Published in final edited form as:

Biometrics. 2013 December ; 69(4): . doi:10.1111/biom.12098.

Omnibus Risk Assessment via Accelerated Failure Time Kernel Machine Modeling

Jennifer A. Sinnott and Tianxi Cai

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA

Jennifer A. Sinnott: jsinnott@hsph.harvard.edu; Tianxi Cai: tcai@hsph.harvard.edu

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 possible complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. Standard approaches for identifying important markers often rely on marginal associations or linearity assumptions and may not capture non-linear or interactive effects. In recent years, 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 can capture nonlinear effects if nonlinear kernels are used (Scholkopf and Smola, 2002; Liu et al., 2007, 2008). For survival outcomes, KM regression modeling and testing procedures have been derived under a proportional hazards (PH) assumption (Li and Luan, 2003; Cai et al., 2011). In this paper, we derive testing and prediction methods for KM regression under the accelerated failure time model, a useful alternative to the PH model. We approximate the null distribution of our test statistic using resampling procedures. When multiple kernels are of potential interest, it may be unclear in advance which kernel to use for testing and estimation. We propose a robust Omnibus Test that combines information across kernels, and an approach for selecting the best kernel for estimation. The methods are illustrated with an application in breast cancer.

Keywords

Accelerated Failure Time Model; Kernel Machines; Omnibus Test; 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 traditional approaches for quantifying this relationship identify individual markers with marginal associations with disease; however, resulting lists of differentially expressed genes can be hard to interpret or replicate, and may not include truly important markers with modest, nonlinear, or interactive effects. An appealing alternative is to leverage biological knowledge by grouping markers into pathways consisting of genes thought to work together.

Supplementary Materials

The Web Appendix referenced in Section 3.2 and the R code referenced in Section 5.1 is available with this paper at the *Biometrics* website on Wiley Online Library.

Working at the pathway level improves interpretability, reduces dimensionality by decreasing the number of statistical hypotheses to be tested, and can improve power to detect associations. Ideally, pathway methods should identify genes which may individually have only modest associations with outcome but which together have a substantial joint effect. They should also allow for complicated relationships among genes in the pathway that could reflect more complex biological signals.

Kernel machine (KM) methods (Scholkopf and Smola, 2002) are attractive tools for relating biological pathways to disease outcomes because they can capture complex effects without explicit specification of the form of those effects, and because they can leverage the within-pathway correlation which is likely to exist in genomic data. For non-censored outcomes, KM regression and testing procedures have been proposed in Liu et al. (2007) and Liu et al. (2008). For survival outcomes, Li and Luan (2003) and Cai et al. (2011) proposed KM testing and estimation procedures under the proportional hazards (PH) model. However, when the PH assumption fails to hold, these procedures may have incorrect size and/or suboptimal power to identify important pathways or accurately predict risk. In this paper, we propose KM methods for survival outcomes under the accelerated failure time (AFT) model (Kalbeisch et al., 1980), a useful alternative to the PH 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. A variety of inference procedures exist for the regression parameters under the AFT model. The rank-based approach (Tsiatis, 1990; Ritov, 1990), which fits the model using a weighted log rank (WLR) estimating equation for various weights, 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.

We propose the use of the AFT KM model to capture potentially complex, non-linear pathway effects on survival. We first develop a pseudo-score test under the KM framework using the WLR estimating function, using resampling procedures for inference. The WLR weight affects the estimation efficiency, and the kernel determines the structure of the pathway effect, but in practice it is often unclear which weight or kernel is most appropriate for a given dataset. To overcome this challenge, we propose omnibus testing procedures that pool information across weights and kernels. We demonstrate the effectiveness of this approach in simulations, where the power lost by using an omnibus test is minimal, but the power gained in some circumstances can be substantial.

When a pathway is associated with outcome, we may hope to use the pathway information to improve risk prediction. Under the AFT KM framework, we propose procedures for estimating the pathway effect in order to construct individual risk scores. 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. We also propose a data-driven approach for selecting the kernel to use in estimation, which works well in practice. Recently, Liu et al. (2010) proposed the usage of the AFT KM model and presented a weighted least squares estimator. However, in the presence of censoring, it is unclear whether this estimator is consistent even under linear effects. Because our procedure is derived using the WLR class of estimating equations, we are guaranteed consistent estimates of the underlying effects when the AFT model is correctly specified.

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 Test. 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

Let T denote the survival time, $\mathbf{Z}$ be a $P \times 1$ vector of genetic measurements in a gene set, and $\mathbf{D}$ be a vector of clinical covariates such as age and gender. Due to censoring of T , we observe $X = \min\{T, C\}$ and $\Delta = I[T \leq C]$, where C is a censoring time that is assumed to be independent of T given $\mathbf{W} = (\mathbf{D}^\top, \mathbf{Z}^\top)^\top$. The observed data consist of n independent and identically distributed (iid) random vectors, $\mathcal{O} = \{(X_i, \Delta_i, \mathbf{W}_i) : i = 1, \dots, n\}$.

To derive a prediction model for T based on $\mathbf{W}$, we consider the KM generalization of the AFT model:

$$\log T_i = \gamma^\top \mathbf{D}_i + h(\mathbf{Z}_i) + E_i, i = 1, \dots, n. \quad (1)$$

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, and $h(\cdot)$ is an unknown smooth function that belongs to $\mathcal{H}_K$, 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* $K(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^\top \mathbf{z}_2$ leads to $h(\mathbf{z}) = \beta^\top \mathbf{z}$, a linear function of the covariates. The *quadratic kernel*

$K(\mathbf{z}_1, \mathbf{z}_2; \rho) = (\rho + \mathbf{z}_1^\top \mathbf{z}_2)^2$ yields a Hilbert space $\mathcal{H}_K$ spanned by basis functions $\{z_j, z_j z_{j'}, j, j' = 1, \dots, p\}$, 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 $\mathcal{H}_K$ is generated by the radial basis functions.

We will be interested in testing the null hypothesis $H_0 : h(\cdot) = 0$, which suggests that the genes in the pathway do not affect survival given the covariates $\mathbf{D}$. Under the null, model (1) reduces to the standard AFT model for the clinical covariates,

$$\log T_i = \gamma^\top \mathbf{D}_i + E_i, i = 1, \dots, n. \quad (2)$$

As described in Jin et al. (2003), the regression parameter γ can be estimated by solving the WLR estimating function, $\mathbf{U}_\phi(\gamma) = n^{-1} \sum_{i=1}^n \phi(\gamma, e_i(0; \gamma)) \Delta_i [\mathbf{D}_i - \bar{\mathbf{D}}(\gamma, e_i(0; \gamma))]$, where $e_i(0; \gamma)$ is the residual $e_i(h; \gamma) = \log X_i - \gamma^\top \mathbf{D}_i - h(\mathbf{Z}_i)$ evaluated at $h = 0$, $\phi(\gamma, t)$ is a weight function, and $\bar{\mathbf{D}}(\gamma, t) = S^{(1)}(\gamma, t)/S^{(0)}(\gamma, t)$ for $S^{(k)}(\gamma, t) = n^{-1} \sum_{j=1}^n \mathbf{I}\{e_j(0; \gamma) \geq t\} \mathbf{D}_j^{\otimes k}$. Here, we use the notation $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\top$ for any vector $\mathbf{a}$. $\mathbf{U}_\phi$ is not component wise monotonic in γ in general, but it is when we use the Gehan weights $\phi = S^{(0)}$, which yield an estimating function $\mathbf{U}_G(\gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i [\mathbf{D}_i - \mathbf{D}_j] \mathbf{I}\{e_j(0; \gamma) \geq e_i(0; \gamma)\}$ which is the gradient of the convex Gehan objective function

$$L_G(\gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(0; \gamma) - e_i(0; \gamma)|_+, \quad (3)$$

where $|a|_+ = a\mathbf{I}\{a > 0\}$. $L_G(\gamma)$ can be minimized to get an estimator $\tilde{\gamma}_G$ which is consistent for γ in (2); Jin et al. (2003) also provides an iterative procedure to find estimators γ_ϕ for general weights ϕ , which may be desirable because the variance of γ_ϕ depends on 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 (3),

$$L_G^R(\gamma, h) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(h; \gamma) - e_i(h; \gamma)|_+ + \frac{c^2}{2} \|h\|_{\mathcal{H}_K}^2, \quad (4)$$

where $\|h\|_{\mathcal{H}_K}$ is the norm of h in $\mathcal{H}_K$. By the representer theorem (Kimeldorf and Wahba, 1970; Scholkopf and Smola, 2002, pp. 90–1), the minimizer of $L_G^R(\gamma, h)$ for h must take a dual form, $\hat{h}(\mathbf{z}) = \sum_{l=1}^n \alpha_l K(\mathbf{Z}_l, \mathbf{z})$, where the α_l are unknown parameters. Here and in the sequel, we suppress ρ from K for ease of presentation, but note that testing and estimation may depend on ρ , and that tuning ρ will be discussed when needed.

