boldsymbol{\epsilon}\}^k$, $\widehat{\Lambda}_0(s) = \sum_{i=1}^n \int_0^s dN_i(u)/\widehat{\mathcal{S}}^{(0)}(u)$, and

$$\widehat{q}(\rho) = n^{-1} \sum_{i=1}^n \int K_{ii}(\rho) Y_i(s) e^{\widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{U}_i} d\widehat{\Lambda}_0(s) - n^{-1} \sum_{i=1}^n \sum_{j=1}^n \int \frac{Y_i(s) Y_j(s) e^{\widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{U}_i + \widehat{\boldsymbol{\gamma}}^{\mathsf{T}} \mathbf{U}_j} K_{ij}(\rho)}{\widehat{\mathcal{S}}^{(0)}(s)} d\widehat{\Lambda}_0(s).$$

In the Appendix, we show that $n^{-1}\widehat{Q}(\rho) = \widehat{\mathcal{W}}(\rho) - q(\rho) + o_p(1)$ which centers around 0, where $q(\rho) = \int [E\{K_{ii}(\rho) \widetilde{Y}_i(t)\} + \omega_{00}(t, t, \rho)/E\{\widetilde{Y}_j(t)\}] d\Lambda_0(t)$ which is the limiting mean of $\widehat{\mathcal{W}}(\rho)$,

$$\begin{aligned} \widehat{\mathcal{W}}(\rho) = & \int\int K(\mathbf{z}_1, \mathbf{z}_2; \rho) d\widehat{\mathbb{W}}_M(\mathbf{z}_1) d\widehat{\mathbb{W}}_M(\mathbf{z}_2) - 2 \int \left\{ \int \omega_0(\mathbf{z}, t, \rho) d\widehat{\mathbb{W}}_\Lambda(t) + \widehat{\mathbf{W}}_\gamma^{\mathsf{T}} \boldsymbol{\omega}_1(\mathbf{z}, \rho) \right\} d\widehat{\mathbb{W}}_M(\mathbf{z}) \\ & + \int\int \omega_{00}(t, s, \rho) d\widehat{\mathbb{W}}_\Lambda(t) d\widehat{\mathbb{W}}_\Lambda(s) + \widehat{\mathbf{W}}_\gamma^{\mathsf{T}} \boldsymbol{\omega}_{11}(\rho) \widehat{\mathbf{W}}_\gamma + 2 \int \widehat{\mathbf{W}}_\gamma^{\mathsf{T}} \boldsymbol{\omega}_{10}(s, \rho) d\widehat{\mathbb{W}}_\Lambda(s) \end{aligned} \tag{3}$$

$\widehat{\mathbb{W}}_M(\mathbf{z}) = n^{-\frac{1}{2}} \sum_{i=1}^n M_i I(\mathbf{Z}_i \leqslant \mathbf{z})$, $M_i = M_i(\infty)$, $M_i(t) = N_i(t) - \int_0^t \widetilde{Y}_j(s) d\Lambda_0(s)$, $\widetilde{Y}_j(s) = Y_j(s) e^{\gamma_0^{\mathsf{T}} \mathbf{U}_j}$, $\omega_0(\mathbf{z}, t, \rho) = E\{K(\mathbf{Z}_i, \mathbf{z}; \rho) \widetilde{Y}_i(t)\}$, $\boldsymbol{\omega}_1(\mathbf{z}, \rho) = E\{K(\mathbf{Z}_i, \mathbf{z}; \rho) \widetilde{\mathbf{U}}_i\}$, $\widetilde{\mathbf{U}}_i = \mathbf{U}_i \int \widetilde{Y}_i(t) d\Lambda_0(t)$, $\omega_{00}(t, s, \rho) = E[K_{ij}(\rho) \widetilde{Y}_i(t) \widetilde{Y}_j(s)]$, $\boldsymbol{\omega}_{11}(\rho) = E[K_{ij}(\rho) \widetilde{\mathbf{U}}_i \widetilde{\mathbf{U}}_j^{\mathsf{T}}]$, and $\boldsymbol{\omega}_{10}(s, \rho) = E\{K_{ij}(\rho) \widetilde{\mathbf{U}}_i \widetilde{Y}_j(s)\}$.

For the linear kernel with $K(\mathbf{Z}_i, \mathbf{Z}_j; \rho) = \mathbf{Z}_i^{\mathsf{T}} \mathbf{Z}_j$, $\widehat{Q}(\rho)$ is equivalent to the test statistic proposed in equation (4) of Goeman et al. (2005). In general, if the orthogonal basis functions $\{\phi_l(\cdot; \rho)\}_{l=1}^L$ for $K(\cdot, \cdot, \rho)$ with a given ρ are known and L is finite, then one may directly derive $\widehat{Q}(\rho)$ based on the primal representation $h(\mathbf{z}) = \sum_{l=1}^L \beta_j \phi_j(\mathbf{z}; \rho)$. Specifically, assuming $\text{var}(\beta_l) = \tau^2$ and $\text{cov}(\beta_l, \beta_{l'}) = 0$, one may obtain the same score statistic $\widehat{Q}(\rho)$ based on arguments given in Goeman et al. (2005, 2006). Furthermore, based on Lemma 4 of Goeman et al. (2006), one may argue that the score test based on $\widehat{Q}(\rho)$ is locally most powerful. However, deriving tests directly from basis functions may not be feasible as $\{\phi_l(\cdot;\rho)\}_{l=1}^{L}$ often cannot be specified explicitly and the parameter ρ is often unknown and not estimable under the null. Furthermore, Goeman et al. (2006) indicated that it is unclear how to approximate the null distribution of the score test statistic with a large L. Note that for many non-linear kernels such as the Gaussian kernel, L could be infinity even if p is fixed. The use of the kernel machine framework avoids the explicit specification of basis functions and allows us to derive the asymptotic null distribution of our proposed test as a process in ρ for a general kernel function.

3.2 *Approximating the Score Test Statistic and Its Null Distribution*

In the Appendix, we show that $n^{-1}\widehat{Q}(\rho) + q(\rho)$ converges weakly to the process $\mathcal{W}(\rho)$ defined in (A.3). For a fixed ρ, one may use the Satterthwaite method to approximate the distribution of $\mathcal{W}(\rho)$ using a rescaled χ^2 distribution, $c_0\chi^2_{d_0}$. The scale parameter c_0 and the degrees of freedom (DF) d_0 can be estimated by matching the first two moments of $\mathcal{W}(\rho)$. The scaled χ^2 distribution has shown to provide a good approximation to the distribution of the score statistic for continuous and binary outcomes (Liu et al., 2007, 2008).

The effective DF d_0 decreases as we increase the correlation between covariates. Furthermore, d_0 increases slowly with p when the correlation is moderate/high, but more rapidly with p when the correlation is low. Thus with the same $h(\mathbf{z})$, a higher correlation among $\mathbf{Z}$ is likely to yield a higher power of the score test and a more gradual power loss over increasing p. This suggests that the kernel machine score test improves the power for testing for the pathway effect by effectively borrowing information across different genes and accounting for between-gene correlation in calculating a data-adaptive DF. By doing so, it often results in a low DF test and increase the test power. Specifically, consider the case $h(\mathbf{z}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{z}$. The null hypothesis for no pathway effect, $H_0 : h(\mathbf{z}) = 0$ is equivalent to $H_0 : \boldsymbol{\beta} = 0$. The traditional testing approach based on a p-DF test would suffer from power loss if p is large. The kernel machine score test effectively accounts for the correlation among the genes and is often based on a much smaller DF d_0. For example, consider the extreme case if there are say 20 genes within a pathway and the genes are highly correlated, say correlation >0.95, the traditional 20-DF test for $H_0 : \boldsymbol{\beta} = 0$ will have little power. One can show that d_0 is close to 1 by effectively accounting for the correlation between genes, and is hence much more powerful.

