Biometrics 000, 000–000 DOI: 000
000 0000

Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines

Arnab Maity

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

email: amaity@ncsu.edu

and

Xihong Lin

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

email: xlin@hsph.harvard.edu

Summary: We propose in this paper a powerful testing procedure for detecting a gene effect on a continuous outcome in the presence of possible gene-gene interactions (epistasis) in a gene set, e.g. a genetic pathway or network. Traditional tests for this purpose require a large number of degrees of freedom by testing the main effect and all the corresponding interactions under a parametric assumption, and hence suffer from low power. In this paper, we propose a powerful kernel machine based test. Specifically, our test is based on a garrote kernel method and is constructed as a score test. Here, the term garrote refers to an extra nonnegative parameter which is multiplied to the covariate of interest so that our score test can be formulated in terms of this nonnegative parameter. A key feature of the proposed test is that it is flexible and developed for both parametric and nonparametric models within a unified framework, and is more powerful than the standard test by accounting for the correlation among genes and hence often uses a much smaller degrees of freedom. We investigate the theoretical properties of the proposed test. We evaluate its finite sample performance using simulation studies, and apply the method to the Michigan prostate cancer gene expression data.

Key words: Garrote; Gene-gene interaction; Kernel machine; Mixed models; Restricted maximum likelihood; Score test; Semiparametric regression.

1. Introduction

In this paper, we consider the problem of testing for the effect of a gene on a continuous response in the presence of possible gene-gene interactions in a gene set, such as a genetic pathway. There is growing evidence that gene-gene interaction or epistasis ubiquitously contributes to complex traits, partly because of the sophisticated regulatory mechanisms encoded in the human genome. Epistasis is a phenomenon whereby the effects of a given gene on a biological trait are masked or enhanced by one or more genes (Bateson, 1909). It plays an important role in the mechanisms of complex diseases (Moore, 2005). Research has shown that it is important to account for gene-gene interactions in the search for susceptibility genes for complex diseases, and ignoring epistasis could explain difficulties in replicating significant findings (De Miglio et al., 2004; Aston et al., 2005). Developing powerful statistical methods for studying the effect of a gene by accounting for possible gene-gene interactions in a gene set is of significant interest in genetic association studies.

The data example that motivated the current research is the Michigan prostate cancer study data (Dhanasekaran et al., 2001). Recently there have been significant breakthroughs in the effort of finding candidate genes related to prostate cancer. Prostate Specific Antigen (PSA) is commonly used as a biomarker for prostate cancer screening. The early results of Dhanasekaran et al. (2001) indicate that certain functional genetic pathways seemed dysregulated in prostate cancer relative to non-cancerous tissues. There is a considerable literature to study the genetic pathway effects on PSA after adjusting for effects of clinical and demographic covariates (Dhanasekaran et al., 2001), and statistical methods have been developed to test for pathway effects (Liu, Lin and Ghosh, 2007; Liu, Ghosh and Lin, 2008). However, little work has been done on identifying the genes associated with PSA accounting for gene-gene interactions in a pathway.

Our goal in this paper is to develop a testing procedure for the effects of individual genes on a continuous outcome while accounting for possible epistasis and other clinical covariates in a regression model. An usual and popular approach to test for individual gene effects in such a context is to fit a simple linear model with main and interaction terms, e.g., two-

way interactions, and conduct an ANOVA based multi-degrees-of-freedom F-test (see for example, Howard et al, 2002; Li et al, 1997, among others). However, this approach has two main disadvantages. First, testing for the effect of a particular gene in the presence of possible gene-gene interactions requires testing for the corresponding interaction effects of the gene under consideration with other genes, resulting in a test using high degrees of freedom and a considerable loss of power, as we will see later. This phenomenon is more pronounced when genes are correlated among each other. In fact, for small sample sizes, such a test for all interaction terms may not be even computationally possible. Second, this also requires a strong parametric assumption, resulting in a power loss if the parametric model is misspecified. In fact, the true form of gene-gene interactions is typically unknown. Modeling the interaction terms using a standard two-way cross-product model may be overly simplistic and the resulting model could be misspecified.

In this paper, we address both issues by considering a general regression problem, where we model the joint effect of genes in a gene set, e.g., pathway or network, using a flexible parametric/nonparametric function. We then propose to test whether the function depends on an individual gene of interest. Specifically, we propose to use the powerful kernel machine framework to develop the test. The kernel machine framework, originally developed in the field of machine learning as a powerful learning technique for multi-dimensional data (Vapnik, 1998; Schølkopf and Smola, 2002), has become popular in genetic studies for dealing with a large number of genes. Popular examples of kernel machine methods include Support Vector Machine (SVM) (Vapnik, 1998), Relevance Vector Machine (RVM) (Tipping, 2000), a probabilistic model whose functional form is equivalent to SVM and Bayesian Gaussian process (Rasmussen and Williams, 2006). Kernel machine methods start with a kernel function which implicitly determines the smoothness property of the unknown function and hence greatly simplify specification of parametric and nonparametric models, especially for multi-dimensional covariates. Liu, Lin and Ghosh (2007) and Liu, Ghosh and Lin (2008) developed kernel machine regression theory for least squares and logistic regression and demonstrated their connections with mixed models, and developed a testing procedure for

the whole pathway effect. In contrast, we address in this paper the problem of testing for the effect of an individual gene of interest on a continuous outcome accounting for plausible gene-gene interactions, instead of testing the effect of the entire pathway/network.

To test for the gene effect in the presence of possible gene-gene interactions in a gene set, we introduce the concept of 'garrote kernel' where we attach an extra garrote parameter (Breiman, 1995) to the gene of interest in the kernel function and reparametrize our testing problem in terms of the new garrote parameter. The key advantage of this approach is that it allows us to reduce the dimensionality of the testing problem to a scalar parameter. We develop a score based testing procedure for the garrote parameter based on the mixed model framework and investigate theoretical properties. Such a score test effectively reduces the degrees of freedom by accounting for the correlation among genes, resulting in a more powerful test compared to the traditional F-test for all the interaction terms. It also allows for complex interactions.

The rest of the paper is organized as follows. Section 2 describes the joint semiparametric kernel machine regression framework for a gene set. Section 3 proposes the garrote kernel machine (GKM) score test for individual gene effects in the presence of possible gene-gene interactions and studies its theoretical properties. Section 4 discusses a multiple testing procedure in the GKM framework. Section 5 evaluates the performance of our test using simulation studies. Section 6 applies the proposed test to the Michigan prostate cancer gene expression data, followed by discussions in Section 7.

2. Semiparametric Model of Gene Effects in a Gene Set

Suppose we observe data from n subjects. For each subject $i = 1, \ldots, n$, let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iM})^{\mathrm{T}}$ be covariates associated with M genes within a gene set, e.g., a genetic pathway/network, Y_i be a continuous response and $\boldsymbol{X}_i$ be a set of clinical covariates. It is important to note that $\boldsymbol{X}$ and $\boldsymbol{Z}$ can contain both continuous and binary variables. We assume the following

model to relate the response to the clinical covariates and the genetic covariates:

$$Y_i \quad = \quad \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + h(Z_{i1}, \ldots, Z_{iM}) + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta}$ is a set of unknown regression coefficients for the clinical covariate effects, $h(\cdot)$ is an unknown function for the gene effects in a gene set, and ϵ_i's are independent random errors having Normal$(0, \sigma^2)$ distribution.

Typically, one assumes a parametric form for $h(\cdot)$ to model the gene effects in a gene set. For example, $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\eta}$ corresponds to a model with only main genetic effects, whereas $h(\boldsymbol{Z}_i) = \sum_{k=1}^{M} Z_{ik}\eta_k + \sum_{j<k} Z_{ij}Z_{ik}\gamma_{jk}$ allows for the first order gene-gene interactions. However, these classical specifications make strong parametric assumptions and are often subject to considerable power loss for testing for individual gene effects in the presence of gene-gene interactions as we will show later.

In this paper, we allow $h(\cdot)$ to be specified parametrically and nonparametrically in a unified kernel machine framework, which is more convenient and powerful to work with for multi-dimensional data. Specifically, rather than specifying the function $h(\cdot)$ using the basis functions as above, we specify $h(\cdot)$ using a positive definite kernel function $K(\cdot, \cdot)$. Mercer's theorem (Cristianini and Shawe-Taylor, 2000) guarantees that under some regularity conditions, the kernel function $K(\cdot, \cdot)$ implicitly specifies a unique function space, say $\mathcal{H}_K$, spanned by a particular set of orthogonal basis functions $\phi_j(\boldsymbol{z}), j = 1 \ldots, J$. Here orthogonality is defined with respect to the L_2 norm. Hence, the function space $\mathcal{H}_K$ has the property that any function $h(\cdot) \in \mathcal{H}_K$ can be represented in two ways: using a set of basis functions as $h(\boldsymbol{z}) = \sum_{j=1}^{J} \phi_j(\boldsymbol{z})\eta_j$ known as the primal or basis representation; or equivalently using the kernel function as $h(\boldsymbol{z}) = \sum_{k=1}^{L} K(\boldsymbol{z}_k^*, \boldsymbol{z}; \rho)\alpha_k$ for some integer L, some constants $\alpha_1, \ldots, \alpha_L$ and some $\{\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_L^*\}$. The later representation is called the dual representation.

Two most commonly used kernels are the d^{th} polynomial kernel and the Gaussian kernel. The d^{th} polynomial kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = (\rho + \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_2)^d$ corresponds to the models with dth-order polynomials including the cross product terms. For example, the first order polynomial kernel $(d = 1)$ corresponds to the model with only main effects $h(\boldsymbol{Z}_i) = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\eta}$, and the second order

polynomial kernel ($d = 2$) corresponds to the model with linear and quadratic main effects and two-way interactions $h(\boldsymbol{Z}_i) = \sum_{k=1}^{M} Z_{ik}\eta_{1k} + \sum_{j<k} Z_{ij}Z_{ik}\gamma_{jk} + \sum_{k=1}^{M} Z_{ik}^2\eta_{2k}$. The Gaussian kernel is $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \exp\{-\sum_{j=1}^{M}(z_{1j}-z_{2j})^2/\rho\}$, where ρ is a tuning parameter. The Gaussian kernel generates the function space spanned by the radial basis functions (Buhmann, 2003) and includes many nonlinear functions. Both dth polynomial kernel ($d > 1$) and Gaussian kernels allow for gene-gene interactions.

