Estimation and Testing for the Effect of a Genetic Pathway on a Disease Outcome Using Logistic Kernel Machine Regression via Logistic Mixed Models

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

[1]Center for Statistical Sciences, Brown University, Providence, RI 02912, USA
[2]Departments of Statistics and Public Health Sciences, Pennsylvania State University, University Park, PA 16802, USA
[3]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA

Email: Dawei Liu*- daweiliu@stat.brown.edu; Debashis Ghosh - ghoshd@psu.edu; Xihong Lin - xlin@hsph.harvard.edu;

*Corresponding author

Abstract

Background: Growing interest on biological pathways has called for new statistical methods for modeling and testing a genetic pathway effect on a health outcome. The fact that genes within a pathway tend to interact with each other and relate to the outcome in a complicated way makes nonparametric methods more desirable. The kernel machine method provides a convenient, powerful and unified method for multi-dimensional parametric and nonparametric modeling of the pathway effect.

Results: In this paper we propose a logistic kernel machine regression model for binary outcomes. This model relates the disease risk to covariates parametrically and to genes within a genetic pathway parametrically or nonparametrically using kernel machines. The nonparametric genetic pathway effect allows for possible interactions among the genes within the same pathway and a complicated relationship of the genetic pathway and the outcome. We show that kernel machine estimation of the model components can be formulated using a logistic mixed model. Estimation hence can proceed within a mixed model framework using standard statistical software. A score test based on a Gaussian process approximation is developed to test for the genetic pathway effect. The methods are illustrated using a prostate cancer data set and evaluated using simulations. An extension to continuous and discrete outcomes using generalized kernel machine models and its connection with generalized linear mixed models is discussed.

Conclusions: Logistic kernel machine regression and its extension generalized kernel machine regression provide a novel and flexible statistical tool for modeling pathway effects on discrete and continuous outcomes. Their close connection to mixed models and attractive performance make them have promising wide applications in bioinformatics and other biomedical areas.

Background

The rapid progress in gene expression array technology in the past decade has greatly facilitated our understanding of the genetic aspect of various diseases. Knowledge-based approaches, such as gene set or pathway analysis, have become increasingly popular. In such gene sets/pathways, groups of genes act in concert to accomplish tasks related to a cellular process and the resulting genetic pathway effects may manifest themselves through phenotypic changes, such as occurrence of disease. Thus it is potentially more meaningful to study the overall effect of a group of genes rather than a single gene, as single-gene analysis may miss important effects on pathways and difficult to reproduce from studies to studies [1]. Researchers have made significant progress in identifying metabolic or signaling pathways based on expression array data [2, 3]. Meanwhile, new tools for identification of pathways, such as GenMAPP [4], Pathway Processor [5], MAPPFinder [6], have made pathway data more widely available. However, It is a challenging task to model the pathway data and test for a potentially complex pathway effect on a disease outcome.

One way to model pathway data is through the linear model approach, where the pathway effect is represented by a linear combination of individual gene effects. This approach has severe limitations. Activities of genes within a pathway are highly complicated, thus a linear model is far from sufficient to capture the relationship between these genes. Furthermore, genes within a pathway tend to interact with each other. The linear model approach also makes it difficult to completely express these interactions.

In this paper we propose a nonparametric approach, the kernel machine regression, to model a pathway effect. The kernel machine method, with the support vector machine (SVM) as a most popular example, has emerged in the last decade as a powerful machine learning technique in high-dimensional settings [7, 8]. This method provides a flexible way to model linear and nonlinear effects of variables and gene-gene

interactions, unifies the model building procedure in both one- and multi-dimensional settings, and shows attractive performance compared to other nonparametric methods such as splines.

Liu et al. [9] proposed a kernel machine-based regression model for continuous outcomes. In this paper, we propose a logistic kernel machine regression model for binary outcomes, where covariate effects are modeled parametrically and the genetic pathway effect is modeled parametrically or nonparametrically using the kernel machine method. A main contribution of this paper is to establish a connection between logistic kernel machine regression and the logistic mixed model. We show that the kernel machine estimator of the genetic pathway effect can be obtained from the estimator of the random effects in the corresponding logistic mixed model. This connection provides a convenient vehicle to connect the powerful kernel machine method with the popular mixed model in the statistical literature. This mixed model connection also provides an unified framework for statistical inference for model parameters, including the regression coefficients, the nonparametric genetic pathway function, and the regularization and kernel scale parameters.

Based on the proposed logistic kernel machine regression model, we develop a new test for the nonlinear pathway effect on disease risk. An appealing feature of the proposed test is that it performs well without the need to correctly specify the functional form of the effects of each gene or of their interactions. This feature has significant practical implication when analyzing genetic pathway data, where the true relationship between the pathway and the disease outcome is often unknown. We extend the results to generalized kernel machine regression for a class of continuous and discrete outcomes and discuss its connection with generalized linear mixed models [10].

Recently, Wei and Li [11] proposed a nonparametric pathway-based regression (NPR) to model pathway data. NPR is a pathway-based gradient boosting procedure, where the base learner is usually a regression or classification tree. It provides a flexible approach in modeling pathways and interactions among genes within a pathway. Michalowski et al. [12] proposed a Bayesian Belief Network approach for pathway data. First of all, neither method is likelihood-based. Thus parameter estimation and inference cannot be casted into a unified likelihood framework. It is hence difficult to estimate and quantify the overall pathway effect on disease risk and assess its statistical uncertainty. Secondly, a primary interest in this paper is to test for the statistical significance of the overall pathway effect on the risk of a disease. Both NPR and Bayesian belief network do not provide such a statistical test for the pathway effect. For example, NPR uses an importance score to rank the relative importance of each pathway. It lacks formal inferential procedure for assessing the statistical significance of a pathway. Further, when considering a single pathway, the

importance score loses its meaning in assessing the importance of a pathway. Our method, on the other hand, is based on penalized likelihood and estimation and inference can be conducted in a systematic manner within the likelihood framework. Moreover, we propose a formal statistical test for the significance of a pathway effect on the risk of a disease.

