Biometrics 000, 000–000 DOI: 000
000 0000

Omnibus Risk Assessment via Accelerated Failure Time Kernel Machine Modeling

Summary: Integrating genomic information with traditional clinical risk factors to improve the prediction of disease outcomes could profoundly change the practice of medicine. However, the large number of potential markers and the complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. Standard approaches to identifying important markers often rely on marginal associations and may not capture non-linear or interactive effects. At the same time, much work has been done to group genes into pathways and networks. Integrating such biological knowledge into statistical learning could potentially improve model interpretability and reliability. One effective approach is to employ a kernel machine (KM) framework, which has been recently extended to analyzing survival outcomes under the Cox model. In this paper, we propose KM regression under an accelerated failure time model. We derive a pseudo score statistic for testing and a risk score for prediction of survival. To approximate the null distribution of our test statistic, we propose resampling procedures which also enable us to develop alternative robust testing procedures that combine information across models and kernels. Numerical studies show that the testing and estimation procedures perform well. The methods are illustrated with an application in breast cancer.

Key words: Accelerated Failure Time Model, Kernel Machine, Resampling, Risk Prediction, Survival Analysis

1. Introduction

Understanding the relationship between genomic markers and complex disease could have a profound impact on biological research, pharmacology, and medicine. Many standard approaches for quantifying the relationship between genomic data and a phenotype of interest identify important genetic markers by assessing their marginal associations with the phenotype; however, such approaches may (i) yield highly ranked genes that are poorly annotated and are in fact downstream genes with limited interpretability and reproducibility (Fortunel et al., 2003; Mootha et al., 2003); or (ii) suffer from low power due to inability to capture non-linear relationships and genes that interact with one another biologically. Furthermore, single marker analyses may not be applicable to rare variants arising from next generation sequencing studies (Wu et al., 2011). At the same time much work has been done to group genetic markers into gene-sets based on genomic position, co-regulation, and putative pathway membership. Throughout, we use the terms "pathways" and "gene-sets" generally to denote collections of genetic markers grouped for joint analysis based on biological knowledge. Integrating such prior biological knowledge into statistical learning has intuitive appeal, and pathway-based results can be more interpretable and potentially more reliable and reproducible than marginal gene methods.

Kernel machine (KM) learning has become an increasingly popular tool for capturing complex effects (Scholkopf and Smola, 2002). Methods for KM modeling of biological pathway effects have been developed for linear and logistic regression (Liu et al., 2007, 2008) for non-censored outcomes. For survival outcomes, Li and Luan (2003) and Cai et al. (2011) proposed KM testing and estimation procedures under the Cox proportional hazards (PH) model. However, when the PH assumption fails to hold, these procedures may have little power to identify important pathways or accurately predict risk. In this paper, we propose KM methods for the testing and estimation of pathway effects on survival outcomes under the accelerated failure time (AFT) model (Kalbfleisch et al., 1980), a useful alternative to the Cox model.

The standard semiparametric AFT model relates covariates to log-survival time through a linear model. This model is appealingly interpretable, but has been used less than the Cox model in part because it can be somewhat challenging to fit in the presence of censoring. Inference procedures for the regression parameters under the AFT model include the inverse probability weighting (IPW), Buckley-James, rank-based, and sieve likelihood (SL) methods (Buckley and James, 1979; Koul et al., 1981; Tsiatis, 1990; Wei, 1992; Zeng and Lin, 2007). The IPW approach requires that the conditional censoring distribution be correctly specified and that the support of the censoring contain that of the failure time, which are both unlikely in practice. The Buckley-James procedure relies on the identifiability of the entire residual distribution, which may not be available in the presence of censoring. The sieve likelihood estimator is fully efficient, but could be computationally challenging because it requires estimating a non-parametric functional. The rank-based approach (Tsiatis, 1990; Ritov, 1990), which fits the model using a weighted log rank (WLR) estimating equation, has advantages including consistency of estimation without additional censoring assumptions, and an effective implementation developed in Jin et al. (2003) for making inference using resampling.

To capture potentially complex pathway effects on survival, we propose the use of the AFT KM model that allows for non-linear effects. We first develop a pseudo KM score test under this framework and propose resampling procedures to approximate the null distribution of our test statistic. In practice, it is often unclear which WLR weight or kernel is most appropriate for a given dataset. To overcome this challenge, we propose omnibus testing procedures that allow the data to make these choices automatically. We demonstrate the effectiveness of this approach in simulations, where the power lost is often minimal, but the power gained over the worst choice can be substantial.

When a pathway is potentially predictive of outcome, it is often of interest to incorporate the pathway information to improve risk prediction. Under the AFT KM framework, we propose procedures for estimating the covariate effects to construct individual risk scores. Our simulation results suggest that these outperform those derived from standard linear effects models when the underlying effects are non-linear, while maintaining similar accuracy when the effects are linear. Recently, Liu et al. (2010) proposed a weighted least squares estimator under this framework; however, it is unclear whether the estimator is consistent even under linear effects in the presence of censoring. Our procedure is derived using the WLR class of estimating equations and hence provides consistent estimates of the underlying effects when the AFT KM model is correctly specified.

The rest of this paper is organized as follows. In Section 2 we introduce the AFT KM model. In Section 3 we present testing and estimation procedures using Gehan weights in the WLR. In Section 4 we present testing and estimation for general weights, as well as our omnibus testing procedure. Simulations are presented in section 5 and our method is illustrated in section 6 in application to breast cancer data. Final remarks are in Section 7.

2. The Kernel Machine Accelerated Failure Time Model

Suppose we are interested in assessing the relationship between a possibly censored survival time and a collection of genetic measurements — for example, genes in a pathway — adjusting for other covariates. Let T denote the survival time, $\mathbf{Z}$ be a $P \times 1$ vector of genetic measurements, and $\mathbf{D}$ be a vector of clinical covariates such as age and gender. Due to censoring, we observe $X = \min\{T, C\}$ and $\Delta = I[T \leqslant C]$, where C is a censoring time that is assumed to be independent of T given $\mathbf{W} = (\mathbf{D}^{\mathsf{T}}, \mathbf{Z}^{\mathsf{T}})^{\mathsf{T}}$. The observed data consist of n independent and identically distributed (iid) random vectors, $\mathcal{O} = \{(X_i, \Delta_i, \mathbf{W}_i) : i = 1, \ldots, n\}$. To derive a prediction model for T based on $\mathbf{W}$, we consider the KM generalization of the AFT model:

$$\log T_i = \gamma^\mathsf{T}\mathbf{D}_i + h(\mathbf{Z}_i) + E_i,\ \ i = 1, \ldots, n. \tag{1}$$

where γ is the unknown covariate effect of $\mathbf{D}_i$, E_i is an iid error term independent of $\mathbf{W}_i$ with completely unspecified distribution, $h(\cdot) \in \mathcal{H}_K$ is an unknown smooth function and $\mathcal{H}_K$ is the Hilbert space generated by a given positive definite kernel $K(\cdot, \cdot; \rho)$. The kernel is a measure of similarity between two vectors of genetic measurements, and may depend on a possibly unknown scaling parameter ρ. Different choices of kernel K will yield different collections of possible functions $h(\cdot)$. For example, the *linear kernel* $K(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^\mathsf{T}\mathbf{z}_2$ leads to $h(\mathbf{z}) = \beta^\mathsf{T}\mathbf{z}$, a linear function of the covariates. The *quadratic kernel* $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = (\rho + \mathbf{z}_1^\mathsf{T}\mathbf{z}_2)^2$ yields a Hilbert space $\mathcal{H}_K$ spanned by basis functions $\{z_j, z_j z_{j'} : j, j' = 1, \ldots, p\}$, which incorporates main effects, quadratic effects, and 2-way interactions. To allow for more complex non-linear effects, one may consider the *Gaussian kernel*, defined by $K(\mathbf{z}_1, \mathbf{z}_2; \rho) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$. The resulting function space $\mathcal{H}_K$ consists of the radial basis functions.

When $h(\cdot) \equiv 0$, model (1) reduces to the standard AFT model for the clinical covariates,

$$\log T_i = \gamma^\mathsf{T}\mathbf{D}_i + E_i,\ \ i = 1, \ldots, n. \tag{2}$$

The regression parameter γ can be estimated by solving the WLR estimating function,

$$\mathbf{U}_\phi(\gamma) = n^{-1}\sum_{i=1}^{n} \Delta_i \phi(\gamma, e_i(0;\gamma))[\mathbf{D}_i - \overline{\mathbf{D}}(\gamma, e_i(0;\gamma))] \tag{3}$$

where $e_i(h;\gamma) = \log X_i - \gamma^\mathsf{T}\mathbf{D}_i - h(\mathbf{Z}_i)$ is the residual, $\overline{\mathbf{D}}(\gamma, t) = S^{(1)}(\gamma, t)/S^{(0)}(\gamma, t)$, $S^{(k)}(\gamma, t) = n^{-1}\sum_{i=1}^{n} \mathbf{I}\{e_i(0;\gamma) \geqslant t\}\mathbf{D}_i^{\otimes k}$, $\phi(\gamma, t)$ is a weight function, and for any vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\mathsf{T}$. The Gehan weights, $\phi = S^{(0)}$, yield an estimating function which simplifies to

$$\mathbf{U}_G(\gamma) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i[\mathbf{D}_i - \mathbf{D}_j] I\{e_i(0;\gamma) \leqslant e_j(0;\gamma)\} \tag{4}$$

which is the gradient of the convex Gehan objective function

$$L_G(\gamma) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i |e_i(0;\gamma) - e_j(0;\gamma)|_+, \tag{5}$$

where $|a|_+ = a\mathbf{I}\{a > 0\}$. A root of $\mathbf{U}_G(\gamma)$ can be found by minimizing $L_G(\gamma)$, and this solution, say $\widetilde{\gamma}_G$, can be immediately used as a consistent estimator of γ in (2). However, the variance of the estimator $\widetilde{\gamma}_G$ depends on the underlying distributions of survival and censoring, so in some settings a better estimate of γ may be obtained from finding a root of $\mathbf{U}_\phi$ in (3) with a different weight ϕ. This is difficult to do directly, but can be accomplished as described in Jin et al. (2003) by using $\widetilde{\gamma}^{(0)} = \widetilde{\gamma}_G$ as an initial value and, for each $k \geqslant 1$, letting $\widetilde{\gamma}^{(k)}$ be the solution to:

$$\widetilde{\mathbf{U}}_\phi(\gamma; \widetilde{\gamma}^{(k-1)}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(\widetilde{\gamma}^{(k-1)}, e_i(0; \widetilde{\gamma}^{(k-1)}))\Delta_i [\mathbf{D}_i - \mathbf{D}_j]\mathbf{I}\{e_i(0;\gamma) \leqslant e_j(0;\gamma)\} \tag{6}$$

where $\psi(\gamma, b) = \phi(\gamma, t)/S^{(0)}(\gamma, t)$. Because the weights $\psi(\widetilde{\gamma}^{(k-1)}, e_i(0; \widetilde{\gamma}^{(k-1)}))$ do not depend on γ, $\widetilde{\mathbf{U}}_\phi$ is the gradient of the convex function:

$$\widetilde{L}_\phi(\gamma; \widetilde{\gamma}^{(k-1)}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(\widetilde{\gamma}^{(k-1)}, e_i(0; \widetilde{\gamma}^{(k-1)}))\Delta_i |e_i(0;\gamma) - e_j(0;\gamma)|_+, \tag{7}$$

so finding a root of $\widetilde{\mathbf{U}}_\phi$ can be accomplished by minimizing $\widetilde{L}_\phi$. Each such $\widetilde{\gamma}^{(k)}$ is consistent for γ and, as k increases, approaches a solution of $\mathbf{U}_\phi(\gamma) = 0$.