For the linear kernel $K(\mathbf{z}_1, \mathbf{z}_2, \rho) = \rho + \mathbf{z}_1^{\mathsf{T}}\mathbf{z}_2$, the score test can be obtained based on $n^{-1}\widehat{Q}(0)$ with $\rho = 0$ since ρ corresponds to the intercept which would be absorbed by the unknown baseline hazard function. For the Gaussian kernel, the score test $n^{-1}\widehat{Q}(\rho)$ may depend on ρ. It is not difficult to see that under H_0, the kernel matrix $\mathbb{K}(\rho)$ disappears and hence the scale parameter ρ is 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 ρ. Here, we propose to take a similar approach by considering the score test statistic

$$\widehat{S} = \sup_{\rho \in \mathcal{I}} \{n^{-1}\widehat{Q}(\rho)/\widehat{\sigma}(\rho)\},$$

where $\mathcal{I}$ is the range of ρ to be considered and $\widehat{\sigma}^2(\rho)$ is a consistent estimator of $\sigma^2(\rho) = \text{var}\{\mathcal{W}(\rho)\}$. To approximate the limiting distributions of $\widehat{Q}(\rho)$ and $\widehat{S}$ in finite samples, we consider a resampling procedure which has been successfully used in the literature (Parzen et al., 1994; Cai et al., 2000; Park and Wei, 2003). Specifically, let $\mathcal{V} = (\mathcal{V}_1, \cdots, \mathcal{V}_n)^{\mathsf{T}}$ be a vector of n i.i.d standard normal random variables generated independent of the data. For each set of $\mathcal{V}$, we obtain

$$\begin{aligned}\widehat{\mathcal{W}}^*(\rho) = &\int\int K(\mathbf{z}_1, \mathbf{z}_2, \rho) d\widehat{\mathbb{W}}_M^*(\mathbf{z}_1) d\widehat{\mathbb{W}}_M^*(\mathbf{z}_2) - 2\int\left\{\int \widehat{\omega}_0(\mathbf{z}, t, \rho) d\widehat{\mathbb{W}}_\Lambda^*(t) + \widehat{\mathbf{W}}_\gamma^{*\mathsf{T}}\widehat{\omega}_1(\mathbf{z}, \rho)\right\} d\widehat{\mathbb{W}}_M^*(\mathbf{z}) \\ &+ \int\int \widehat{\omega}_{00}(t, s, \rho) d\widehat{\mathbb{W}}_\Lambda^*(t) d\widehat{\mathbb{W}}_\Lambda^*(s) + \widehat{\mathbf{W}}_\gamma^{*\mathsf{T}}\widehat{\omega}_{11}(\rho)\widehat{\mathbf{W}}_\gamma^* + 2\int \widehat{\mathbf{W}}_\gamma^{*\mathsf{T}}\widehat{\omega}_{10}(s, \rho) d\widehat{\mathbb{W}}_\Lambda^*(s)\end{aligned}$$

where $\widehat{\omega}_0(\mathbf{z}, t, \rho)$, $\widehat{\boldsymbol{\omega}}_1(\mathbf{z}, \rho)$, $\widehat{\omega}_{00}(t, s, \rho)$, $\widehat{\boldsymbol{\omega}}_{10}(s, \rho)$ and $\widehat{\boldsymbol{\omega}}_{11}(\rho)$ are the empirical counterparts of $\omega_0(\mathbf{z}, t, \rho)$, $\boldsymbol{\omega}_1(\mathbf{z}, \rho)$, $\omega_{00}(t, s, \rho)$, $\boldsymbol{\omega}_{10}(s, \rho)$ and $\boldsymbol{\omega}_{11}(\rho)$, $\widehat{\mathbb{W}}_M^*(\mathbf{z}) = n^{-\frac{1}{2}} \sum_{i=1}^n \widehat{M}_i I(\mathbf{Z}_i \leqslant \mathbf{z})\mathcal{V}_i$, $\widehat{\mathbb{W}}_\Lambda^*(t) = n^{-\frac{1}{2}} \sum_{i=1}^n \widehat{W}_{\Lambda i}\mathcal{V}_i$, $\widehat{\mathbf{W}}_\gamma^* = n^{-\frac{1}{2}} \sum_{i=1}^n \widehat{\mathbf{W}}_{\gamma i}\mathcal{V}_i$, $\widehat{W}_{\Lambda i}$ and $\widehat{\mathbf{W}}_{\gamma i}$ are the empirical counterparts of $W_{\Lambda i}$ and $\mathbf{W}_{\gamma i}$ which are defined in (A.1). It follows from similar arguments as given in Cai et al. (2000) that $\widehat{\mathcal{W}}^*(\rho)$ conditional on the data converges weakly to $\mathcal{W}(\rho)$. To calculate the p-value for testing H_0, one may generate a large number, say B, realizations of $\mathcal{V}$ and obtain realizations of $\widehat{\mathcal{W}}^*(\rho)$, $\vec{\mathcal{W}}^*(\rho) = \{\widehat{\mathcal{W}}^*_{(b)}(\rho), b = 1, \cdots, B\}$. For any fixed ρ, the p-value of the test can be obtained as $\sum_{b=1}^{B} I\{\widehat{\mathcal{Q}}^*_{(b)}(\rho) > n^{-1}\widehat{Q}(\rho)\}/B$, where $\widehat{\mathcal{Q}}^*_{(b)}(\rho) = \widehat{\mathcal{W}}^*_{(b)}(\rho) - \bar{q}^*(\rho)$ and $\bar{q}^*(\rho)$ is the empirical mean of $\vec{\mathcal{W}}^*(\rho)$. Alternatively, one may estimate the first two moments of $\mathcal{W}(\rho)$ based on $\vec{\mathcal{W}}^*(\rho)$ and then calculate the p-value based on the χ^2 approximation. When ρ is unknown in practice, one may test H_0 based on the sup-statistic. The null distribution of $\widehat{S}$ can be approximated by the empirical distribution of $\{\widehat{S}^*_{(b)} = \sup_{\rho\in\mathcal{I}}\{\widehat{\mathcal{Q}}^*_{(b)}(\rho)/\widehat{\sigma}(\rho)\}, b = 1, ..., B\}$ given the data, where $\widehat{\sigma}^2(\rho)$ is the empirical variance of $\vec{\mathcal{W}}^*(\rho)$. To assess the significance for testing H_0, one may the p-value based on $\widehat{S}$ as $\sum_{b=1}^{B} I(\widehat{S}^*_{(b)} > \widehat{S})/B$.

3.3 *Kernel PCA Approximation*

An appropriate selection of the kernel function could potentially lead to improved power in detecting the signal. However, especially in finite sample, the complexity of the feature space corresponding to the kernel function may lead to a larger number of nuisance parameters and thus result in a loss in power. On the other hand, there often exists some correlations among the covariates and thus dimensionality reduction or so-called *feature extraction* may allow us to restrict the entire feature space to a sub-space of lower dimensionality. One approach to achieving such a goal is through kernel principal component analysis (PCA) (Scholkopf et al., 1998). We propose to investigate the effect of dimension reduction on the power of the score test. To this end, we take a singular value decomposition of the matrix $\mathbb{K}(\rho)$: $\mathbb{K}(\rho) = \sum_{\ell=1}^{n} \nu_\ell(\rho)\mathbf{e}_\ell(\rho)\mathbf{e}_\ell(\rho)^\intercal$, where $\nu_1(\rho) \geqslant \nu_2(\rho) \geqslant \cdots \geqslant \nu_n(\rho) \geqslant 0$ are the eigenvalues and $\{\mathbf{e}_\ell(\rho), \ell = 1, ..., n\}$ are the corresponding eigenvectors. We assume that with a properly chosen ρ, the eigenvalues decay quickly and thus there exists an ℓ_0 such that $\sum_{l=1}^{\ell_0} \nu_l(\rho)/\sum_{l=1}^{n} \nu_l(\rho) \geqslant \mathfrak{p}_0$, where $\mathfrak{p}_0 \in (0, 1)$ is a pre-specified constant. Consider the kernel PCA approximation to the original kernel matrix $\mathbb{K}(\rho)$, $\widetilde{\mathbb{K}}(\rho) = \sum_{\ell=1}^{\ell_0} \nu_\ell(\rho)\mathbf{e}_\ell(\rho)\mathbf{e}_\ell(\rho)^\intercal$. Let $\widetilde{Q}(\rho)$ and $\widetilde{\mathcal{W}}(\rho)$ be $\widehat{Q}(\rho)$ and $\widehat{\mathcal{W}}(\rho)$ with $\mathbb{K}(\rho)$ replaced by $\widetilde{\mathbb{K}}(\rho)$, respectively. It is not difficult to show that $n^{-1}\widetilde{\mathcal{W}}(\rho) = \sum_{\ell=1}^{\ell_0} \widetilde{W}_\ell(\rho)^2$, where $\widetilde{W}_\ell(\rho) =$