A few additional kernels are useful for modeling gene-gene interactions. For example, we can define the product linear kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \prod_{k=1}^{M}(1+z_{1k}z_{2k})$, which corresponds to the model $h(\boldsymbol{z}) = \beta_0 + \sum_k \beta_{1k}z_k + \sum_{j<k} \beta_{2jk}z_j z_k + \ldots + \beta_M z_1 \ldots z_M$, which includes all the multiplicative interactions up to the order M. If one wishes to include only two-way interactions in the model along with the main effects, i.e., the primal representation is $h(\boldsymbol{z}) = \beta_0 + \sum_k \beta_{1k}z_k + \sum_{j<k} \beta_{2jk}z_j z_k$, then the corresponding kernel can be specified as $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = 1+\sum_{k=1}^{M} z_{1k}z_{2k} + \sum_{j<k} z_{1j}z_{2j}z_{1k}z_{2k}$. We call this kernel the two-way interaction kernel. Examples of other choices of kernel functions include the sigmoid and neural network kernels, and the B-spline kernel (Scholkopf and Smola, 2002).

In theory, given any basis functions $\boldsymbol{\phi}(\boldsymbol{z}) = \{\phi_1(\boldsymbol{z}), \ldots, \phi_J(\boldsymbol{z})\}$ in the primal representation, one can construct the corresponding kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sum_{j=1}^{J} \phi_j(\boldsymbol{z}_1)\phi_j(\boldsymbol{z}_2)$ to facilitate the dual representation, and vice versa. For high-dimensional data, it is more convenient to work with the dual representation for $h(\cdot)$ using the kernel function $K(\cdot, \cdot)$, as will be done in this paper. The estimation and testing procedure will be described below.

3. Testing for a Gene Effect in the Presence of Gene-gene Interactions

In this section, we develop under model (1) a score based test for an individual gene effect accounting for gene-gene interactions in a gene set. Without loss of generality, consider testing for the effect of Z_1. Then our hypothesis is

$$H_0 : h(z_1, z_2, \ldots, z_M) = h(z_2, \ldots, z_M),$$

that is, the function $h(\cdot)$ does not depend on z_1. Note that the above formulation is quite general and covers a broad range of models, including the common parametric models for gene-gene interactions. For example, under the main effects only model $h(\boldsymbol{Z}_i) = \sum_{j=1}^{M} Z_{ij}\eta_j$, H_0 corresponds to the problem of testing for $\eta_1 = 0$. Under the classical two-way interaction model

$$h(\boldsymbol{Z}_i) = \sum_{k=1}^{M} Z_{ik}\eta_k + \sum_{j<k} Z_{ij}Z_{ik}\gamma_{jk}, \tag{2}$$

H_0 corresponds to testing for $\eta_1 = \gamma_{12} = \ldots = \gamma_{1M} = 0$, which requires M degrees of freedom and is subject to considerable loss of power.

The central idea of our approach is to use the dual representation of model (1) and introduce a garrote parameter in the kernel function to develop a more powerful test by reducing the degrees of freedom and borrowing information across genes within a gene set. To get the idea across, consider the traditional two-way interaction model (2), which can be specified using the two-way interaction kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = 1 + \sum_{k=1}^{M} z_{1k}z_{2k} + \sum_{j=1}^{M-1}\sum_{k=j+1}^{M} z_{1j}z_{2j}z_{1k}z_{2k}$, and can be rewritten as $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = (1 + z_{11}z_{21})(1 + \sum_{k=2}^{M} z_{1k}z_{2k}) + \sum_{j=2}^{M-1}\sum_{k=j+1}^{M} z_{1j}z_{2j}z_{1k}z_{2k}$. To test for the effect of Z_1, we introduce a garrote parameter δ to the terms involving Z_1, and define the garrote two-way interaction kernel as $K_{\mathrm{g}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \delta) = (1 + \delta z_{11}z_{21})(1 + \sum_{k=2}^{M} z_{1k}z_{2k}) + \sum_{j=2}^{M-1}\sum_{k=j+1}^{M} z_{1j}z_{2j}z_{1k}z_{2k}$. One can easily see that setting the null hypothesis $H_0 : \delta = 0$, we have $K(\boldsymbol{z}_1, \boldsymbol{z}_2; \delta = 0) = 1 + \sum_{k=2}^{M} z_{1k}z_{2k} + \sum_{j=2}^{M-1}\sum_{k=j+1}^{M} z_{1j}z_{2j}z_{1k}z_{2k}$, which corresponds the two-way interaction model with only $Z_2, \ldots, Z_M$. This formulation allows us to test for the effect of Z_1 using a single parameter instead of M parameters under the two-way interaction model, suggesting that we could have considerable power gain.

More generally, to test for a gene effect in presence of gene-gene interactions, it is relatively easy to construct a garrote version of any particular kernel that model interactions. For example, the garrote polynomial kernel is given by $K_{\mathrm{g}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \delta, \rho) = (\rho + \delta z_{11}z_{21} + \sum_{j=2}^{M} z_{1j}z_{2j})^d$, and the garrote Gaussian kernel function is $K_{\mathrm{g}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \delta, \rho) = \exp\{-\delta(z_{11} - z_{21})^2/\rho - \sum_{j=2}^{M}(z_{1j} - z_{2j})^2/\rho\}$. One can easily see that under $H_0 : \delta = 0$, $h(\cdot)$ corresponds to the model without Z_1 under the dth polynomial kernel and the Gaussian kernel respectively.

In general, given a kernel, the garrote kernel can be obtained by simply multiplying the garrote $\delta^{1/2}$ to the first component of the covariate vector and constructing the usual kernel using this modified covariates. In this formulation we now test for

$$H_0 : \delta = 0.$$

This formulation has two main advantages, as we will observe later. First, the testing problem is reduced to a one dimensional problem from a possibly high or infinite dimensional (in the case $h(\cdot)$ is modeled nonparametrically) problem. Second, in contrast to the usual F-test that uses M degrees of freedom to test the effect of Z_1, our method uses a simple scaled chi-squared distribution $m\chi^2_d$ with much smaller degrees of freedom that are estimated from the data by accounting for correlation among genes, as we will discuss in Section 3.2. As shown in the simulation studies, d is much smaller than M and hence our test gains more power than the F-test.

3.1 *Derivation of the score test statistic*

As discussed before, we assume the function $h(\cdot)$ belongs to a functional space $\mathcal{H}_K$ with a kernel $K(\cdot,\cdot)$. Under the full model, $\boldsymbol{\beta}$ and $h(\cdot)$ are estimated by maximizing the penalized likelihood function, apart from a constant

$$J(\beta, h) \;=\; -\frac{1}{2}\sum_{i=1}^{n}\{Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - h(\boldsymbol{Z}_i)\}^2 - \lambda_*^{-1}||h||^2_{\mathcal{H}_K}/2, \tag{3}$$

where λ_* is a tuning parameter which controls the trade off between goodness of fit and complexity of the model, and $||h||_{\mathcal{H}_K}$ denotes the functional norm in the functional space $\mathcal{H}_K$. In particular, if we consider the primal representation $h(\boldsymbol{z}) = \sum_{j=1}^{J}\phi_j(\boldsymbol{z})\eta_j$, where $\phi_j, j = 1,\ldots,J$ are orthonormal basis functions, then $||h||^2_{\mathcal{H}_K}$ is written as $\sum_{j=1}^{J}\eta_j^2$. However, we will use the dual representation of $h(\cdot)$. Liu et al (2007) discussed the estimation procedure using this penalized loglikelihood function via linear mixed models in detail. Specifically, using the representer theorem (Kimeldorf and Wahba, 1970), the solution of (3) can be written as $h(\cdot) = \sum_{i=1}^{n}\alpha_i K(\cdot, \boldsymbol{Z}_i)$, where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)^{\mathrm{T}}$ is an unknown parameter vector. Instead of using the original kernels, we propose to use the garrote kernels $K_{\mathrm{g}}(\cdot,\cdot)$ to express

$h(\cdot)$. As noted in Liu et al (2007), maximization of (3) is equivalent to maximizing

$$\begin{aligned} J(\boldsymbol{\beta}, h) &= -\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{K}_{\mathrm{g}}(\delta,\rho)\boldsymbol{\alpha}\}^{\mathrm{T}}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{K}_{\mathrm{g}}(\delta,\rho)\boldsymbol{\alpha}\}/2 \\ &\quad -\lambda_*^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_{\mathrm{g}}(\delta,\rho)\boldsymbol{\alpha}/2, \end{aligned} \tag{4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and $\boldsymbol{K}_{\mathrm{g}}(\delta, \rho)$ is an $n \times n$ matrix with the $(i,j)^{\mathrm{th}}$ element given by $\boldsymbol{K}_{\mathrm{g},ij}(\delta,\rho) = K_{\mathrm{g}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \delta, \rho)$. We reiterate that $K(\cdot,\cdot)$ and $K_{\mathrm{g}}(\cdot,\cdot)$ denote the kernel and the corresponding garrote kernel functions, and $\boldsymbol{K}(\cdot)$ and $\boldsymbol{K}_{\mathrm{g}}(\cdot,\cdot)$ denote $n \times n$ matrices resulting from applying the kernel functions to the covariates $\boldsymbol{Z}$.

It is important to note that the criterion function $J(\boldsymbol{\beta}, h)$ can be viewed as a penalized log-likelihood function of a linear mixed model (Harville, 1977; Laird and Ware, 1982). Specifically, Liu et al (2007) show that one can consider an equivalent linear mixed model regression problem

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + h_i + \epsilon_i, \tag{5}$$

with $\boldsymbol{h} = (h_1, \ldots, h_n)^{\mathrm{T}}$ is a vector of random effects following a Normal$\{0, \lambda\boldsymbol{K}_{\mathrm{g}}(\delta,\rho)\}$ distribution where $\lambda = \lambda_*\sigma^2$. In this setup, the parameters λ and ρ can be considered as variance components. Using this formulation, the parameters can be estimated using the standard mixed model theory, see Liu et al (2007) for details.