Jin et al. (2003) further provide an inference procedure using perturbation resampling which we will mimic in our setting, and so describe briefly here. Specifically, a vector $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_n)$ of iid random variables $\mathcal{V}_i$ with mean 1 and variance 1 is generated, and $\widetilde{\gamma}_G^*$ is found as the root of

$$\mathbf{U}_G^*(\gamma) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i [\mathbf{D}_i - \mathbf{D}_j]\mathbf{I}\{e_i(0;\gamma) \leqslant e_j(0;\gamma)\}\mathcal{V}_i\mathcal{V}_j. \tag{8}$$

It can then be shown that the distribution of $n^{\frac{1}{2}}(\widetilde{\gamma}_G - \gamma_0)$, which is asymptotically 0-mean normal, may be approximated by the distribution of $n^{\frac{1}{2}}(\widetilde{\gamma}_G^* - \widetilde{\gamma}_G)$ conditional on the data. Analogous results hold for perturbations $\widetilde{\gamma}^{(k)*}$ of $\widetilde{\gamma}^{(k)}$ for more general weights ϕ.

When we do not assume the pathway effect $h(\cdot)$ is identically 0 in model (1), we may obtain estimators for γ and h by minimizing the penalized Gehan objective function analogous to equation (5),

$$L_G(\gamma, h) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i |e_i(h;\gamma) - e_j(h;\gamma)|_+ + \frac{c}{2} \|h\|^2_{\mathcal{H}_K}, \tag{9}$$

where $\|h\|_{\mathcal{H}_K}$ is the norm of h under $\mathcal{H}_K$. By the representer theorem (Kimeldorf and Wahba, 1970), the minimizer of $L_G(\gamma, h)$ for h must take a dual form, $\widehat{h}(\mathbf{z}) = \sum_{k=1}^{n} \alpha_k K(\mathbf{Z}_k, \mathbf{z})$, where the α_k are unknown parameters. Here and in the sequel, unless noted otherwise, we suppress ρ from K and other related quantities for the ease of presentation but note that all the testing and estimation procedures may depend on the tuning parameter ρ in the kernel function. Then minimization of (9) is equivalent to the minimization of

$$L_G^R(\alpha;\gamma) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i |e_i(\alpha;\gamma) - e_j(\alpha;\gamma)|_+ + \frac{c}{2} \alpha^{\mathsf{T}} \mathbb{K} \alpha \tag{10}$$

where $\mathbb{K}$ is the kernel matrix, whose $(i,j)^{th}$ entry is $K(\mathbf{Z}_i, \mathbf{Z}_j)$, and, with a slight abuse of notation, $e_i(\alpha;\gamma) = \log X_i - \gamma^{\mathsf{T}} \mathbf{D}_i - \alpha^{\mathsf{T}} \mathbf{K}_i$, for $\mathbf{K}_i$ the i^{th} row of $\mathbb{K}$. Hence, a plug-in estimator for h may be obtained by minimizing $L_G^R(\alpha;\gamma)$ with respect to α and γ. Moreover, we may view (10) as the Gehan objective function arising from the random effects AFT model

$$\log T_i = \gamma^{\mathsf{T}} \mathbf{D}_i + \alpha^{\mathsf{T}} \mathbf{K}_i + E_i, \quad \alpha = \tau\epsilon, \quad E(\epsilon) = 0, \quad \text{var}(\epsilon) = \mathbb{K}^{-}, \tag{11}$$

with ϵ multivariate normal and $c^{-1} = \tau$. Here $\mathbb{K}^{-}$ is the Moore-Penrose generalized inverse of $\mathbb{K}$. Analogous connections between penalized KM models and the mixed model framework were used successfully to fit KM regression in other models (Liu et al., 2007, 2008; Cai et al., 2011).

3. Testing and Estimation with Gehan Weights

3.1 *The Pseudo Score Statistic*

An important step in constructing risk scores for the prediction of survival is identifying pathways that are associated with T. Here, we propose the use of the AFT KM framework and derive testing procedures for

$$H_0 : h(\cdot) = 0.$$

in model (1). Specifically, we derive a KM pseudo score test of H_0 using the WLR estimating function with Gehan weights, then in section 4 show how to extend our test to more general weights.

By using the mixed effects formulation (11) as a working model, we see that this hypothesis is equivalent to testing $H_0 : \tau = 0$ and so we can derive a KM pseudo score test procedure by writing the penalized Gehan objective function (10) as a function of τ conditional on random effects ϵ, $L_{G,\epsilon}(\tau;\gamma) + \frac{\tau}{2}\epsilon^{\mathsf{T}}\mathbb{K}\epsilon$, where

$$L_{G,\epsilon}(\tau;\gamma) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i|e_i(\tau\epsilon;\gamma) - e_j(\tau\epsilon;\gamma)|_{+} \tag{12}$$

is the Gehan objective function conditional on ϵ. If γ is known, a pseudo score statistic can be obtained as

$$\widehat{Q}(\gamma) = E_\epsilon\left[\left\{\nabla_\tau L_{G,\epsilon}(\tau;\gamma)|_{\tau=0}\right\}^2\Big|\,\mathcal{O}\right],$$

where ∇_τ is the partial derivative with respect to τ. When γ is unknown, as would be the case in practice, we obtain the estimate $\widetilde{\gamma}_G$ by minimizing (5), the Gehan-weighted objective function from the model under H_0. Then, we define our pseudo score statistic as $\widehat{Q} = \widehat{Q}(\widetilde{\gamma}_G)$. Since $\nabla_\tau L_{G,\epsilon}(\tau;\gamma)|_{\tau=0} = \epsilon^{\mathsf{T}}\mathbb{K}\widehat{\mathbf{R}}(\gamma)$, where $\widehat{\mathbf{R}}(\gamma) = (\widehat{R}_1(\gamma),\ldots,\widehat{R}_n(\gamma))^{\mathsf{T}}$, and

$$\widehat{R}_k(\gamma) = n^{-2}\sum_{j=1}^{n}\left[\Delta_k\mathbf{I}\{e_k(0;\gamma) \leqslant e_j(0;\gamma)\} - \Delta_j\mathbf{I}\{e_j(0;\gamma) \leqslant e_k(0;\gamma)\}\right],$$

we have

$$\widehat{Q}(\gamma) = E_\epsilon[\widehat{\mathbf{R}}(\gamma)^\intercal \mathbb{K}\epsilon\epsilon^\intercal \mathbb{K}\widehat{\mathbf{R}}(\gamma)|\mathcal{O}] = \widehat{\mathbf{R}}(\gamma)^\intercal \mathbb{K}\widehat{\mathbf{R}}(\gamma) = \widehat{\mathbf{U}}_G(\gamma)^\intercal \widehat{\mathbf{U}}_G(\gamma) \tag{13}$$

where

$$\widehat{\mathbf{U}}_G(\gamma) = \widetilde{\mathbb{B}}_n\widehat{\mathbf{R}}(\gamma) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\widetilde{\mathbf{B}}_{ni} - \widetilde{\mathbf{B}}_{nj})\Delta_i\mathbf{I}\{e_i(0;\gamma) \leqslant e_j(0;\gamma)\} \tag{14}$$

for $\widetilde{\mathbb{B}}_n$ is the square root matrix of the symmetric non-negative definite matrix $\mathbb{K}$, $\mathbb{K} = \widetilde{\mathbb{B}}_n^\intercal\widetilde{\mathbb{B}}_n$, and $\widetilde{\mathbf{B}}_{ni}$ is the ith row of $\widetilde{\mathbb{B}}_n$. $\widehat{\mathbf{U}}_G(\gamma)$ is analogous to the standard Gehan WLR function (4).

3.2 *Kernel Principal Components Analysis Approximation*

A motivation for using the KM approach is to gain power to detect signal in the data; however, by using KMs, we may introduce a large number of parameters that cause a loss in power. However, especially when there is some correlation among the covariates, the feature space defined by the kernel and the effects of $\mathbf{Z}$ on T may be captured well by the space spanned by the first few leading eigenfunctions of $\mathcal{H}_K$. When this is the case, we can benefit from kernel principal components analysis (PCA) for dimension reduction (Blanchard et al., 2007; Mika et al., 1999). Kernel PCA aims to approximate the underlying feature space $\mathcal{H}_K$ by a small number of eigenfunctions. Specifically, the basis spanning $\mathcal{H}_K$ consists of $\{\mathscr{B}_l(\cdot) = \sqrt{\lambda_l}\zeta_l(\cdot), l = 1, 2, \ldots\}$, where ζ_l is the eigenfunction corresponding to the lth eigenvalue λ_l, with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots$, and $K(\mathbf{z}, \mathbf{z}') = \sum_{l=1}^{\infty}\mathscr{B}_l(\mathbf{z})\mathscr{B}_l(\mathbf{z}')$. We could use this basis to express h in its primal representation $h(\mathbf{z}) = \sum_{l=1}^{\infty}\beta_l\mathscr{B}_l(\mathbf{z})$. If the eigenvalues decay quickly, the feature space $\mathcal{H}_K$ can be approximated by the space spanned by $\{\mathscr{B}_1(\cdot), \ldots, \mathscr{B}_{r_0}(\cdot)\}$ and $h(\mathbf{z})$ can be approximated well by $h_{r_0}(\mathbf{z}) = \sum_{l=1}^{r_0}\beta_l\mathscr{B}_l(\mathbf{z})$, for some r_0 chosen so that $\sum_{l=1}^{r_0}\lambda_l/\sum_{l=1}^{\infty}\lambda_l \geqslant \mathfrak{p}$ and $\mathfrak{p} \in [0, 1]$ is a pre-specified threshold value.