$\nu_\ell(\rho)^{\frac{1}{2}} n^{-\frac{1}{2}} \sum_{i=1}^{n} [e_{\ell i}(\rho) M_i - \int \omega_{0\ell}(t,\rho) dW_{\Lambda i}(t) - \boldsymbol{\omega}_{1\ell}(\rho)^\intercal \mathbf{W}_{\gamma i}$, $e_{\ell i}(\rho)$ is the ith element of $\mathbf{e}_\ell(\rho)$, $\omega_{0\ell}(t,\rho)$ and $\boldsymbol{\omega}_{1\ell}(\rho)$ are the respective limit of $n^{-1}\sum_{i=1}^{n} e_{\ell i}(\rho)\widetilde{Y}_i(t)$ and $n^{-1}\sum_{i=1}^{n} e_{\ell i}(\rho)\widetilde{\mathbf{U}}_i$. Based on the convergence properties of the eigenvectors to the eigenfunctions (Bengio et al., 2004; Zwald and Blanchard, 2006), one may view $\widetilde{\mathbb{K}}(\rho)$ as an empirical kernel matrix corresponding to $K^{[\ell_0]}(\rho)$, the ℓ_0-degenerated approximation of K (Braun, 2005). Thus, treating $\widetilde{K}(\rho)$ as a new kernel matrix, the same arguments as given in the Appendix can be used to establish the limiting distribution of $\widetilde{\mathcal{W}}(\rho)$ and justify the resampling method. For a given range of $\mathcal{I}$ of ρ, the final score test based on $\widetilde{\mathbb{K}}$ could be obtained as $\widetilde{S} = \sup_{\rho \in \mathcal{I}}\{n^{-1}\widetilde{Q}(\rho)/\widetilde{\sigma}(\rho)\}$, where $\widetilde{\sigma}(\rho)^2$ is the estimated variance of $\widetilde{\mathcal{W}}(\rho)$.

The kernel PCA approximation is equivalent to approximating $h(\mathbf{z}) = \sum_{l=1}^{L} \beta_l \phi_l(\mathbf{z},\rho)$ by $\widetilde{h}(\mathbf{z}) = \sum_{l=1}^{\ell_0} \beta_l \phi_l(\mathbf{z},\rho)$. Based on truncated set of basis functions, one may approximate the leading term of the score statistic by $\widetilde{\mathcal{W}}^\dagger(\rho) = \sum_{l=1}^{\ell_0} \lambda_l(\rho)\{\sum_{i=1}^{n} \psi_l(\mathbf{Z}_i,\rho)\widehat{M}_i\}^2$. Based on Satterthwaite method, one may approximate the distribution of $\widetilde{\mathcal{W}}^\dagger(\rho)$ as $c_0(\ell_0)\chi^2_{d_0(\ell_0)}$ under H_0 and as $c_0(\ell_0)\chi^2_{d_0(\ell_0)}\{\delta(\ell_0)\}$ under the alternative. Both the DF $d_0(\ell_0)$ and the non-centrality parameter $\delta(\ell_0)$ are dominated by the larger eigenvalues. Thus, we expect that the majority of the information about h to be captured by $\widetilde{\mathbb{K}}$. Similar findings based on eigenvalue decomposition have been discussed in Goeman et al. (2006) for the linear model.

3.4 *Practical Issues with the Choice of ρ*

Although our simulation results show that the score test is not sensitive to the choice of ρ within a reasonable range $\mathcal{I}$. However, when $\rho \to 0$ or $\rho \to \infty$, the kernel matrix may become degenerated and thus it may become infeasible to draw inference. For example, when K is the Gaussian kernel, $\rho \to 0$ corresponds to no similarity between subjects and $\rho \to \infty$ corresponds to no heterogeneity between subjects. For both extreme cases, one may be unable to draw conclusion about the association between T and $\mathbf{Z}$ based on the given kernel. It is important to note that the eigenvalues of the kernel matrix does not decay when $\rho \to 0$ and would be all 0 except for the first one when $\rho \to \infty$.

For the Gaussian kernel, to prevent $\mathbb{K}(\rho)$ from being degenerated while including a sufficiently wide range of ρ, one may as determine the range of ρ based on the kernel PCA. Specifically, let $\widehat{\ell}_\rho$ denote the smallest ℓ such that $\sum_{l=1}^{\ell} \nu_l(\rho)/\sum_{l=1}^{n} \nu_l(\rho) \geqslant \mathfrak{p}_0$. Then the range of ρ can be chosen as $\mathcal{I} = [\min\{\rho : \widehat{\ell}_\rho \leqslant \ell_0\}, \max\{\rho : \widehat{\ell}_\rho \geqslant 2\}]$. In practice, we recommend $\mathcal{I}$ with $\ell_0 = \sqrt{n}$. The constants $\{2, \sqrt{n}\}$ are chosen to ensure that the rank of $\mathbb{K}(\rho)$ is not close to 1 and the eigenvalue of $\mathbb{K}(\rho)$ has a reasonable decay rate. See more discussions on the choice of ℓ_0 in the discussion section.

4. Simulation Studies

We conducted simulation studies to assess the performance of the proposed score test. Throughout, we generated $\mathbf{Z}$ from a multivariate normal with mean 0, unit variance and correlation $\wp$. We considered $\wp = 0.5$, 0.2, and 0 to represent a moderate, weak, and zero correlation among the $\mathbf{Z}$'s, respectively. For simplicity, no additional covariates were considered for numerical studies. The censoring was generated from an exponential with mean μ_C. For each configuration, we generated 2000 datasets to calculate the empirical size and 1000 datasets to calculate the empirical power. For each simulated dataset, we carried out the score test for two types of kernels: (1) Gaussian kernel with $K_G(\mathbf{z}_1, \mathbf{z}_2; \rho) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$; and (2) Linear kernel with $K_L(\mathbf{z}_1, \mathbf{z}_2; \rho) = \mathbf{z}_1^\intercal \mathbf{z}_2$. For comparison, we evaluated the performance of the score test using $\mathbb{K}$ and the approximated kernel matrix $\widetilde{\mathbb{K}}$ with $\mathfrak{p} = 0.90$. We considered testing based on (a) sup statistic, $\widehat{S}$ and $\widetilde{S}$, with K_G; (b) $\widehat{Q}(\rho_2)$ and $\widetilde{Q}(\rho_2)$ with K_G and ρ_2 being the upper bound of $\mathcal{I}$; and (c) $\widehat{Q}$ and $\widetilde{Q}$ with K_L. For the resampling procedure, we let $B = 1000$. We considered $n = 100$ and 200; censoring proportion of 25% ($\mu_C = 3$) and 50% ($\mu_C = 1$), and $p = 5$, 10, and 100.

First, to examine the validity of our proposed testing procedure in finite samples, we generated data under the null model to assess the size of the score test. Specifically, we generated $\log T$ from an extreme value distribution and thus the survival time is independent of the covariates. The empirical sizes of the proposed tests are summarized in Table 1 for $\wp = 0.5$ and Table 2 for $\wp = 0.2$.

We evaluated the performance of testing procedures based on various approximations to the null distribution: (i) the resampling procedure; (ii) the scaled χ^2 approximation; and (iii) the normal approximation as suggested in the literature. Across all the configurations, the empirical sizes for $\widehat{Q}(\rho_2)$, $\widetilde{Q}(\rho_2)$, $\widehat{Q}$ and $\widetilde{Q}$, are close to the nominal levels when the null is approximated based on the resampling and the scaled χ^2. The PCA based test appears to perform slightly better when p is large. This is in part due to the reduction in the DF. On the contrary, the normal approximations do not appear to work well in many of the settings. For example, when $n = 100$, $p = 5$, $\wp = 0.5$ with 50% of censoring, the empirical sizes for $\{\widehat{Q}(\rho_2), \widetilde{Q}(\rho_2)\}$ were $(7.4\%, 8.4\%)$ based on the normal approximation, $(5.8\%, 5.5\%)$ based on the resampling procedure and $(5.5\%, 6.2\%)$ based on the χ^2 approximation. Similar patterns were observed for $\wp = 0.2$ and larger p's.

[Table 1 about here.]

[Table 2 about here.]

Results reported in Tables 1 and 2 suggest that one may use the χ^2 approximation to assess the DF for the score test. For example, when the correlation is weak with $\wp = 0.2$, the effective DF of $\widehat{Q}$ are about 3.9, 6.1, and 11.4 for $p = 5$, 10, and 100, respectively. When $\wp = 0.5$, the DFs are 2.5, 2.9 and 3.7, respectively. Thus, as we increase the correlation among $\mathbf{Z}$, the effective DFs of the score test are much lower and do not increase much with the covariate dimension. This suggests that the kernel machine test effectively borrows information across genes within a pathway by accounting for their correlation to construct for a low-DF test and increase the power of the test. This shows an attractive feature of the kernel machine test as a powerful approach to detecting the pathway effect in high-dimensional data problems.