The main focus of this paper is to test for the main effect of a gene in the presence of gene-gene interactions using the garrote kernel framework, that is, testing for $H_0 : \delta = 0$. Specifically, we propose a variance component test for $H_0 : \delta = 0$ by using the mixed model formulation (5) and derive a score statistic. Specifically, the marginal covariance matrix of $\boldsymbol{Y}$ under (5) is given by $\boldsymbol{V}(\boldsymbol{\theta}) = \sigma^2 I + \lambda\boldsymbol{K}_{\mathrm{g}}(\delta,\rho)$, where $\boldsymbol{\theta} = (\delta, \rho, \lambda, \sigma^2)$. The log-likelihood for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is then given by

$$L_{\mathrm{ML}} = -\log|\boldsymbol{V}(\boldsymbol{\theta})|/2 - (\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\theta})(\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})/2.$$

One can estimate the variance components λ and ρ by maximizing L_{ML}. However, a major disadvantage of the maximum likelihood approach is that it does not take into account the loss of degrees of freedom due to estimation of the fixed effects $\boldsymbol{\beta}$, and the resulting estimators

for the variance components are usually biased (Harville, 1977). We therefore propose to use the restricted maximum likelihood (REML)

$$\begin{aligned} L_{\text{REML}} &= -\log|\boldsymbol{V}(\boldsymbol{\theta})|/2 - \log|\boldsymbol{X}\boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X}^{\mathrm{T}}|/2 \\ &\quad -(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\theta})(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})/2. \end{aligned}$$

Under $H_0 : \delta = 0$, we set $\boldsymbol{\theta}_0 = (0, \rho, \lambda, \sigma^2)$ and

$$\boldsymbol{P}_0(\boldsymbol{\theta}_0) = \boldsymbol{V}^{-1}(\boldsymbol{\theta}_0) - \boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}^{\mathrm{T}}\{\boldsymbol{X}\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}^{\mathrm{T}}\}^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}(\boldsymbol{\theta}_0).$$

The REML based score function of δ evaluated at H_0 is

$$\begin{aligned} S_{\delta,n} &= (\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta}_0)^{-1}\{\lambda\boldsymbol{K}_{\mathrm{g},\delta}(0,\rho)\}\boldsymbol{V}(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})/2 \\ &\quad -\mathrm{tr}\{\lambda\boldsymbol{K}_{\mathrm{g},\delta}(0,\rho)\boldsymbol{P}_0\}/2, \end{aligned}$$

where $\boldsymbol{K}_{\mathrm{g},\delta}(\delta,\rho)$ denotes the derivative of $\boldsymbol{K}_{\mathrm{g}}(\delta,\rho)$ with respect to δ. To test for $H_0 : \delta = 0$, we propose to use the score-based test statistic

$$T_n = (\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}\{\widehat{\lambda}\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})\}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})/2, \tag{6}$$

where $\widehat{\boldsymbol{\theta}}_0 = (0, \widehat{\rho}, \widehat{\lambda}, \widehat{\sigma}^2)$ and $\widehat{\boldsymbol{\beta}}$ are estimators of $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}$ constructed under the null hypothesis $H_0 : \delta = 0$.

It is important to note that a major advantage of our testing procedure is that one only needs to estimate the parameters under the null model. Recall that the full model is given in (1). Under $H_0 : \delta = 0$, the reduced model is

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + h(Z_{2i}, \ldots, Z_{Mi}) + \epsilon_i.$$

Define $\boldsymbol{Z}_{i0} = (Z_{2i}, \ldots, Z_{Mi})^{\mathrm{T}}$. The model components for this reduced model can be estimated following the procedure of Liu et al (2007, Section 4) by fitting the linear mixed model formulation

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + h_i + \epsilon_i, \tag{7}$$

where $h_i \sim N\{0, \lambda\boldsymbol{K}_0(\rho)\}$ and $\boldsymbol{K}_0$ is the kernel matrix constructed under H_0 using the genes $Z_2, \cdots, Z_M$. We can estimate $\boldsymbol{\beta}$ and $h(\cdot)$ using the best linear unbiased predictors (BLUPs), and estimate the variance components using REML under the null working linear mixed

model (7). The numerical properties of BLUPs and REML estimates are well known in the mixed model literature, see for example Harville (1977). We refer to Liu et al (2007) for more details about fitting the null model.

3.2 *The Null distribution of the test statistic*

To test for $H_0 : \delta = 0$, we need to derive the distribution of the score-based test statistics T_n under H_0. To start, we first note that the information matrix of $\boldsymbol{\theta}$ based on the REML function is given by

$$\boldsymbol{I} = \begin{bmatrix} I_{\delta\delta} & \boldsymbol{I}_{\delta\boldsymbol{\theta}_0}^{\mathrm{T}} \\ \boldsymbol{I}_{\delta\boldsymbol{\theta}_0} & \boldsymbol{I}_{\boldsymbol{\theta}_0\boldsymbol{\theta}_0} \end{bmatrix},$$

where

$$\boldsymbol{I}_{\theta_j\theta_k} = \mathrm{tr}\Big\{\boldsymbol{P}_0(\boldsymbol{\theta})\frac{\partial \boldsymbol{V}(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{P}_0(\boldsymbol{\theta})\frac{\partial \boldsymbol{V}(\boldsymbol{\theta})}{\partial \theta_k}\Big\}/2.$$

Hence the efficient information of δ accounting for the fact that $\widehat{\lambda}, \widehat{\sigma}^2, \widehat{\rho}$ are estimated is given by

$$\widehat{I}_{\delta\delta} = I_{\delta\delta} - \boldsymbol{I}_{\delta\boldsymbol{\theta}_0}^{\mathrm{T}}\boldsymbol{I}_{\boldsymbol{\theta}_0\boldsymbol{\theta}_0}^{-1}\boldsymbol{I}_{\delta\boldsymbol{\theta}_0},$$

where all the quantities are evaluated at $\widehat{\boldsymbol{\theta}}_0$ computed under H_0. Note that we have used the fact that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are orthogonal in linear mixed models and there is no loss of information for δ when estimating $\boldsymbol{\beta}$.

Note that our proposed test statistic T_n is a quadratic form in $\boldsymbol{Y}$ and follows a mixture of chi-squared distribution under $H_0 : \delta = 0$. We use the Satterthwaite approximation to approximate its distribution by a scaled chi-squared distribution $m\chi_d^2$. The parameters m and d are calculated equating the mean and variance of T_n to those of $m\chi_d^2$. Specifically, we solve $md = \widehat{\lambda}\,\mathrm{tr}\{\boldsymbol{P}_0(\widehat{\boldsymbol{\theta}}_0)\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})\}/2$ and $2m^2d = \widehat{I}_{\delta\delta}$, which in turn gives us the solution

$$\begin{aligned} \widehat{m} &= \widehat{I}_{\delta\delta}/[\widehat{\lambda}\,\mathrm{tr}\{\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\theta}}_0)\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})\}]; \\ \widehat{d} &= [\widehat{\lambda}\,\mathrm{tr}(\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\theta}}_0)\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho}))]^2/(2\widehat{I}_{\delta\delta}). \end{aligned}$$

Now the p-value of the test can be obtained using the tail probabilities of the $\widehat{m}\chi^2_{\widehat{d}}$ distribution. Note that the degrees of freedom d is estimated from the data and accounts for

the correlation among genes. We will demonstrate in the simulation study section that d is often much smaller than M and hence the garrote kernel machine test often has considerably higher power than the M-df F-test.

Note that $\widehat{m}$ and $\widehat{d}$ depend on $\boldsymbol{K}_{\mathrm{g},\delta}(\cdot,\cdot)$, the derivative of the garrote kernel with respect to the garrote parameter δ. For many frequently used kernels, this matrix can be easily computed. For example, in the case of polynomial kernel, the (i,j)-th element of the derivative matrix is $\boldsymbol{K}_{\mathrm{g},\delta,ij}(\delta=0,\rho) = dZ_{i1}Z_{j1}(\rho+\sum_{k=2}^{M} Z_{ik}Z_{jk})^{d-1}$; for Gaussian kernel, $\boldsymbol{K}_{\mathrm{g},\delta,ij}(\delta=0,\rho) = \{-(Z_{i1}-Z_{j1})^2/\rho\}\exp\{-\sum_{k=2}^{M}(Z_{ik}-Z_{jk})^2/\rho\}$, and for product linear kernel, $\boldsymbol{K}_{\mathrm{g},\delta,ij}(\delta=0) = Z_{i1}Z_{j1}\prod_{k=2}^{M}(1+Z_{ik}Z_{jk})$.

3.3 *Principal Component Analysis (PCA) Based Testing Procedure*

Recall that the proposed test statistic T_n given in (6) is a quadratic form and hence follows a weighted mixture of χ^2-distribution under H_0. Specifically, let $\zeta_1,\ldots,\zeta_n$ be the eigenvalues of $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ with corresponding eigenvectors $\boldsymbol{U}=[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n]$. Then we can write

$$T_n = \widehat{\lambda}\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{\psi}, \tag{8}$$

where $\boldsymbol{\psi} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{D} = diag(\zeta_1,\ldots,\zeta_n)$. The Satterthwaite approximation of the distribution of T_n derived in Section 3.2 inherently assumes that the eigenvalues $\zeta_1,\ldots,\zeta_n \geq 0$, that is, the derivative matrix $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ is positive semi-definite. If this condition is violated then T_n can potentially take negative values and the chi-squared approximation becomes invalid.

Note that this problem depends on the choice of the kernel function $K(\cdot,\cdot)$. For many usual kernel functions, this problem does not arise at all. For example, in the case of the linear kernel, polynomial kernel, product linear kernel and the two-way interaction kernel, one can easily show that the matrix $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ is in fact positive semi-definite and hence the chi-squared approximation in Section 3.2 is valid and the testing procedure remains valid.

One notable exception is the case where one uses the Gaussian kernel $K_{\mathrm{g}}(\boldsymbol{z}_1,\boldsymbol{z}_2;\delta,\rho) =$

$\exp\{-\delta(z_{11}-z_{21})^2/\rho - \sum_{j=2}^{M}(z_{1j}-z_{2j})^2/\rho\}$. In this case, we have

$$\boldsymbol{K}_{\mathrm{g},\delta} = \boldsymbol{M}_1 \odot \boldsymbol{M}_2,$$

where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are matrices such that $\boldsymbol{M}_{1,ij} = -(Z_{i1}-Z_{j1})^2/\rho$ and $\boldsymbol{M}_{2,ij} = \exp\{-\sum_{k=2}^{M}(z_{ik}-z_{jk})^2/\rho\}$, and $\odot$ denotes the element-wise product or Schur product of two matrices. In this case, the matrix $\boldsymbol{K}_{\mathrm{g},\delta}$ can be an indefinite matrix and hence the chi-squared approximation in Section 3.2 may not be valid.