For properly chosen kernels, ζ_l and λ_l are unknown, but can be estimated by employing a spectral decomposition for $\mathbb{K}$, $\mathbb{K} = \sum_{l=1}^{n}\widehat{\lambda}_l\widehat{\boldsymbol{\zeta}}_l\widehat{\boldsymbol{\zeta}}_l^\intercal$ where $\widehat{\lambda}_1 \geqslant \cdots \geqslant \widehat{\lambda}_n$ are the ordered eigenvalues and $\widehat{\boldsymbol{\zeta}}_l = [\widehat{\zeta}_l(\mathbf{Z}_1), ..., \widehat{\zeta}_l(\mathbf{Z}_n)]^\intercal$ is the eigenvector associated with $\widehat{\lambda}_l$. It has been shown that the eigenvalue and eigenvectors obtained based on $\mathbb{K}$ can be used to consistently estimate the underlying true eigenvalues and eigenfunctions corresponding to $K(\cdot,\cdot)$ (Koltchinskii and Giné, 2000; Braun, 2005). For a specified $\mathfrak{p}$, we choose the smallest possible r so that $\sum_{i=1}^{r}\widehat{\lambda}_i/\sum_{i=1}^{n}\widehat{\lambda}_i \geqslant \mathfrak{p}$, and approximate $\mathbb{K}$ by $\widetilde{\mathbb{K}}_r = \widetilde{\mathbb{B}}_r^{\intercal}\widetilde{\mathbb{B}}_r$, where $\widetilde{\mathbb{B}}_r = \left[\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{\zeta}}_1 \cdots \sqrt{\widehat{\lambda}_r}\widehat{\boldsymbol{\zeta}}_r\right]$, an $n \times r$ matrix. In many situations, r is significantly smaller than n, and fitting the AFT KM model can be simplified by reformulating it using the approximated primal form with $\widetilde{\mathbb{B}}_r$. Model (11), with $\widetilde{\mathbb{K}}_r$ in place of $\mathbb{K}$, becomes:

$$\log \mathbf{T} = \mathbf{D}\gamma + \widetilde{\mathbb{K}}_r\alpha + \mathbf{E}, \quad \alpha = \tau\epsilon, \quad E(\epsilon) = 0, \quad \mathrm{var}(\epsilon) = \widetilde{\mathbb{K}}_r^{-}$$

By applying the variable transformation $\beta = \widetilde{\mathbb{B}}_r\alpha$, this model simplifies to:

$$\log \mathbf{T} = \mathbf{D}\gamma + \widetilde{\mathbb{B}}_r\beta + \mathbf{E}, \quad E(\beta) = 0, \quad \mathrm{Var}(\beta) = \mathbb{I}_{r\times r}. \tag{15}$$

Mimicking the derivation based on $\mathbb{K}$, the score statistic with kernel PCA is

$$\widetilde{Q} = \widehat{\mathbf{R}}(\widetilde{\gamma}_G)^{\intercal}\widetilde{\mathbb{K}}_r\widehat{\mathbf{R}}(\widetilde{\gamma}_G) = \widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)^{\intercal}\widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)$$

where

$$\widetilde{\mathbf{U}}_G(\gamma) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\widetilde{\mathbf{B}}_{ri} - \widetilde{\mathbf{B}}_{rj})\Delta_i\mathbf{I}\{e_i(0;\gamma) \leqslant e_j(0;\gamma)\} \tag{16}$$

and $\widetilde{\mathbf{B}}_{ri}$ is the i^{th} row of $\widetilde{\mathbb{B}}_r$. The form of $\widetilde{\mathbf{U}}_G(\gamma)$ is again analogous to the standard Gehan WLR function (4);

3.3 *Approximating the Null Distribution of the Score Statistic*

In the appendix, we outline the derivation of the asymptotic distribution $\mathcal{Q}$ of $n\widehat{Q}$. The asymptotic distribution generally does not have an explicit form and hence is difficult to estimate explicitly. To approximate its null distribution in finite samples, we implement a perturbation approach. For ease of presentation, we focus on the kernel PCA based test statistic (16) which is most convenient for implementation, while noting that the original $\widehat{Q}$ corresponds to $\widetilde{Q}$ when $\mathfrak{p} = 1$.

Let $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_n)$, and let the $\{\mathcal{V}_i\}$ be iid random variables with mean 1 and variance 1. We first find $\widetilde{\gamma}_G^*$, the root of (8). Then, we calculate the perturbation $\widetilde{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*)$ of $\widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)$ as:

$$\widetilde{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\widetilde{\mathbf{B}}_{ri} - \widetilde{\mathbf{B}}_{rj}) \Delta_i \mathbf{I}\{e_i(0; \widetilde{\gamma}_G^*) \leqslant e_j(0; \widetilde{\gamma}_G^*)\} \mathcal{V}_i \mathcal{V}_j. \tag{17}$$

Using similar arguments as given in Jin et al. (2001), we can show that under H_0, the distribution of $n^{\frac{1}{2}} \widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)$ can be approximated by the distribution of $n^{\frac{1}{2}} \{\widetilde{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) - \widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)\}$ conditional on $\mathcal{O}$. Thus, the null distribution of $n\widetilde{Q}$ can be approximated by the distribution of $n\widetilde{Q}^* = n\widetilde{Q}^*(\widetilde{\gamma}_G^*) = n\{\widetilde{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) - \widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)\}^{\mathsf{T}} \{\widetilde{\mathbf{U}}_G^*(\widetilde{\gamma}_G^*) - \widetilde{\mathbf{U}}_G(\widetilde{\gamma}_G)\}$ given $\mathcal{O}$.

To calculate a p-value for $\widetilde{Q}$, we can compare it to its null distribution approximated by perturbations. We generate a large number of realizations of $\mathcal{V}$, say $\mathcal{V}_{(1)}, \ldots, \mathcal{V}_{(B)}$, and use them to generate $\widetilde{Q}_{(1)}^*, \ldots, \widetilde{Q}_{(B)}^*$ Then, for any fixed ρ, the p-value of the test can be obtained by

$$\widehat{p} = \#\{\widetilde{Q}_{(b)}^* \geqslant \widetilde{Q}\}/B.$$

Alternatively, we may use the Satterthwaite method to approximate the null distribution using a scaled χ^2 distribution, $c_0 \chi_{d_0}^2$. We estimate c_0 and d_0 by matching moments with the estimated null $\{\widetilde{Q}_{(b)}^*\}$, and calculate $\widehat{p}_{\chi^2} = 1 - F(\widetilde{Q})$, where F is the distribution function of a $\widehat{c}_0 \chi_{\widehat{d}_0}^2$ random variable. This approximation, previously shown to perform well for other models (Liu et al., 2007, 2008; Cai et al., 2011), also works well in our setting.

For the linear kernel, which does not rely on an additional parameter ρ, we may use the p-values $\widehat{p}_{\text{obs}} = \widehat{p}$ or $\widehat{p}_{\chi^2,\text{obs}} = \widehat{p}_{\chi^2}$ directly. For the quadratic and Gaussian kernels, the test statistic $\widetilde{Q}_{\text{obs}}$ may depend on the parameter ρ. Since the kernel matrix $\mathbb{K}(\rho)$ drops out of the model under the null, the parameter ρ is not estimable (Davies, 1987). Instead, we propose to use one of several test statistics:

$$\widehat{S}_{\mathcal{I}} = \sup_{\rho \in \mathcal{I}} \left\{ \widetilde{Q}(\rho)/\widehat{\sigma}(\rho) \right\}; \widehat{T}_{\mathcal{I},1} = \inf_{\rho \in \mathcal{I}} \{\widehat{p}(\rho)\}; \text{ or } \widehat{T}_{\mathcal{I},2} = \inf_{\rho \in \mathcal{I}} \{\widehat{p}_{\chi^2}(\rho)\}, \tag{18}$$

where $\widetilde{Q}(\rho)$, $\widehat{p}(\rho)$, and $\widehat{p}_{\chi^2}(\rho)$ denote respectively the test statistic, p-value from perturbation, and p-value from χ^2 approximation, derived under kernel function $K(\cdot,\cdot;\rho)$, $\widehat{\sigma}(\rho)$ is the estimated standard error of $\widetilde{Q}(\rho)$ obtained from the perturbations, and $\mathcal{I}$ is an appropriately chosen range for ρ. The final p-value is calculated by comparing the chosen test statistic to its approximate null distribution calculated using the perturbations. For example, in simulations we show $\widehat{p}_{\text{obs}} = \#\{\widehat{T}^*_{\mathcal{I},2,(b)} \leqslant \widehat{T}_{\mathcal{I},2,\text{obs}}\}/B$.

For the Gaussian and quadratic kernels, we determine the range $\mathcal{I}$ of ρ using kernel PCA. Specifically, let $\widehat{r}_\rho$ be the smallest r so that $\sum_{i=1}^r \widehat{\lambda}_i / \sum_{i=1}^n \widehat{\lambda}_i \geqslant \mathfrak{p}$; that is, $\widehat{r}_\rho$ is the number of eigenvectors needed to capture $\mathfrak{p}$ of the eigenvalues. For the Gaussian kernel, we let $\mathcal{I} = [\min\{\rho : \widehat{r}_\rho \leqslant \sqrt{n}\}, \max\{\rho : \widehat{r}_\rho \geqslant 3\}]$. For the quadratic kernel, we let $\mathcal{I} = [\min\{\rho : \widehat{r}_\rho \leqslant \sqrt{n} + p\}, \max\{\rho : \widehat{r}_\rho \geqslant \widehat{r}_{\text{lin}}\}]$, where $\widehat{r}_{\text{lin}}$ is the number of eigenvectors needed to explain $\mathfrak{p}$ of the eigenvalues in the linear kernel matrix.