Using the dual representation, minimization of (4) is equivalent to minimization of

$$L_G^R(\alpha; \gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(\alpha; \gamma) - e_i(\alpha; \gamma)|_+ + \frac{c^2}{2} \alpha^\top \mathbb{K} \alpha \quad (5)$$

where $\mathbb{K}$ is the kernel matrix, whose $(i, j)^{\text{th}}$ entry is $K(\mathbf{Z}_i, \mathbf{Z}_j)$, and, with a slight abuse of notation, $e_i(\alpha; \gamma) = \log X_i - \gamma^\top \mathbf{D}_i - \alpha^\top \mathbf{K}_i$, 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 (5) as the Gehan objective function arising from the random effects AFT model

$$\log T_i = \gamma^\top \mathbf{D}_i + \alpha^\top \mathbf{K}_i + E_i, \quad \alpha = \tau \varepsilon, \quad E(\varepsilon) = 0, \quad \text{var}(\varepsilon) = \mathbb{K}^-, \quad (6)$$

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

3. Testing and Estimation with Gehan Weights

3.1 The Pseudo-Score Statistic

To identify pathways associated with survival, we propose the use of the AFT KM framework and derive testing procedures for $H_0 : h(\cdot) = 0$ in model (1). Here we derive a KM pseudo-score test of H_0 using the WLR estimating function with Gehan weights; in Section 4 we show how to extend the test to more general weights.

By using the mixed effects formulation (6) as a working model, we see that the hypothesis $H_0 : h(\cdot) = 0$ 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 (5) as a function of τ conditional on random effects ε , $L_{G,\varepsilon}(\tau; \gamma) + \varepsilon^\top \mathbb{K} \varepsilon$, where

$$L_{G,\varepsilon}(\tau; \gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(\tau \varepsilon; \gamma) - e_i(\tau \varepsilon; \gamma)|_+. \quad (7)$$

If γ is known, a pseudo-score statistic can be obtained as

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

where ∇_{τ} is the partial derivative with respect to τ . When γ is unknown, as would be the case in practice, we estimate γ_G under the null as described in Section 2, and define our pseudo-score statistic as $\hat{Q} = \hat{Q}(\hat{\gamma}_G)$. Since $\nabla_{\tau} L_{G,\varepsilon}(\tau; \gamma) \Big|_{\tau=0} = \varepsilon^{\top} \mathbb{K} \mathbf{R}(\gamma)$, where $\mathbf{R}(\gamma) = (R_1(\gamma), \dots, R_n(\gamma))^{\top}$ and $\hat{R}_k(\gamma) = n^{-2} \sum_{j=1}^n [\Delta_k \mathbf{I}\{e_j(0; \gamma) \geq e_k(0; \gamma)\} - \Delta_j \mathbf{I}\{e_k(0; \gamma) \geq e_j(0; \gamma)\}]$, we can see that $\hat{Q} = \mathbf{R}(\hat{\gamma}_G)^{\top} \mathbb{K} \mathbf{R}(\hat{\gamma}_G)$. We may further rewrite this by employing a spectral decomposition for $\mathbb{K}$. If we let the eigenvalues and associated eigenvectors of $\mathbb{K}$ be λ_l and ζ_l respectively, for $l = 1, \dots, n$, where we assume that $\lambda_1 \geq \dots \geq \lambda_n$ and that the ζ_l have norm 1, then we may write $\mathbb{K} = \tilde{\mathbb{B}}_n \tilde{\mathbb{B}}_n^{\top}$, where $\tilde{\mathbb{B}}_n = \left(\sqrt{\lambda_1} \hat{\zeta}_1 \dots \sqrt{\lambda_n} \hat{\zeta}_n \right)$. This allows us to reexpress $\hat{Q}(\gamma)$ as:

$$\hat{Q}(\gamma) = \hat{\mathbf{U}}_G(\gamma)^{\top} \hat{\mathbf{U}}_G(\gamma)$$

where

$$\hat{\mathbf{U}}_G(\gamma) = \tilde{\mathbb{B}}_n^{\top} \hat{\mathbf{R}}(\gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i (\tilde{\mathbf{B}}_{ni} - \tilde{\mathbf{B}}_{nj}) \mathbf{I}\{e_j(0; \gamma) \geq e_i(0; \gamma)\}. \quad (8)$$

Here $\tilde{\mathbf{B}}_{ni}$ is the i^{th} row of $\tilde{\mathbb{B}}_n$. The form of $\hat{\mathbf{U}}_G$ mimics that of the standard Gehan WLR function $\mathbf{U}_G$ defined in Section 2, with the $\mathbf{B}_{ni}$ taking the role of the covariates $\mathbf{D}_i$.

3.2 Approximating the Null Distribution of the Score Statistic

In the Web Appendix, we outline the derivation of the asymptotic distribution of $n\hat{Q}$. This distribution generally does not have an explicit form, so to approximate the null distribution of $\hat{Q}$ in finite samples, we propose a perturbation approach similar to that used in Jin et al. (2003) for the linear effects AFT model.

Let $\mathcal{V} = (\mathcal{V}_1, \dots, \mathcal{V}_n)$, let the $\{\mathcal{V}_i\}$ be iid random variables with mean 1 and variance 1, and let $\mathcal{V} = \sum_{i=1}^n \mathcal{V}_i$. We first find $\tilde{\gamma}_G^*$, the minimizer of $L_G^*(\gamma) = (\mathcal{V} \cdot)^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_j(0; \gamma) - e_i(0; \gamma)|_+ \mathcal{V}_i \mathcal{V}_j$. Then, we calculate the perturbation $\hat{\mathbf{U}}_G^*(\tilde{\gamma}_G^*)$ of $\hat{\mathbf{U}}_G(\tilde{\gamma}_G)$ as:

$$\hat{\mathbf{U}}_G^*(\tilde{\gamma}_G^*) = (\mathcal{V} \cdot)^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i (\tilde{\mathbf{B}}_{ni} - \tilde{\mathbf{B}}_{nj}) \mathbf{I}\{e_j(0; \tilde{\gamma}_G^*) \geq e_i(0; \tilde{\gamma}_G^*)\} \mathcal{V}_i \mathcal{V}_j. \quad (9)$$

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

$$\hat{Q}^* = \hat{Q}^*(\tilde{\gamma}_G^*) = \{\hat{\mathbf{U}}_G^*(\tilde{\gamma}_G^*) - \hat{\mathbf{U}}_G(\tilde{\gamma}_G)\}^{\top} \{\hat{\mathbf{U}}_G^*(\tilde{\gamma}_G^*) - \hat{\mathbf{U}}_G(\tilde{\gamma}_G)\} \text{ given } \mathcal{O}$$

To calculate a p-value for testing H_0 , we may generate a large number of realizations of $\mathcal{V}$ say $\mathcal{V}_{(1)}, \dots, \mathcal{V}_{(B)}$, and use them to generate $\hat{Q}_{(1)}^*, \dots, \hat{Q}_{(B)}^*$. Then the p-value of the test can

be obtained by $\hat{p} = \#\{\hat{Q}_{(b)}^* \geq \hat{Q}\}/B$. 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 $\{\hat{Q}_{(b)}^*\}$, and calculate $p_{\chi^2} = 1 - F(\hat{Q})$, where F is the distribution function of a $c_0\chi_{d_0}^2$ random variable. Similar approximations have performed well for other models (Liu et al., 2007, 2008; Cai et al., 2011). However, such an approximation may not perform well for extremely low tail probabilities and hence needs to be used with caution when very small p-values are of interest.