Goeman et al. [13] proposed a linear mixed model to relate the pathway effect with a continuous outcome. They modeled the pathway effect using a linear function with each gene entering into the model as a regressor. They assumed the regression coefficients of the gene as random from a common distribution with mean 0 and an unknown variance. The pathway effect can then be tested through a variance component test for random effects. Our approach is different from theirs in three aspects. First of all, we model the pathway effect using a nonparametric model rather than a parametric one. As we commented earlier, the highly complicated nature of activities of genes within a pathway makes the linear model assumption untenable. Secondly, since genes in a pathway act in concert in a cellular process, the independence assumption of random effects of genes used in [13] is also tenuous. Our model does not make this assumption. Thirdly, the kernel function used in kernel machine regression usually contains unknown tuning parameters. The parameter is present under the alternative hypothesis but disappears under null hypothesis. This makes tests as proposed in [13, 14] not applicable. Our proposed test, on the other hand, works quite well under this scenario. Further, Goeman, et al (2006) extended their linear model results to discrete outcomes using basis functions. A key advantage of the kernel machine approach over this basis approach for modeling multi-gene effects is that one does not need to specify bases explicitly, which is often difficult for high-dimensional data especially when interactions are modeled.

Methods

The Logistic Kernel Machine Model

Throughout the paper we assume that gene expression data have been properly normalized. Suppose the data consist of n samples. For subject i ($i = 1, \cdots, n$), y_i is a binary disease outcome taking values either 0 (non-disease) or 1 (disease), $\boldsymbol{x}_i$ is a $q \times 1$ vector of covariates, $\boldsymbol{z}_i$ is a $p \times 1$ vector of gene expression measurements in a pathway/gene set. We assume that an intercept is included in $\boldsymbol{x}_i$. The binary outcome y_i depends on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ through the following semiparametric logistic regression model:

$$\text{logit}(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{z}_i), \tag{1}$$

where $\mu_i = P(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{z}_i)$, $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, and $h(\boldsymbol{z}_i)$ is an unknown centered smooth function.

In model (1), covariate effects are modeled parametrically, while the multi-dimensional genetic pathway effect is modeled parametrically or nonparametrically. A nonparametric specification for $h(\cdot)$ reflects our limited knowledge of genetic functional forms. Note that $h(\cdot) = 0$ means genes in the pathway have no association with the disease risk. If $h(\boldsymbol{z}) = \gamma_1 z_1 + \ldots + \gamma_p z_p$, the model becomes the linear model considered by Goeman et al. [13].

In nonparametric modeling, such as smoothing splines, the unknown function is usually assumed to lie in a certain function space. For the kernel machine method, this function space, denoted by $\mathcal{H}_K$, is generated by a given positive definite kernel function $K(\cdot, \cdot)$. The mathematical properties of $\mathcal{H}_K$ imply that any unknown function $h(\boldsymbol{z})$ in $\mathcal{H}_K$ can be written as a linear combination of the given kernel function $K(\cdot, \cdot)$ evaluated at each sample point. Two popular kernel functions are the dth polynomial kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = (\boldsymbol{z}_1^T \boldsymbol{z}_2 + \rho)^d$ and the Gaussian Kernel $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \exp\{-||\boldsymbol{z}_1 - \boldsymbol{z}_2||^2/\rho^2\}$, where $||\boldsymbol{z}_1 - \boldsymbol{z}_2||^2 = \sum_{k=1}^{p}(z_{1k} - z_{2k})^2$ and ρ is an unknown parameter. The first and second degree polynomial kernels ($d = 1, 2$) correspond to assuming $h(\cdot)$ to be linear and quadratic in z's, respectively. The choice of a kernel function determines which function space one would like to use to approximate $h(\boldsymbol{z})$. The unknown parameter of a kernel function plays a critical role in function approximation. It remains a challenging problem to optimally estimate it using data. In the machine learning literature, this parameter is usually pre-fixed at some values based on some ad-hoc methods. In this paper, we show that we can optimally estimate it using data based on a mixed model framework.

The Estimation Procedure

Assuming $h(\cdot) \in \mathcal{H}_K$, the function space generated by a kernel function $K(\cdot, \cdot)$, we can estimate $\boldsymbol{\beta}$ and $h(\cdot)$ by maximizing the penalized log-likelihood function

$$\begin{aligned} J(h, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \log(1-\mu_i) \right\} - \frac{1}{2}\lambda \|h\|^2_{\mathcal{H}_K} \\ &= \sum_{i=1}^{n} \left(y_i \{\boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{z}_i)\} - \log[1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{z}_i)\}] \right) - \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_K}, \end{aligned} \tag{2}$$

where λ is a regularization parameter that controls the tradeoff between goodness of fit and complexity of the model. When $\lambda = 0$, it fits a saturated model, and when $\lambda = \infty$, the model reduces to a simple logistic model $\text{logit}(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$. Note that there are two tuning parameters in the above likelihood function, the regularization parameter λ and kernel parameter ρ. Intuitively, λ controls the magnitude of the unknown function while ρ mainly governs the smoothness property of the function.

By the representer theorem [15], the general solution for the nonparametric function $h(\cdot)$ in (2) can be expressed as

$$h(\boldsymbol{z}_i) = \sum_{i'=1}^{n} \alpha_{i'} K(\boldsymbol{z}_i, \boldsymbol{z}_{i'}) = \boldsymbol{k}_i^T \boldsymbol{\alpha}, \tag{3}$$

where $\boldsymbol{k}_i = \{K(\boldsymbol{z}_i, \boldsymbol{z}_1), \ldots, K(\boldsymbol{z}_i, \boldsymbol{z}_n)\}^T$ and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^T$, an $n \times 1$ vector of unknown parameters. Substituting (3) into (2) we have

$$J(\boldsymbol{\beta}, \boldsymbol{\alpha}) \quad = \quad \sum_{i=1}^{n} \left[y_i (\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{k}_i^T \boldsymbol{\alpha}) - \log \left\{ 1 + \exp \left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{k}_i^T \boldsymbol{\alpha} \right) \right\} \right] - \frac{1}{2} \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}, \tag{4}$$

where $\boldsymbol{K} = \boldsymbol{K}(\rho)$ is an $n \times n$ matrix whose (i, i')th element is $K(\boldsymbol{z}_i, \boldsymbol{z}_{i'})$ and often depends on a scale parameter ρ.