3.4 *Estimation*

To construct a risk score $\gamma^\intercal \mathbf{D} + h(\mathbf{Z})$ for predicting T, we can minimize the penalized objective function (10) to get estimates for γ and α. The parameter c in (10) controls the smoothness of the resulting estimator for h. We may improve the estimation by employing kernel PCA and minimizing the ridge penalized Gehan-objective function with covariates $\mathbf{D}_i$ and $\widetilde{\mathbf{B}}_{ri}$,

$$L_G^R(\beta,\gamma) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \Delta_i |\widetilde{e}_i(\beta;\gamma) - \widetilde{e}_j(\beta;\gamma)|_+ + \frac{c}{2}\beta^\intercal \beta \tag{19}$$

where $\widetilde{e}_i(\beta;\gamma) = \log X_i - \gamma^\intercal \mathbf{D}_i - \beta^\intercal \widetilde{\mathbf{B}}_{ri}$. This formulation essentially uses kernel PCA to estimate the bases of the underlying feature space $\mathcal{H}_K$ and estimate h in its approximated primal form. An additional advantage of this approach comes from the computation efficiency since the dimension of β is often much smaller than n. Let $(\widehat{\gamma}^\intercal, \widehat{\beta}^\intercal)$ be the resulting estimators for $(\gamma^\intercal, \beta^\intercal)$. For a future subject with predictors $\mathbf{W}_0 = (\mathbf{D}_0^\intercal, \mathbf{Z}_0^\intercal)^\intercal$, we may construct a risk score using the Nyström approximation method (Rasmussen and Williams, 2006) as

$$\mathbf{D}_0^{\intercal}\widehat{\gamma}+\sum_{l=1}^{r}\widehat{\beta}_l\mathbf{K}_{\mathbf{Z}_0}^{\intercal}\widehat{\boldsymbol{\zeta}}_l\widehat{\lambda}_l^{-\frac{1}{2}},$$

where $\mathbf{K}_{\mathbf{Z}_0}=[K(\mathbf{Z}_0,\mathbf{Z}_1),\ldots,K(\mathbf{Z}_0,\mathbf{Z}_n)]^{\intercal}$.

For kernels that depend on a tuning parameter ρ, we use the value of ρ that minimized $\widehat{T}_{\mathcal{I},1}$ in (18). This choice is intuitively appealing because this ρ corresponds to $\mathbb{K}$ that produced the most evidence that $h(\mathbf{z})\neq 0$. We choose the shrinkage parameter c by cross-validation.

As in testing, the choice of kernel may not be obvious *a priori*. We consider two approaches for determining which kernel to use for the estimation of h. First, we apply our KM test to the pathway and use the kernel which yields to the smallest p-value to construct the estimate $\widehat{h}$. Alternatively, we estimate $\widehat{h}_L$, $\widehat{h}_Q$, and $\widehat{h}_G$ using the linear, quadratic, and Gaussian kernels, then choose one as our final estimate based on its prediction accuracy. For example, we could estimate the C-statistic for each one, using the method in Uno et al. (2011), and choose the estimate of h that maximizes this statistic.

4. General WLR and Selection of Kernel

4.1 *WLR KM Testing and Prediction with General Weights*

The performance of the testing and risk prediction made using the WLR with Gehan weights will vary depending on the underlying distributions of survival and censoring times. To optimize the power and prediction accuracy for a given dataset, it will be desirable to consider the WLR procedures with a general weight $\phi(\gamma,t)$. To this end, we first consider the estimation of h and γ iteratively, using an approach similar to Jin et al. (2003). We present results based on the kernel PCA approximation which is most convenient for computation, but results corresponding to the original kernel can be obtained by setting $\mathfrak{p}=1$. For any given initial estimator $\{\widehat{\gamma}^{(k-1)},\widehat{\beta}^{(k-1)}\}$, we may obtain an updated estimator for $\{\gamma,\beta\}$,

$\{\widehat{\gamma}^{(k)}, \widehat{\beta}^{(k)}\}$, as the minimizer of

$$\widetilde{L}_\phi(\gamma, \beta) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi\{\widehat{\gamma}^{(k-1)}, e_i(\widehat{\beta}^{(k-1)}; \widehat{\gamma}^{(k-1)})\} \Delta_i |e_i(\beta; \gamma) - e_j(\beta; \gamma)|_+ + \frac{c}{2} \beta^\intercal \beta \qquad (20)$$

Using similar arguments as given in Jin et al. (2003), we may show that as $k \to \infty$, $\{\widehat{\gamma}^{(k)}, \widehat{\beta}^{(k)}\}$ converges to $\{\widehat{\gamma}_\phi, \widehat{\beta}_\phi\}$, a solution to $\widetilde{\mathbf{U}}_\phi(\beta; \gamma) + c\beta = 0$, where

$$\widetilde{\mathbf{U}}_\phi(\beta; \gamma) = n^{-1} \sum_{i=1}^{n} \Delta_i \phi(\gamma, e_i(\beta; \gamma)) \left[\begin{pmatrix} \mathbf{D}_i \\ \widetilde{\mathbf{B}}_{ri} \end{pmatrix} - \frac{S^{(1)}(\beta, \gamma, e_i(\beta; \gamma))}{S^{(0)}(\beta, \gamma, e_i(\beta; \gamma))} \right]$$

$$\text{and} \quad S^{(m)}(\beta, \gamma; t) = n^{-1} \sum_{j=1}^{n} I\{e_j(\beta; \gamma) \geqslant t\} \begin{pmatrix} \mathbf{D}_j \\ \widetilde{\mathbf{B}}_{rj} \end{pmatrix}^{\otimes m}.$$

For estimation, we can start with $\{\widehat{\gamma}^{(0)}, \widehat{\beta}^{(0)}\} = \{\widehat{\gamma}_G, \widehat{\beta}_G\}$ where both are estimated as described in section (3.4), and iteratively minimize (20) until convergence. For testing $H_0 : h(\cdot) = 0$ with a general weight, we may derive the score test using the same arguments as given in section 3 but add weights $\psi\{\widehat{\gamma}^{(k-1)}, e_i(\widehat{\beta}^{(k-1)}; \widehat{\gamma}^{(k-1)})\}$ to (16), where $\{\widehat{\gamma}^{(k-1)}, \widehat{\beta}^{(k-1)}\}$ are some initial estimates of $\{\gamma, \beta\}$. When testing $h(\cdot) = 0$, convenient choices are $\widehat{\beta}^{(k-1)}$ is 0, its value under the null, and $\widehat{\gamma}^{(k-1)} = \widetilde{\gamma}_\phi$, the WLR estimator under H_0. This yields our general class of WLR KM test statistics

$$\widetilde{Q}_\phi = \widetilde{\mathbf{U}}_\phi(0; \widetilde{\gamma}_\phi)^\intercal \widetilde{\mathbf{U}}_\phi(0; \widetilde{\gamma}_\phi).$$

It is interesting to note that when $\phi = 1$, $\widetilde{Q}_\phi$ reduces to the KM score statistic proposed in Cai et al. (2011) in the absence of clinical covariates.

4.2 *Combining P-Values Across Models and Kernels*

In a real data analysis, it is unlikely that the researcher has a strong prior belief about which weights to use in the WLR estimating function, or which of several kernels will best capture the covariate feature space. Luckily, by generating the null distribution using perturbations, we can very easily combine information across weights and kernels by defining test statistics that combine p-values in the observed data, and generating null distributions for those test statistics by using perturbed nulls calculated from the same vectors $\mathcal{V}_{(b)}$ across tests. For example, for the WLR with Gehan weights, we can define a test statistic $\widehat{A}_{\text{obs}}$ combining p-values across the three kernels by:

$$\widehat{A}_{\text{obs}} = -2\left[\log \widehat{p}^{\,\text{Gaussian}}_{\chi^2,\text{obs}} + \log \widehat{p}^{\,\text{linear}}_{\chi^2,\text{obs}} + \log \widehat{p}^{\,\text{quadratic}}_{\chi^2,\text{obs}}\right],$$

and approximate its null distribution by computing, for $b = 1, \ldots, B$,

$$\widehat{A}^*_{(b)} = -2\left[\log \widehat{p}^{\,*\,\text{Gaussian}}_{\chi^2,\text{(b)}} + \log \widehat{p}^{\,*\,\text{linear}}_{\chi^2,\text{(b)}} + \log \widehat{p}^{\,*\,\text{quadratic}}_{\chi^2,\text{(b)}}\right].$$

Then our combined overall p-value is $\widehat{p}_{\text{combined}} = \#\{\widehat{A}^*_{(b)} \geqslant \widehat{A}_{\text{obs}}\}/B$. Similarly, we can combine p-values across different weights in the WLR. Since some of the candidate kernel or weights might be poor choices and hence the inclusion of them may result in a power loss. To overcome such difficulties, we propose the use of truncated product p-values (Zaykin et al., 2002); i.e., by replacing each $\log \widehat{p}$ by $I(\widehat{p} \leqslant p_0) \log \widehat{p}$ in the above test statistic and its null, where p_0 is a pre-specified p-value threshold.

5. Simulation Studies

5.1 *Testing*

We conducted simulation studies to assess the performance of the proposed testing and estimation procedures. We generated the pathway covariates $\mathbf{Z}$ from a multivariate normal distribution with mean 0 and compound symmetry covariance structure with variance 1 and correlation $\wp$. We considered pathways of size $P = 5$ and 20, and correlations $\wp = 0.8, 0.5$ and 0.2 to represent strong, moderate, and weak within-pathway correlation. For simplicity we did not include any additional covariates. We simulated different types of underlying signals $h(\mathbf{z})$ to understand the performance of our procedures in different settings. For a given $h(\mathbf{z})$, we generated survival times according to the model $\log T = h(\mathbf{Z}) + E$. We compared two different types of error distributions for E_i, generating E_i either from the extreme value distribution (EVD) or from an exponential distribution $\text{Exp}(\theta_E)$, where θ_E was chosen so that the medians of the two error distributions are approximately equal. The censoring was generated from a uniform distribution with range chosen so that approximately 25% of the individuals were censored.