For the linear kernel, which does not rely on an additional parameter ρ , we may use the p-values $p_{\text{obs}} = p$ or $p_{\chi^2, \text{obs}} = p_{\chi^2}$ directly. For the quadratic and Gaussian kernels, the test statistic Q_{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 as a test statistic:

$$\hat{T}_{\mathcal{J}} = \inf_{\rho \in \mathcal{J}} \{\hat{p}_{\chi^2}(\rho)\}, \quad (10)$$

where $p_{\chi^2}(\rho)$ denotes the p-value from the χ^2 approximation derived under kernel function $K(\cdot, \cdot; \rho)$ and $\mathcal{J}$ is an appropriately chosen range for ρ . The final p-value then is

$\hat{p}_{\text{obs}} = \#\{\hat{T}_{\mathcal{J}, (b)}^* \leq \hat{T}_{\mathcal{J}, \text{obs}}\}/B$. For the Gaussian and quadratic kernels, we determine the range $\mathcal{J}$ of ρ by requiring that the associated kernel matrices $\mathbb{K}(\rho)$ have eigenvalues $\hat{\lambda}_i$ that decay at a polynomial rate $O(i^{-\alpha})$ for some range of $\alpha > 1$. This approach is motivated by work in Braun (2005) which bounds the error due to projecting the feature space onto the first r principal components using terms whose behavior depends on the decay properties of the eigenvalues. By experimentation we found that $\alpha \in [1.75, 6]$ yielded a reasonable range of feature space complexity. To implement this, for each ρ , we regressed the logarithms of the top 95% of eigenvalues on their indices and used the regression coefficient as an estimate of α ; we considered values ρ whose associated α fell in the specified range.

3.3 Estimation and the Kernel Principal Components Analysis Approximation

To construct a risk score $\gamma^\top \mathbf{D} + h(\mathbf{Z})$ for predicting T , we can minimize the penalized Gehan objective function (5) to get estimates for γ and α . The parameter c in (5) controls the smoothness of the estimator of h , and can be chosen by cross-validation to minimize, for example, the unpenalized Gehan objective function. For kernels that depend on a tuning parameter ρ , we use the value of ρ that minimized $T_{\mathcal{J}}(10)$. This choice is appealing because this ρ corresponds to the kernel $K(\cdot, \cdot; \rho)$ that produced the most evidence that $h(\cdot) \neq 0$.

While estimating (γ, α) in this manner is possible, it can be computationally demanding because α has n components. Computation time can be decreased with little loss of accuracy by using a kernel PCA approximation for dimension reduction (Schölkopf et al., 1998; Mika et al., 1999). In fact, employing kernel PCA may even improve estimation efficiency (Braun, 2005). To carry out kernel PCA in practice, we again employ the spectral decomposition of $\mathbb{K}$ but retain only the first r principal components, where r is the smallest number for which $\sum_{i=1}^r \hat{\lambda}_i / \sum_{i=1}^n \hat{\lambda}_i \geq p$, where p is some prespecified fraction. In many situations, r is significantly smaller than n . Ideally, the included eigenvectors encode aspects of maximal variability in the data, while the excluded eigenvectors capture noise. The best choice for p would require estimation of the eigenvalue decay rate (Braun, 2005). One could use cross-validation to choose an optimal p as a tuning parameter together with other tuning parameters in the estimation procedure. For practical purposes, we have found that for

estimation of $h(\cdot)$, $p \approx 1 - \log(n)/n$ works well in practice, but further research could identify a better method for choosing p

Writing $\tilde{\mathbb{B}}_r = \left(\sqrt{\hat{\lambda}_1} \hat{\zeta}_1 \cdots \sqrt{\hat{\lambda}_r} \hat{\zeta}_r \right)$ and $\tilde{\mathbb{K}}_r = \tilde{\mathbb{B}}_r \tilde{\mathbb{B}}_r^\top$, we may consider an approximate working model analogous to (6) with $\tilde{\mathbb{K}}_r$ in place of $\mathbb{K}$. In fact, our pseudo-score test may be derived exactly as above with $\tilde{\mathbb{K}}_r$ and $\tilde{\mathbb{B}}_r$ in place of $\mathbb{K}$ and $\mathbb{B}_n$ to yield a slightly faster test procedure, which is what we use in practice. For testing, it is not difficult to see that the test statistic is dominated by contributions from those $\sqrt{\hat{\lambda}_l} \hat{\zeta}_l$ with larger $\hat{\lambda}_l$ and hence increasing p typically does not negatively affect the power; in practice, we find that $p \approx 1 - 1/n$ provides stable testing results. For estimation, we may apply the variable transformation $\beta = \tilde{\mathbb{B}}_r^\top \alpha$ and rewrite the working model as:

$$\log \mathbf{T} = \mathbf{D} \gamma + \tilde{\mathbb{B}}_r \beta + \mathbf{E}, \quad E(\beta) = 0, \quad \text{Var}(\beta) = \tau^2 \mathbb{I}_{r \times r}. \quad (11)$$

Then, instead of estimating (γ, α) , we estimate (γ, β) by minimizing

$$L_G^R(\beta, \gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Delta_i |\tilde{e}_j(\beta, \gamma) - \tilde{e}_i(\beta, \gamma)|_+ + \frac{c^2}{2} \beta^\top \beta. \quad (12)$$

where $\tilde{e}_i(\beta; \gamma) = \log X_i - \gamma^\top \mathbf{D}_i - \beta^\top \tilde{\mathbf{B}}_{ri}$, for $\tilde{\mathbf{B}}_{ri}$ the i th row of $\tilde{\mathbb{B}}_r$. The parameter β has r components, so when the sample size is not small and the eigenvalues decay quickly, kernel PCA has the computational advantage of greatly reducing the number of unknown parameters to be estimated. Minimization of (12) can be accomplished by linear programming, as discussed in Jin et al. (2003); alternatively, one may adopt minimization procedures such as the Nelder-Mead simplex minimization (Nelder and Mead, 1965) that do not rely on the smoothness of the objective function.

Formulating the estimation of h in this way is also appealing because it relates to the primal representation of h in the Hilbert space $\mathcal{H}_K$. The space $\mathcal{H}_K$ has a (possibly infinite) basis $\{\sqrt{\lambda_l} \zeta_l(\cdot), l=1, 2, \dots\}$ made up of eigenfunctions and eigenvalues $\zeta_l(\cdot)$ and λ_l of the integral transform $T: \mathcal{H}_K \rightarrow \mathcal{H}_K$ defined by $[Tf](\mathbf{z}) = \int K(\mathbf{z}, \mathbf{z}') f(\mathbf{z}') d\mathbf{z}'$, where $\lambda_1 \geq \lambda_2 \geq \dots$. Any function $h \in \mathcal{H}_K$ may be written in its primal representation, $h(\mathbf{z}) = \sum_{l=1}^{\infty} \beta_l \sqrt{\lambda_l} \zeta_l(\mathbf{z})$, and if the eigenvalues decay quickly, can be well-approximated by a truncated sum $h_{r_0}(\mathbf{z}) = \sum_{l=1}^{r_0} \beta_l \sqrt{\lambda_l} \zeta_l(\mathbf{z})$ for some sufficiently large r_0 . It has been shown that the eigenvalues and eigenvectors obtained based on $\mathbb{K}$ can be used to consistently estimate the underlying true eigenvalues and eigenfunctions (Koltchinskii and Giné, 2000; Braun, 2005). Thus, kernel PCA uses the eigendecomposition of $\mathbb{K}$ to approximate the basis of $\mathcal{H}_K$, and uses that to estimate h in its approximate primal form.