Since $J(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in (4) is a nonlinear function of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, one can use the Fisher scoring or Newton-Raphson iterative algorithm to maximize (4) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Let (k) denote the k^{th} iteration step, then it can be shown (for details see Appendix A.3) that the $(k+1)^{\text{th}}$ update for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ solves the following normal equation:

$$\begin{bmatrix} \boldsymbol{X}^T \boldsymbol{D}^{(k)} \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{D}^{(k)} \boldsymbol{K} \\ \boldsymbol{D}^{(k)} \boldsymbol{X} & \tau^{-1} \boldsymbol{I} + \boldsymbol{D}^{(k)} \boldsymbol{K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{(k+1)} \\ \boldsymbol{\alpha}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{D}^{(k)} \tilde{\boldsymbol{y}}^{(k)} \\ \boldsymbol{D}^{(k)} \tilde{\boldsymbol{y}}^{(k)} \end{bmatrix}. \tag{5}$$

where $\tilde{\boldsymbol{y}}^{(k)} = \boldsymbol{X} \boldsymbol{\beta}^{(k)} + \boldsymbol{K} \boldsymbol{\alpha}^{(k)} + {\boldsymbol{D}^{(k)}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}^{(k)})$, $\tau = 1/\lambda$, $\boldsymbol{h}^{(k)} = \boldsymbol{K} \boldsymbol{\alpha}^{(k)}$, and $\boldsymbol{D}^{(k)} = \text{Diag}\{\mu_i^{(k)}(1 - \mu_i^{(k)})\}$. The estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{h}}$ at convergence are the kernel machine estimators that maximize (4).

The Connection of Logistic Kernel Machine Regression to Logistic Mixed Models

Generalized linear mixed models (GLMMs) have been used to analyze correlated categorical data and have gained much popularity in the statistical literature [10]. Logistic mixed models are a special case of GLMMs. We show in this section that the kernel machine estimator in the semiparametric logistic regression model (1) corresponds to the Penalized Quasi-Likelihood (PQL) estimator from a logistic mixed model, and the regularization parameter $\tau = 1/\lambda$ and kernel scale parameter ρ can be treated as variance components and estimated simultaneously from the corresponding logistic mixed model. Specifically, consider the following logistic mixed model:

$$\text{logit}(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + h_i, \tag{6}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed effects, and $\boldsymbol{h} = (h_1, \ldots, h_n)$ is a $n \times 1$ vector of subject-specific random effects following $\boldsymbol{h} \sim N\{\boldsymbol{0}, \tau \boldsymbol{K}(\rho)\}$, and the covariance matrix $\boldsymbol{K}(\rho)$ is the $n \times n$ kernel matrix defined in Section 2.2.

As $\boldsymbol{K}$ is not diagonal or block-diagonal, the random effects h_i's across all subjects are assumed to be correlated. The i^{th} mean response μ_i depends on other random effects $h_{i'}$ $(i' \neq i)$ through the correlations of h_i with other random effects. To estimate the unknown parameters in the logistic mixed model (6), we estimate $\boldsymbol{\beta}$ and $\boldsymbol{h}$ by maximizing the penalized quasi-likelihood (PQL) [10], which can be viewed as a joint log likelihood of $(\boldsymbol{\beta}, \boldsymbol{h})$,

$$\sum_{i=1}^{n} \left[y_i(\boldsymbol{x}_i^T\boldsymbol{\beta} + h_i) - \log\left\{1 + \exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + h_i\right)\right\}\right] - \frac{1}{2\tau}\boldsymbol{h}^T\boldsymbol{K}^{-1}\boldsymbol{h}. \tag{7}$$

Setting $\tau = 1/\lambda$ and $\boldsymbol{h} = \boldsymbol{K}\boldsymbol{\alpha}$, one can easily see that equations (4) and (7) are identical. It follows that the logistic kernel machine estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{h}}$ can be obtained by fitting the logistic mixed model representation (1) using PQL. In fact, examination of the kernel machine normal equations (5) shows that they are identical to the PQL normal equations obtained from the PQL (7) (see Breslow and Clayton, 1993), where $\tilde{\boldsymbol{y}}$ in (5) is in fact the PQL working vector and $\boldsymbol{D}$ is the PQL working weight matrix.
Note that the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{h}$ depend on the unknown regularization parameter τ and the kernel scale parameter ρ. Within the PQL framework, we can estimate these parameters $\boldsymbol{\delta} = (\tau, \rho)$ by maximizing the approximate REML likelihood

$$\ell_R(\hat{\beta}(\boldsymbol{\delta}), \boldsymbol{\delta}) \approx -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{8}$$

where $\boldsymbol{V} = \boldsymbol{D}^{-1} + \tau\boldsymbol{K}$, and $\tilde{\boldsymbol{y}}$ is the working vector as defined above. The estimator of $\boldsymbol{\delta}$ can be obtained by solving the first derivative of (8) with respect to $\boldsymbol{\delta}$, and its standard error can be obtained using the expected information matrix calculated using the second derivative of (8) with respect to $\boldsymbol{\delta}$.
These calculations show that we can fit the logistic kernel machine model by iteratively fitting the following working linear mixed model estimating $(\boldsymbol{\beta}, \boldsymbol{h})$ using BLUPs and (τ, ρ) using REML, until convergence

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{h} + \boldsymbol{\epsilon},$$

where $\tilde{\boldsymbol{y}}$ is the working vector defined below equation (4), $\boldsymbol{h}$ is a random effect vector following $N\{0, \tau\boldsymbol{K}(\rho)\}$, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{D})$
Denote the PQL/kernel machine estimator by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{h}})$ and the REML estimator by $\hat{\boldsymbol{\delta}} = (\hat{\tau}, \hat{\rho})^T$. The covariance of $\hat{\boldsymbol{\beta}}$ is estimated by $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$, and the covariance of $\boldsymbol{h}$ is estimated by $\tau\boldsymbol{K} - \tau\boldsymbol{K}\boldsymbol{P}\boldsymbol{K}$, where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$ and $\boldsymbol{V} = \boldsymbol{V}(\hat{\boldsymbol{\delta}})$. The square roots of the diagonal elements of the estimated covariance matrices give the standard errors of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{h}}$. Our results in this section show that we can easily fit the logistic kernel machine regression using the existing PQL-based mixed model software, such as SAS GLIMMIX and R GLMMPQL.

Test for the Genetic Pathway Effect

It is of significant practical interest to test the overall genetic pathway effect $H_0 : h(\boldsymbol{z}) = 0$. Assuming $h(\boldsymbol{z}) \in \mathcal{H}_k$, one can easily see from the logistic mixed model representation (6) that $H_0 : h(\boldsymbol{z}) = 0$ vs $H_1 : h(\boldsymbol{z}) \neq 0$ is equivalent to testing the variance component τ as $H_0 : \tau = 0$ vs $H_1 : \tau > 0$. Note that the null hypothesis places τ on the boundary of the parameter space. Since the kernel matrix $\boldsymbol{K}$ is not block diagonal, unlike the standard case considered by Self and Liang [16], the likelihood ratio for $H_0 : \tau = 0$ does not follow a mixture of χ^2_0 and χ^2_1 distribution. We consider instead a score test in this paper.