To assess the performance of the testing procedure in each setting, we simulated 2000 data sets for empirical size and 1000 for empirical power, and compared the AFT weighted log-rank KM score test procedures with Gehan weights (WLR_{Ge}) and log-rank weights (WLR_{LR}) for the Gaussian, quadratic, and linear kernels. Note that the WLR_{LR} test is equivalent to the Cox KM test proposed by Cai et al. (2011) in this simulation setting without other covariates. Kernel PCA was used to reduce dimensionality, with $\mathfrak{p} = 0.95$. All null distributions were generated using $B = 2000$ perturbations. For the linear kernel which needs no tuning, we consider the perturbation p-value $\widehat{p}$ and the χ^2-approximated p-value $\widehat{p}_{\chi^2}$ directly. For the kernels which rely on a tuning parameter ρ, we considered the test based on the supremum statistic $\widehat{S}_{\mathcal{I}}$, as well as the tests based on the minimum perturbed and χ^2-approximated p-values, $\widehat{T}_{\mathcal{I},1}$ and $\widehat{T}_{\mathcal{I},2}$. The results for the different tests used for each kernel were nearly identical; those based on the χ^2- approximations are shown in the tables. We also considered five different combined p-values as described in section 4.2 with $p_0 = 0.05$: the two p-values combined across kernels within WLR weight; the three p-values combined across weights within kernel, and the p-value combined across all six weight-kernel combinations.

For comparison to our KM-based tests, we also considered a standard method for assigning a p-value to a pathway based on the individual associations between survival and each gene in the pathway. For each gene Z_i, $i = 1, \ldots, P$, in the pathway, we calculated a marginal p-value $\widehat{p}_i^{\text{marg}}$ from the Wald test from a standard univariate Cox model relating that gene to survival. Then we calculated the minimum of these P p-values, $\widehat{p}_{\min}^{\text{marg}} = \min\{\widehat{p}_1^{\text{marg}}, \ldots, \widehat{p}_P^{\text{marg}}\}$, and adjusted for multiple testing using the effective number of tests, M_{eff}, as described in

Nyholt (2004). That is, we define $\widehat{p}^{\text{marg}}_{\text{pathway}} = 1 - [1 - \widehat{p}^{\text{marg}}_{\text{min}}]^{M_{\text{eff}}}$, where $M_{\text{eff}} = 1 + (P-1)(1 - \text{Var}(\widehat{\ell}_1, \ldots, \widehat{\ell}_P)/P)$ and the $\widehat{\ell}_j$ are the observed eigenvalues from the covariance matrix of the genes in the pathway.

To examine the validity of the test procedure in finite samples, we generated data under the null setting $h(\mathbf{z}) = 0$. The empirical sizes at Type I error rate of 0.05 are shown in Table 1 for $n = 200$ and 400. When $n = 200$, the empirical sizes of the WLR_{Ge} tests tend to be slightly below the nominal level, while the empirical sizes of the WLR_{LR} tests are closer to the nominal level. The five combined p-values maintain their nominal level. The empirical sizes for the WLR_{Ge} tests are closer to their nominal level when the sample size is increased to $n = 400$.

To assess the power of the proposed tests, we considered (1) a linear signal, $h_1(\mathbf{z}) = c(z_1 + z_2 + z_3 + z_4 + z_5)$ with $c = 0.05$, and (2) a nonlinear signal, $h_2(\mathbf{z}) = c[z_1 + 4z_1^2 + z_2 + 4z_2^2 - 2z_1z_2 + g(z_3)(4z_4 + 4z_5) + (1 - g(z_3))(-3z_4 - 3z_5 + 4z_4z_5)]$, with $c = 1.5$ and $g(z) \sim \text{Bernoulli}(e^{-|z|})$. This function was chosen to have different types of nonlinear signal: the linear, quadratic, and interactive effects of z_1 and z_2, and the latent classes defined by z_3 with differential signal defined by z_4 and z_5. Results are shown in Table 2.

When the true signal is linear, one would expect the linear kernel to outperform the other kernels, but interestingly, all three kernels yield tests with competitive power. This can in part be attributed to the fact that both the Gaussian and quadratic kernels can capture primarily linear effects at certain values of their tuning parameters. For all tests, the power decreases somewhat when we increase the number of covariates from $P = 5$ to $P = 20$, as we would expect. However, the power loss is small when the correlation is high due to the low effective degrees of freedom in such settings. This highlights one of the advantages of the KM based tests. The tests combined across kernels within models perform well, usually exhibiting power close to that of the most powerful kernel. This suggests that such combined tests could be quite robust and useful for a wide range of settings. The tests based on marginal procedures tend to have less power compared to the score tests in most settings, particularly when compared to the omnibus tests. The Cox model holds when the error comes from the extreme value distribution, and in that setting the $\mathrm{WLR}_{\mathrm{LR}}$ test outperforms the $\mathrm{WLR}_{\mathrm{Ge}}$ equivalent. The reverse is true when the error comes from an exponential distribution.

When the true signal is nonlinear, the linear kernel performs poorly, particularly in the high correlation setting. In comparison, the quadratic kernel has somewhat higher power, while the Gaussian kernel has sometimes much higher power. The tests combined across kernels within models are also quite competitive, though in this setting are almost uniformly beat by the Gaussian kernel. The Gaussian kernel tests and the test combined across models and weights tend to outperform the marginal gene methods. In other settings when the quadratic effect is strong (results not shown), the quadratic kernel based tests substantially outperform tests based on other kernels, and the omnibus test performs robustly in most cases without significant power loss compared to that of the best kernel. However, when $P = 20$ and correlation is 0.8, the omnibus test pays a substantial price for selecting the optimal weight and kernel.

5.2 *Estimation*

To assess the performance of the estimation procedure when the true signal is linear (h_1 from above, with $c = 0.4$) or nonlinear (h_2 with $c = 2$), we ran 500 simulations in the same configurations of correlations, pathway sizes, and error distributions. Each simulation consisted of two data sets: a training data set with sample size $n_{\mathrm{train}} = 100$ on which to build the estimates $\widehat{h}$, and a validation data set with sample size $n_{\mathrm{test}} = 1000$ on which to assess the predictive performance of each estimate $\widehat{h}$. To assess predictive ability, we use the C-statistic, which quantifies how well $\widehat{h}$ predicts survival up until some time t_0 (Uno et al., 2011). In these simulations, we selected a reference time t_0 that was near the 70$^{\text{th}}$ percentile of followup time.

Using the training data, we built estimators $\widehat{h}_L, \widehat{h}_Q$, and $\widehat{h}_G$ from the linear, quadratic, and Gaussian kernels, working in only the WLR_{Ge} for simplicity. We fit standard full Cox and AFT models in each training data set for comparison. We applied these estimators and models to the validation data and calculated the C-statistic. We considered two methods of choosing which kernel to use for the final estimate $\widehat{h}$; in one, we choose the kernel with the smallest p-value from the pseudo-score test in the training data; in the other, we choose the kernel yielding the largest estimated C-statistic in the training data. These two methods yielded almost identical response, so only results from kernel selection using the C-statistic in the training data are shown. Results are presented in Table 3.

In the linear setting, the C-statistics are quite similar across the three kernels. The kernel models tend to have slightly better predictive ability than the full models due to the use of kernel PCA, and using the kernel selected by the C-statistic tends to pick the model with optimal predictive ability. Here, choosing kernel based on the estimated C-statistic in the training data yields good results. In the nonlinear setting, we see huge gains over the full model by using nonlinear kernel functions. In this setting, the predictive ability of the three kernels varies more, so it is more meaningful to select a kernel. The model using the selected kernel has almost uniformly better predictive ability than all other models. Thus, we recommend selecting kernel based on the estimated C-statistic.

6. Example: Breast Cancer Gene Expression Study

Genomic information has already improved our understanding of breast cancer. The mutations found in the BRCA1 and BRCA2 genes identify women at high risk of developing breast cancer (Narod and Foulkes, 2004), and a number of gene expression signatures have been introduced into clinical practice to better identify cancers with high and low risk of recurrence (Desmedt et al., 2011). Despite recent progress 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 interested in assessing the association between recurrence-free survival and 32 candidate pathways from the molecular signature database, such as AKAP13, EGFR_SMRTE and p53. Overexpression of cAMP-dependent protein kinase A (PKA) is a hallmark of many human cancers including breast cancer. EGFR is a receptor tyrosine kinase expressed in a wide variety of epithelial malignancies (Kuwahara et al., 2004; Nicholson et al., 2001). p53 mutation has been previously found to be associated with more aggressive disease and worse overall survival in breast cancer (Gasco et al., 2002).

The 32 pathways range from 7 to 238 genes with a median of 24 genes. To assess the effects of these pathways on breast cancer progression, we applied our pseudo-score test to each of the pathways in a training set of 286 lymph node negative breast cancer patients from the Netherlands with gene expression on an Affymetrix U133a Gene Chip (Wang et al., 2005). A total of 107 recurrences were observed, with follow-up time ranging between 2 months and 14.3 years (median 7.2 years); 63% of observations were censored. For illustration of our method, we implement the test with multiple kernels in the $\mathrm{WLR}_{\mathrm{Ge}}$ framework. Figure 1 shows the results of the testing procedure. For each pathway, we compare the p-value from the marginal gene based method to the p-values for the $\mathrm{WLR}_{\mathrm{Ge}}$ KM test with Gaussian kernel and the $\mathrm{WLR}_{\mathrm{Ge}}$ test combined across Gaussian, linear, and quadratic kernels. 66% of pathways are significant at the nominal 0.05 level using the marginal gene method; 81% are significant using the $\mathrm{WLR}_{\mathrm{Ge}}$ test with Gaussian kernel, and 88% using the $\mathrm{WLR}_{\mathrm{Ge}}$ test combined across kernels. In 26 pathways, both $\mathrm{WLR}_{\mathrm{Ge}}$ tests have p-values smaller than that from the marginal approach. These results suggest that the $\mathrm{WLR}_{\mathrm{Ge}}$ KM tests could potentially have more power compared to the marginal approach.

For one of the pathways declared extremely significant using our $\mathrm{WLR}_{\mathrm{Ge}}$ KM test, the

TNFR1 pathway consisting of 28 genes, we estimated the pathway effect on recurrence-free survival, and applied that estimate to an independent validation set of 149 lymph node negative breast cancer patients from Sweden and the UK, also with gene expression on a U133a Gene Chip (Sotiriou et al., 2006). 33 recurrences were observed, with follow-up time ranging between 2 months and 14.5 years (median 7.3 years); 77% of observations were censored.

Fitting the standard Cox model to this pathway yielded a C-statistic for recurrence-free survival up to 5 years of 59% [95% CI 47 - 70%]. The AFT KM estimation procedure selected the quadratic kernel based on the estimated C-statistic in the training data; when applied to the validation set, the estimated C-statistic was 65% [95% CI 54 - 76%]. These results suggest potential for improvement in risk prediction using AFT KM modeling.