To apply the risk score to a future subject with predictors $\mathbf{W}_0 = (\mathbf{D}_0^\top, \mathbf{Z}_0^\top)^\top$, we may use the Nyström approximation method (Rasmussen and Williams, 2006) which relates these predictors to those in the training data. Letting $(\hat{\gamma}^\top, \hat{\beta}^\top)$ be the estimators for $(\gamma^\top, \beta^\top)$, the risk score for the future subject is $\mathbf{D}_0^\top \hat{\gamma} + \sum_{l=1}^r \hat{\beta}_l \mathbf{K}_{\mathbf{Z}_0}^\top \hat{\zeta}_l \hat{\lambda}_l^{-\frac{1}{2}}$, where $\mathbf{K}_{\mathbf{Z}_0} = (K(\mathbf{Z}_0, \mathbf{Z}_1), \dots, K(\mathbf{Z}_0, \mathbf{Z}_n))^\top$.

4. General WLR and Kernel Selection

4.1 Testing and Prediction with General Weights

Although parameter estimates using the Gehan-weighted WLR are consistent, their performance in finite samples will vary depending on the underlying distributions of survival and censoring. Thus, it is desirable to extend these procedures to incorporate a general weight $\phi(\beta; \gamma, t)$. Since the WLR does not in general correspond to an objective function, we follow the same iterative strategy as proposed in Jin et al. (2003) to derive estimation and testing procedures for a general weight.

We first discuss estimation, using the computationally convenient kernel PCA approximation which uses a pre-specified proportion of the eigenvalues; results corresponding to the original kernel can be obtained by setting $p = 1$. We initialize $\{\gamma^{(0)}, \beta^{(0)}\} = \{\gamma_G, \beta_G\}$, which are estimated using the Gehan weights as described in Section 3.3. Then at step k , from given estimates $\{\gamma^{(k-1)}, \beta^{(k-1)}\}$, we may obtain an updated estimator $\{\gamma^{(k)}, \beta^{(k)}\}$ as the minimizer of

$$\tilde{L}_\phi^R(\beta; \gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \psi(\hat{\beta}^{(k-1)}; \hat{\gamma}^{(k-1)}, e_i(\hat{\beta}^{(k-1)}; \hat{\gamma}^{(k-1)})) \Delta_i |\tilde{e}_j(\beta; \gamma) - \tilde{e}_i(\beta; \gamma)|_+ + \frac{c_k^2}{2} \beta^\top \beta \quad (13)$$

where the c_k are a sequence of regularizing constants which converge to some c , and $\psi(\beta; \gamma, t) = \phi(\beta; \gamma, t)/S^{(0)}(\beta; \gamma, t)$ for $S^{(k)}(\beta; \gamma, t) = n^{-1} \sum_{j=1}^n I\{e_j(\beta; \gamma) \geq t\} \begin{pmatrix} \mathbf{D}_j \\ \tilde{\mathbf{B}}_{rj} \end{pmatrix}^{\otimes k}$. As $k \rightarrow \infty$, $\{\gamma^{(k)}, \beta^{(k)}\}$ converges to $\{\gamma_\phi, \beta_\phi\}$, a solution to $\tilde{\mathbf{U}}_\phi(\beta; \gamma) + c^2 \beta = 0$, where

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

For testing with a general weight, we may derive the score test using the same arguments as in Section 3 but add weights $\psi(\beta^{(k-1)}; \gamma^{(k-1)}, e_i(\beta^{(k-1)}; \gamma^{(k-1)}))$ to (8), where $\{\gamma^{(k-1)}, \beta^{(k-1)}\}$ are some initial estimates of $\{\gamma, \beta\}$. When testing $H_0: h(\cdot) = 0$, it is convenient to set $\beta^{(k-1)} = 0$, its value under the null, and let $\gamma^{(k-1)} = \gamma_\phi$, the WLR estimator under H_0 which can be found using the iterative method in Jin et al. (2003). With these choices, no iteration is needed to calculate the test statistic after γ_ϕ has been found. Our WLR KM test statistic for a general weight ϕ is thus $Q_\phi = \hat{\mathbf{U}}_\phi(\gamma_\phi)^\top \hat{\mathbf{U}}_\phi(\gamma_\phi)$ where

$$\hat{\mathbf{U}}_\phi(\gamma) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \psi(0; \gamma, e_i(0; \gamma)) \Delta_i (\tilde{\mathbf{B}}_{ni} - \tilde{\mathbf{B}}_{nj}) I\{e_j(0; \gamma) \geq e_i(0; \gamma)\}.$$

Perturbations used to approximate the null take the form $\hat{\mathbf{Q}}_\phi^* = (\hat{\mathbf{U}}_\phi^*(\tilde{\gamma}_\phi^*) - \hat{\mathbf{U}}_\phi(\tilde{\gamma}_\phi))^T (\hat{\mathbf{U}}_\phi^*(\tilde{\gamma}_\phi^*) - \hat{\mathbf{U}}_\phi(\tilde{\gamma}_\phi))$ where $\tilde{\gamma}_\phi^*$ is a perturbation of γ_ϕ under H_0 associated with a particular realization of $\mathcal{V}$ and where:

$$\hat{\mathbf{U}}_\phi^*(\gamma) = (\mathcal{V} \cdot)^{-2} \sum_{i=1}^n \sum_{j=1}^n \psi^*(0; \gamma, e_i(0; \gamma)) \Delta_i (\tilde{\mathbf{B}}_{ni} - \tilde{\mathbf{B}}_{nj}) I\{e_j(0; \gamma) \geq e_i(0; \gamma)\} \mathcal{V}_i \mathcal{V}_j \text{ with}$$

$\psi^*(0; \gamma, t) = 1/S^{(0)*}(0; \gamma, t)$ for $S^{(0)*}(0; \gamma, t) = (\mathcal{V} \cdot)^{-1} \sum_{j=1}^n I\{e_j(0; \gamma) \geq t\} \mathcal{V}_j$. When $\phi = 1$, Q_ϕ reduces to the Cox 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 practice, little prior information is available on the best choice of kernel or WLR weight for a given dataset. Hence, data-driven methods to combine information from multiple

weights and kernels are desirable. We propose a test statistic $\hat{A}$ which combines the test statistics for several weight-kernel combinations of interest. To generate the null distribution of $\hat{A}$, we use a common collection of realizations $\{\mathcal{V}_{(1)}, \dots, \mathcal{V}_{(B)}\}$ of the n -vector $\mathcal{V} = (\mathcal{V}_1, \dots, \mathcal{V}_n)$ of iid mean-1 variance-1 random variables $\mathcal{V}_i$ to generate null distributions of each weight-kernel combination of interest, thus obtaining their joint null distribution. We can then generate the null distribution of $\hat{A}$ by calculating its value, $\hat{A}_{(b)}^*$, associated with each $\mathcal{V}_{(b)}$, and generate a P-value by comparing $\hat{A}_{\text{obs}}$ to $\{\hat{A}_{(1)}^*, \dots, \hat{A}_{(B)}^*\}$ – i.e.,

$$\hat{p}_{\text{combined}} = \#\{\hat{A}_{(b)}^* \geq \hat{A}_{\text{obs}}\} / B.$$

We propose an Omnibus Test by adapting the method developed in Xu et al. (2003) for combining dependent Z-statistics to our KM test setting. To mimic a Z-statistic, we transform the P-values into z-scores by $Z_k = \Phi^{-1}(1 - p_k)$, yielding a vector of observed z-scores $\mathbf{Z}_{\text{obs}}$ and its associated null $\{\mathbf{Z}_{(1)}^*, \dots, \mathbf{Z}_{(B)}^*\}$. The $\mathbf{Z}_{(b)}^*$ are used to estimate Γ , the covariance of $\mathbf{Z}$, and then the function $W(c) = \mathbf{Z}_c^T \Gamma^{-1} \mathbf{Z}$ is computed for $c \in [0, 5]$, where $\mathbf{Z}_c$ is a vector with entries $\max\{Z_k, c\}$. By calculating $W(c)$ for both the observed $\mathbf{Z}_{\text{obs}}$ and the null $\mathbf{Z}_{(b)}^*$, an estimated P-value $p_{\text{Omnib}}(c)$ can be calculated for each c , and the final test statistic whose value on the observed data is compared to its value on the null is $\hat{A}_{\text{Omnib}} = \min_{c \in [0, 5]} p_{\text{Omnib}}(c)$.