To address this problem, we propose a principal component based solution. Specifically, utilizing the form of the test statistic in (8), we propose the following modified test statistic

$$T_n^* = \widehat{\lambda}\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{D}^*\boldsymbol{\psi},$$

where $\boldsymbol{D}^* = diag(|\zeta_1|,\ldots,|\zeta_n|)$ is constructed using the absolute values of the eigenvalues. In other words, the T_n^* can be written as

$$T_n^* = (\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}\{\widehat{\lambda}\boldsymbol{K}^*_{\mathrm{g},\delta}(0,\widehat{\rho})\}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})/2, \tag{9}$$

where $\boldsymbol{K}^*_{\mathrm{g},\delta}(0,\widehat{\rho}) = \boldsymbol{U}\boldsymbol{D}^*\boldsymbol{U}^{\mathrm{T}}$ is a reconstruction of $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ using absolute eigenvalues. Note that when $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ is non-negative definite, $T_n = T_n^*$.

Another option is to consider

$$T_n^{**} = (\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}\{\widehat{\lambda}\boldsymbol{K}^{**}_{\mathrm{g},\delta}(0,\widehat{\rho})\}\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_0)^{-1}(\boldsymbol{Y}-\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})/2, \tag{10}$$

where $\boldsymbol{K}^{**}_{\mathrm{g},\delta}(0,\widehat{\rho}) = \boldsymbol{U}\boldsymbol{D}^{**}\boldsymbol{U}^{\mathrm{T}}$ is a reconstruction of $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ using $\boldsymbol{D}^{**} = diag[\zeta_1 1(\zeta_1 > 0),\ldots,\zeta_n 1(\zeta_n > 0)]$, where one replaces the negative eigenvalues with zero.

It is important to note that T_n^* and T_n^{**} are valid test statistics, and $\boldsymbol{K}^*_{\mathrm{g},\delta}(0,\widehat{\rho})$ and $\boldsymbol{K}^{**}_{\mathrm{g},\delta}(0,\widehat{\rho})$, by construction, are positive semi-definite matrices. Hence we can now apply the Satterthwaite approximation on T_n^* or T_n^{**} to obtain the null distribution as a scaled chi-squared distribution as demonstrated in Section 3.2, where one uses $\boldsymbol{K}^*_{\mathrm{g},\delta}(0,\widehat{\rho})$ or $\boldsymbol{K}^{**}_{\mathrm{g},\delta}(0,\widehat{\rho})$ instead of $\boldsymbol{K}_{\mathrm{g},\delta}(0,\widehat{\rho})$ in all the expressions.

It is worth mentioning that in our simulation study, the proportion of time T_n assumed negative values for the Gaussian kernel case is negligible (0.5-5%) and hence did not pose any real problem. However, the PCA based reconstruction approach is practically very useful

when one applies our method to a real data set simply because a negative test statistic in that case will result in p-value being 1.0 whereas the modified test will produce a valid and sensible p-value.

REMARK 1: It is worth noting that one can derive the exact distribution of an indefinite quadratic form in multivariate normal variables as in (6), see for example Johnson and Kotz (1970), Section 7. However, the exact distribution and hence the p-value are numerically very difficult to compute and require to perform contour integration on a complex plane. Thus we prefer the PCA based approach for its easy accessibility. If one wants to keep the original indefinite quadratic form, critical values can be determined using simulations.

4. Multiple Comparison Procedure

The score test proposed in Section 3 is developed to test for a single gene at a time. However, when testing for genes in a gene set, the p-values from individual tests can be misleading and may result in false discoveries. Specifically, when comparing M genes in a pathway/network, we are interested in simultaneously testing M null hypotheses each corresponding to one gene. Let the individual p-values be given by $p_1, \ldots, p_M$. A typical approach is that instead of using the individual p-values, one modifies them to obtain modified p-values adjusted for multiple testing. The simplest example is the Bonferroni correction where the adjusted p-values are given by $p_k^{Bon} = p_k M$, $k = 1, \ldots, M$. However, this approach is overly conservative and does not account for the correlation among the genes, and often leads to a very small number of discoveries. An alternative approach is based on false discovery rates (Benjamini and Hochberg, 1995). However, FDR requires the hypotheses are independent or block-independent. This assumption is not appropriate in many cases, specially if one is interested in testing for genes in a particular pathway where there are gene-gene interactions present.

We instead propose a permutation-based procedure for multiple comparisons that can be used with our proposed score test. We present the procedure for the case where there are no clinical covariates $\boldsymbol{X}$, and indicate afterward how the clinical covariates can be handled.

Recall that our model is

$$Y_i = h(Z_{i1}, \ldots, Z_{iM}) + \epsilon_i, \tag{11}$$

where we test each of the M genes one by one using our testing procedure. Let the individual p-values be $p_1, \ldots, p_M$. The multiple comparison procedure is as follows:

(1) Fit the full model (11) and obtain the residual vector $\widehat{\epsilon}$.

(2) For any given $m \in \{1, \ldots, M\}$,

   (a) Fit the null model $Y_i = h(Z_{i1}, \ldots, Z_{i,m-1}, Z_{i,m+1}, \ldots, Z_{iM}) + \epsilon_i$, and estimate the effect under H_0, that is $\widehat{h}_{-m}(\cdot) = \widehat{h}(Z_{i1}, \ldots, Z_{i,m-1}, Z_{i,m+1}, \ldots, Z_{iM})$

   (b) Permute the residual vector $\widehat{\epsilon}$ randomly to obtain $\widehat{\epsilon}^*$ and compute $Y^* = \widehat{h}_{-m}(\cdot) + \widehat{\epsilon}^*$

   (c) Perform our proposed score test using Y^* as observations and Z as the covariates and obtain the p-value p^*_m.

   (d) Repeat the steps (b) and (c) for a large number, say $b = 1, \ldots, B$, of times and obtain p-values for the m^{th} gene $p^*_{m1}, \ldots, p^*_{mB}$.

(3) Repeat step 2 for each $m = 1, \ldots, M$.

(4) Derive the distribution of $p^* = \min(p^*_1, \ldots, p^*_M)$ and compute the adjusted p-value of the m^{th} gene as the tail probability of this distribution, that is, the p-value of the m^{th} gene is

$$p^{\text{Adj}}_m = B^{-1} \sum_{b=1}^{B} I[p_m > \min\{p^*_{1b}, \ldots, p^*_{Mb}\}].$$

(5) The m^{th} hypothesis is rejected al level α if $p^{\text{Adj}}_m < \alpha$.

Note that in presence of extra variables $\boldsymbol{X}$, one just needs to include $\boldsymbol{X}$ as a linear predictor while fitting the full and null models. The estimation procedure will then be same as described in Section 3.1. The rest of the procedure and computation of the p-values remain the same as above.

5. Simulation Study

We performed a simulation study to evaluate the finite-sample performance of our garrote kernel machine (GKM) test and compare its performance with the usual F-test. The data sets were simulated from the true model

$$Y_i \;=\; \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\eta} + h(Z_{i1},\ldots,Z_{iM}) + \epsilon_i,$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2})^{\mathrm{T}}$ were generated from a standard bivariate normal distribution and the true value of $\boldsymbol{\eta} = (0.2, 0.2)^{\mathrm{T}}$. The random errors ϵ_i's were generated from a standard normal distribution. We generated $Z_1,\ldots,Z_M$ from a normal distribution with the compound symmetry correlation structure with marginal variance 1.0 and correlation r. We considered different values of $r = 0, 0.2$ and 0.5, and four different settings for sample sizes and number of genes: $(n, M) = (50, 5)$, $(n, M) = (100, 8)$, $(n, M) = (200, 10)$ and $(n, M) = (200, 20)$. For each of these settings, we considered four different functions $h(\cdot)$ as follows. First define

$$g(Z_{i1},\ldots,Z_{iM}) = 1 + \sum_{k=1}^{M} Z_{ik}\beta_k + \sum_{k=2}^{m} Z_{i1}Z_{ik}\gamma_k,$$

where $m \leq M$. Here M is the number of genes in a pathway/gene set and m is the number of genes that are interacting with Z_1. We test for the effect of Z_1. For example, $m = M$ implies that Z_1 is interacting with all other variables in the model, and $m = 2$ refers to the case where Z_1 interacts only with Z_2. The four different functions of $h(\cdot)$ are

(1) $h_1(Z_{i1},\ldots,Z_{iM}) = g(Z_{i1},\ldots,Z_{iM})$ with $m = 2$,

(2) $h_2(Z_{i1},\ldots,Z_{iM}) = g(Z_{i1},\ldots,Z_{iM})$ with $m = M$,

(3) $h_3(Z_{i1},\ldots,Z_{iM}) = sign\{g(\cdot)\}|g(\cdot)|^{1/2}$ with $m = 2$,

(4) $h_4(Z_{i1},\ldots,Z_{iM}) = sign\{g(\cdot)\}|g(\cdot)|^{1/2}$ with $m = M$.

Note that case (1) and (2) are linear interaction functions and (3) and (4) are nonlinear.

For each of these cases, we applied our testing procedure to test for the effect of Z_1. For type I error and power calculations, we set the true values of $\beta_2 = \ldots = \beta_M = 0.7$ and $\beta_1 = c/10$, $\gamma_2 = \ldots = \gamma_m = c/20$, where we varied $c = 0, 2, 4, 6, 8$. Here $c = 0$ corresponds to the null hypothesis of no effect of Z_1 and was used to study for the size of the test.