7. Discussion

In this paper, we proposed KM based procedures to test for the effect of pathways on survival in an AFT model framework. Our proposed test could have high power in capturing non-linear effects compared to standard procedures based on linearity assumptions. When interest lies in risk prediction, our proposed risk scores from the AFT KM framework may be easily used for risk assessment. These risk scores, by capturing non-linear effects, could have higher prediction accuracy compared to those derived under linear assumptions. This is indeed reflected by the results from our simulation studies when the underlying effects are truly non-linear.

In analysis of real data, it is unlikely that the researcher knows which weight or kernel is most appropriate for his or her data set. Thus, we have proposed omnibus testing procedures that allow the data to choose the weight and kernel automatically. The perturbation procedures used to generate the null distributions for our AFT KM tests enable us to efficiently combine information across kernels and weights. Based on simulation results for testing, the power lost is often minimal, but the power gained over the worst choice can be quite large. In some settings with larger P, there does appear to be a substantial power loss due to searching for the optimal weight and kernel. Optimal combination of tests in these more difficult settings warrants further research. On the other hand, for risk prediction, it appears that the omnibus estimation procedure has negligible loss in prediction accuracy when compared to the optimal kernel.

We were motivated to develop these kernel machine methods for the AFT model because of an attractive property of linear models that when two groups of covariates are independent, their marginal effects on outcome are equal to their joint effects. In the context of gene-sets, this would mean that if we are considering high-dimensional genetic data which we can divide into a large number of independent pathways, we may assess the effect of each pathway individually, and then combine their marginal effects additively into a joint model. Future work will explore this application to large data sets, and investigate our abilities to combine information across both independent and correlated pathways. When the underlying signal is sparse with only a few genes in a pathway associated with survival, it would also be interesting to extend our proposed procedure to allow for feature selection under the KM framework.

8. Appendix: Asymptotic Distribution for the Test Statistic

In this section, we derive the asymptotic null distribution of our test statistic. Throughout, we assume that the covariates $\mathbf{D}_i$ and the genomic marker values $\mathbf{Z}_i$ are bounded by $\mathbf{z}_m$. We assume that the true value of γ, γ_0, is an interior point of a compact set Ω and without loss of generality, we also assume that X has a finite support $[0, \tau]$. For the ease of notation, unless noted otherwise, the supremum is always taken over $(-\infty, \log(\tau)]$ for the index t, $[-\mathbf{z}_m, \mathbf{z}_m]$ for $\mathbf{Z}$ and Ω for γ. We also require the same set of assumptions given in Jin et al. (2003) for the iterative WLR estimation procedures. For simplicity, we focus on the Gehan weight but note that similar arguments can be used for the general weight. From Jin et al. (2001, 2003), we have $n^{\frac{1}{2}}(\widetilde{\gamma}-\gamma_0) = n^{-\frac{1}{2}}\sum_{i=1}^n \mathcal{U}_{\gamma i} + o_p(1)$ for some independent and identically distributed random variables $\mathcal{U}_{\gamma i}$, where $\widetilde{\gamma} = \widetilde{\gamma}_G$ in the text. We will derive the null distribution of $\widehat{Q}(\widetilde{\gamma};\rho)$ and demonstrate its convergence as a process in ρ. The kernel function K is assumed to be continuously differentiable and the regularity conditions required in Braun (2005) are assumed to hold for the convergence of the empirical eigenvalues and eigenfunctions of $\mathcal{H}_K$. Under these regularity conditions, the convergence of the null distribution of $\widetilde{Q}(\widetilde{\gamma};\rho)$ can be derived using the convergence of $\widetilde{\mathbb{K}}$ to the kernel matrix corresponding to a truncated kernel, which spans $\mathcal{H}_{K_{r_0}} = \text{span}\{\mathscr{B}_1(\cdot),...,\mathscr{B}_{r_0}(\cdot)\}$ (Braun, 2005).

The test statistic in section 3.1 takes the form $\widehat{Q}(\widetilde{\gamma},\rho) = \widehat{\mathbf{R}}(\widetilde{\gamma})^{\mathsf{T}}\mathbb{K}(\rho)\widehat{\mathbf{R}}(\widetilde{\gamma})$, where $\widehat{\mathbf{R}}(\gamma) = (\widehat{R}_1(\gamma),\ldots,\widehat{R}_n(\gamma))^{\mathsf{T}}$,

$$\widehat{R}_i(\gamma) = n^{-2}\sum_{j=1}^n\left\{\int Y_j(t;\gamma)dN_i(t;\gamma) - \int Y_i(t;\gamma)dN_j(t;\gamma)\right\},$$

$N_i(t;\gamma) = \Delta_i\mathbf{I}[e_i(0;\gamma)\leqslant t]$ and $Y_i(t;\gamma) = I[e_i(0;\gamma)\geqslant t]$. We can write $n\widehat{Q}(\widetilde{\gamma},\rho)$ as:

$$\begin{aligned} n\widehat{Q}(\widetilde{\gamma},\rho) &= n\sum_{i=1}^n\sum_{j=1}^n K(\mathbf{Z}_i,\mathbf{Z}_j;\rho)\widehat{R}_i(\widetilde{\gamma})\widehat{R}_j(\widetilde{\gamma}) \\ &= \int\int K(\mathbf{u},\mathbf{v};\rho)d\left\{\sqrt{n}\,\widehat{\mathbb{W}}_R(\mathbf{u};\widetilde{\gamma})\right\}d\left\{\sqrt{n}\,\widehat{\mathbb{W}}_R(\mathbf{v};\widetilde{\gamma})\right\}. \end{aligned}$$

where
$$\begin{aligned} \widehat{\mathbb{W}}_R(\mathbf{u};\widetilde{\gamma}) &= \sum_{i=1}^n \mathbf{I}[\mathbf{Z}_i\leqslant\mathbf{u}]\widehat{R}_i(\widetilde{\gamma}). \\ &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n\mathbf{I}[\mathbf{Z}_i\leqslant\mathbf{u}]\left\{\int Y_j(t;\widetilde{\gamma})dN_i(t;\widetilde{\gamma}) - \int Y_i(t;\widetilde{\gamma})dN_j(t;\widetilde{\gamma})\right\} \\ &= \int\widehat{\pi}(\mathbf{z}_m,t;\widetilde{\gamma})\widehat{\mathbf{W}}_N(\mathbf{u},dt;\widetilde{\gamma}) - \int\widehat{\pi}(\mathbf{u},t;\widetilde{\gamma})\widehat{\mathbf{W}}_N(\mathbf{z}_m,dt;\widetilde{\gamma}), \\ \widehat{\pi}(\mathbf{u},t;\gamma) &= n^{-1}\sum_{i=1}^n\mathbf{I}[\mathbf{Z}_i\leqslant\mathbf{u}]Y_i(t;\gamma), \qquad \widehat{\mathbf{W}}_N(\mathbf{u},t;\gamma) = n^{-1}\sum_{i=1}^n\mathbf{I}[\mathbf{Z}_i\leqslant\mathbf{u}]N_i(t;\gamma), \end{aligned}$$

and $\mathbf{I}(\mathbf{Z}_i\leqslant\mathbf{u}) = I(Z_{i1}\leqslant u_1,...,Z_{ip}\leqslant u_p)$. Let $\pi(\mathbf{u},t;\gamma) = E\left[\mathbf{I}[\mathbf{Z}\leqslant\mathbf{u}]Y(t;\gamma)\right]$, $\mu(\mathbf{u},t;\gamma) = E\left[\mathbf{I}[\mathbf{Z}\leqslant\mathbf{u}]N(t,\gamma)\right]$, $\widehat{\mathbb{W}}_N(\mathbf{u},t;\gamma) = \widehat{\mathbf{W}}_N(\mathbf{u},t;\gamma) - \mu(\mathbf{u},t;\gamma)$, and $\widehat{\mathbb{W}}_\pi(\mathbf{u},t;\gamma) = \widehat{\pi}(\mathbf{u},t;\gamma) -$

$\pi(\mathbf{u}, t; \gamma)$. Then we can expand $\sqrt{n}\,\widehat{\mathbb{W}}_R(\mathbf{u}; \widetilde{\gamma})$ as:

$$\int \widehat{\mathbb{W}}_\pi(\mathbf{z}_m, t; \widetilde{\gamma}) \left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, dt; \widetilde{\gamma})\right\} - \int \widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \widetilde{\gamma}) \left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{z}_m, dt; \widetilde{\gamma})\right\} \tag{21}$$

$$+ \int \pi(\mathbf{z}_m, t; \widetilde{\gamma}) \left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, dt; \widetilde{\gamma})\right\} - \int \pi(\mathbf{u}, t; \widetilde{\gamma}) \left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{z}_m, dt; \widetilde{\gamma})\right\} \tag{22}$$

$$+ \int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{z}_m, t; \widetilde{\gamma})\right\} \mu(\mathbf{u}, dt; \widetilde{\gamma}) - \int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \widetilde{\gamma})\right\} \mu(\mathbf{z}_m, dt; \widetilde{\gamma}) \tag{23}$$

$$+ \sqrt{n}\left\{\int \pi(\mathbf{z}_m, t; \widetilde{\gamma})\mu(\mathbf{u}, dt; \widetilde{\gamma}) - \int \pi(\mathbf{u}, t; \widetilde{\gamma})\mu(\mathbf{z}_m, dt; \widetilde{\gamma})\right\} \tag{24}$$

We first show that the first pair of integrals (21) is $o_p(1)$. To this end, we note that by a functional central limit theorem (FCLT) (Pollard, 1990), $\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, t, \gamma)$ converges weakly to a Gaussian process in $(\mathbf{u}, t, \gamma)$, denoted by $\mathbb{W}_N(\mathbf{u}, t; \gamma)$. It follows that

$$\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, t; \widetilde{\gamma}) = \sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, t, \gamma_0) + o_p(1) \tag{25}$$

by stochastic equicontinuity. On the other hand, by a uniform law of large numbers (ULLN) (Pollard, 1990), $\sup_{\mathbf{u},t,\gamma} |\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \gamma)| = o_p(1)$ which implies that $\sup_{\mathbf{u},t} |\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \widetilde{\gamma})| = o_p(1)$. This, together with Lemma A.3 of Bilias et al. (1997) and the strong representation theorem, implies that $(21) = o_p(1)$ uniformly in $\mathbf{u}$.