As in testing, the best choice of kernel and weight for risk prediction may not be obvious. We consider two automated approaches for determining which kernel-weight combination to use. In the first approach, we apply the associated KM tests to the pathway and use the kernel-weight pair yielding the smallest p-value. In the second approach, we choose the kernel-weight pair which optimizes prediction accuracy, as assessed by a C-statistic which captures how well the order of the associated risk predictions corresponds to the order of the true survival times during a pre-specified follow-up period $(0, \tau)$, $C_\tau = P(\gamma^\top \mathbf{D}_i + \hat{h}(\mathbf{Z}_i) > \gamma^\top \mathbf{D}_j + \hat{h}(\mathbf{Z}_j) | T_i > T_j, T_j < \tau)$. We estimate C_τ using the nonparametric estimator proposed in Uno et al. (2011), and select the kernel-weight combination which maximizes C_τ .

5. Simulation Studies

5.1 Testing

We conducted simulation studies to assess the performance of the proposed testing and estimation procedures. R code for implementing the procedures is available with this article at the Biometrics website on Wiley Online Library. 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 ρ . We considered pathways of size $P = 5$ and 20, and correlations $\rho = 0.5$ and 0.2 to represent strong and moderate 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, generating E from either the extreme value distribution (EVD) or a logistic distribution $\text{Logistic}(\mu, \sigma)$, where μ and σ were chosen so that the survival times resulting from both choices of E were of comparable magnitude. The censoring was generated from a uniform distribution with range chosen so that either 25% or 50% of the individuals were censored.

To assess the performance of the testing procedures in each setting, we simulated 2000 data sets for empirical size calculations and 1000 for empirical power, and compared the AFT KM pseudo-score tests with Gehan weights (WLR_{Ge}) and log-rank weights (WLR_{LR}) for

the Gaussian, quadratic, and linear kernels. For comparison, we also performed the Cox KM tests for these three kernels, following the method in Cai et al. (2011). All null distributions were generated using $B = 2000$ perturbations, and kernel PCA was used to speed computation time, with $p = 0.99$. We also considered the Omnibus Test described in Section 4.2. For comparison, we also considered a truncated version of the Omnibus Test, which we call Omnibus-Truncated, where we only combine p-values that are nominally significant, basing the combined test on $\hat{A}_{\text{Omni Trunc}} = \min_{c \in [1.96, 5]} p_{\text{Omni}}(c)$. We also compared three other methods for combining p-values. For combining K p-values $p_1, \dots, p_K$, we define the

Fisher statistic, $\hat{A}_{\text{Fisher}} = -2 \sum_{k=1}^K \log p_k$ (Fisher, 1925); the Fisher-Truncated statistic,

$\hat{A}_{\text{Fisher Trunc}} = -2 \sum_{k=1}^K \mathbf{I}\{p_k \leq 0.05\} \log p_k$ (Zaykin et al., 2002); and the Min-P statistic,

$\hat{A}_{\text{min-p}} = \min_k \{p_k\}$. For each $\hat{A}$, the final p-value was calculated by comparing the observed value of $\hat{A}$ to the approximate null distribution $\{\hat{A}_{(1)}^*, \dots, \hat{A}_{(B)}^*\}$ found by calculating $\hat{A}$ using the p-values from the perturbations, as described in Section 4.2. The p-values combined across all six weight-kernel combination are shown in the tables.

For comparison to our KM-based tests, we considered a method for assigning a p-value to a pathway based on the marginal associations between the genes in the pathway and survival. For each gene Z_i , $i = 1, \dots, P$, in the pathway, we calculated a marginal p-value $\hat{p}_i^{\text{marg}}$ from the Wald test from a standard univariate Cox model, and then calculated the minimum of these P p-values, $\hat{p}_{\min}^{\text{marg}} = \min\{\hat{p}_1^{\text{marg}}, \dots, \hat{p}_P^{\text{marg}}\}$. We adjusted this p-value for multiple testing using the effective number of tests, M_{eff} , as described in Nyholt (2004). That is, we define $\hat{p}_{\text{pathway}}^{\text{marg}} = 1 - [1 - \hat{p}_{\min}^{\text{marg}}]^{M_{\text{eff}}}$, where for observed eigenvalues $\hat{\ell}_j$ from the correlation matrix of the genes in the pathway, $M_{\text{eff}} = 1 + (P - 1)(1 - \text{Var}(\hat{\ell}_1, \dots, \hat{\ell}_P)/P)$. In addition to this method based on marginal associations, we also applied the Global Test proposed in Goeman et al. (2005), which tests for an overall effect in a joint linear model; the formulation of this test is very similar to the Cox KM and WLR_{LR} tests with linear kernel, except that the null distribution is approximated differently.

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 5% are shown in Table 1 for $n = 200$ and 400. All tests are close to nominal size when $n = 200$, and the sizes improve when $n = 400$. To assess the power of the proposed tests, we considered for $n = 200$ (1) a linear signal, $h_1(\mathbf{z}) = C(z_1 + z_2 + z_3 + z_4 + z_5)$ with $C = 0.07$, and (2) a nonlinear signal, $h_2(\mathbf{z}) = C\{g(z_1)(z_2 + z_3 + z_4 + z_5) + 2[1 - g(z_1)](z_2 z_3 - z_4 - z_5)\}$, with $C = 0.35$ and $g(z) \sim \text{Bernoulli}\{\exp[\sin(z + 1)/2]\}$. This function was chosen to produce two latent classes with differential covariate effects. 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 is because both the Gaussian and quadratic kernels have a tuning parameter ρ which controls model complexity, and both kernels approximate the linear kernel for large ρ . The KM-based tests tend to be more powerful than the marginal gene method and the Global Test. Because the WLR_{LR} test statistic is the same as the proposed Cox KM test statistic in the absence of covariates, the power for the Cox KM and WLR_{LR} tests are very similar; differences are due to differences in the methods for estimating the null distribution of the test statistic. In this setting, there is little to distinguish the five methods for combining p-values from one another. In each case, the power of a combined test tends to lie in the middle of the power levels of the component tests, but closer to the optimal one.

When the true signal is nonlinear, there is more to distinguish the tests; the tests based on linear effects, such as the marginal gene method, the Global Test, and the KM tests with linear kernel, tend to have low power. The tests that permit nonlinear effects, such as the KM tests with Gaussian and quadratic kernels, have substantially more power. Comparing the proposed procedure to existing methods, we can see that our proposed Omnibus Test stands out in its ability to combine information across other tests – it is not strongly pulled down when one of its components (such as a KM test with linear kernel) has low power, and in several settings it outperforms its component tests.

For all tests, the power decreases somewhat when we increase the number of covariates from $P = 5$ to $P = 20$; however, the power loss is smaller 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 – that they benefit from within-pathway correlation.

Among the five methods for combining p-values considered, $\hat{A}_{\text{Omni}}$ appears to most robustly maintain power across different settings. Its performance is very slightly weaker than the other tests combining p-values when the true signal is linear, while in the nonlinear setting, it tends to be more powerful. In several of the nonlinear settings, its power is greater than all individual weight-kernel tests, suggesting it is gaining power when several tests have nearly significant results; this interpretation is supported because it is at times noticeably more powerful than the Omnibus-Truncated method we considered. This points to the robustness of the Omnibus Test, which could make it useful in a wide range of settings.