When conducting statistical tests for pathways, two types of tests could be formulated. The first is called the competitive test and the second the self-contained test [17]. The competitive test compares an interested gene set to all the other genes on a gene chip. An example of the competitive test is the gene set enrichment analysis (GSEA) [1], where an enrichment score of a gene set is defined and a permutation test is used to test for the significance of the gene set based on the enrichment score. The self-contained test compares the gene set to an internal standard which does not involve any genes outside the gene set considered. In other words, the self-contained test examines the null hypothesis that a pathway has no effect on the outcome versus the alternative hypothesis that the pathway has an effect. The variance component test of [13] for the linear pathway effect is a self-contained test. Goeman and Bühlmann [17] pointed out that the self-contained test has a higher power than a competitive test and that its statistical formulation is also consistent for both single gene tests and gene set tests, and the statistical sampling properties of the competitive test can be difficult to interpret.

Our pathway effect hypothesis $H_0 : h(\boldsymbol{z}) = 0$ vs $H_1 : h(\boldsymbol{z}) \neq 0$ is a self-contained hypothesis. We propose in this paper a self-contained test for the pathway effect by developing a kernel machine variance component score test for $H_0 : \tau = 0$ vs $H_0 : \tau > 0$. The proposed test allows for both linear and nonlinear pathway effects and includes the tests by Goeman et al. [13, 14] as a special case. A key advantage of our kernel-based test is that we do not need to explicitly specify the basis functions for $h(\cdot)$, which is often difficult for modeling the joint effects of multiple genes, and we all let the data to estimate the best curvature of $h(\cdot)$.

Zhang and Lin [18] proposed a score test for $H_0 : \tau = 0$ to compare a polynomial model with a smoothing spline. Goeman et al. [14] also proposed a global test against a high dimensional alternative under the empirical Bayesian framework. The variance-covariance matrix used in these tests do not involve any unknown parameters. However, the kernel function $K(\cdot, \cdot)$ in a kernel machine model usually depends on some unknown parameter ρ. One can easily see from the mixed model representation (6) that under

$H_0 : \tau = 0$, the kernel matrix $\boldsymbol{K}$ disappears. This makes the parameter ρ inestimable under the null hypothesis and therefore renders the above tests inapplicable.

Davies [19, 20] studied the problem of a parameter disappearing under H_0 and proposed a score test by treating the score statistic as a Gaussian process indexed by the nuisance parameter and then obtaining an upper bound to approximate the p-value of the score test. We adopt this line of approaches for our proposed score test.

Using the derivative of (8) with respect to τ, we propose the following score test statistic for $H_0 : \tau = 0$ as,

$$S(\rho) = \frac{Q_\tau(\widehat{\boldsymbol{\beta}_0}, \rho) - \mu_Q}{\sigma_Q}, \tag{9}$$

where

$$Q_\tau(\widehat{\boldsymbol{\beta}}_0, \rho) = (\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0)^T \boldsymbol{D}\boldsymbol{K}(\rho)\boldsymbol{D}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)^T \boldsymbol{K}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0),$$

where $\widehat{\boldsymbol{\beta}}_0$ is the MLE of $\boldsymbol{\beta}$ under $H_0 : \tau = 0$, $\hat{\boldsymbol{\mu}}_0 = logit^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)$, $\mu_Q = \mathrm{tr}\{\boldsymbol{P}_0\boldsymbol{K}(\rho)\}$, $\sigma_Q^2 = 2\mathrm{tr}\{\boldsymbol{P}_0\boldsymbol{K}(\rho)\boldsymbol{P}_0\boldsymbol{K}(\rho)\}$, and $\boldsymbol{P}_0 = \boldsymbol{D}_0 - \boldsymbol{X}\boldsymbol{D}_0(\boldsymbol{X}^T\boldsymbol{D}_0\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}_0$, where $\boldsymbol{D}_0 = \mathrm{diag}\{\hat{\mu}_{i0}(1 - \hat{\mu}_{i0})\}$. Note that under $H_0 : \tau = 0$, model (1) reduces to the simple logistic model $logit(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Hence the $\hat{\boldsymbol{\beta}}_0$ is the MLE of $\boldsymbol{\beta}$ under this null logistic model.

If the Gaussian kernel is used, then an arbitrary nonlinear pathway effect is implicitly assumed. Our proposed test, which is derived to test for any nonlinear effect, is therefore more powerful than tests based on a parametric assumption. We show in Appendix A.1 that when ρ is large in the Gaussian kernel, our test statistic reduces asymptotically to the one based on linearity assumption of genetic effects. Hence our test includes linear model based test as a special case. From (9) it is also clear that our test is invariant to the relative scaling of the kernel function $K(\cdot, \cdot)$.

Under appropriate regularity conditions similar to those specified in [21], $S(\rho)$ under the null hypothesis can be considered as an approximate Gaussian process indexed by ρ. Using this formulation, we can then apply Davies' results [19, 20] to obtain the p-value of the test. Since a large value of $Q_\tau(\widehat{\boldsymbol{\beta}}, \rho)$ would lead to the rejection of H_0, the p-value of the test corresponds to the up-crossing probability. Following Davies [20], the p-value is upper-bounded by

$$\Phi(-M) + W\exp(-\frac{1}{2}M^2)/\sqrt{8\pi}, \tag{10}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function, M is the maximum of $S(\rho)$ over the range of ρ, $W = |S(\rho_1) - S(L)| + |S(\rho_2) - S(\rho_1)| + \ldots + |S(U) - S(\rho_m)|$, L and U are the lower and upper bound of ρ respectively and ρ_l, $l = 1, \ldots, m$ are the m grid points between L and U. Davies [19] points out that this

bound is sharp. For the Gaussian kernel, we suggest to set the bound of ρ as $L = 0.1 \min_{i \neq j} \sum_{l=1}^{p} (z_{il} - z_{jl})^2$ and $U = 100 \max_{i \neq j} \sum_{l=1}^{p} (z_{il} - z_{jl})^2$. For justifications, see the Appendix A.2.