The integrals in (22) have the same limiting distribution as

$$\int \pi(\mathbf{z}_m, t; \gamma_0) \left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, dt; \gamma_0)\right\} - \int \pi(\mathbf{u}, t; \gamma_0) \left\{\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{z}_m, dt; \gamma_0)\right\}.$$

We may see this by adding and subtracting $\pi(\mathbf{u}, t; \gamma_0)$ to and from the integrands $\pi(\mathbf{u}, t; \widetilde{\gamma})$, and using the fact that $\pi(\mathbf{u}, t; \widetilde{\gamma}) - \pi(\mathbf{u}, t; \gamma_0) \xrightarrow{P} 0$, and by using (25) to replace the integrating functions by $\sqrt{n}\,\widehat{\mathbb{W}}_N(\mathbf{u}, dt; \gamma_0)$.

The integrals in (23) have the same limiting distribution as

$$\int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{z}_m, t; \gamma_0)\right\} \mu(\mathbf{u}, dt; \gamma_0) - \int \left\{\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \gamma_0)\right\} \mu(\mathbf{z}_m, dt; \gamma_0)$$

To see this, note that a FCLT implies that $\sqrt{n}\,\widehat{\mathbb{W}}_\pi(\mathbf{u}, t, \gamma)$ converges weakly to a zero-mean Gaussian process $\mathbb{W}_\pi(\mathbf{u}, t, \gamma)$. Hence, $\sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \widetilde{\gamma}) = \sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \gamma_0) + o_p(1)$, which allows us to replace the integrands $\sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \widetilde{\gamma})$ by $\sqrt{n}\widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \gamma_0)$. Further, we can expand

$$\begin{aligned}\int \left\{\sqrt{n}\, \widehat{\mathbb{W}}_\pi(\mathbf{u}_1, t; \gamma_0)\right\} \mu(\mathbf{u}_2, dt; \widetilde{\gamma}) \;&=\; \int \left\{\sqrt{n}\, \widehat{\mathbb{W}}_\pi(\mathbf{u}_1, t; \gamma_0)\right\} \mu(\mathbf{u}_2, dt; \gamma_0) &(26)\\ &\quad + \int \left\{\sqrt{n}\, \widehat{\mathbb{W}}_\pi(\mathbf{u}_1, t; \gamma_0)\right\} \{\mu(\mathbf{u}_2, dt; \widetilde{\gamma}) - \mu(\mathbf{u}_2, dt; \gamma_0)\} &(27)\end{aligned}$$

and see that (27) $= o_p(1)$ by another application of Lemma A.3 of Bilias et al. (1997) and the strong representation theorem, because $\mu(\mathbf{u}, dt; \widetilde{\gamma}) - \mu(\mathbf{u}, dt; \gamma_0) \xrightarrow{P} 0$.

Finally, we can write (24) as:

$$\begin{aligned}\sqrt{n}\Bigg\{ \quad & \int \pi(\mathbf{z}_m, t; \gamma_0)\mu(\mathbf{u}, dt; \gamma_0) - \int \pi(\mathbf{u}, t; \gamma_0)\mu(\mathbf{z}_m, dt; \gamma_0) &(28)\\ & + \int \pi(\mathbf{z}_m, t; \widetilde{\gamma}) \{\mu(\mathbf{u}, dt; \widetilde{\gamma}) - \mu(\mathbf{u}, dt; \gamma_0)\} &(29)\\ & + \int \{\pi(\mathbf{z}_m, t; \widetilde{\gamma}) - \pi(\mathbf{z}_m, t; \gamma_0)\} \mu(\mathbf{u}, dt; \gamma_0) &(30)\\ & - \int \pi(\mathbf{u}, t; \widetilde{\gamma}) \{\mu(\mathbf{z}_m, dt; \widetilde{\gamma}) - \mu(\mathbf{z}_m, dt; \gamma_0)\} &(31)\\ & - \int \{\pi(\mathbf{u}, t; \widetilde{\gamma}) - \pi(\mathbf{u}, t; \gamma_0)\} \mu(\mathbf{z}_m, dt; \gamma_0) \quad \Bigg\} &(32)\end{aligned}$$

The first line (28) is identically 0 because $\int \mathbf{I}[\mathbf{Z} \leqslant \mathbf{u}]Y(s; \gamma_0)\lambda_0(s)ds$ is the compensator of $\mathbf{I}[\mathbf{Z} \leqslant \mathbf{u}]N(s; \gamma_0)$, where $\lambda_0(s)$ is the common hazard function of E_i so that line (28) is exactly:

$$\int E\left[Y(s; \gamma_0)\right] E\left[\mathbf{I}[\mathbf{Z} \leqslant \mathbf{u}]Y(s; \gamma_0)\right] \lambda_0(s)ds - \int E\left[\mathbf{I}[\mathbf{Z} \leqslant \mathbf{u}]Y(s; \gamma_0)\right] E\left[Y(s; \gamma_0)\right] \lambda_0(s)ds = 0.$$

Then, it follows from a Taylor series expansion and the expansion of $n^{\frac{1}{2}}(\widetilde{\gamma} - \gamma_0)$,

$$(24) = \sqrt{n}(\widetilde{\gamma} - \gamma_0)^{\intercal}\mathbf{A} + o_p(1) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{A}^{\intercal}\mathcal{U}_{\gamma i} + o_p(1),$$

where $\mathbf{A} = \int \pi(\mathbf{z}_m, t; \gamma_0)\dot{\mu}(\mathbf{z}_m, dt, \gamma_0) + \int \dot{\pi}(\mathbf{z}_m, t, \gamma_0)\mu(\mathbf{u}, dt; \gamma_0) - \int \pi(\mathbf{u}, t; \gamma_0)\dot{\mu}(\mathbf{z}_m, dt, \gamma_0) - \int \dot{\pi}(\mathbf{u}, t, \gamma_0)\mu(\mathbf{z}_m, dt; \gamma_0)$, $\dot{\mu}(\mathbf{z}, t, \gamma) = \partial\mu(\mathbf{z}, t, \gamma)/\partial\gamma$ and $\dot{\pi}(\mathbf{z}, t, \gamma) = \partial\pi(\mathbf{z}, t, \gamma)/\partial\gamma$.

Putting all the aforementioned expansions together, we have $\sqrt{n}\, \widehat{\mathbb{W}}_R(\mathbf{u}; \widetilde{\gamma})$ asymptotically equivalent to

$$\bar{\mathbb{W}}(\mathbf{u}) = \sqrt{n} \int \Big\{ \pi(\mathbf{z}_m, t; \gamma_0) \widehat{\mathbb{W}}_N(\mathbf{u}, dt; \gamma_0) - \pi(\mathbf{u}, t; \gamma_0) \widehat{\mathbb{W}}_N(\mathbf{z}_m, dt; \gamma_0) \\ + \widehat{\mathbb{W}}_\pi(\mathbf{z}_m, t; \gamma_0) \mu(\mathbf{u}, dt; \gamma_0) - \widehat{\mathbb{W}}_\pi(\mathbf{u}, t; \gamma_0) \mu(\mathbf{z}_m, dt; \gamma_0) \Big\} + n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{A}^\intercal \mathcal{U}_{\gamma i}.$$

It then follows from another application of FCLT that $\bar{\mathbb{W}}(\mathbf{u})$ converges weakly to a zero mean Gaussian process $\mathbb{W}(\mathbf{u})$. This together with the smoothness of K, implies that $n\widehat{Q}(\widetilde{\gamma}, \rho)$ converges weakly as a process to

$$\mathcal{Q}(\rho) = \int \int K(\mathbf{u}, \mathbf{v}; \rho) d\mathbb{W}_R(\mathbf{u}) d\mathbb{W}_R(\mathbf{v}).$$

[Figure 1 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

References

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

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning* **66,** 259–294.

Braun, M. (2005). *Spectral properties of the kernel matrix and their application to kernel methods in machine learning*. PhD thesis, PhD thesis, University of Bonn.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429.

Cai, T., Tonini, G., and Lin, X. (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics* .

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

Desmedt, C., Michiels, S., Haibe-Kains, B., Loi, S., and Sotiriou, C. (2011). Time to move forward from first-generation prognostic gene signatures in early breast cancer. *Breast Cancer Research and Treatment* pages 1–3.

Fortunel, N., Otu, H., Ng, H., Chen, J., Mu, X., Chevassut, T., Li, X., Joseph, M., Bailey, C., Hatzfeld, J., et al. (2003). Comment on" 'stemness': transcriptional profiling of embryonic and adult stem cells" and" a stem cell molecular signature"(i). *Science* **302,** 393.

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

Jin, Z., Lin, D., Wei, L., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90,** 341.

Jin, Z., Ying, Z., and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88,** 381.

Kalbfleisch, J., Prentice, R., and Kalbfleisch, J. (1980). *The statistical analysis of failure* *time data*, volume 5. Wiley New York.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41,** 495–502.

Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* pages 113–167.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics* pages 1276–1288.

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.

Li, H. and Luan, Y. (2003). Kernel cox regression models for linking gene expression profiles to censored survival data. *Pac Symp Biocomput* pages 65–76.

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

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* **63,** 1079–88.

Liu, Z., Chen, D., Tan, M., Jiang, F., and Gartenhaus, R. B. (2010). Kernel based methods for accelerated failure time model with ultra-high dimensional data. *BMC Bioinformatics* **11,** 606.

Mika, S., Schölkopf, B., Smola, A., Müller, K., Scholz, M., and Rätsch, G. (1999). Kernel pca and de-noising in feature spaces. *Advances in neural information processing systems* **11,** 536–542.

Mootha, V., Lindgren, C., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34,** 267–273.

Narod, S. and Foulkes, W. (2004). Brca1 and brca2: 1994 and beyond. *Nature Reviews Cancer* **4,** 665–676.

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.

Nyholt, D. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics* **74,** 765–769.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR.

Rasmussen, C. and Williams, C. (2006). Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA* **38,** 715–719.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics* pages 303–328.

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

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* **98,** 262.

Tsiatis, A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* pages 354–372.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* **30,** 1105–17.

Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365,** 671–679.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat Med* **11,** 1871–9.

Wu, M., Lee, S., Tianxi, C., Yun, L., Boehnke, M., and Xihong, L. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* **89,** 82–93.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining p-values. *Genet Epidemiol* **22,** 170–85.

Zeng, D. and Lin, D. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association* **102,** 1387–1396.

Figure 1. $\log_{10}$ p-values for testing the overall effect of 32 pathways on breast cancer survival. Black squares represent the pathway p-value from the marginal-gene approach, and the pathways are ordered by this p-value. Blue filled diamonds represent the p-values from the WLR_{Ge} test with the Gaussian kernel, and red unfilled diamonds represent the p-value from the WLR_{Ge} test combined across all three kernels. Results are based on $B = 5000$ perturbations.