5.2 Estimation

To assess the performance of the estimation procedure when the true signal is linear (h_1 from above, with $C = 0.18$) or nonlinear (h_2 with $C = 1.5$), 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 $n_{\text{train}} = 100$ on which to build the estimates $\hat{h}$, and a validation data set with sample size $n_{\text{test}} = 1000$ on which to assess how well each estimate predicted survival up until some time t_0 using the C-statistic (Uno et al., 2011). In these simulations, we selected a reference time t_0 that was near the 70th percentile of followup time. We used kernel PCA to reduce dimensionality, with $p = 0.95$.

Using the training data, we built estimators for the Gehan and LR weights using linear, quadratic, and Gaussian kernels. Minimization of the penalized objective function (12) was done using Nelder-Mead simplex minimization. We also fit standard full Cox and AFT models for comparison. We applied these estimators and models to the validation data and calculated the C-statistic. We considered two methods for choosing which kernel to use in the final estimate $\hat{h}$; 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. Results are presented in Table 3.

In the linear setting, the C-statistics are nearly identical across all models. In the nonlinear setting, we see 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 two methods for selecting the kernel have very similar C-statistics in all settings; however, in our experience, the method based on the C-statistic has more stable predictive performance in simulation and data analysis. 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. A number of gene expression signatures have been introduced into clinical practice to better identify cancers with high and low risk of recurrence, but despite these advances, approximately 60% of patients with early-stage breast cancer are given adjuvant therapy in addition to local treatment, while only a small proportion are thought to benefit (Reis-Filho and Pusztai, 2011). Better markers of aggressive disease would help physicians predict which patients could safely avoid adjuvant therapy and its negative side-effects. One promising hypothesis-driven approach is to investigate the effects of candidate pathways on patient survival. We examined the associations between recurrence-free survival and 32 candidate pathways.

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 454 lymph node negative breast cancer patients; 286 patients were from the study described in Wang et al. (2005) and had tumor gene expression assessed using the Affymetrix U133A chip, while 168 were from the study described in Ivshina et al. (2006) with expression assessed using the Affymetrix U133A&B chip. We standardized each gene to have mean 0 and variance 1 within study, restricted to the genes common to both chips, and then merged the two data sets; all tests were performed controlling for data set and ER status. A total of 119 deaths or recurrences were observed, with follow-up ranging between 1 month and 14.3 years (median 7.7 years); 66% of observations were censored. Figure 1(a) shows the results of the testing procedure; the KM results are based on $B = 10000$ perturbations. For each pathway, we compare the p-value from the Global Test (Goeman et al., 2005) to the p-value from our Omnibus Test combining KM tests for all six weight-kernel combinations. There were 24 pathways significant at the nominal 0.05 level using the Global Test, and 25 were significant using the Omnibus Test.

Since multiple pathways were under consideration, we needed to adjust for multiple comparisons. We adjusted the Global Test p-values using a Bonferroni-type correction (Holm, 1979) to maintain a family-wise error rate of 0.05. To adjust the KM-based p-values, we used a common set of perturbation random variables $\{\mathcal{V}_{(1)}, \dots, \mathcal{V}_{(B)}\}$ across pathways to construct a joint null distribution, and used this joint null to construct the null distribution of the *minimal* p-value across the 32 pathways; we adjusted the observed pathway p-values by comparing them to this null distribution. Results are shown in Figure 1(b). After adjustment, 5 pathways were significant at the family-wise 0.05 level using the Global Test, while 10 pathways were significant using the Omnibus Test.

When we compared results from the Omnibus Test to the Cox KM test with Gaussian kernel as proposed in Cai et al. (2011), we found that the Cox KM test was somewhat more powerful in this data analysis, identifying 27 pathways significant at the nominal 0.05 level and 16 pathways significant at the family-wise 0.05 level. This may suggest that the Cox KM model is a reasonable choice for this dataset. However, the Omnibus Test, by considering more possible models, could be more robust in settings where the PH assumption is violated.

For the ten pathways declared most significant by our KM test, we estimated each pathway's effect on recurrence-free survival in the training set using the kernel selected by the C-statistic criterion; we also estimated the pathway effect using a standard Cox model, and a Cox KM model with a Gaussian kernel. We applied these models to an independent validation set of 119 lymph node negative patients with gene expression assessed on the Affymetrix U133A chip (Sotiriou et al., 2006). In this validation set, 27 deaths or

recurrences were observed, with follow-up time ranging between 2 months and 14.5 years (median 7.7 years); 77% of observations were censored. We assessed the accuracy of each model using the C-statistic and calculated 95% confidence intervals (Uno et al., 2011). Results are shown in Figure 2. In this example, the AFT KM model with kernel selection frequently selected the linear kernel as the optimal choice, suggesting that for many of the pathways considered, a linear model is a good fit. This may explain why in many cases the AFT KM model with kernel selection has prediction performance that is better than the Cox KM model with Gaussian kernel, which may be assuming too complex a model for some of these pathways. In the smaller pathways, the AFT KM model and the standard Cox model have similar performance with respect to prediction, which is also likely because the linear kernel is frequently selected. For the larger pathways, we see a gain in prediction using the KM-based fit, which is likely due to the dimension reduction from the PCA. In one of the top pathways, ATRBRCAPATHWAY, the Gaussian kernel is selected and predictiveness in the testing data improves over the Cox models; this improvement could be due to dimension reduction together with the incorporation of non-linear effects.

7. Discussion

In this paper, we proposed KM based procedures to test for the effect of pathways on survival in an AFT model framework. By taking a pathway based approach, we take advantage of pre-existing biological knowledge and privilege groups of genes believed to work together. By working with both linear and nonlinear kernels, we can detect linear effects well, while improving power to detect nonlinear effects when compared to methods 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 have the potential to capture nonlinear effects which can improve over methods assuming linearity, as demonstrated in simulation. For simplicity, we have relied on the C-statistic to assess prediction because it is a very robust measure of discrimination ability, but other accuracy measures that capture calibration could be useful in practice.

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 and estimation procedures that allow the data to drive weight and kernel choice. 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. Our perturbation procedures also allow for straightforward control of multiple testing that is not overly stringent when there is between-pathway correlation. Moreover, for risk prediction, it appears that the estimation procedure using kernel selection has negligible loss in prediction accuracy when compared to the optimal kernel.

We employ a kernel PCA approach to improve computational efficiency in both testing and estimation. In testing, the eigenvectors with large eigenvalues drive the statistical significance, and inclusion of eigenvectors with smaller eigenvalues does not change the resulting p-value much; in our experience, the kernel PCA threshold chosen does not strongly impact results, but we found that choosing a large fraction of eigenvalues performs well. In estimation, however, when we include r eigenvectors, we must estimate a

coefficient for each included term $\sqrt{\hat{\lambda}_j} \hat{\zeta}_j$, $j=1, \dots, r$, where $\hat{\zeta}_j$ is the j^{th} eigenvector of the kernel matrix with associated eigenvalue $\hat{\lambda}_j$. As the eigenvalue $\hat{\lambda}_j$ decreases, the variability in $\sqrt{\hat{\lambda}_j} \hat{\zeta}_j$ decreases, and the variance of the associated coefficient estimate increases. Because we estimate the coefficients with a ridge penalty, our procedure essentially combines both a hard thresholding and a smooth shrinkage for the β 's. As a result, a slight difference in the

number of included eigenvectors doesn't strongly affect results, but the difference in estimation accuracy between $r \ll n$ and $r = n$ is substantial. To truly optimize the estimation procedure, a data-driven method to select p should be developed.

The Cox PH model is used in practice much more frequently than the AFT model, and the Cox KM model has much to recommend it, including computational ease; moreover, in our experience, the approximation of the null distribution of the test statistic derived in the Cox framework in Cai et al. (2011) is more accurate in smaller sample sizes than the approximation derived here in the AFT framework, possibly due to the smoothness of the loss function for the Cox model. Thus, in settings where the PH assumption is reasonable, using a KM procedure under the Cox model is a reasonable choice. An Omnibus testing and kernel selection method identical to ours can be implemented in the Cox KM model, and would be useful for settings where the best choice of kernel is not known in advance. However, when the proportional hazards assumption is severely violated, the test derived under the Cox model may be suboptimal, and hence an alternative model such as the AFT model could be valuable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

JAS was supported by the National Institutes of Health (NIH) grants T32 GM074897 and T32 CA09001. TC was supported by the NIH grant R01 GM079330 and the National Science Foundation (NSF) grant DMS-0854970. The authors thank the editor and the reviewers for their helpful comments for improving the article.