Extension to generalized kernel machine model

For simplicity, we focus in this paper on logistic regression for binary outcomes. The proposed semiparametric model (1) can be easily extended to other types of continuous and discrete outcomes, such as normal, count, skewed data, whose distributions are in the exponential family [22]. In this section, we briefly discuss how to generalize our estimation and testing procedures for binary to other data types within the generalized kernel machine framework and discuss its fitting using generalized linear mixed models. Suppose the data consist of n independent subjects. For subject i ($i = 1, \ldots, n$), y_i is a response variable, $\boldsymbol{x}_i$ is a $q \times 1$ vector of covariates, $\boldsymbol{z}_i$ is a $p \times 1$ vector of gene expressions within a pathway. Suppose y_i follows a distribution in the exponential family with density [22]

$$p(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - a(\theta_i)}{\phi / m_i} + c(y_i, \phi) \right\}, \tag{11}$$

where θ_i is the canonical parameter, $a(\cdot)$ and $c(\cdot)$ are known functions, ϕ is a dispersion parameter, and m_i is a known weight. The mean of y_i satisfies $\mu_i = E(y_i) = a'(\theta_i)$ and $Var(y_i) = \phi m_i a''(\theta_i)$. The generalized kernel machine model is an extension of the generalized linear model [22] by allowing the pathway effect to be modeled nonparametrically using kernel machine as

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{z}_i), \tag{12}$$

where $g(\cdot)$ is a known monotone link function, and $h(\cdot)$ is an unknown centered smooth function lying in the function space $\mathcal{H}_K$ generated by a positive definite kernel function $K(\cdot, \cdot)$. For binary data, setting $g(\mu) = \text{logit}(\mu) = \log \frac{\mu}{1-\mu}$ gives the logistic kernel machine model (1); for count data, $g(\mu) = \log(\mu)$ gives the Poisson kernel machine model; for Gaussian data, $g(\mu) = \mu$ gives linear kernel machine model [9]. The regression coefficients $\boldsymbol{\beta}$ and the nonparametric function $h(\cdot)$ in (12) can be obtained by maximizing the penalized log-likelihood function

$$J(h, \boldsymbol{\beta}) = \sum_{i=1}^{n} \ell\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, h(\cdot)\} - \frac{1}{2} \lambda \|h\|^2_{\mathcal{H}_K} \tag{13}$$

where $\ell(\cdot) = ln(p)$ is the log-likelihood, and λ is a tuning parameter. Using the kernel expression of $h(\cdot)$ in (3), the generalized kernel machine model (12) can be written as

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{k}_i^T \boldsymbol{\alpha},$$

and the penalized likelihood can be written

$$J(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) - \frac{1}{2}\lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}, \tag{14}$$

where $\boldsymbol{K}$ is an $n \times n$ matrix whose (i, j)th element is $K(\boldsymbol{z}_i, \boldsymbol{z}_j)$.

One can use the Fisher scoring iteration to solve for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The procedure is virtually the same as that described in Section "The Estimation Procedure". The normal equation takes the same form as (5), except that now μ_i is specified under (12) and $\boldsymbol{D} = diag\{var(y_i)\}$ under (11). Similar calculations to those in Section "The Connection of Logistic Kernel Machine Regression to Logistic Mixed Models" show that model (12) can be fit using the generalized linear mixed model [10] via PQL

$$g(\mu_i^{\boldsymbol{b}}) = \boldsymbol{x}_i^T \boldsymbol{\beta} + h_i,$$

where $\tau = 1/\lambda$, and $\boldsymbol{h} = (h_1 \ldots, h_n)$ is an $n \times n$ random vector with distribution $N\{0, \tau \boldsymbol{K}(\rho)\}$. The same PQL statistical software, such as SAS PROC GLIMMIX and R GLMMPQL, can be used to fit this model and obtain the kernel machine estimators of $\boldsymbol{\beta}$ and $h(\cdot)$.

The score test (9) also has a straightforward extension. The only change is that the elements in matrix $\boldsymbol{D}$ in (9) be replaced by appropriate variance function $var(y_i)$ under the assumed parametric distribution of y_i.

Results

Analysis of prostate cancer data

In this section, we apply the proposed logistic kernel machine regression model (1) to the analysis of a prostate cancer data set. The data came from the Michigan prostate cancer study [23]. This study involved 81 patients with 22 diagnosed as non-cancerous and 59 diagnosed with local or advanced prostate cancer. Besides the clinical and demographic covariates such as age, cDNA microarray gene expressions were also available for each patient. The early results of Dhanasekaran et al. [23] indicate that certain functional genetic pathways seemed dys-regulated in prostate cancer relative to non-cancerous tissues. We are interested in studying how a genetic pathway is related to the prostate cancer risk, controlling for the covariates. We focus in this analysis on the cell growth pathway, which contains 5 genes. The pathway we describe was annotated by the investigator (A. Chinnaiyan) and is simply used to illustrate the methodology. Of course, one could take the pathways stored in commercial databases such as Ingenuity Pathway Analysis (IPA) and use the proposed methodology based on those gene sets.

The outcome was the binary prostate cancer status and the covariate includes age. Since the functional relationship between the cell growth pathway and the prostate cancer risk is unknown, the kernel machine

method provides a convenient and flexible framework for the evaluation of the pathway effect on the prostate cancer risk. Specifically, we consider the following semiparametric logistic model

$$\text{logit}(P(y=1)) = \beta_0 + \beta_1 \text{age} + h(\text{gene}_1, \ldots, \text{gene}_5), \tag{15}$$

where $h(\cdot)$ is a nonparametric function of 5 genes within the cell growth pathway. We fit this model using the kernel machine method via the logistic mixed model representation and using the Gaussian kernel function in estimating $h(\cdot)$. Under the mixed model representation, we estimated (β_0, β_1) and $h(\cdot)$ using PQL, and estimated the smoothing parameter τ and the Gaussian kernel scale parameter ρ simultaneously by treating them as variance components. The results are presented in Table 1.

The test for the cell growth pathway effect on the prostate cancer status $H_0 : h(\boldsymbol{z}) = 0$ vs $H_1 : h(\boldsymbol{z}) \neq 0$, was conducted using the proposed score test. For the purpose of comparison, we also conducted the global test proposed by Goeman et al. [13] that assumed a linear pathway effect. Note that our test allows a nonlinear pathway effect and gene-gene interactions. Table 1 gives the p-values for both tests. The p-value of our test suggests that cell growth pathway has a highly significant effect on the disease status, while the test from Goeman et al.'s [13] indicates only marginal significance of the growth pathway effect.

Simulation Study for the Parameter Estimates

We conducted a simulation study to evaluate the performance of the parameter estimates of the proposed logistic kernel machine regression by using the logistic mixed model formulation. We considered the following model

$$\text{logit}(P(y_i = 1)) = x_i + h(z_{i1}, \cdots, z_{ip}), \tag{16}$$

where the true regression coefficient $\beta = 1$. We consider $p = 5$ and set
$h(z_1, \ldots, z_5) = 2\{\sin(z_1) - z_2^2 + z_1 \exp(-z_3) - \sin(z_2)\cos(z_3) + z_4^2 + \sin(z_4)\cos(z_1) + z_5^2 + z_3 z_5\}$. To allow x_i and $(z_{i1}, \cdots, z_{ip})$ to be correlated, x_i was generated as $x_i = \sin(z_{i1}) + 2u_i$, where u_i and z_{ij} $(j = 1, \cdots, p)$ follow independent Uniform$(-0.5, 0.5)$. The Gaussian kernel was used. All simulations ran 300 times. Settings 1, 2, and 3 correspond to sample size $n = 100,\ 200$, and 300, respectively.