We generated 2,000 data sets for each case to compute sizes and powers of our test at a nominal level of 0.05. To test for the effect of Z_1, we used the GKM test with the Gaussian kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \exp\{-\sum_{j=1}^{M}(z_{1j}-z_{2j})^2/\rho\}$ and the two-way interaction kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = 1+\sum_{k=1}^{M} z_{1k}z_{2k}+\sum_{j<k} z_{1j}z_{2j}z_{1k}z_{2k}$. For the Gaussian kernel, we computed the p-values based on both the scaled chi-squared approximation using the test statistic in (6) as in Section 3.2 and also the PCA based modified test statistics (9) and (10) as described in Section 3.3. Note that for the two-way interaction kernel, the PCA based method and the scaled chi-squared approximation are identical, because $\boldsymbol{K}_{\mathrm{g},\delta}(\cdot)$ is positive definite. Hence we only report results for one test. We also compared the power of our GKM test to that of the usual ANOVA based F-test where one fits a linear model with all main and two-way interaction effects. For linear cases ($h_1(\cdot)$ and $h_2(\cdot)$), we also present the results for the classical score test assuming the true direction under the alternative $(\beta_1, \gamma_2, ..., \gamma_m) = c(0.10, 0.05, ..., 0.05)$ is known, that is, one only tests for $c = 0$ against $c \neq 0$. This of course is not possible in real data situations but in our simulation study this test serves as an ideal benchmark. We do not include the results of the classical score tests for the nonlinear models ($h_3(\cdot)$ and $h_4(\cdot)$) because in our simulations we found that the parametric estimation of these nonlinear models, even under H_0, are unstable and practically infeasible, especially for large p. In fact, this is one of the reasons for adopting a kernel machine based test for nonlinear models.

The results are given in Tables 1 - 4. These results show that all the tests have the size close to the nominal value $\alpha = 0.05$. The M-df F-test loses power in all the cases, specially when there is correlation among the covariates ($r = 0.2$) or ($r = 0.5$). For the two-way interaction models h_1 and h_2, when $r = 0$, the F-test has a similar performance to that of the two-way interaction kernel but both are outperformed by the Gaussian kernel. When $r = 0.2$, the two-way interaction kernel-based test improves the power over the F-test considerably, while both are outperformed by the Gaussian-kernel based test. Note that when $r = 0$, although the covariates $(Z_1, \cdots, Z_m)$ are independent, however, the interaction terms $Z_1Z_2, \cdots, Z_1Z_m$ are correlated. The GKM test is able to take such correlation into account when testing for the effect of Z_1 and hence is more powerful. For the nonlinear cases h_3 and h_4, the Gaussian

kernel-based test performs much better than both the F-test and the two-way interaction kernel test, as expected.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

For the Gaussian kernel, we also compared the power of the GKM test with the PCA based tests. It is evident that in all the cases considered, the GKM test and the two PCA based tests perform very similarly. This is simply because the proportion of times the test statistic using the Gaussian kernel takes a negative value is very small in all the cases, ranging between 0.5% to 5% and hence their impact on the power of the test is negligible. However, we recommend the PCA based test for practical purposes when one applies the test to a real data set simply because the original GKM test based on the Satterthwaite approximation may result in a negative test statistic giving a p-value 1.0. In contrast, the PCA based test will provide a valid test and produce a valid and sensible p-value.

Comparing the results from the GKM tests with the ideal score test by hypothetically assuming the true direction of the alternative were known, we see that the GKM test performs closely to the ideal score test for h_1, especially for $(n, p) = (50, 5), (100, 8)$ and $(200, 10)$ with low to moderate correlation $(r = 0, 0.2)$. However, in other cases, the GKM test loses some power compared to the best power possible. This result is expected because the ideal score test assumes that the direction of the alternative is known in advance. This is not feasible in practice. It is also interesting to note that given a known direction of the alternative $(\beta_1, \gamma_2, \ldots, \gamma_m) = c(\beta_1^*, \gamma_2^*, \ldots, \gamma_m^*)$, and the knowledge of the true form of $h(\cdot)$, one can rewrite $h(z_1, z_2, \ldots, z_m) = h(z_1^*, z_2, \ldots, z_m)$, where $z_1^* = z_1\beta_1^* + \sum_{k=2}^m z_1 z_k \gamma_m^*$. In case of linear functions such as $h_1(\cdot)$ and $h_2(\cdot)$, testing for Z_1 then amounts to using a linear kernel with the covariates $Z_1^*, Z_2, \ldots, Z_m$ and testing for Z_1^*. However for nonlinear cases, one still needs to use Gaussian kernel to capture the nonlinearity of the function. Nevertheless, these tests

are still hypothetical as neither the true direction of the alternative or the actual functional form of $h(\cdot)$ are known in practice.

To gain insight about the improvement in power for the Gaussian kernel and two-way interaction kernel over F-test, we summarize the degrees of freedom used for tests based on both kernels. The boxplots of degrees of freedom used in the chi-squared approximation for the case $n = 50$ and $M = 5$ with $c = 6$ in displayed in Figure 1. Note that while the F-test uses 5 degrees of freedom, the Gaussian-kernel based test uses an average 1.5 DF while the two-way interaction kernel test uses an average 4 DF, implying considerable power gain. The Gaussian kernel uses much less degrees of freedom than the two-way interaction kernel and hence performs better than the interaction kernel. As the correlation r among $Z_1, \ldots, Z_M$ increases, the degrees of freedom used by the two-way interaction kernel decreases and hence gaining more power over the F-test. This is also supported by the results in Table 1.

[Figure 1 about here.]

6. The Prostate Cancer Pathway Data Example

We applied our proposed GKM testing procedure to analyze the prostate cancer genetic pathway data from the Michigan prostate cancer study (Dhanasekaran, et al, 2001). The data set contained 59 patients who were clinically diagnosed with local or advanced prostate cancer. An objective of the study is to evaluate whether a gene of interest in a genetic pathway has an overall effect on pre-surgery prostate specific antigen (PSA) after adjusting for covariates and accounting for interactions with other genes in the pathway. For each patient, cDNA microarray gene expressions were collected. Liu et al (2007) considered this data set and the cell growth pathway which contains 5 genes. They tested for the joint effect of the whole pathway and found the effect to be statistically significant. We are interested in testing for the effects of the individual genes accounting for gene-gene interactions on PSA level. We took a log transformation of PSA to make the normality assumption more plausible. We included two covariates: age and Gleason score, a well established histological

grading system for prostate cancer, in our model,

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + h(Z_{i1}, \ldots, Z_{i5}) + \epsilon_i,$$

where Y denotes the log-PSA level, X_1 and X_2 denote age and Gleason score, and $Z_{i1}, \ldots, Z_{i5}$ denote the gene expression levels of the 5 genes in the cell growth pathway. The symbols and description of the five genes in the pathway are provided in Table 5.

We started by fitting a main effects only linear model, that is, setting $h(Z_{i1}, \ldots, Z_{i5}) = \eta_0 + Z_{i1}\eta_1 + \ldots + Z_{i5}\eta_5$, and tested for each gene separately. For comparison purposes, we fit a model including all the main effects and two-way interactions, that is, we set $h(Z_{i1}, \ldots, Z_{i5}) = \eta_0 + \sum_{k=1}^{5} Z_{ik}\eta_k + \sum_{k=1}^{5}\sum_{j>k} \eta_{jk} Z_{ij} Z_{ik}$. In this setting, we performed the usual F-test for each gene. The individual p-values are reported in Table 5, columns 3 and 4, respectively. Note that the main-effect-only naive test suggests that the gene FGF7 has a significant effect on PSA levels. However, when we incorporated all the two-way interaction, this effect disappears using the 5-df F-test.

We next applied our score testing procedure to test for individual gene effect accounting for gene-gene interactions (epistasis). We used the Gaussian kernel and the two-way interaction kernel to model the function $h(\cdot)$. For the Gaussian kernel, we implemented both the original GKM scaled chi-squared approximation (Section 3.2) and the PCA based modified test based on T_n^* (Section 3.3). Note that these two tests are identical when the two-way linear kernel is used and hence we only report one p-value for the garrote two-way interaction kernel test. The results are presented in Table 6. The Gaussian kernel based test found the gene FGF7 was significant. However, the two-way linear kernel did not capture this result. This seems to imply that the interaction structure between the genes may not be simply multiplicative and the nonparametric modeling of the covariates is in fact needed. Further as shown in the simulation study, the Gaussian-kernel based test has more power.

[Table 5 about here.]

[Table 6 about here.]

However, individual p-values may be misleading and we need to take into account the

multiple testing issue. We employed the multiple comparison procedure described in Section 4 to obtain adjusted p-values. We used $B = 10000$ permutations to generate the p-value distribution. The distribution of p^* as described in Section 4, step 4 is displayed in Figure 2. The adjusted p-values are displayed in columns 5-7 of Table 6. It is evident that FGF7 is still found to be significant by the Gaussian kernel based tests.

[Figure 2 about here.]

7. Discussion

We have proposed in this paper a kernel machine framework to test for individual gene effects on a continuous outcome in the presence of possible gene-gene interactions in a gene set/pathway using the garrote kernel (GKM) test. We have developed a garrote kernel based score test by introducing a garrote parameter in the kernel function. This framework does not require us to explicitly test for all interaction terms and allows for modeling the gene effects both parametrically and nonparametrically. We have proposed to calculate the p-value of the GKM test using an easy-to-compute scaled chi-square test, where the degrees of freedom of the test are estimated from the data and account for correlation among the covariates. Our approach reduces the dimensionality of the testing problem to testing for one parameter and provides a significant power gain over the usual F-test by using a considerably smaller DF than the M DFs used by the F-test, especially when the Gaussian kernel is used, as shown in our simulation studies.

To correct for multiple comparisons of testing for individual gene effects in a gene set/pathway in the presence of possible gene-gene interactions, we have proposed a permutation test, which properly accounts for the correlation among the hypotheses, i.e. the same data are used for testing for different hypotheses. It would be useful to develop in the future an analytic test that can effectively accounts for the correlation among the tests. Some discussions about this issue using FDRs can be found in Schwartzman and Lin (2010).

Our permutation method can be easily extended to the case where a gene belongs to

multiple pathways. Specifically, if a gene belongs to multiple pathways, to test for its significance, we analyze one pathway at a time by accounting for possible gene-gene interactions under model (1) and calculate the p-value for the gene. We then apply the same permutation procedure in Section 4 to calculate the p-value across multiple pathways, where the minimum of the p-values in step 4 is calculated across all the pathways.

We have mainly focused in this paper on continuous outcomes and a Gaussian regression model. The fundamental idea of introducing a garrote parameter in the kernel framework to conduct powerful tests can still be applied to more general models such as logistic regression or exponential class models. We are currently pursuing the testing problem in this generalized linear model setting. The results will be reported elsewhere.