The tests based on $\widehat{S}$ and $\widetilde{S}$ also perform well in maintaining the size. As expected, the performance improves as we increase n and/or decrease the censoring proportion. When $p = 100$ and $\wp = 0.2$, the empirical sizes based on $\widehat{S}$ tends to be slightly lower than the nominal level. For such settings, the effective DF is large relative to n, and thus the standard large sample distribution theory may not hold for $\widehat{S}$ and $\widehat{Q}$ with a general K. On the other hand, the use of kernel PCA seems to reduce the effective DF and thus lead to $\widetilde{S}$ with reasonable empirical sizes even with $p = 100$.

To assess the power of the proposed tests, we generated data from a proportional hazards model $\log T = h(\mathbf{Z}) + \varepsilon$, where ε follows an extreme value distribution. Two forms of $h(\cdot)$ was considered: (1) $h(\mathbf{Z}) = 0.1 \sum_{l=1}^{5} Z_l$ corresponding to a standard Cox model with linear effects, and (2) a complex functional form with $h(\mathbf{Z}) = Z_1^2 + Z_2^2 + \sin(3Z_3) + \sin(3Z_4) + \sin(3Z_5)$. For each model, we generated p covariates with $p = 5, 10, 100$, and the event time T is only determined by the first 5 covariates based on the model. The empirical power was summarized in Table 3 for the standard Cox model and in Table 4 for the non-linear model.

[Table 3 about here.]

[Table 4 about here.]

First, consider the case with moderate correlation $\wp = 0.5$. For the standard Cox model, the best strategy would be based on K_L with $p = 5$ (i.e., using the true 5 genes). For a $n = 100$ with 50% of censoring, this best strategy achieves about 74% of power. When we fit the model with $p = 100$ by including additional 95 noise covariates, the power decreases to 66%. Thus, the score test appears to have reasonable power in detecting the signal even in the presence of a large number of noise features. The use of kernel PCA does not appear to affect the power for linear kernel. When K_G is used, the score test $\widehat{S}$ achieves about 67% of power with $p = 5$ and with $p = 100$. Applying kernel PCA test $\widetilde{S}$ by thresholding the eigenvalues at $\mathfrak{p} = 90\%$, the power remains almost identical. There appears to be little loss in power for using K_G compared to $K_L$when $\widehat{S}$ and $\widetilde{S}$ are considered. This can in part be attributed to the fact that K_G is approximately K_L when ρ is relatively large (Liu et al, 2008). When $h(\mathbf{z})$ is non-linear, the use of K_G resulted in a substantial improvement in power when compared to K_L. For example, with $p = 5$, $n = 100$ and 50% of censoring, the power is about 94% for $\widehat{S}$ and 8% with the linear kernel. Even when $p = 100$ and $n = 200$, the power was 75% based on $\widehat{S}$ and merely 8% based on $\widehat{Q}$. This suggests drastically amount of gain in power for employing K_G over K_L when the true effects are highly non-linear. Again, the power does not appear to decrease dramatically as we increase the number of noise covariates. This is likely due to the fact that the kernel machine test effectively accounts for the correlation between the covariates which results in little increase in the effective DF as p increases.

We next summarize how the results may be affected by the covariate correlation $\wp$. As we decrease $\wp$, the effective DF of the score statistics under the null increases which is likely to yield a power loss. This is consistent with the results shown in Table 3 and 4. For example, under the standard Cox model, the power of $\widehat{S}$ decreased from 83% to 49% when $\wp$ was decreased from 0.5 to 0.2, when $p = 10$, $n = 100$ and 25% censoring. Moreover, the power appears to decrease more rapidly with p when $\wp$ is small. When $n = 100$ and 25% censoring, as p increases from 5 to 100, the power of $\widehat{S}$ decreases from 83% to 82% when $\wp = 0.5$ and from 57% to 37% when $\wp = 0.2$. This is also expected since when $\wp$ is small, the effective DF increases more rapidly with p. One cannot borrow information across genes to improve power. In the extreme case that all genes are independent, the best test would need to be based on p DF. In pathway analysis, as genes are often correlated with pathway, one would expect a high power using the kernel machine test as it effectively borrow information across genes. It is also worth noting that when $\wp = 0.2$ and the underlying effects are non-linear, there appears to be a substantial power loss by using $\widetilde{S}$ which only includes PCAs that account for 90% of the total variation. Thus, for such settings, a higher threshold values should be considered to minimize the information loss. When the threshold value was increased to 99%, the difference between $\widehat{S}$ and $\widetilde{S}$ becomes minimal with respect to power.

5. Example: Breast Cancer Gene Expression Study

Findings from genomic research hold great potential to improve disease outcomes for breast cancer patients. For example, the discovery that mutations in BRCA1 and BRCA2 genes increase the risk of breast cancer has radically transformed our understanding of the genetic basis of breast cancer, leading to improved management of high-risk women. A number of genomic biomarkers have been developed for clinical use, and increasingly, pharmacogenetic end points are being incorporated into clinical trial design (Olopade et al., 2008). Despite recent advances in understanding genetic susceptibility to breast cancer, it remains important to identify and understand molecular pathways of pathogenesis (Nathanson et al., 2001). Here, we are particularly interested in assessing whether various canonical pathways from the molecular signature database are related to breast cancer survival. Examples of these pathways include AKAP13, EGFR_SMRTE and p53. Genetic alterations within AKAP13 are expected to provoke a constitutive Rho signaling, thereby facilitating the development of cancer (Wirtenberger et al., 2006). Epidermal growth factor receptor (EGFR) is a receptor tyrosine kinase and is expressed in a wide variety of epithelial malignancies including non-small-cell lung cancer, head and neck cancer, and breast cancer (Kuwahara et al., 2004; Nicholson et al., 2001). EGFR activation promotes tumor growth by increasing cell proliferation, motility, or angiogenesis, and by blocking apoptosis (Holbro et al., 2003). p53 mutation remains the most common genetic change identified in human neoplasia and is associated with more aggressive disease and worse overall survival in breast cancer (Gasco et al., 2002).

Here, we apply the proposed score test to assess the overall effect of various canonical pathways using a recent breast cancer study by van de Vijver et al. (2002). The primary goal of this study was to evaluate the performance of prognostic rules constructed based on microarray gene expression data from Van't Veer et al. (2002). There are a total of 260 patients with primary breast carcinomas from the Netherlands Cancer Institute. The median followup time was 8.8 years and there was about

75% of censoring. For illustration, we consider 70 pathways that are potentially related to breast cancer survival. The number of genes contained in a pathway ranges from 2 to 235 with a median of 22. We applied the the proposed score test with both the Gaussian based on $\widehat{S}_{\mathcal{I}_2}$ and the linear kernel based on $\widehat{Q}$. The p-value for the aforementioned 70 pathways are shown in Figure 1. The score test identifies 56 (80%) pathways as significantly associated with breast cancer survival with p-value < 0.05 when the Gaussian kernel is used; and 46 (66%) when the linear kernel is used. For example, the AKAP13 pathway is significant with a p-value of 0.015 based on the Gaussian kernel but not significant with a p-value of 0.21 based on the linear kernel. For this pathway, there are a total of 10 genes and their effects appear to be non-linear. When comparing the fitted Cox model with linear effects only to the model with both linear and quadratic effects, the inclusion of quadratic effects yields an increase of 19.5 in the log partial likelihood with a p-value of 0.035. This also suggests the underlying effects of these genes are potentially non-linear.

[Figure 1 about here.]