References

- Braun, M. PhD thesis, PhD thesis. University of Bonn; 2005. Spectral properties of the kernel matrix and their application to kernel methods in machine learning.
- Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*. 2011
- Davies R. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*. 1987; 74:33–43.
- Fisher, RA., et al. Statistical methods for research workers. Edinburgh, London: Oliver and Boyd; 1925.
- Goeman JJ, Oosting J, Cleton-Jansen A-M, Anninga JK, Van Houwelingen HC. Testing association of a pathway with survival using gene expression data. *Bioinformatics*. 2005; 21:1950–1957. [PubMed: 15657105]
- Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. 1979:65–70.
- Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*. 2006; 66:10292–10301. [PubMed: 17079448]
- Jin Z, Lin D, Wei L, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003; 90:341.
- Jin Z, Ying Z, Wei L. A simple resampling method by perturbing the minimand. *Biometrika*. 2001; 88:381.
- Kalbeisch, J.; Prentice, R.; Kalbeisch, J. The statistical analysis of failure time data. Vol. volume 5. New York: Wiley; 1980.
- Kimeldorf G, Wahba G. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*. 1970; 41:495–502.

- Koltchinskii V, Giné E. Random matrix approximation of spectra of integral operators. *Bernoulli*. 2000;113–167.
- Li H, Luan Y. Kernel cox regression models for linking gene expression profiles to censored survival data. *Pac Symp Biocomput*. 2003;65–76. [PubMed: 12603018]
- Liu D, Ghosh D, Lin X. 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*. 2008; 9:292. [PubMed: 18577223]
- Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*. 2007; 63:1079–1088. [PubMed: 18078480]
- Liu Z, Chen D, Tan M, Jiang F, Gartenhaus RB. Kernel based methods for accelerated failure time model with ultra-high dimensional data. *BMC Bioinformatics*. 2010; 11:606. [PubMed: 21176134]
- Mika S, Schölkopf B, Smola A, Müller K, Scholz M, Rätsch G. Kernel pca and de-noising in feature spaces. *Advances in neural information processing systems*. 1999; 11:536–542.
- Nelder JA, Mead R. A simplex method for function minimization. *The computer journal*. 1965; 7:308–313.
- Nyholt D. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*. 2004; 74:765–769.
- Rasmussen, C.; Williams, C. Gaussian processes for machine learning. Vol. 38. Cambridge, MA, USA: The MIT Press; 2006. p. 715-719.2006.
- Reis-Filho J, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*. 2011; 378:1812–1823.
- Ritov Y. Estimation in a linear regression model with censored data. *The Annals of Statistics*. 1990:303–328.
- Scholkopf, B.; Smola, A. Learning with kernels. Cambridge, Mass: MIT Press; 2002.
- Schölkopf B, Smola A, Müller K. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*. 1998; 10:1299–1319.
- Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*. 2006; 98:262. [PubMed: 16478745]
- Tsiatis A. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*. 1990:354–372.
- Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011; 30:1105–1117. [PubMed: 21484848]
- Wang Y, Klijn J, Zhang Y, Sieuwerts A, Look M, Yang F, Talantov D, Timmermans M, Meijer-van Gelder M, Yu J, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*. 2005; 365:671–679.
- Xu X, Tian L, Wei L. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics*. 2003; 4:223–229. [PubMed: 12925518]
- Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining p-values. *Genet Epidemiol*. 2002; 22:170–185. [PubMed: 11788962]

Figure 1.

Presented are $\log_{10}$ p-values for testing the overall effect of 32 pathways on breast cancer survival. White squares represent the p-values from the Global Test, and the pathways are ordered by this p-value. Black diamonds represent the Omnibus Test p-values. Panel (a) shows the unadjusted p-values, while the panel (b) shows the p-values adjusted for multiple testing.

Figure 2.

Presented are point estimates and confidence intervals for the C-statistics for a full standard Cox model (Cox Full), a Cox KM model with Gaussian kernel (Cox KM), and our KM model with kernel selection (Omnibus KM) for the 10 pathways with smallest Omnibus Test p-value. Models were built on training data ($n = 454$) and applied to testing data ($n = 119$), where the C-statistics and confidence intervals were estimated. Below each pathway is the kernel and AFT weight selected using the C-statistic criterion in the training data and the number of genes in the pathway.

Empirical sizes (%) at nominal Type I error rate of 5% when $n = 200$ and $n = 400$. Testing was performed using the Gaussian, linear, and quadratic 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 p-value based on the χ^2 min p statistic ($T_{\hat{g}}$ in the text) is presented. Results shown use kernel PCA with 99% of eigenvalues used. Also shown are the empirical sizes of the five methods for combining p-values – the Omnibus Test, Fisher, Fisher-Truncated, Min-P, and Omnibus-Truncated – across all 6 weight-kernel combinations. Shown for comparison are results for the marginal gene method, the Global Test, and the KM tests based on the Cox model. Here, “E” denotes the extreme value distribution and “L” the logistic distribution, used for the error in the AFT model.

Table 1

% Censoring	25%						50%					
	0.2			0.5			0.2			0.5		
	5	20	E	5	20	E	5	20	E	5	20	E
Pathway Size	E	L	E	L	E	L	E	L	E	L	E	L
Error	E	L	E	L	E	L	E	L	E	L	E	L
<i>n = 200</i>												
Marginal Gene	6	5	4	5	6	5	5	5	6	6	5	5
Global Test	4	4	3	3	3	3	3	3	4	4	3	3
Gaussian	5	5	4	4	5	5	5	5	5	5	5	5
Linear	5	5	4	5	5	5	5	6	5	5	5	5
Quadratic	5	5	4	4	5	5	5	5	6	5	5	5
Cox KM	5	5	4	5	5	5	5	5	6	5	5	5
Gaussian	4	4	3	3	4	5	5	5	4	4	3	4
Linear	4	4	2	2	4	5	5	4	4	3	2	4
Quadratic	5	5	4	4	5	5	5	5	4	4	3	5
WLR _{Ge}	5	5	4	4	5	5	5	5	4	4	3	5
Gaussian	6	6	4	4	7	6	6	6	6	5	6	6
Linear	6	5	3	4	6	6	6	6	5	4	5	6
Quadratic	6	6	5	5	6	6	6	6	6	5	6	6
Omnibus	5	5	3	3	6	6	6	6	4	5	3	5
Fisher	6	5	4	4	6	6	6	6	5	5	4	5
Fisher-Truncated	5	5	3	4	6	6	6	6	5	5	4	5
Min-P	5	5	3	3	5	5	6	5	5	4	3	5
Omnibus-Truncated	6	5	3	4	6	6	6	6	5	5	4	5