Acknowledgment

Our research is supported by R37CA76404, P01-CA134294, a Pilot Project funding from the HSPH-NIEHS Center for Environmental Health (ES000002) and Award Number K99ES017744 from the National Institute Of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of Environmental Health Sciences or the National Institutes of Health. We thank Dr. Dawei Liu for providing the Michigan prostate cancer data used in this paper.

References

Aston, C.E., Ralph, D.A., Lalo, D.P., Manjeshwar, S., Gramling, B.A., Defreese, D.C., West, A.D., Branam, D.E., Thompson, L.F., Craft, M.A., Mitchell, D.S., Shimasaki, C.D., Mulvihill, J.J. and Jupe, E.R. (2005). Oligogenic combinations associated with breast cancer risk in women under 53 years of age. *Human Genetics* **116,** 208-221.

Barlassina, C., Lanzani, C., Manunta, P., Bianchi, G. (2002). Genetics of essential hypertension: from families to genes. *Journal of the American Society of Nephrology,* Suppl 3 **13,** S155-S164.

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics* **37,** 373–384.

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

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

De Miglio, M.R., Pascale, R.M., Simile, M.M., Muroni, M.R., Virdis, P., Kwong, K.M., Wong, L.K., Bosinco, G.M., Pulina, F.R., Calvisi, D.F., Frau, M.,Wood, G.A., Archer, M.C., Feo, F. (2004). Polygenic control of hepatocarcinogenesis in Copenhagen × F344 rats. *International Journal of Cancer* **111,** 9-16

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

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* **72,** 320–338.

Howard, T. D., Koppelman, G. H., Xu, J., Zheng, L. S., Postma, D. S., Meyers, D. A., and Bleecker, E. R. (2002). Gene-Gene Interaction in Asthma: IL4RA and IL13 in a Dutch Population with Asthma. *The American Journal of Human Genetics* **70,** 230-236.

Johnson, N. L. and Kotz, S. (1970). *Continuous univariate distributions - 2*. New York: Wiley.

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

Li, Z., Pinson, R. S., Park, W. D., Paterson, A. H., and Stansel, J. W. (1997). Epistasis for Three Grain Yield Components in Rice. *Genetics* **145,** 453-465.

Liu, D., Lin, X. and Ghosh, D. (2007) Semiparametric Regression of Multi-Dimensional

Genetic Pathway Data: Least Squares Kernel Machines and Linear Mixed Models. *Biometrics* **63,** 1079–1088.

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.

Moore, J.H. (2005). A global view of epistasis. *Nature Genetics* **37,** 13–14.

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

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

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

Schwartzman, A. and Lin, X. (2010) The Effect of Correlation in False Discovery Rate Estimation. *Biometrika*, revision invited.

Tipping, M. (2000). The relevance vector machine. In *Neural Information Processing Systems, NIPS Vol 12, ed. S. Solla, T. Leen and K. Muller*, 652–658. Massachusetts: MIT Press.

Tripodis, N., Hart, A.A., Fijneman, R.J. and Demant, P. (2001). Complexity of lung cancer modifiers: mapping of thirty genes and twenty-five interactions in half of the mouse genome. *Journal of the National Cancer Institute* **93,** 1484-1491.

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

Williams, S.M., Addy, J.H., Phillips, J.A. 3rd, Dai, M., Kpodonu, J., Afful, J., Jackson, H., Joseph, K., Eason, F., Murray, M.M., Epperson, P., Aduonum, A., Wong, L.J., Jose, P.A., Felder, R.A. (2000). Combinations of variations in multiple genes are associated with hypertension. *Hypertension* **36,** 2-6.

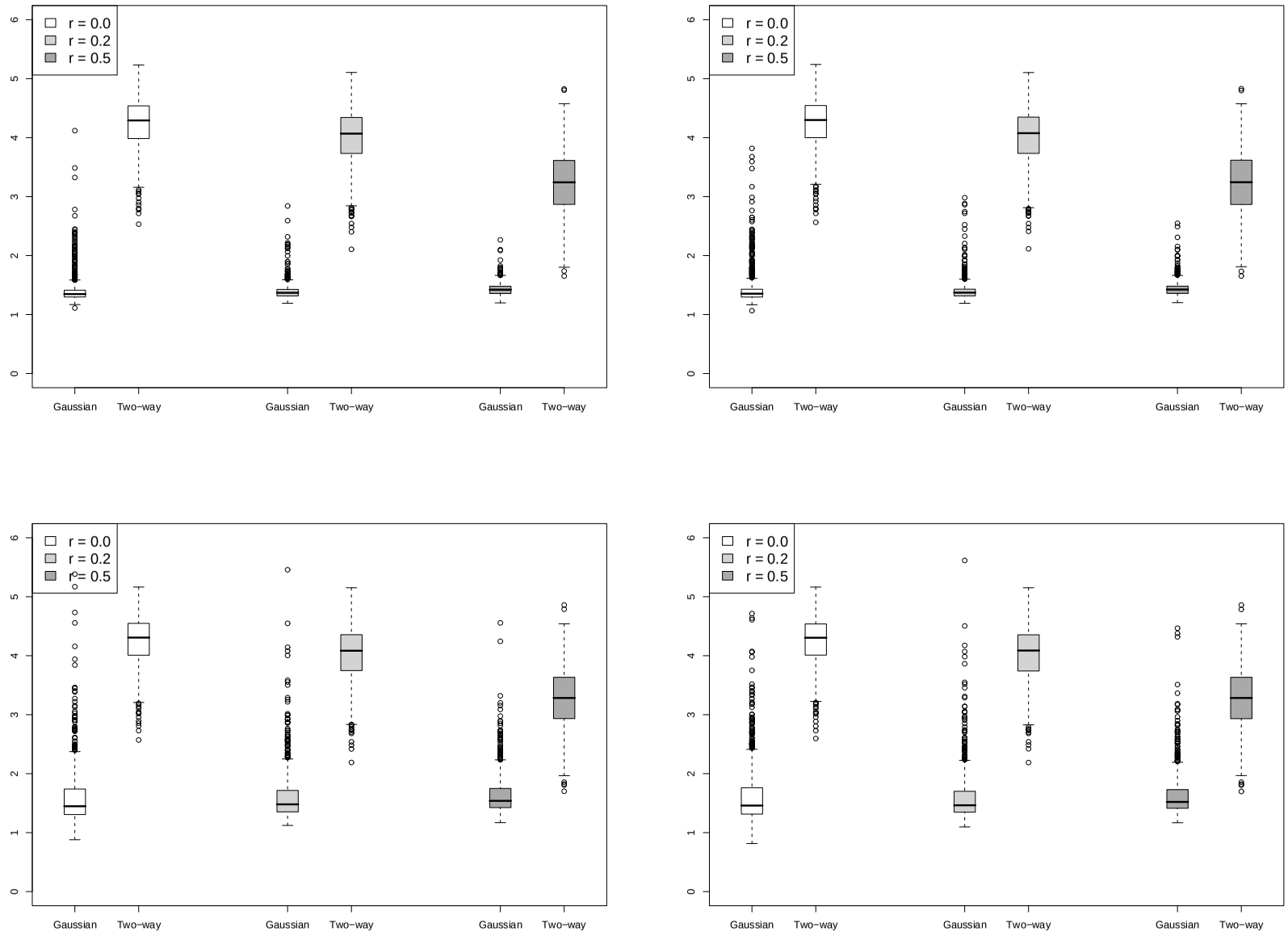

Figure 1. Results of the simulation study described in Section 5. Presented are the boxplots for the estimated degrees of freedom of our proposed GKM test for the cases $h_1(\cdot)$ (top-left), $h_2(\cdot)$ (top-right), $h_3(\cdot)$ (bottom-left) and $h_4(\cdot)$ (bottom-right) for $n = 50$ and $p = 5$ and $c = 6$. In each boxplot, the three pairs of boxes are for different values of correlation r among the covariates. In each pair, the degrees of freedom are plotted using the Gaussian (left box) and two-way interaction kernels (right box).

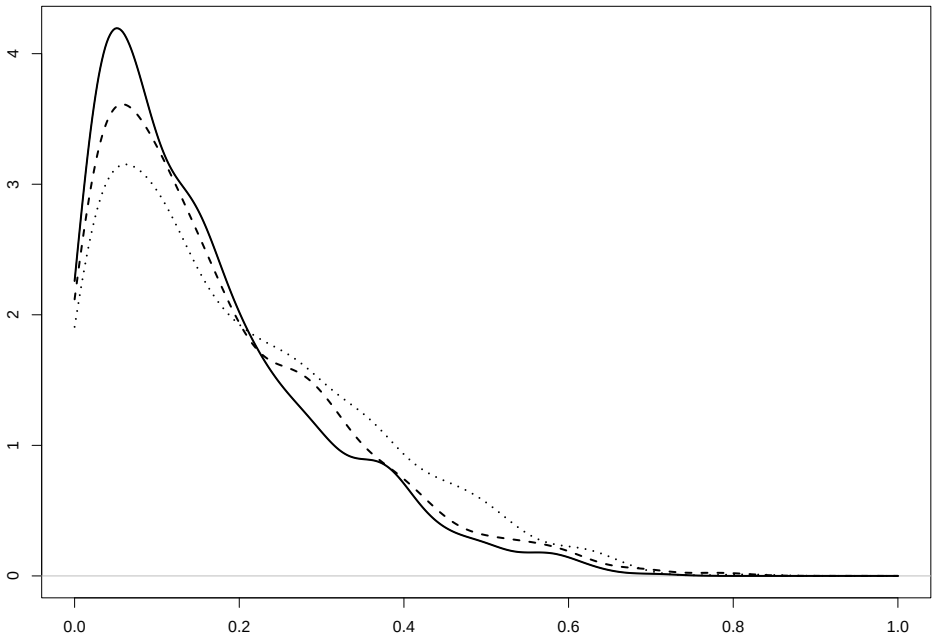

Figure 2. Results from data analysis. Displayed is the distribution of p-values for the multiple comparison procedure: solid line represents the Gaussian kernel test, dashed line represents PCA based Gaussian kernel test and dotted line represents the two-way interaction kernel.