In settings where multiple pathways are under examination, it is important to adjust for multiple comparison. The proposed procedure can easily be extended to incorporate such an adjustment. Specifically, let $\widehat{S}(j)$ denote the observed score statistic for the jth pathway and $\widehat{S}^*(j)$ represents its null counterpart. Then the adjusted p-value can be obtained by comparing the observed score statistic $\widehat{S}(j)$ to the distribution of $\widehat{S}^*_{\max} = \max\{\widehat{S}^*(1), \cdots, \widehat{S}^*(J)\}$. In practice, realizations of $\widehat{S}^*_{\max}$ can be easily generated via the aforementioned resampling procedure. For $b = 1, ..., B$, we generate a set of standard normal random variable $\mathcal{V}_{(b)}$ to calculate $\widehat{S}^*_{(b)}(1), \cdots, \widehat{S}^*_{(b)}(J)$ simultaneously and obtain $\widehat{S}^*_{\max(b)}$. The adjusted p-value for $\widehat{S}(j)$ can be calculated as $B^{-1}\sum_{b=1}^{B} I\{\widehat{S}^*_{\max(b)} > \widehat{S}(j)\}$. After adjusting for multiple comparisons, 23 pathways remain significant when Gaussian kernel is used and 17 when Linear kernel is used. For example, without adjusting for multiple comparisons, the EGFR pathway is significantly associated with breast cancer survival with p-value 0.0008 based on the Gaussian kernel and 0.014 based on the linear kernel. However, after accounting for the fact that we tested 70 pathways, it remains significant with a p-value of 0.038 when the Gaussian kernel is used but no longer significant when the linear kernel is used.

6. Discussions

We develop a powerful kernel machine test to test for the parametric/nonparametric pathway effect for survival data. A key advantage of the kernel machine test is that it can effectively borrow information across genes within a pathway by accounting for between-gene correlation and yield a powerful test with a low DF test if the genes within a pathway are moderately correlated. Its power is shown not affected when the number of genes including noisy genes increases within a pathway as long as the genes are correlated. When the underlying effects of $\mathbf{Z}$ is complex, testing based on the Gaussian kernel could potentially lead to a substantial power gain when compared to testing with linear kernel which corresponds to the Goeman et al (2005) procedure. The R code for implementing the proposed testing procedures will be available upon request.

The scaled χ^2 approximation to the null distribution of $\widehat{Q}(\rho)$ provides us a venue to assess the effective DF for the proposed test. As one expects, the DF does not increase with p quickly when $\mathbf{Z}$ are moderately or highly correlated and thus under such settings, the proposed test works well even for large p. On the other hand, when there is little correlation between the genes, the effective DF would be of similar order as p and thus the test may perform poorly when p is in the similar magnitude as n. In this case, resampling methods such as the permutation procedure could be used to estimate the null distribution of the test.

For the Gaussian kernel, the selection of ρ plays an important role in the score test. Large sample approximation to the null distribution may perform poorly when ρ is too small or too large. We determine the range of ρ based on kernel PCA. It is important to note the trade-off between selecting different values of ℓ_0. A small ℓ_0 might result in a significant power loss due to approximating $\mathbb{K}$

with $\widetilde{\mathbb{K}}$ while a large ℓ_0 may result in a higher effective DF and thus may also lead to efficiency loss. From Theorem 5.541 of Braun (2005), the projection error due to kernel PCA is approximately of order $O(\ell_0^{-a})$ for some $a > 0$ when the eigenvalues of the kernel function K decay at a polynomial rate and of order $O(e^{-b\ell_0})$ for some $b > 0$ when the eigenvalues decay exponentially. It would be interesting to investigate the optimal thresholding value ℓ_0. Our simulation results suggest that in most cases, the score test with $\rho \in \mathcal{I}$ with respect to power and size.

Our proposed test adjust for covariates $\mathbf{U}$ under a linearity assumption on the effects of $\mathbf{U}$. When the true effect of $\mathbf{U}$ is non-linear, the proposed test derived under the linearity assumption may have incorrect size. Thus, it is important to examine the appropriateness of the linearity assumption. In practice, since $\mathbf{U}$ involves clinical variables that are typically well studied, prior information is often available to pre-specified potentially non-linear functional forms for $\mathbf{U}$. More robust procedures such as the regression spline could also be used to incorporate non-linear effects in $\mathbf{U}$.

Acknowledgement

Cai's research is supported by grants R01-GM079330-03 from the National Institute of General Medical Sciences and DMS 0854970 from the National Science Foundation. Lin's research is supported by the grants R37-CA076404 and P01-CA134294 from the National Cancer Institute. The authors thank the editor and the reviewers for their helpful comments for improving the manuscript.

Appendix: Asymptotic Distribution for the Score Test Statistic

Throughout, we assume that $\rho \in [\mathfrak{p}_l, \mathfrak{p}_r] \in (-\infty, \infty)$, $\mathbf{Z}$ is bounded, $K(\mathbf{z}_1, \mathbf{z}_2, \rho)$ is continuously differentiable with respect to all of its arguments and is symmetric about $\mathbf{z}_1$ and $\mathbf{z}_2$, the conditional density of C given $\mathbf{Z}$ is continuous and bounded and the marginal density of C is bounded away from 0 on $[0, \tau]$. We assume that the upper bound of the support of X, denoted by τ, is finite.

To derive the asymptotic distribution for $\widehat{Q}(\rho)$, we first note that $\sup_{t \leqslant \tau} |\widehat{\Lambda}_0(t) - \Lambda_0(t)| + \|\widehat{\boldsymbol{\gamma}} -$

$\boldsymbol{\gamma}_0\| = o_p(1)$ (Fleming and Harrington, 1991), $n^{\frac{1}{2}}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{W}_{\gamma i} + O_p(n^{-\frac{1}{2}})$, $\widehat{\mathbb{W}}_\Lambda(t) = n^{\frac{1}{2}}\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\} = n^{-\frac{1}{2}} \sum_{i=1}^n W_{\Lambda i}(t) + O_p(n^{-\frac{1}{2}})$, where $\boldsymbol{\gamma}_0$ is the true value of $\boldsymbol{\gamma}$,

$$\mathbf{W}_{\gamma i} = \int \mathbb{A}^{-1} \left\{ \mathbf{U}_i - \frac{\mathfrak{S}^{(1)}(t)}{\mathfrak{S}^{(0)}(t)} \right\} dM_i(t), \quad W_{\Lambda i}(t) = \int_0^t \frac{dM_i(s) - \mathbf{W}_{\gamma i}^\intercal \mathfrak{S}^{(1)}(s) d\Lambda_0(s)}{\mathfrak{S}^{(0)}(s)} \tag{A.1}$$

$\mathbb{A} = \int \{\mathfrak{S}^{(2)}(t) - \mathfrak{S}^{(1)}(t)^{\otimes 2}\} \mathfrak{S}^{(0)}(t)^{-2} dE\{N_i(t)\}$, $\mathfrak{S}^{(k)}(t) = E\{\widetilde{Y}_i(t) \mathbf{U}_i^{\otimes k}\}$, $\widetilde{Y}_i(t) = Y_i(t) e^{\boldsymbol{\gamma}_0^\intercal \mathbf{U}_i}$, for any vector $\mathbf{u}$, $\mathbf{u}^{\otimes 0} = 1$, $\mathbf{u}^{\otimes 1} = \mathbf{u}$, and $\mathbf{u}^{\otimes 2} = \mathbf{u}\mathbf{u}^\intercal$. It follows that

$$\widehat{M_i} = M_i - n^{-\frac{1}{2}} \int \widetilde{Y}_i(t) d\widehat{\mathbb{W}}_\Lambda(t) - (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal \widetilde{\mathbf{U}}_i + O_p(n^{-1}). \tag{A.2}$$

We next write $n^{-1}\widehat{Q}(\rho) = \widehat{\mathscr{Q}}_1(\rho) + \widehat{\mathscr{Q}}_2(\rho) + \widehat{\mathscr{Q}}_3(\rho) - \widehat{q}(\rho)$, where

$$\widehat{\mathscr{Q}}_1(\rho) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n K_{ij}(\rho) M_i M_j = \int K(\mathbf{z}_1, \mathbf{z}_2, \rho) d\widehat{\mathbb{W}}_M(\mathbf{z}_1) d\widehat{\mathbb{W}}_M(\mathbf{z}_2)$$

$$\widehat{\mathscr{Q}}_2(\rho) = 2n^{-1} \sum_{i=1}^n \sum_{j=1}^n K_{ij}(\rho)(\widehat{M_i} - M_i) M_j, \quad \widehat{\mathscr{Q}}_3(\rho) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n K_{ij}(\rho)(\widehat{M_i} - M_i)(\widehat{M_j} - M_j)$$