The simulation results are shown in Table 2. Due to the multi-dimensional nature of the variables $\boldsymbol{z}$, it is difficult to visualize the fitted curve $\hat{h}(\boldsymbol{z})$. We hence summarized the goodness-of-fit of $h(\cdot)$ in the following way. For each simulated data set, we regressed the true h on the fitted value $\hat{h}$, both evaluated at the design points. We then empirically summarized the goodness- of-fit of $\hat{h}(\cdot)$ by reporting the average intercepts, slopes and R^2's obtained from these regressions over the 300 simulations. If the kernel machine

method fits the nonparametric function well, then we would expect the intercept to be close to 0, the slope close to 1, and R^2 also close to 1.

Our results show that even when the sample size is as low as 100, estimation of the regression coefficient and nonparametric function only has small bias. If ρ is estimated, these biases tend to be small compared with those when ρ is fixed. With the increase of sample size, the estimates of β and h become closer to the true values especially when ρ is estimated. There are still some bias when ρ is fixed at values farther away from the estimated one. Table 3 compares the estimated standard errors of $\hat{\beta}$ with the empirical standard errors. Our results show that they agree to each other well even when ρ is estimated.

Simulation Study of the Score Test for the Pathway Effect

We next conducted a simulation study to evaluate the performance of the proposed variance component score test for the pathway effect $H_0 : h(\cdot) = 0$ vs $H_1 : h(\cdot) \neq 0$. In order to compare the performance of our test with the linearity-based global test proposed by Goeman, et al. [13], both tests were applied to each simulated data set. Nonlinear and linear functions of $h(\boldsymbol{z})$ were both considered. For the nonlinear pathway effect, the true model is $\text{logit}(y) = x + ah(\boldsymbol{z})$, where
$h(\boldsymbol{z}) = 2(z_1 - z_2)^2 + z_2 z_3 + 3\sin(2z_3)z_4 + z_5^2 + 2\cos(z_4)z_5$. For the linear pathway effect, the true model is $\text{logit}(y) = x + ah(\boldsymbol{z})$, where $h(\boldsymbol{z}) = 2z_1 + 3z_2 + z_3 + 2z_4 + z_5$. All z's were generated from the standard normal distribution, and $a = 0, 0.2, 0.4, 0.6, 0.8$. To allow x and $(z_{i1}, \cdots, z_{ip})$ to be correlated, x was generated as $x = z_1 + e/2$ with e being independent of z_1 and following $N(0, 1)$. We studied the size of the test by generating data under $a = 0$, and studied the power by increasing a. The sample size was 100. For the size calculations, the number of simulations was 2000; whereas for the power calculations, the number of runs was 1000. Based on the discussions in Section "Test for the genetic Pathway Effect", the bound of ρ is set up by interval $[\min_{i\neq j}\sum_{l=1}^{5}(z_{il} - z_{jl})^2/5, 10\max_{i\neq j}\sum_{l=1}^{5}(z_{il} - z_{jl})^2]$, and the interval is divided by 500 equally spaced grid points. All simulations were conducted using R 2.5.0, and the package "globaltest" v4.6.0 was used for the test proposed by Goeman, et al. [13] as a comparison.

Table 4 reports the empirical size ($a = 0$) and power ($a > 0$) of the variance component score test for the no pathway effect H_0. When the true function $h(\boldsymbol{z})$ is non-linear in $\boldsymbol{z}$, our results show that the size of our test was very close to the nominal value 0.05, while the size of the global test of Goeman, et al. [13] test is inflated, and our test had a much higher power. This was not surprising since the test of Goeman et al. [13] was based on a linearity assumption of the pathway effect. When the true underlying model is far from linear, the linearity assumption breaks down and the test quickly loses power. Our results also show

that the proposed test works well for moderate sample sizes. When the pathway effect is linear, our results show that the size of both tests were very close to the nominal value 0.05 and their power were also very close. This demonstrates that our test is as powerful as the global test when the true underlying $h(\boldsymbol{z})$ is linear. Therefore our test could be used as a universal test for testing the overall effect of a set of variables without the need to specify the true functional forms of each variable. This feature is especially desirable for genetic pathway data, because the relationship between genes and clinical outcome is often unknown.

Conclusions and Discussion

In this paper, we developed a logistic kernel machine regression model for binary outcomes, where the covariate effects are modeled parametrically and the genetic pathway effect is modeled nonparametrically using the kernel machine method. This method provides an attractive way to model the pathway effect, without the need to make strong parametric assumptions on individual gene effects or their interactions. Our model also allows for parametric pathway effects if a parametric kernel, such as the first-degree polynomial kernel, is used.

A key result of this paper is that we have established a close connection between the generalized kernel machine regression and generalized linear mixed models, and show that the kernel machine estimators of regression coefficients and the nonparametric multi-dimensional pathway effect can be easily obtained from the corresponding generalized linear mixed models using PQL. The mixed model connection provides a unified framework for estimation and inference and can be easily implemented in existing software, such as SAS PROC GLIMMIX or R GLMMPQL. The mixed model connection also makes it possible to test for the overall pathway effect through the proposed variance component test. A key advantage of the proposed score test for the pathway effect is that it does not require an explicit functional specification of individual gene effects and gene-gene interactions. This feature is of practical significance as the pathway effect is often complex. Our simulation study shows the proposed test performs well for moderate sample size. It has similar power to the linearity-based pathway test of Goeman, et al [13] when the effect is linear, but much higher power when the effect is nonlinear.

We have considered in this paper a single pathway. One could generalize the proposed semiparametric model to incorporate multiple pathways by fitting an additive model:

$$\text{logit}(P(y=1)) = \boldsymbol{x}^T\boldsymbol{\beta} + h_1(\boldsymbol{z}_1) + \cdots + h_m(\boldsymbol{z}_m),$$

where $\boldsymbol{z}_j$ $(j = 1, \cdots)$ denotes a $p_j \times 1$ vector of genes in the jth pathway and $h_j(\cdot)$ denotes the

nonparametric function associated with the jth genetic pathway.