Table 1

Results of the simulation study described in Section 5. Displayed are the empirical size ($c = 0$) and power ($c = 2, \ldots, 8$) of the proposed garrote kernel machine (GKM) test using the two-way interaction kernel (K_{Int}), the Gaussian kernel (K_{Gauss}), the PCA based test statistics T_n^* and T_n^{**} ($K_{\text{Gauss,PCA1}}$ and $K_{\text{Gauss,PCA2}}$) using the Gaussian kernel as in Section 3.3, and the usual F-test for different settings with $n = 50$ and $M = 5$. Also presented, for the linear cases ($h_1(\cdot)$ and $h_2(\cdot)$), are the results for the ideal classical score tests assuming the direction of the alternative is known. The nominal level was set be $\alpha = 0.05$.

	Test	$r = 0.0$					$r = 0.2$					$r = 0.5$				
		$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$
$h_1(\cdot)$	F-test	0.049	0.142	0.409	0.723	0.940	0.042	0.109	0.356	0.703	0.931	0.047	0.095	0.223	0.494	0.765
	K_{Int}	0.049	0.141	0.435	0.757	0.933	0.042	0.164	0.446	0.771	0.955	0.049	0.127	0.294	0.572	0.819
	K_{Gauss}	0.048	0.233	0.685	0.926	0.991	0.050	0.274	0.678	0.941	0.990	0.051	0.236	0.541	0.847	0.977
	$K_{\text{Gauss,PCA1}}$	0.049	0.228	0.665	0.915	0.993	0.055	0.267	0.661	0.927	0.990	0.050	0.197	0.502	0.794	0.948
	$K_{\text{Gauss,PCA2}}$	0.049	0.238	0.675	0.917	0.992	0.058	0.277	0.671	0.924	0.991	0.055	0.207	0.552	0.804	0.958
	Score test	0.054	0.279	0.726	0.940	0.993	0.054	0.255	0.685	0.925	0.994	0.051	0.221	0.594	0.864	0.970
$h_2(\cdot)$	F-test	0.049	0.143	0.417	0.793	0.945	0.042	0.134	0.446	0.794	0.963	0.047	0.107	0.332	0.661	0.881
	K_{Int}	0.049	0.154	0.474	0.807	0.921	0.042	0.208	0.582	0.857	0.967	0.049	0.189	0.509	0.777	0.925
	K_{Gauss}	0.048	0.248	0.665	0.933	0.982	0.050	0.263	0.682	0.917	0.987	0.051	0.230	0.548	0.808	0.938
	$K_{\text{Gauss,PCA1}}$	0.049	0.242	0.658	0.920	0.988	0.055	0.278	0.658	0.909	0.982	0.050	0.227	0.580	0.828	0.945
	$K_{\text{Gauss,PCA2}}$	0.049	0.248	0.665	0.927	0.982	0.058	0.267	0.681	0.926	0.981	0.055	0.237	0.582	0.824	0.932
	Score test	0.054	0.354	0.881	0.972	0.991	0.054	0.453	0.889	0.983	0.997	0.051	0.637	0.954	0.996	1.000
$h_3(\cdot)$	F-test	0.058	0.083	0.178	0.356	0.544	0.049	0.081	0.137	0.249	0.404	0.056	0.059	0.077	0.149	0.217
	K_{Int}	0.044	0.081	0.213	0.377	0.562	0.046	0.077	0.187	0.330	0.487	0.046	0.067	0.113	0.192	0.290
	K_{Gauss}	0.047	0.112	0.351	0.615	0.821	0.051	0.131	0.319	0.581	0.793	0.050	0.117	0.258	0.423	0.599
	$K_{\text{Gauss,PCA1}}$	0.052	0.128	0.341	0.583	0.801	0.052	0.132	0.325	0.545	0.749	0.050	0.107	0.232	0.380	0.539
	$K_{\text{Gauss,PCA2}}$	0.052	0.130	0.344	0.593	0.820	0.052	0.130	0.328	0.555	0.759	0.050	0.116	0.242	0.401	0.559
$h_4(\cdot)$	F-test	0.058	0.076	0.134	0.293	0.483	0.049	0.071	0.116	0.225	0.389	0.056	0.067	0.083	0.107	0.194
	K_{Int}	0.044	0.066	0.145	0.314	0.527	0.046	0.082	0.152	0.301	0.493	0.046	0.074	0.100	0.155	0.279
	K_{Gauss}	0.047	0.100	0.268	0.571	0.804	0.051	0.122	0.274	0.540	0.734	0.050	0.108	0.225	0.357	0.510
	$K_{\text{Gauss,PCA1}}$	0.052	0.112	0.254	0.541	0.795	0.052	0.122	0.273	0.498	0.702	0.050	0.101	0.205	0.322	0.498
	$K_{\text{Gauss,PCA2}}$	0.052	0.111	0.264	0.551	0.799	0.052	0.121	0.264	0.501	0.713	0.050	0.102	0.209	0.337	0.508

Table 2

Results of the simulation study described in Section 5. Displayed are the empirical size ($c = 0$) and power ($c = 2, \ldots, 8$) of the proposed garrote kernel machine (GKM) test using the two-way interaction kernel (K_{Int}), the Gaussian kernel (K_{Gauss}), the PCA based test statistics T_n^* and T_n^{**} ($K_{\text{Gauss,PCA1}}$ and $K_{\text{Gauss,PCA2}}$) using the Gaussian kernel as in Section 3.3, and the usual F-test for different settings with $n = 100$ and $M = 8$. Also presented, for the linear cases ($h_1(\cdot)$ and $h_2(\cdot)$), are the results for the ideal classical score tests assuming the direction of the alternative is known. The nominal level was set be $\alpha = 0.05$.

	Test	$r = 0.0$					$r = 0.2$					$r = 0.5$				
		$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$
$h_1(\cdot)$	F-test	0.057	0.160	0.606	0.934	0.996	0.054	0.173	0.546	0.911	0.995	0.053	0.116	0.372	0.739	0.940
	K_{Int}	0.053	0.198	0.658	0.952	0.993	0.052	0.233	0.691	0.954	0.997	0.055	0.149	0.429	0.773	0.948
	K_{Gauss}	0.050	0.402	0.943	1.000	1.000	0.051	0.443	0.938	0.998	1.000	0.051	0.332	0.803	0.983	0.999
	$K_{\text{Gauss,PCA1}}$	0.051	0.396	0.912	0.998	1.000	0.058	0.385	0.895	0.994	1.000	0.051	0.315	0.796	0.954	0.998
	$K_{\text{Gauss,PCA2}}$	0.051	0.406	0.922	0.998	1.000	0.054	0.405	0.904	0.996	1.000	0.051	0.321	0.799	0.961	0.991
	Score test	0.050	0.441	0.947	1.000	1.000	0.054	0.454	0.943	0.996	1.000	0.048	0.335	0.862	0.988	0.998
$h_2(\cdot)$	F-test	0.057	0.197	0.713	0.981	0.999	0.054	0.259	0.808	0.993	1.000	0.053	0.228	0.736	0.970	1.000
	K_{Int}	0.053	0.223	0.758	0.967	0.998	0.052	0.431	0.916	0.998	1.000	0.055	0.453	0.93	0.998	1.000
	K_{Gauss}	0.050	0.398	0.921	0.995	0.999	0.051	0.456	0.918	0.995	0.998	0.051	0.469	0.816	0.983	0.998
	$K_{\text{Gauss,PCA1}}$	0.051	0.389	0.910	0.992	1.000	0.058	0.420	0.901	0.992	0.999	0.051	0.426	0.857	0.991	0.999
	$K_{\text{Gauss,PCA2}}$	0.051	0.407	0.912	0.997	1.000	0.054	0.445	0.914	0.991	1.000	0.051	0.421	0.849	0.983	0.992
	Score test	0.050	0.556	0.961	0.998	1.000	0.054	0.805	0.995	0.999	1.000	0.048	0.991	1.000	1.000	1.000
$h_3(\cdot)$	F-test	0.051	0.095	0.214	0.462	0.701	0.057	0.069	0.138	0.248	0.443	0.051	0.053	0.085	0.123	0.203
	K_{Int}	0.045	0.099	0.234	0.513	0.758	0.046	0.103	0.213	0.380	0.612	0.043	0.069	0.093	0.167	0.280
	K_{Gauss}	0.052	0.174	0.535	0.848	0.968	0.050	0.196	0.474	0.748	0.931	0.051	0.155	0.332	0.559	0.740
	$K_{\text{Gauss,PCA1}}$	0.053	0.173	0.548	0.820	0.966	0.051	0.197	0.443	0.706	0.914	0.052	0.135	0.295	0.567	0.720
	$K_{\text{Gauss,PCA2}}$	0.053	0.171	0.534	0.833	0.971	0.051	0.189	0.451	0.712	0.921	0.052	0.143	0.292	0.541	0.731
$h_4(\cdot)$	F-test	0.051	0.068	0.162	0.360	0.662	0.057	0.069	0.115	0.244	0.414	0.051	0.060	0.09	0.132	0.219
	K_{Int}	0.045	0.067	0.180	0.428	0.700	0.046	0.097	0.196	0.407	0.639	0.043	0.084	0.119	0.260	0.409
	K_{Gauss}	0.052	0.143	0.454	0.798	0.963	0.050	0.162	0.389	0.654	0.870	0.051	0.134	0.259	0.464	0.638
	$K_{\text{Gauss,PCA1}}$	0.053	0.143	0.402	0.744	0.947	0.051	0.142	0.344	0.603	0.828	0.052	0.116	0.234	0.415	0.602
	$K_{\text{Gauss,PCA2}}$	0.053	0.142	0.442	0.787	0.952	0.051	0.136	0.371	0.623	0.837	0.052	0.121	0.255	0.485	0.646

Table 3

Results of the simulation study described in Section 5. Displayed are the empirical size ($c = 0$) and power ($c = 2, \ldots, 8$) of the proposed garrote kernel machine (GKM) test using the two-way interaction kernel (K_{Int}), the Gaussian kernel (K_{Gauss}), the PCA based test statistics T_n^* and T_n^{**} ($K_{\text{Gauss,PCA1}}$ and $K_{\text{Gauss,PCA2}}$) using the Gaussian kernel as in Section 3.3, and the usual F-test for different settings with $n = 200$ and $M = 10$. Also presented, for the linear cases ($h_1(\cdot)$ and $h_2(\cdot)$), are the results for the ideal classical score tests assuming the direction of the alternative is known. The nominal level was set be $\alpha = 0.05$.