From (A.2), $\widehat{\mathscr{Q}}_2(\rho) = -2n^{-1} \int [\int \{\sum_{i=1}^n K(\mathbf{Z}_i, \mathbf{z}, \rho) \widetilde{Y}_i(t)\} d\widehat{\mathbb{W}}_\Lambda(t) - \widehat{\mathbf{W}}_\gamma^\intercal \{\sum_{i=1}^n K(\mathbf{Z}_i, \mathbf{z}, \rho) \widetilde{\mathbf{U}}_i\}] d\widehat{\mathbb{W}}_M(\mathbf{z}) + o_p(1)$. By a uniform law of large numbers (ULLN, Pollard, 1990), $\sup_{\mathbf{z},\rho,t} |n^{-1} \sum_{i=1}^n K(\mathbf{Z}_i, \mathbf{z}, \rho) \widetilde{Y}_i(t) - \omega_0(\mathbf{z}, t, \rho)| + \sup_{\mathbf{z},\rho} \|n^{-1} \sum_{i=1}^n K(\mathbf{Z}_i, \mathbf{z}, \rho) \widetilde{\mathbf{U}}_i - \omega_1(\mathbf{z}, \rho)\| = o_p(1)$. This, together with Lemma A1 of Bilias et al (1997), implies that

$$\widehat{\mathscr{Q}}_2(\rho) = -2 \int \left\{ \int \omega_0(\mathbf{z}, t, \rho) d\widehat{\mathbb{W}}_\Lambda(t) + \widehat{\mathbf{W}}_\gamma^\intercal \omega_1(\mathbf{z}, \rho) \right\} d\widehat{\mathbb{W}}_M(\mathbf{z}) + o_p(1)$$

Furthermore, $\widehat{\mathscr{Q}}_3(\rho)$ is asymptotically equivalent to

$$\widehat{\mathbf{W}}_\gamma^\intercal \left\{ n^{-2} \sum_{i=1}^n \sum_{j=1}^n K_{ij}(\rho) \widetilde{\mathbf{U}}_i \widetilde{\mathbf{U}}_j \right\} \widehat{\mathbf{W}}_\gamma + 2\widehat{\mathbf{W}}_\gamma^\intercal \int \left\{ n^{-2} \sum_{i=1}^n \sum_{j=1}^n K_{ij}(\rho) \widetilde{\mathbf{U}}_i \widetilde{Y}_j(s) \right\} d\widehat{\mathbb{W}}_\Lambda(s)$$
$$\int\int \left\{ n^{-2} \sum_{i=1}^n \sum_{j=1}^n K_{ij}(\rho) \widetilde{Y}_i(t) \widetilde{Y}_j(s) \right\} d\widehat{\mathbb{W}}_\Lambda(t) d\widehat{\mathbb{W}}_\Lambda(s).$$

This, together with a ULLN for U-processes (Nolan and Pollard, 1987) and Lemma A1 of Bilias et al. (1997), implies that $\widehat{\mathscr{Q}}_3(\rho) = \int\int \omega_{00}(t, s, \rho) d\widehat{\mathbb{W}}_\Lambda(t) d\widehat{\mathbb{W}}_\Lambda(s) + \widehat{\mathbf{W}}_\gamma^\intercal \omega_{11}(\rho) \widehat{\mathbf{W}}_\gamma + 2\widehat{\mathbf{W}}_\gamma^\intercal \int \omega_{10}(s, \rho) d\widehat{\mathbb{W}}_\Lambda(s)$. Lastly, by a ULLN and the consistency of $\widehat{\boldsymbol{\gamma}}$, $\widehat{\Lambda}_0(t)$, $\sup_\rho |\widehat{q}(\rho) - q(\rho)| = o_p(1)$. This, together with the approximations for $\widehat{\mathscr{Q}}_2$ and $\widehat{\mathscr{Q}}_3$, implies that $\widehat{\mathscr{Q}}(\rho) = \widehat{\mathcal{W}}(\rho) - q(\rho) + o_p(1)$.

We next derive the asymptotic distribution of $\widehat{\mathcal{W}}(\rho)$. First, both $I(\mathbf{Z} \leqslant \mathbf{z})$ and $M_i(t)$ have finite pseudo-dimensions t$\widehat{\mathbb{W}}_\Lambda(t)$, implies that $\{\widehat{\mathbb{W}}_M(\mathbf{z}), \widehat{\mathbb{W}}_\Lambda(t), \widehat{\mathbf{W}}_\gamma\}$ converge jointly to zero-mean Gaussian processes $\mathbb{W}_M(\mathbf{z})$, $\mathbb{W}_\Lambda(t)$ and $\mathbf{W}$. Since $\widehat{\mathcal{W}}(\rho)$ is a smooth functional of $\mathbb{W}_M(\mathbf{z})$, $\mathbb{W}_\Lambda(t)$ and $\mathbf{W}$, it then follows from a strong representation theorem (Pollard, 1990) and Lemma A1 of Bilias et al (1997) that that $\widehat{\mathcal{W}}(\rho)$ converges weakly to the process

$$\begin{aligned}\mathcal{W}(\rho) = &\int\int K(\mathbf{z}_1, \mathbf{z}_2, \rho) d\mathbb{W}_M(\mathbf{z}_1) d\mathbb{W}_M(\mathbf{z}_2) - 2\int \left\{\int \omega_0(\mathbf{z}, t, \rho) d\mathbb{W}_\Lambda(t) + \mathbf{W}_\gamma^\intercal \omega_1(\mathbf{z}, \rho)\right\} d\mathbb{W}_M(\mathbf{z}) \\ &+ \int\int \omega_{00}(t, s, \rho) d\mathbb{W}_\Lambda(t) d\mathbb{W}_\Lambda(s) + \mathbf{W}_\gamma^\intercal \omega_{11}(\rho) \mathbf{W}_\gamma + 2\mathbf{W}_\gamma^\intercal \int \omega_{10}(s, \rho) d\mathbb{W}_\Lambda(s) \qquad (A.3)\end{aligned}$$

References

Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J., Vincent, P., and Ouimet, M. (2004). Learning Eigenfunctions Links Spectral Embedding and Kernel PCA. *Neural Computation* **16,** 2197–2219.

Bilias, Y., Gu, M., and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics* **25,** 662–682.

Braun, M. (2005). *Spectral Properties of the Kernel Matrix and their Application to Kernel Methods in Machine Learning.* PhD thesis, University of Bonn.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9–25.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* **97,** 262–267.

Buhmann, M. (2003). *Radial Basis Functions: Theory and Implementations.* Cambridge University Press.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* **2,** 121–167.

Cai, T., Wei, L., and Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika* **87,** 867–878.

Commenges, D. and Andersen, P. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1,** 145–156.

Cox, D. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **34,** 187–220.

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

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

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis.* NY John Wiley and Sons.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16,** 906–914.

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. *Breast Cancer Research* **4,** 70–76.

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

Goeman, J., Van De Geer, S., De Kort, F., and Van Houwelingen, H. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20,** 93–99.

Goeman, J., Van De Geer, S., and Van Houwelingen, H. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 477–493.

Holbro, T., Civenni, G., and Hynes, N. (2003). The ErbB receptors and their role in cancer progression. *Experimental cell research* **284,** 99–110.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist* **41,** 495–502.

Kuwahara, Y., Hosoi, H., Osone, S., Kita, M., Iehara, T., Kuroda, H., and Sugimoto, T. (2004). Antitumor activity of gefitinib in malignant rhabdoid tumor cells in vitro and in vivo. *Clinical Cancer Research* **10,** 5940–5948.

Lee, Y. and Lee, C. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics* **19,** 1132–1139.

Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Biocomputing* page 65.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics* **9,** 292–2.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics* **63,** 1079–1088.

Nathanson, K., Wooster, R., and Weber, B. (2001). Breast cancer genetics: what we know and what we need. *Nature Medicine* **7,** 552–556.

Nicholson, R., Gee, J., and Harper, M. (2001). EGFR and cancer prognosis. *European Journal of Cancer* **37,** 9–15.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics* pages 780–799.

Olopade, O., Grushko, T., Nanda, R., and Huo, D. (2008). Advances in Breast Cancer: Pathways to Personalized Medicine. *Clinical Cancer Research* **14,** 7988.

Park, Y. and Wei, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90,** 717–723.

Parzen, M., Wei, L., and Ying, Z. (1994). A resampling method based on pivotal functions. *Biometrika* **81,** 341–350.

Pollard, D. (1990). *Empirical processes: theory and applications.* Institute of Mathematical Statistics.

Ramaswamy, S., Tamayo, P., Rifkin, R., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences* **98,** 15149.

Scholkopf, B. and Smola, A. (2002). *Learning with kernels.* MIT Press Cambridge, Mass.