Machine learning is a powerful tool in advancing bioinformatics research. Our effort helps to build an attractive bridge between m kernel machine methodss and traditional statistical mixed models. This connection will undoubtedly provide a new and convenient tool for the bioinformatics community and opens a door for future research.

Availability

Our algorithm is available at *http://www.biostat.harvard.edu/ xlin*.

Acknowledgements

Liu and Lin's research was supported by a grant from the National Cancer Institute (R37 CA–76404). Ghosh's research was supported by a grant from the National Institute of Health (R01 GM–72007).

Authors' contributions

DL performed the statistical analysis. All authors participated in the preparation of the manuscript. All authors read and approved the final manuscript.

References

1. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, Mesirov J: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**:15545–15550.
2. Eisenberg D, Graeber TG: **Bioinformatic identification of potential autorine signaling loops in cancers from gene expression profiles**. *Nature Genetics* 2001, **29**:295–300.
3. Raponi M, Belly R, Karp J, Lancet J, Atkins D, Wang Y: **Microarray analysis reveals genetic pathways modulated by tipifarnib in acute myeloid leukemia**. *BMC Cancer* 2004, **4**:56.
4. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways**. *Nature Genetics* 2002, **31**:19–20.
5. Grosu P, Twonsend JP, Hartl DL, Cavalieri D: **Pathway Processor: A tool for integrating whole-genome expression results into metabolic networks**. *Genome Research* 2002, **12**:1121–1126.
6. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biology* 2003, **4**:R7.
7. Vapnik V: *Statistical Learning Theory*. New York: Wiley 1998.
8. Schölkopf B, Smola A: *Learning with Kernels*. Cambridge, Massachusetts: MIT press 2002.
9. Liu D, Lin X, Ghosh D: **Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models**. *Biometrics* 2007, **63**(4):1079–1088.
10. Breslow N, Clayton D: **Approximate inference in generalized linear mixed models**. *Journal of the American Statistical Association* 1993, **88**:9 – 25.

11. Wei Z, Li H: **Nonparametric pathway-based regression models for analysis of genomic data**. *Biostatistics* 2007, **8**(2):265–284.

12. Sprague R (Ed): *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, Los Alamitos: IEEE 2006. [CD ROM version].

13. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* 2004, **20**:93–99.

14. Goeman JJ, van de Geer SA, van Houwelingen HC: **Testing against a high dimensional alternative**. *Journal of the Royal Statistical Society: Series B* 2006, **68**:477–493.

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

16. Self SG, Liang KY: **Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions**. *Journal of the American Statistical Association* 1987, **82**:605–610.

17. Goeman JJ, Bühlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues**. *Bioinformatics* 2007, **23**:980–987.

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

19. Davies R: **Hypothesis testing when a nuisance parameter is present only under the alternative**. *Biometrika* 1977, **64**:247–254.

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

21. le Cessie S, van Houwelingen J: **Goodness of fit tests for generalized linear models based on random effect models**. *Biometrics* 1995, **51**(2):600–614.

22. McCullagh P, Nelder J: *Generalized Linear Models*. New York: Chapman & Hall 1989.

23. Dhanasekaran S, Barrette T, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta K, Rubin M, Chinnaiyan A: **Delineation of prognostic biomarkers in prostate cancer**. *Nature* 2001, **412**(6849):822–6.

Tables

Table 1 - Analysis of prostate cancer data

Parameter estimates and score test of the logistic kernel machine regression model for the genetic pathway effect applied to the prostate cancer data. In the table, KM stands for Kernel machine method using the Gaussian kernel, and GT for global test of Geoman et al [13] assuming linearity.

Covariate	Estimate	S.E.	P-value
Intercept	0.9893	2.7552	0.7205
Age	-0.0140	0.0425	0.7430
τ	4.7362	3.6190	
ρ	1.9093	0.6603	

Score test for the genetic pathway effect $H_0 : h(\boldsymbol{z}) = 0$:

Test	P-value
KM	< 0.0001
GT	0.0661

Table 2 - Simulation results on estimation

This table shows the simulation results of estimated regression coefficients $\boldsymbol{\beta}$ and the nonparametric function $h(\cdot)$ in model $\text{logit}(\pi) = x\beta + h(\boldsymbol{z})$ for binary outcomes based on 300 runs. True $\beta = 1$. In the table, ^{a} is the average of the estimated $\hat{\rho}$ from 300 simulations.

				Model Parameter Estimates		Reg of h on $\hat{h}$		
setting	true # z	used # z	n	β	ρ	Intercept	Slope	R^2
1	5	5	100	1.10	71.50 ^{a} (estimated)	-0.06	1.06	0.82
			100	1.14	1.00 (fixed)	-0.28	1.48	0.79
			100	1.08	20.00 (fixed)	-0.08	1.15	0.84
2	5	5	200	0.99	90.03 (estimated)	0.01	1.04	0.87
			200	1.05	1.00 (fixed)	-0.01	1.13	0.84
			200	0.96	20.00 (fixed)	-0.00	1.07	0.87
3	5	5	300	0.98	111.76 (estimated)	-0.01	1.04	0.90
			300	1.03	1.00 (fixed)	-0.02	1.10	0.87
			300	0.97	20.00 (fixed)	-0.01	1.06	0.90

Table 3 - Simulation results on standard errors

This table shows the simulation study results of standard error estimates of $\hat{\beta}$ in model $\text{logit}(\pi) = x\beta + h(\boldsymbol{z})$ for binary outcomes based on 300 simulations.

				Standard Errors of $\hat{\beta}$		
setting	true # z	used # z	n	Empirical SE	Model-based SE	ρ
1	5	5	100	0.49	0.48	71.50 (estimated)
			100	0.45	0.47	1.00 (fixed)
			100	0.48	0.47	20.00 (fixed)
2	5	5	200	0.32	0.32	90.03 (estimated)
			200	0.32	0.32	1.00 (fixed)
			200	0.33	0.32	20.00 (fixed)
3	5	5	300	0.26	0.26	111.76 (estimated)
			300	0.25	0.26	1.00 (fixed)
			300	0.26	0.26	20.00 (fixed)

Table 4 - Simulation results on score test

This table shows the simulation study results of standard error estimates of $\hat{\beta}$ in model $\text{logit}(\pi) = x\beta + h(\boldsymbol{z})$ for binary outcomes based on 300 simulations.

$h(\boldsymbol{z})$	Method	Size	Power		
		$a=0$	$a=0.2$	$a=0.4$	$a=0.8$
Nonlinear	KM	0.054	0.142	0.896	1.000
	GT	0.068	0.098	0.110	0.156
Linear	KM	0.055	0.265	0.896	1.000
	GT	0.065	0.302	0.900	1.000

Appendix

A.1 Proof of the relationship of the proposed score test and that of Goeman, et al [13] under the linearity assumption

We show in this section when the scale parameter ρ is large, the proposed nonparametric variance component test for the pathway effect using the Gaussian kernel reduces to the linearity-based global test of Goeman, et al. [13].