	Test	$r = 0.0$					$r = 0.2$					$r = 0.5$				
		$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$
$h_1(\cdot)$	F-test	0.050	0.325	0.940	1.000	1.000	0.052	0.302	0.900	0.998	1.000	0.058	0.183	0.694	0.966	1.000
	K_{Int}	0.054	0.371	0.953	1.000	1.000	0.052	0.355	0.935	0.999	1.000	0.052	0.170	0.646	0.961	0.999
	K_{Gauss}	0.052	0.706	1.000	1.000	1.000	0.051	0.668	0.998	1.000	1.000	0.051	0.487	0.948	1.000	1.000
	$K_{\text{Gauss,PCA1}}$	0.049	0.716	0.996	1.000	1.000	0.049	0.644	0.996	1.000	1.000	0.048	0.453	0.953	0.999	1.000
	$K_{\text{Gauss,PCA2}}$	0.049	0.721	0.999	1.000	1.000	0.049	0.654	0.995	1.000	1.000	0.048	0.477	0.963	1.000	1.000
	Score test	0.051	0.799	1.000	1.000	1.000	0.055	0.756	0.998	1.000	1.000	0.049	0.644	0.990	1.000	1.000
$h_2(\cdot)$	F-test	0.050	0.476	0.990	1.000	1.000	0.052	0.587	0.998	1.000	1.000	0.058	0.589	0.997	1.000	1.000
	K_{Int}	0.054	0.494	0.986	1.000	1.000	0.052	0.829	1.000	1.000	1.000	0.052	0.893	1.000	1.000	1.000
	K_{Gauss}	0.052	0.715	1.000	1.000	1.000	0.051	0.701	0.999	1.000	1.000	0.051	0.611	0.988	1.000	1.000
	$K_{\text{Gauss,PCA1}}$	0.049	0.730	0.997	1.000	1.000	0.049	0.744	0.997	1.000	1.000	0.048	0.719	0.998	1.000	1.000
	$K_{\text{Gauss,PCA2}}$	0.049	0.737	0.996	1.000	1.000	0.049	0.774	0.999	1.000	1.000	0.048	0.709	0.999	1.000	1.000
	Score test	0.051	0.884	0.999	1.000	1.000	0.055	0.992	1.000	1.000	1.000	0.049	1.000	1.000	1.000	1.000
$h_3(\cdot)$	F-test	0.043	0.112	0.414	0.817	0.956	0.044	0.070	0.176	0.419	0.713	0.054	0.065	0.102	0.196	0.305
	K_{Int}	0.044	0.117	0.435	0.837	0.964	0.045	0.094	0.274	0.572	0.841	0.044	0.064	0.099	0.216	0.326
	K_{Gauss}	0.049	0.274	0.828	0.989	1.000	0.051	0.253	0.690	0.922	0.987	0.052	0.168	0.443	0.715	0.894
	$K_{\text{Gauss,PCA1}}$	0.046	0.293	0.803	0.981	0.999	0.051	0.234	0.641	0.912	0.986	0.051	0.153	0.439	0.712	0.866
	$K_{\text{Gauss,PCA2}}$	0.046	0.295	0.819	0.983	0.999	0.051	0.244	0.652	0.914	0.984	0.051	0.171	0.459	0.721	0.881
$h_4(\cdot)$	F-test	0.043	0.091	0.301	0.678	0.944	0.044	0.065	0.168	0.421	0.725	0.054	0.063	0.120	0.207	0.380
	K_{Int}	0.044	0.090	0.328	0.696	0.957	0.045	0.098	0.336	0.661	0.895	0.044	0.089	0.214	0.435	0.676
	K_{Gauss}	0.049	0.190	0.700	0.963	0.996	0.051	0.201	0.560	0.857	0.970	0.052	0.133	0.349	0.634	0.835
	$K_{\text{Gauss,PCA1}}$	0.046	0.226	0.691	0.964	0.996	0.051	0.177	0.524	0.848	0.972	0.051	0.128	0.331	0.619	0.844
	$K_{\text{Gauss,PCA2}}$	0.042	0.215	0.690	0.963	0.996	0.053	0.174	0.534	0.853	0.977	0.051	0.124	0.342	0.606	0.889

Table 4

Results of the simulation study described in Section 5. Displayed are the empirical size ($c = 0$) and power ($c = 2, \ldots, 8$) of the proposed garrote kernel machine (GKM) test using the two-way interaction kernel (K_{Int}), the Gaussian kernel (K_{Gauss}), and the PCA based test statistics T_n^* and T_n^{**} ($K_{\text{Gauss,PCA1}}$ and $K_{\text{Gauss,PCA2}}$) using the Gaussian kernel as in Section 3.3 for different settings with $n = 200$ and $M = 20$. Also presented, for the linear cases ($h_1(\cdot)$ and $h_2(\cdot)$), are the results for the ideal classical score tests assuming the direction of the alternative is known. The nominal level was set be $\alpha = 0.05$. Note that in this case, F-test is not computable.

	Test	$r = 0.0$					$r = 0.2$					$r = 0.5$				
		$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$	$c = 0$	$c = 2$	$c = 4$	$c = 6$	$c = 8$
$h_1(\cdot)$	K_{Int}	0.047	0.068	0.147	0.253	0.436	0.039	0.060	0.123	0.217	0.361	0.049	0.060	0.103	0.193	0.310
	K_{Gauss}	0.055	0.578	0.988	1.000	1.000	0.055	0.456	0.948	0.999	1.000	0.051	0.348	0.821	0.988	0.999
	$K_{\text{Gauss,PCA1}}$	0.055	0.558	0.982	1.000	1.000	0.053	0.425	0.925	1.000	1.000	0.051	0.286	0.779	0.983	0.998
	$K_{\text{Gauss,PCA2}}$	0.055	0.578	0.987	1.000	1.000	0.052	0.448	0.937	1.000	1.000	0.049	0.314	0.811	0.985	0.999
	Score test	0.055	0.768	1.000	1.000	1.000	0.053	0.712	0.996	1.000	1.000	0.052	0.607	0.978	1.000	1.000
$h_2(\cdot)$	K_{Int}	0.047	0.120	0.376	0.670	0.847	0.039	0.812	1.000	1.000	1.000	0.049	1.000	1.000	1.000	1.000
	K_{Gauss}	0.055	0.582	0.967	0.999	1.000	0.055	0.942	1.000	1.000	1.000	0.051	0.998	1.000	1.000	1.000
	$K_{\text{Gauss,PCA1}}$	0.055	0.578	0.963	0.999	1.000	0.053	0.969	1.000	1.000	1.000	0.051	0.999	1.000	1.000	1.000
	$K_{\text{Gauss,PCA2}}$	0.055	0.578	0.967	1.000	1.000	0.052	0.967	1.000	1.000	1.000	0.049	0.999	1.000	1.000	1.000
	Score test	0.055	0.827	0.992	0.999	1.000	0.053	0.996	1.000	1.000	1.000	0.052	1.000	1.000	1.000	1.000
$h_3(\cdot)$	K_{Int}	0.038	0.051	0.07	0.090	0.133	0.050	0.045	0.045	0.047	0.051	0.047	0.045	0.057	0.067	0.087
	K_{Gauss}	0.055	0.192	0.594	0.896	0.987	0.054	0.193	0.435	0.695	0.912	0.055	0.127	0.272	0.443	0.623
	$K_{\text{Gauss,PCA1}}$	0.055	0.188	0.581	0.889	0.986	0.054	0.169	0.379	0.644	0.857	0.055	0.099	0.198	0.372	0.540
	$K_{\text{Gauss,PCA2}}$	0.055	0.194	0.604	0.901	0.987	0.054	0.184	0.414	0.672	0.894	0.055	0.106	0.239	0.407	0.600
$h_4(\cdot)$	K_{Int}	0.038	0.044	0.173	0.217	0.312	0.05	0.077	0.104	0.153	0.242	0.047	0.060	0.106	0.138	0.153
	K_{Gauss}	0.055	0.112	0.375	0.784	0.946	0.054	0.153	0.438	0.880	0.994	0.055	0.152	0.529	0.955	1.000
	$K_{\text{Gauss,PCA1}}$	0.055	0.110	0.383	0.782	0.946	0.054	0.105	0.396	0.708	0.905	0.055	0.109	0.323	0.697	0.889
	$K_{\text{Gauss,PCA2}}$	0.055	0.110	0.391	0.798	0.952	0.054	0.150	0.390	0.811	0.906	0.055	0.114	0.311	0.694	0.919

Table 5
Results of the data example described in Section 6. Displayed are the F-test based p-values of individual gene effects using the linear model with only main effects (3rd column) and using the linear model with main and all the two-way interaction effects (4th column).

Name	Description	Main effects	F-test
MYBL2	v-myb myeloblastosis viral oncogene homolog	0.059	0.111
FGF2	fibroblast growth factor 2	0.077	0.335
FGF7	**fibroblast growth factor 7 (keratinocyte growth factor)**	**0.006**	0.063
IGFBP1	insulin-like growth factor binding protein 1	0.889	0.381
IGFBP2	insulin-like growth factor binding protein 2, 36kDa	0.353	0.298

Table 6

Results of the prostate cancer data example described in Section 6 for testing for the individual gene effects in the presence of possible gene-gene interactions (epistasis) in the cell growth pathway. Displayed are the p-values of the proposed garrote kernel machine test using the Gaussian kernel: (p_{Gauss} denotes the original GKM test and $p_{\text{Gauss,PCA}}$ denotes the PCA based modified test) and the two-way interaction kernel ($p_{\text{Two-way}}$). Also reported in columns 5-7 are the adjusted p-values for multiple comparison using our permutation method proposed in Section 4.

Name	p_{Gauss}	$p_{\text{Gauss,PCA}}$	$p_{\text{Two-way}}$	p^*_{Gauss}	$p^*_{\text{Gauss,PCA}}$	$p^*_{\text{Two-way}}$
MYBL2	0.287	0.335	0.066	0.832	0.861	0.235
FGF2	0.428	0.593	0.122	0.953	0.988	0.416
FGF7	**0.004**	**0.003**	0.062	**0.020**	**0.017**	0.227
IGFBP1	0.654	0.729	0.488	0.999	0.998	0.943
IGFBP2	0.339	0.590	0.428	0.883	0.987	0.902