Scholkopf, B., Smola, A., and Muller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **10,** 1299–1319.

van de Vijver, M., He, Y., van't Veer, L., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347,** 1999–2009.

Van't Veer, L., Dai, H., Van de Vijver, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530.

Vapnik, V. (1998). Statistical learning theory. *NY Wiley* .

Vo, T., Phan, J., Huynh, K., and Wang, M. (2007). Reproducibility of Differential Gene Detection across Multiple Microarray Studies. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 4231–4234.

Wirtenberger, M., Tchatchou, S., Hemminki, K., et al. (2006). Association of genetic variants in the Rho guanine nucleotide exchange factor AKAP13 with familial breast cancer. *Carcinogenesis* **27,** 593.

Young, R. (2000). Biomedical discovery with DNA arrays. *Cell* **102,** 9–15.

Zwald, L. and Blanchard, G. (2006). On the Convergence of Eigenspaces in Kernel Principal Component Analysis. *Advances In Neural Information Processing Systems* **18,** 1649.

Figure 1. $\log_{10}$ Pvalue for testing the overall effect of the 70 genetic pathways on breast cancer survival based on the kernel machine score test $\widehat{S}$ with Gaussian kernel and $\widehat{Q}$ with the linear kernel. The crosses and the squares represent p-values before and after adjusting for multiple comparisons, respectively. The results are based on $B = 5000$ perturbations.

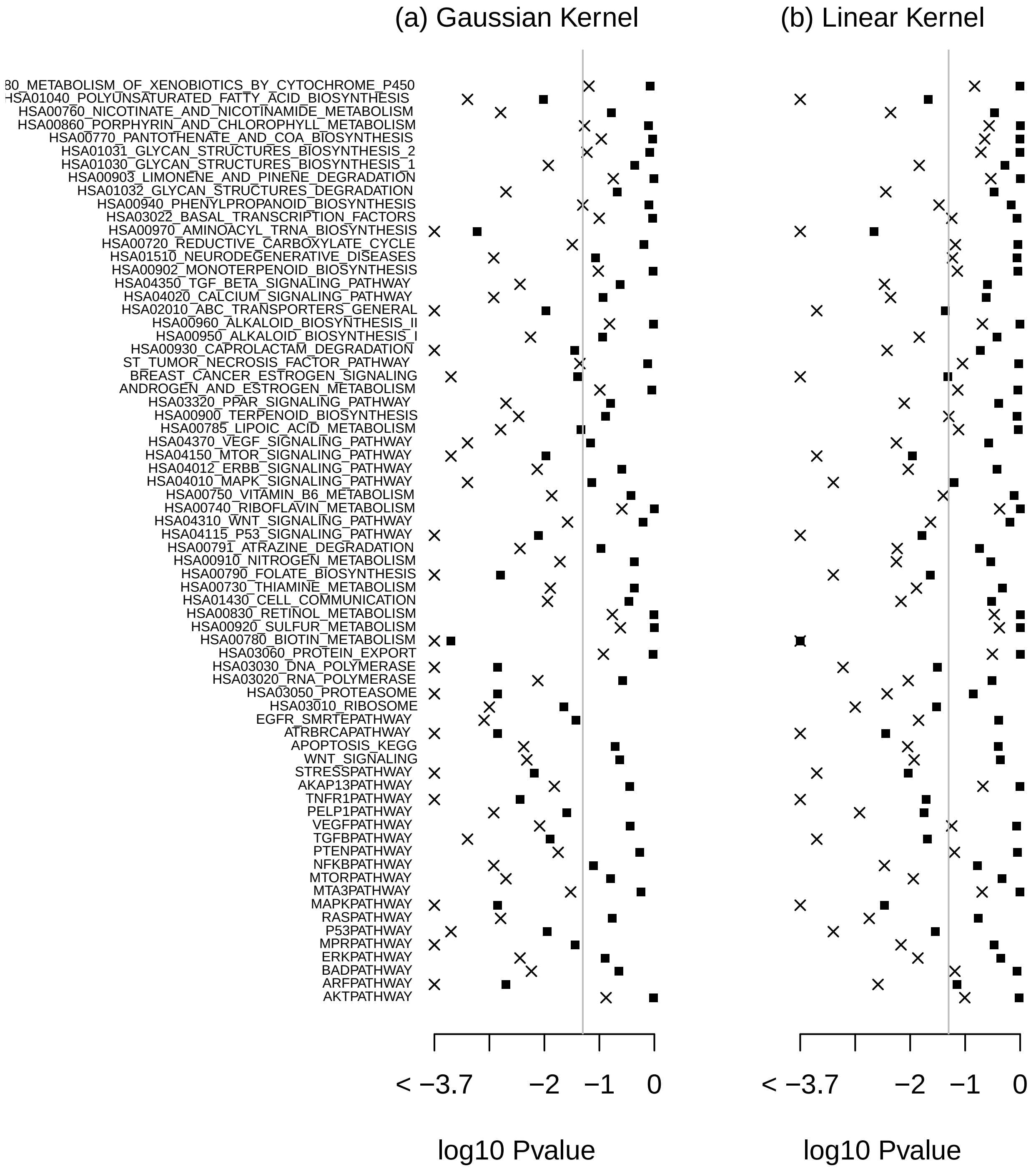

Table 1

Empirical sizes at type I error rate of 0.05 for the case when the genes are moderately correlated ($\wp = 0.5$). Testing was performed based on both the Gaussian kernel and the Linear kernel. For Gaussian kernel, we compared (i) sup-statistic with the original kernel (subscript $\widehat{S}$); (ii) sup-statistic with the kernel PCA including 90% of eigenvalues (subscript $\widetilde{S}$); (iii) score test with ρ fixed at ρ_2 (upper bound of $\mathcal{I}$) with original kernel (subscript $\widehat{Q}$); (iv) score test at ρ_2 with kernel PCA (subscript $\widetilde{Q}$). For the linear kernel, the testing was performed based on $\widehat{Q}$ and $\widetilde{Q}$. The null distributions were generated based on (a) the resampling procedure (indexed by P); (b) the χ^2 approximation (indexed by χ^2); and (c) the normal approximation (indexed by N).

			$p = 5$				$p = 10$				$p = 100$			
Censoring %			50%		25%		50%		25%		50%		25%	
Kernel		n	100	200	100	200	100	200	100	200	100	200	100	200
Gaussian	sup	$P_{\widehat{S}}$	5.8	4.1	4.5	4.2	4.0	5.2	4.0	4.3	5.1	5.2	5.1	4.0
		$P_{\widetilde{S}}$	5.3	4.6	5.1	4.4	4.4	5.7	4.2	4.8	5.4	5.4	5.8	4.6
		$N_{\widehat{S}}$	7.8	6.4	6.4	6.2	6.0	7.6	5.4	6.1	6.4	6.8	7.0	5.6
		$N_{\widetilde{S}}$	8.7	7.0	7.4	6.9	6.8	8.1	6.0	6.8	7.0	7.0	7.9	6.0
	ρ_2	$P_{\widehat{Q}}$	5.8	4.5	5.0	4.5	4.0	5.4	4.0	4.6	5.0	5.2	5.0	4.1
		$P_{\widetilde{Q}}$	5.5	4.7	5.4	4.6	4.8	6.0	4.4	4.9	5.6	5.2	5.9	4.6
		$N_{\widehat{Q}}$	7.4	5.9	6.4	5.8	5.8	7.3	5.2	5.8	6.4	6.8	7.1	5.6
		$N_{\widetilde{Q}}$	8.4	7.0	7.1	6.4	6.2	7.8	6.6	6.9	7.5	7.1	8.0	6.3
		$\chi^2_{\widehat{Q}}$	5.5	4.4	4.8	4.4	4.0	5.5	4.1	4.6	5.2	5.2	5.2	4.4
		$\chi^2_{\widetilde{Q}}$	6.2	4.8	5.4	4.5	4.6	5.9	4.4	4.9	5.4	5.2	6.0	4.7
Linear		$P_{\widehat{Q}}$	4.5	6.0	3.9	5.6	5.2	4.0	4.7	4.8	5.8	4.9	6.0	4.6
		$P_{\widetilde{Q}}$	4.2	5.9	4.4	5.3	5.2	4.2	4.8	4.6	5.8	5.0	6.1	4.6
		$N_{\widehat{Q}}$	6.0	7.2	5.8	6.8	7.0	5.2	6.2	6.4	7.5	6.9	7.6	6.0
		$N_{\widetilde{Q}}$	5.7	7.5	5.8	6.7	6.8	5.4	6.2	6.8	7.6	7.0	7.8	6.0
		$\chi^2_{\widehat{Q}}$	4.5	5.9	3.8	5.6	5.2	4.1	4.6	4.6	6.0	5.2	6.2	4.9
		$\chi^2_{\widetilde{Q}}$	4.2	5.8	4.3	5.2	5.2	4.2	4.7	4.7	6.0	5.2	6.2	5.0

Table 2

Empirical sizes at type I error rate of 0.05 for the case when the genes are weakly correlated ($\wp = 0.2$). Testing was performed based on (i) sup-statistic with the original kernel (subscript $\widehat{S}$); (ii) sup-statistic with the kernel PCA including 90% of eigenvalues (subscript $\widetilde{S}$); (iii) score test with ρ fixed at ρ_2 (upper bound of $\mathcal{I}$) with original kernel (subscript $\widehat{Q}$); (iv) score test at ρ_2 with kernel PCA (subscript $\widetilde{Q}$). The null distributions were generated based on (a) the resampling procedure (indexed by P); (b) the χ^2 approximation (indexed by χ^2); and (c) the normal approximation (indexed by N).