% Censoring Correlation Pathway Size Error	25%												50%											
	0.2				0.5				0.2				0.5				0.2				0.5			
	5	L	E	L	5	L	E	L	5	L	E	L	5	L	E	L	5	L	E	L	5	L	E	L
<i>n</i> = 400																								
Marginal Gene	5	6	6	5	6	6	6	6	5	5	5	5	6	6	5	5	6	6	5	5	6	5	6	5
Global Test	4	3	3	3	3	3	3	3	2	2	3	4	3	2	2	2	3	4	3	2	2	2	2	3
Cox KM	Gaussian	5	5	5	5	6	5	6	6	4	5	6	5	4	5	4	5	6	5	4	4	5	5	5
	Linear	5	5	5	5	6	5	6	6	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	Quadratic	5	5	5	5	6	5	6	6	4	5	6	5	5	6	5	5	6	5	5	5	5	5	5
WLR _{Ge}	Gaussian	4	4	4	4	4	5	5	5	3	4	3	3	4	3	3	4	3	3	4	4	5	5	5
	Linear	4	4	3	3	4	5	4	5	3	4	2	3	4	2	3	4	2	3	4	4	4	4	4
	Quadratic	4	5	4	4	5	5	5	5	4	5	3	4	5	3	4	5	3	4	5	4	5	5	5
WLR _{LR}	Gaussian	5	5	4	4	6	5	6	6	4	5	5	4	5	5	4	5	5	4	5	5	6	5	5
	Linear	4	5	3	4	6	5	6	5	4	5	4	5	4	5	4	5	4	5	4	5	5	5	5
	Quadratic	5	5	4	5	6	5	6	6	4	5	5	4	5	5	4	5	5	4	5	5	6	6	6
Omnibus	4	4	3	4	5	5	5	5	5	3	4	3	4	3	4	3	4	3	4	3	4	5	5	5
Fisher	5	5	4	4	5	5	6	6	6	4	5	4	3	5	5	5	5	5	5	5	5	5	5	5
Fisher-Truncated	4	4	4	4	5	5	5	5	6	3	5	4	3	4	3	4	5	4	3	4	5	5	5	5
Min-P	4	4	3	3	5	5	5	5	5	3	5	3	3	4	3	3	4	3	3	4	4	5	5	5
Omnibus-Truncated	4	5	4	4	5	5	5	5	6	4	5	4	3	5	4	3	5	4	3	5	5	5	5	5

Table 2

Empirical power (%) for linear and nonlinear signal at Type I error rate of 5% when $n = 200$: Tests shown are identical to those in Table 1.

% Censoring	25%					50%				
	0.2		0.5			0.2		0.5		
	5	20	5	20	5	5	20	5	20	5
Pathway Size	E	L	E	L	E	L	E	L	E	L
Error	E	L	E	L	E	L	E	L	E	L
linear $h(z)$										
Marginal Gene	49	39	34	26	85	71	77	65	37	34
Global Test	52	33	40	26	82	56	74	49	37	31
Gaussian	61	47	49	38	89	75	84	73	46	43
Linear	64	49	50	39	90	77	86	74	48	44
Quadratic	63	49	49	39	89	77	85	73	48	45
Gaussian	50	51	35	38	78	81	74	79	37	41
Linear	49	51	31	34	79	81	73	78	35	39
Quadratic	52	53	35	39	79	82	75	80	39	43
Gaussian	64	50	50	38	90	78	87	75	49	45
Linear	64	49	46	36	91	78	86	74	49	44
Quadratic	67	52	51	39	91	79	87	76	51	47
Omnibus	59	50	42	37	88	81	85	79	43	42
Fisher	61	54	44	39	88	83	85	80	46	46
Fisher-Truncated	62	53	43	38	89	83	85	80	47	46
Min-P	60	51	43	38	88	81	84	79	45	44
Omnibus-Truncated	62	53	43	38	90	83	85	79	47	46
nonlinear $h(z)$										
Marginal Gene	26	23	15	13	22	22	16	16	21	21
Global Test	24	21	13	11	20	22	17	19	19	16
Cox KM	80	71	10	11	100	100	64	63	55	55
Gaussian										

% Censoring		25%						50%					
Correlation		0.2			0.5			0.2			0.5		
Pathway Size		5		20	5		20	5		20	5		20
Error		E	L	E	L	E	L	E	L	E	L	E	L
Linear		20	20	10	9	18	19	15	17	22	18	9	8
		6										8	7
Quadratic		34	32	10	9	77	74	26	29	31	28	9	8
		54	10									52	54
Gaussian		56	57	7	7	94	96	36	42	36	40	6	6
		72	80									17	21
WLR _{Ge}	Linear	21	21	6	5	7	8	6	7	19	18	4	7
		4										7	4
Quadratic		32	32	7	6	56	59	12	13	26	26	6	6
		43	8									40	43
Gaussian		77	69	10	10	100	100	67	67	53	54	10	9
		94	32									95	32
WLR _{LR}	Linear	19	19	8	7	18	19	15	17	22	19	8	7
		7										8	7
Quadratic		33	31	10	9	76	74	28	31	31	29	9	8
		57	12									56	57
Omnibus		82	78	11	10	100	100	92	92	59	60	8	6
		97	66									76	
Fisher		41	40	9	7	80	79	21	24	33	33	8	7
		57	10									52	57
Fisher-Truncated		39	38	9	8	85	86	25	30	31	30	7	6
		64	59	9	8	99	99	54	54	43	42	7	6
Min-P		91	21									90	91
		27										21	27
Omnibus-Truncated		45	42	9	8	94	91	23	27	34	34	7	6
		74	10									70	74

Empirical C-statistic (%) for predicting survival up to time t_0 for linear and nonlinear signal, where t_0 is approximately the 70th percentile of follow-up time. For each approach, the function $\hat{h}(\cdot)$ is estimated in a training set of size $n_{\text{train}} = 100$; and its prediction performance is assessed by estimating the C-statistic in an independent testing set of size $n_{\text{test}} = 1000$. For kernel methods, we use kernel PCA with 95% of the eigenvalues used, and present results using each kernel-weight combination, as well as results from two approaches for kernel selection: Omnibus-P, which uses the kernel-weight combination yielding the smallest p-value for its AFT-KM test in the training data; and Omnibus-C, which uses the kernel-weight combination yielding the largest C-statistic estimated in the training data. For comparison, we present the Cox model and the AFT model fit with Gehan and log-rank weights – these are fit with all pathway variables as linear covariates. Here, “E” denotes the extreme value distribution and “L” the logistic distribution, used for the error in the AFT model.

Table 3

% Censoring		25%						50%								
Pathway Size	Error	0.2		0.5		0.2		0.5		0.2		0.5				
		5		20		5		20		5		20				
		E	L	E	L	E	L	E	L	E	L	E	L			
linear $h(\mathbf{z})$																
Standard	AFT-G	64	63	59	59	67	67	63	62	63	63	58	59	67	61	62
	AFT-LR	64	63	60	58	68	66	63	62	63	63	59	58	67	67	62
	Cox	64	62	60	58	68	66	63	62	63	63	59	58	67	67	62
WLR _{Ce} Kernel	Gaussian	65	64	62	60	68	67	67	66	64	65	61	61	68	68	66
	Linear	64	63	62	60	68	67	67	65	64	64	61	61	68	68	66
	Quadratic	65	64	61	60	68	67	65	64	64	64	61	60	68	68	65
WLR _{LR} Kernel	Gaussian	65	64	62	60	69	68	67	66	65	65	62	61	68	68	66
	Linear	64	63	62	60	69	68	67	66	64	64	62	61	68	69	66
	Quadratic	65	63	61	60	68	67	66	64	64	64	61	60	68	68	65
Omnibus- P		65	63	62	60	68	67	67	65	64	64	62	61	68	68	66
Omnibus- C		64	63	62	60	68	67	66	65	64	64	61	61	68	68	65
nonlinear $h(\mathbf{z})$																
Standard	AFT-G	55	55	53	53	52	52	51	51	56	55	54	54	53	53	52
	AFT-LR	54	54	53	53	52	52	51	51	55	55	53	53	53	53	52
	Cox	54	54	53	53	52	52	51	51	55	55	53	53	53	53	52

% Censoring Correlation	Pathway Size Error	25%						50%					
		0.2			0.5			0.2			0.5		
		E	L	E	L	E	L	E	L	E	L	E	L
WLR _{Ge} Kernel	Gaussian	63	63	53	53	69	69	62	63	61	62	53	53
	Linear	55	54	53	53	52	52	51	51	56	55	54	54
	Quadratic	57	57	54	54	68	68	63	63	56	56	54	54
WLR _{LR} Kernel	Gaussian	63	63	53	53	69	69	64	64	63	63	53	53
	Linear	54	54	53	52	52	52	51	50	55	55	53	53
	Quadratic	58	58	54	54	68	68	63	63	56	56	54	54
Omnibus- <i>P</i>		63	62	53	53	69	69	63	63	62	62	53	53
Omnibus- <i>C</i>		64	64	54	54	69	69	64	64	63	63	54	54