Suppose $K(\cdot)$ is the Gaussian kernel. It can be shown that the score statistic for testing $H_0 : \tau = 0$ satisfies

$$Q_\tau(\widehat{\boldsymbol{\beta}}, \rho) = (\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0)^T \mathbf{D}\boldsymbol{K}(\rho)\mathbf{D}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_)^T \boldsymbol{K}(\rho)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0), \tag{17}$$

where $\hat{\boldsymbol{\mu}}_0$ is the MLE of $\boldsymbol{\mu}$ under H_0. The test statistic of Goeman, et al. (2004) takes the form

$$(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{R}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{18}$$

where $\boldsymbol{R} = \boldsymbol{Z}\boldsymbol{Z}^T$. We now show when ρ is large relative to $\max_{i\neq j}\sum_{l=1}^p (z_{il} - z_{jl})^2$,

$$\frac{\rho}{2}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}(\rho)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \approx (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{R}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}). \tag{19}$$

Simple Taylor expansions show that

$$\begin{aligned}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}(\rho)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= \sum_{i=1}^n \sum_{j=1}^n (y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) \exp\{-\sum_{l=1}^p (z_{il} - z_{jl})^2/\rho\} \\ &= \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 + \sum_{i\neq j} (y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) \exp\{-\sum_{l=1}^p (z_{il} - z_{jl})^2/\rho\}.\end{aligned}$$

When $\max_{i\neq j}\sum_{l=1}^p (z_{il} - z_{jl})^2/\rho$ is small, i.e., when ρ is large relative to $\max_{i\neq j}\sum_{l=1}^p (z_{il} - z_{jl})^2$, we have that $\exp\{-\sum_{l=1}^p (z_{il} - z_{jl})^2/\rho\} \approx 1 - \sum_{l=1}^p (z_{il} - z_{jl})^2/\rho$ for any $i \neq j$. Hence

$$\begin{aligned}&(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}(\rho)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ = \;& \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 + \sum_{i\neq j} (y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) - \frac{1}{\rho}\sum_{i\neq j} (y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) \sum_{l=1}^p (z_{il} - z_{jl})^2.\end{aligned}$$

Since $\sum_{j=1}(y_j - \hat{\mu}_j) = 0$ under the PQL, we have $\sum_{j\neq i}(y_j - \hat{\mu}_j) = -(y_i - \hat{\mu}_i)$. Hence

$$\begin{aligned}
& (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{K}(\rho)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\
\approx & \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 - \frac{1}{\rho}\sum_{i\neq j}(y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j)\sum_{l=1}^{p}(z_{il}^2 - 2z_{il}z_{jl} + z_{jl}^2) \\
= & \frac{2}{\rho}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\sum_{l=1}^{p} z_{il}^2 + \frac{2}{\rho}\sum_{i\neq j}(y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j)\sum_{l=1}^{p} z_{il}z_{jl} \\
= & \frac{2}{\rho}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T \boldsymbol{R}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}).
\end{aligned}$$

This proves the approximate relation (19).

A.2 Calculations of the lower and upper bounds of ρ

Although in theory ρ could take any positive values up to infinity, for computational purpose we would require ρ to be bounded. For the proposed test statistic (9), its value in fact only depends on a finite range of ρ values. We describe why this is the case and how to find this range. For a given data set, the proof in the Appendix A.1 shows that when ρ is sufficiently large, the quantity $0.5\rho Q_\tau(\widehat{\boldsymbol{\beta}}_0, \rho)$ converges to $S_0 = (\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}_0)^T \boldsymbol{R}(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}_0)$, which is free of ρ.

These arguments suggest that for numerical evaluation, it is not necessary to consider all ρ values up to infinity. Instead, a moderately large enough value would suffice. Now the questions come down to how to decide on appropriate upper and lower bounds for ρ. The proof in the Appendix A.1 requires $\max_{i\neq j}\sum_{l=1}^{p}(z_{il} - z_{jl})^2/\rho$ is close to 0. Let C_1 be some large positive number such that $1/C_1 \approx 0$. Then if we take the upper bound of ρ to be $C_1 \max_{i\neq j}\sum_{l=1}^{p}(z_{il} - z_{jl})^2$, then this condition would be approximately satisfied. In practice we suggest take $C_1 = 100$, which would give good approximation. Using a similar idea, we can find a lower bound for ρ. It is clear that when $\min_{i\neq j}\sum_{l=1}^{p}(z_{il} - z_{jl})^2/\rho \to \infty$ any non-diagonal element of $\boldsymbol{K}(\rho)$ will be 0 and the kernel matrix reduces to an identity matrix. Hence, if we pick a small enough number C_2 such that $1/C_2 \to \infty$, we can effectively set the lower bound of ρ to be $C_2 \min_{i\neq j}\sum_{l=1}^{p}(z_{il} - z_{jl})^2$. In practice we suggest take $C_2 = 0.1$, which yields a good approximation.

A.3 derivation of normal equation (5)

Taking partial derivative of (4) with respect to $\boldsymbol{\beta}$ and writing in matrix notation, we have $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu})$. Similarly for $\boldsymbol{\alpha}$, we have $\boldsymbol{K}(\boldsymbol{y} - \boldsymbol{\mu}) - \lambda\boldsymbol{K}\boldsymbol{\alpha}$. The gradient vector is thus

$$\boldsymbol{q} = \begin{bmatrix} \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) \\ \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{\mu}) - \lambda\boldsymbol{K}\boldsymbol{\alpha} \end{bmatrix}. \tag{20}$$

Taking derivative with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we can get the following hessian matrix

$$\boldsymbol{H} = -\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{K} \\ \boldsymbol{K}\boldsymbol{D}\boldsymbol{X} & \lambda\boldsymbol{K} + \boldsymbol{K}\boldsymbol{D}\boldsymbol{K} \end{bmatrix}, \tag{21}$$

where $\boldsymbol{D} = \text{Diag}\{\mu_i(1-\mu_i)\}$. The Newton-Raphson iteration states that the parameter value at the $(k+1)^{\text{th}}$ iteration can be updated by the following relationship

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\delta}^{(k)} - (H^{(k)})^{-1}\boldsymbol{q}^{(k)}, \tag{22}$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$. Substitute (20) and (21) into (22), we arrive at normal equation (5).