			$p = 5$				$p = 10$				$p = 100$			
Censoring %			50%		25%		50%		25%		50%		25%	
Kernel		n	100	200	100	200	100	200	100	200	100	200	100	200
Gaussian	sup	$P_{\widehat{S}}$	4.8	3.6	3.4	4.0	3.3	4.5	3.0	3.6	3.5	4.6	3.0	3.0
		$P_{\widetilde{S}}$	5.4	4.5	4.9	4.6	4.2	5.4	4.2	4.1	5.2	5.3	5.4	4.4
		$N_{\widehat{S}}$	6.9	5.4	5.4	6.1	4.8	6.5	4.1	4.9	4.3	5.4	4.1	4.0
		$N_{\widetilde{S}}$	10.	7.4	8.2	7.8	7.1	8.2	7.1	7.3	7.0	7.1	7.6	6.2
	ρ_2	$P_{\widehat{Q}}$	5.4	3.7	3.8	4.0	3.5	4.4	3.2	4.0	3.6	4.6	3.0	3.2
		$P_{\widetilde{Q}}$	6.2	4.7	5.3	4.5	4.4	5.8	4.6	4.8	5.6	5.3	5.8	4.7
		$N_{\widehat{Q}}$	7.2	5.3	5.5	6.0	4.8	6.4	4.6	4.9	4.4	5.3	4.0	4.1
		$N_{\widetilde{Q}}$	8.4	6.7	7.0	6.5	6.2	7.2	6.2	6.8	7.4	7.2	8.0	6.2
		$\chi^2_{\widehat{Q}}$	5.2	3.6	3.6	4.0	3.4	4.8	3.3	4.0	3.6	4.8	3.2	3.3
		$\chi^2_{\widetilde{Q}}$	6.0	4.6	5.3	4.4	4.4	6.0	4.6	4.8	5.5	5.3	5.9	5.0
Linear		$P_{\widehat{Q}}$	4.1	5.6	4.2	5.1	4.3	3.4	3.2	3.6	4.0	4.3	4.0	3.4
		$P_{\widetilde{Q}}$	4.4	5.5	4.1	5.4	4.4	3.7	3.7	4.1	4.4	4.5	4.5	3.4
		$N_{\widehat{Q}}$	6.0	7.1	5.8	6.7	5.8	4.9	5.1	4.8	5.4	5.4	5.0	4.4
		$N_{\widetilde{Q}}$	6.2	7.3	5.9	6.6	6.2	5.2	5.3	5.4	5.6	5.4	5.4	4.6
		$\chi^2_{\widehat{Q}}$	4.0	5.4	4.2	4.8	4.2	3.3	3.3	3.6	4.4	4.6	4.2	3.4
		$\chi^2_{\widetilde{Q}}$	4.2	5.3	4.2	5.2	4.4	3.7	3.8	4.0	4.6	4.8	4.4	3.6

Table 3

Empirical power when $h(\mathbf{Z})$ is linear at the type I error rate of 0.05 with p=5, 10, and 100 genes are used to fit the model while the true number of genes is 5. Testing was performed based on (i) sup-statistic with the original kernel (subscript $\widehat{S}$); (ii) sup-statistic with the kernel PCA including 90% of eigenvalues (subscript $\widetilde{S}$); (iii) score test with ρ fixed at ρ_2 (upper bound of $\mathcal{I}$) with original kernel (subscript $\widehat{Q}$); (iv) score test at ρ_2 with kernel PCA (subscript $\widetilde{Q}$). The null distributions were generated based on the resampling procedure (indexed by P).

			$p=5$				$p=10$				$p=100$			
Censoring %			50%		25%		50%		25%		50%		25%	
Kernel		n	100	200	100	200	100	200	100	200	100	200	100	200
			Correlation = 0.5											
Gaussian	sup	$P_{\widehat{S}}$	67	94	83	99	68	94	83	99	67	92	82	98
		$P_{\widetilde{S}}$	68	94	85	99	69	94	85	99	68	92	83	99
	ρ_2	$P_{\widehat{Q}}$	69	95	86	99	69	94	84	100	67	92	82	98
		$P_{\widetilde{Q}}$	72	96	88	99	71	94	86	100	69	92	84	99
Linear		$P_{\widehat{Q}}$	74	97	88	100	71	96	86	100	66	94	81	99
		$P_{\widetilde{Q}}$	75	96	88	100	72	96	87	100	66	94	81	99
			Correlation = 0.2											
Gaussian	sup	$P_{\widehat{S}}$	42	72	57	89	37	66	49	83	28	57	37	73
		$P_{\widetilde{S}}$	49	78	65	92	42	70	56	86	34	61	48	76
	ρ_2	$P_{\widehat{Q}}$	43	74	59	90	38	67	50	84	28	57	37	73
		$P_{\widetilde{Q}}$	54	82	69	94	45	72	60	87	34	62	48	76
Linear		$P_{\widehat{Q}}$	44	77	59	89	39	69	54	86	29	58	39	73
		$P_{\widetilde{Q}}$	44	78	61	90	40	70	54	86	30	58	40	74

Table 4

Empirical power when $h(\mathbf{Z})$ is non-linear at the type I error rate of 0.05 *with p =5, 10, and 100 genes used to fit the model while the true number of genes is 5. Testing was performed based on (i) sup-statistic with the original kernel (subscript $\widehat{S}$); (ii) sup-statistic with the kernel PCA including 90% of eigenvalues (subscript $\widetilde{S}$); (iii) score test with ρ fixed at ρ_2 (upper bound of $\mathcal{I}$) with original kernel (subscript $\widehat{Q}$); (iv) score test at ρ_2 with kernel PCA (subscript $\widetilde{Q}$). The null distributions were generated based on the resampling procedure (indexed by P).*

			$p = 5$				$p = 10$				$p = 100$			
Censoring %			50%		25%		50%		25%		50%		25%	
Kernel		n	100	200	100	200	100	200	100	200	100	200	100	200
			Correlation = 0.5											
Gaussian	sup	$P_{\widehat{S}}$	94	100	94	100	78	100	76	100	30	75	29	73
		$P_{\widetilde{S}}$	84	100	85	100	70	99	69	98	31	75	31	73
	ρ_2	$P_{\widehat{Q}}$	54	97	56	96	32	81	34	82	21	55	22	56
		$P_{\widetilde{Q}}$	19	36	21	40	13	21	17	26	10	17	12	18
Linear		$P_{\widehat{Q}}$	8	10	9	10	9	11	10	14	8	10	9	10
		$P_{\widetilde{Q}}$	7	9	9	10	9	11	9	13	8	10	9	10
			Correlation = 0.2											
Gaussian	sup	$P_{\widehat{S}}$	65	99	61	99	21	60	19	54	5	7	4	6
		$P_{\widetilde{S}}$	30	89	29	87	10	20	10	16	6	6	5	6
	ρ_2	$P_{\widehat{Q}}$	17	35	16	33	8	15	9	13	5	6	4	6
		$P_{\widetilde{Q}}$	11	16	12	15	8	9	8	9	5	6	5	6
Linear		$P_{\widehat{Q}}$	6	8	7	8	8	7	7	7	4	6	4	6
		$P_{\widetilde{Q}}$	7	8	7	8	8	8	7	7	4	7	5	6