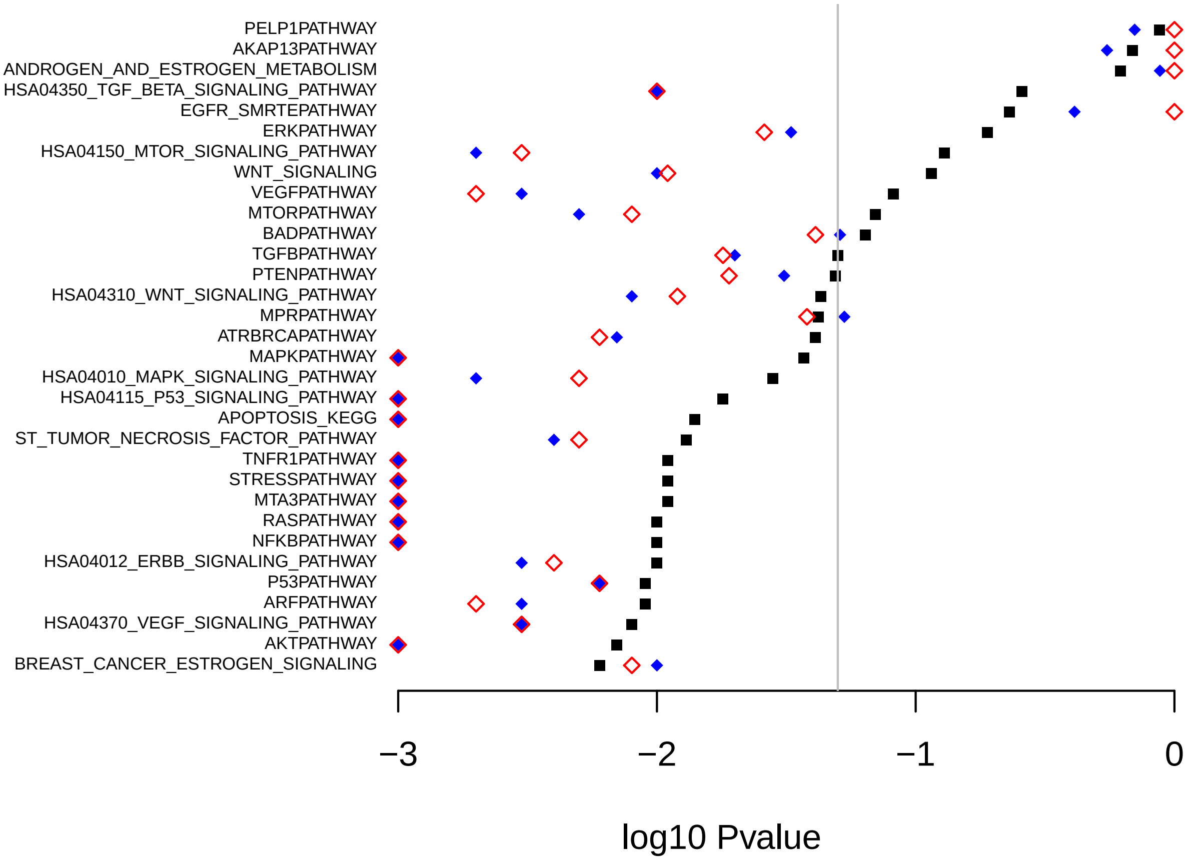

Table 1

Empirical sizes at Type I error rate of 0.05 when $n = 200$. Testing was performed using the Gaussian, quadratic, and linear kernels using both the WLR_{Ge} and WLR_{LR} tests. For the linear kernel, the χ^2 approximated p-value is presented; for the Gaussian and quadratic kernels, the perturbed χ^2 $\min p$ statistic ($\widehat{T}_{2,\mathcal{I}}$ in the text) was used. Results shown use kernel PCA with 95% of eigenvalues used. Also shown empirical sizes of the omnibus tests: across the kernels within each weight (Gehan or Log-Rank), then across weights within each kernel, and finally across all 6 weight-kernel combinations. Shown for comparison are results for the marginal gene method.

Correlation		0.2				0.5				0.8			
Pathway Size		5		20		5		20		5		20	
Error		EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp
						$n = 200$							
Marginal Gene		6	6	5	6	6	6	5	6	7	7	5	5
WLR_{Ge}	Gaussian	4	4	4	3	5	4	4	4	5	4	4	4
	Quadratic	4	4	3	3	5	4	4	4	5	4	4	4
	Linear	4	4	3	3	5	4	4	4	5	4	4	4
	All	4	4	3	3	5	4	4	4	5	4	4	4
WLR_{LR}	Gaussian	5	5	4	5	6	5	5	5	6	6	6	5
	Quadratic	5	5	4	4	6	5	5	4	6	5	5	5
	Linear	6	5	4	5	6	5	5	5	6	5	5	5
	All	5	5	4	5	6	5	5	5	6	5	6	5
Both	Gaussian	5	4	4	4	5	4	4	4	5	4	5	5
	Quadratic	4	4	3	4	5	4	4	4	5	5	5	4
	Linear	5	4	3	4	5	4	4	4	5	5	4	5
	All	5	4	4	4	5	4	4	4	6	5	5	5
						$n = 400$							
Marginal Gene		5	6	5	5	5	6	5	5	6	7	5	4
WLR_{Ge}	Gaussian	5	5	4	4	5	4	5	4	5	4	4	4
	Quadratic	5	4	4	4	5	5	4	4	5	5	4	4
	Linear	5	4	4	3	5	5	5	4	5	5	4	4
	All	5	5	4	3	5	5	4	4	5	5	4	4
WLR_{LR}	Gaussian	5	5	5	5	5	6	5	5	6	5	5	5
	Quadratic	5	5	4	5	5	6	5	5	5	5	5	5
	Linear	5	6	4	5	5	5	6	5	5	6	5	5
	All	5	5	4	5	5	6	5	5	5	6	4	5
Both	Gaussian	5	5	4	4	5	5	5	4	5	5	4	4
	Quadratic	5	5	4	4	5	5	4	5	5	5	4	4
	Linear	5	5	4	4	5	5	5	4	5	5	4	5
	All	5	5	4	4	5	5	5	5	5	5	4	5

Table 2

Empirical power (%) for linear and nonlinear signal at the Type I error rate of 0.05 when $n = 200$. Tests shown are identical to those in Table 1.

Correlation		0.2				0.5				0.8			
Pathway Size		5		20		5		20		5		20	
Error		EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp
						linear $h(\mathbf{z})$							
Marginal Gene		26	25	20	19	57	44	50	37	80	65	77	61
WLR_{Ge}	Gaussian	22	57	20	44	47	78	46	74	57	87	62	90
	Quadratic	22	55	17	40	48	80	43	72	65	92	65	90
	Linear	22	55	18	41	50	80	46	73	65	92	65	91
	All	21	55	18	42	49	80	45	73	64	91	64	90
WLR_{LR}	Gaussian	34	31	29	26	62	48	57	44	74	58	77	61
	Quadratic	33	31	26	24	63	49	55	42	79	64	79	63
	Linear	34	31	28	25	64	50	59	44	80	65	81	65
	All	34	31	28	25	63	49	57	43	79	64	80	63
Both	Gaussian	30	48	26	37	57	71	54	66	69	82	74	84
	Quadratic	28	48	23	34	58	72	52	63	76	87	76	86
	Linear	30	48	24	35	60	74	54	66	76	87	77	86
	All	30	48	24	35	60	73	54	66	75	86	76	87
						nonlinear $h(\mathbf{z})$							
Marginal Gene		51	50	38	39	41	36	24	28	33	34	23	24
WLR_{Ge}	Gaussian	99	100	42	42	100	100	43	46	100	100	76	74
	Quadratic	84	84	42	41	72	74	49	51	22	21	38	37
	Linear	49	49	38	38	31	30	22	24	5	6	4	5
	All	97	97	41	40	98	97	44	46	97	96	61	58
WLR_{LR}	Gaussian	95	95	32	32	93	96	22	28	94	94	29	33
	Quadratic	82	81	34	35	54	56	29	32	16	15	18	22
	Linear	45	43	31	32	24	22	14	17	9	8	8	8
	All	92	92	32	34	83	87	25	28	73	73	23	25
Both	Gaussian	98	99	37	37	100	99	36	40	100	100	69	64
	Quadratic	83	82	36	38	66	66	41	45	18	16	30	30
	Linear	48	47	33	34	29	27	19	22	7	7	7	7
	All	90	89	36	36	86	87	30	35	90	89	29	32

Table 3

Empirical C-statistic (%) for predicting survival up to time t_0 for linear and nonlinear signal, where t_0 is approximately the 70^{th} percentile of follow-up time. $\widehat{h}(\mathbf{z})$ is built in a training data set with $n_{train} = 100$. All C-statistics are calculated by applying $\widehat{h}(\mathbf{z})$ to a testing data set with $n_{test} = 1000$. For kernel methods, we present the C-statistic for $\widehat{h}$ estimated from each of the three kernels, and the C-statistic if we select kernel based on $\widehat{C}$, the estimated C-statistic in the training data. For comparison, we present the Cox and AFT full models, which are the models fit with all pathway variables as linear covariates.

Correlation		0.2				0.5				0.8			
Pathway Size		5		20		5		20		5		20	
Error		EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp	EVD	Exp
						linear $h(\mathbf{z})$							
Full	AFT	74	78	72	76	79	82	77	80	82	85	80	81
	Cox	74	76	73	73	80	81	78	76	82	84	81	79
WLR_{Ge} Kernel	Gaussian	74	78	70	73	80	82	78	80	82	85	83	83
	Linear	74	78	72	76	80	82	79	80	83	85	83	83
	Quadratic	74	78	72	76	80	82	79	80	83	85	83	83
	Omnibus	74	78	72	76	80	82	79	80	83	85	83	83
						nonlinear $h(\mathbf{z})$							
Full	AFT	54	55	53	53	54	56	54	54	54	54	54	52
	Cox	53	54	53	53	54	55	53	54	54	54	54	52
WLR_{Ge} Kernel	Gaussian	70	71	54	54	69	71	55	57	68	68	63	62
	Linear	54	56	54	54	55	56	54	55	53	52	52	51
	Quadratic	70	71	55	56	65	66	59	60	60	61	60	58
	Omnibus	71	72	55	56	69	71	59	60	68	68	62	61