Published in final edited form as:

J Am Stat Assoc. 2015 April 22; 110(509): 393–404. doi:10.1080/01621459.2014.908778.

Risk Classification with an Adaptive Naive Bayes Kernel Machine Model

Jessica Minnier,

Assistant Professor, Department of Public Health & Preventive Medicine, Oregon Health & Science University, Portland, OR 97239

Ming Yuan,

Professor, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706

Jun S. Liu, and

Professor, Department of Statistics, Harvard University, Cambridge, MA 02138

Tianxi Cai

Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Jessica Minnier: minnier@ohsu.edu; Ming Yuan: myuan@stat.wisc.edu; Jun S. Liu: jliu@stat.harvard.edu; Tianxi Cai: tcgai@hsph.harvard.edu

Abstract

Genetic studies of complex traits have uncovered only a small number of risk markers explaining a small fraction of heritability and adding little improvement to disease risk prediction. Standard single marker methods may lack power in selecting informative markers or estimating effects. Most existing methods also typically do not account for non-linearity. Identifying markers with weak signals and estimating their joint effects among many non-informative markers remains challenging. One potential approach is to group markers based on biological knowledge such as gene structure. If markers in a group tend to have similar effects, proper usage of the group structure could improve power and efficiency in estimation. We propose a two-stage method relating markers to disease risk by taking advantage of known gene-set structures. Imposing a naive bayes kernel machine (KM) model, we estimate gene-set specific risk models that relate each gene-set to the outcome in stage I. The KM framework efficiently models potentially non-linear effects of predictors without requiring explicit specification of functional forms. In stage II, we aggregate information across gene-sets via a regularization procedure. Estimation and computational efficiency is further improved with kernel principle component analysis. Asymptotic results for model estimation and gene set selection are derived and numerical studies suggest that the proposed procedure could outperform existing procedures for constructing genetic risk models.

Keywords

Genetic association; Gene-set analysis; Genetic pathways; Kernel machine regression; Kernel PCA; Principal component analysis; Risk prediction

1. INTRODUCTION

Accurate risk prediction is an essential step towards personalized, tailored medicine. To realize the goals of personalized medicine, significant efforts have been made towards building risk prediction models based on markers associated with the disease outcome. For example, statistical models for predicting individual risk have been developed for various types of diseases (Gail et al. 1989; Wolf et al. 1991; D'Agostino et al. 1994; Chen et al. 2006; Thompson et al. 2006; Cassidy et al. 2008). However, these models, largely based on traditional clinical risk factors, have limitations in their clinical utilities (Spiegelman et al. 1994; Gail and Costantino 2001; Vasan 2006). For example, the predictive accuracy as measured by the C-statistics (Pepe 2003) was only about 0.70 for the Framingham stroke models (Wolf et al. 1991; D'Agostino et al. 1994) and about 0.60 for the breast cancer Gail model (Gail et al. 1989).

To improve risk prediction for complex diseases, incorporating genotype information into disease risk prediction has been considered an eventuality of modern molecular medicine (Yang et al. 2003; Janssens and van Duijn 2008; Wray et al. 2008; Johansen and Hegele 2009). Microarray, genome-wide association studies (GWAS) as well as next generation sequencing studies provide attractive mechanisms for identifying important genetic markers for complex diseases (McCarthy et al. 2008; Pearson and Manolio 2008; Mardis 2008). Despite the initial success of GWAS, these studies focus primarily on the discovery of genetic variants associated with risk. A common approach to incorporate genotype information into risk prediction is to perform genome-wide univariate analysis to identify genetic markers associated with disease risk and then construct a genetic score from the total number of risk alleles. Such a genetic score is then included as a new variable in the risk prediction model and assessed for its incremental value in risk prediction. However, adding such simple risk scores to the prediction model has led to little improvement in risk prediction accuracy (Gail 2008; Meigs et al. 2008; Purcell et al. 2009; Lee et al. 2012). This is in part due to the fact that non-linear and interactive effects that may contribute to disease risk have not yet been identified or incorporated (Marchini et al. 2005; McKinney et al. 2006; Wei et al. 2009). Furthermore, existing findings have shown that the top ranked genetic variants reaching genome-wide significance often explain a small portion of genetic heritability of complex diseases and suggest that numerous genes may simultaneously affect the disease risk (Visscher et al. 2008; Paynter et al. 2010; Wacholder et al. 2010; Machiela et al. 2011; Makowsky et al. 2011). Therefore, to achieve optimal accuracy, one must incorporate such complex effects from multiple genes into the new risk prediction model.

Statistical procedures for combining markers to improve risk prediction have been proposed for linear additive effects with a small number of markers (Su and Liu 1993; McIntosh and Pepe 2002; Pepe et al. 2006). However, relatively little statistical research has been done on risk prediction in the presence of high dimensional markers with complex non-linear effects. Current literature on studying non-linear effects focuses primarily on testing for the significance of interactions (Umbach and Weinberg 1997; Yang and Khoury 1997; Chatterjee and Carroll 2005; Murcray et al. 2009). Traditional statistical methods that include explicit interaction terms in regression are not well suited for detecting or quantifying such interactive and nonlinear effects, especially when the number of predictors

is not very small and when higher order and non-linear interactions are present. To overcome such difficulties, we propose to employ a kernel machine (KM) regression framework which has emerged as a powerful technique to incorporate complex effects (Cristianini and Shawe-Taylor 2000; Scholkopf and Smola 2002). KM regression is a machine learning method related to the support vector machine which has been shown to be useful in building accurate risk prediction models with genetic, imaging, and other complex data (Wei et al. 2009; Casanova et al. 2013; Wei et al. 2013). KM regression allows for flexibility in the objective function, can be used to model probabilities, and can be studied within the familiar penalized regression framework. Recently, statistical procedures for making inference about model parameters under KM regression framework have been proposed (Li and Luan 2003; Liu et al. 2007, 2008). The KM models implicitly specify the underlying complex functional form of covariate effects via knowledge-based similarity measures that define the distance between two sets of covariates. These procedures, while useful in capturing nonlinear effects, may not be efficient when the underlying model is too complex. The lack of efficiency is even more pronounced when the number of candidate markers is large, with the possibility that many such markers are unrelated to the risk.

To achieve a good balance between model complexity and estimation efficiency, we propose a multi-stage adaptive estimation procedure when the genomic markers are partitioned into M gene-sets based on prior knowledge. In the first stage, by imposing an adaptive blockwise naive bayes KM (ANBKM) model, the marker effects within a gene-set are allowed to be complex and interactive while the total effects from the M gene-sets are assumed to be aggregated additively. Within each gene-set, we propose to improve the estimation via an KM principal component analysis (PCA) (Scholkopf and Smola 2002; Bengio et al. 2004; Braun 2005) which effectively reduces the dimension of the feature space. In the second stage, we recalibrate our estimates adaptively via a blockwise variable selection procedure to account for the fact that some of the gene-sets may be unrelated to the risk and the model imposed in the first stage may not be optimal. We provide theoretical justification for the root-n consistency of our proposed ANBKM estimators and the selection consistency of the gene-sets. One appealing feature of our proposed approach is that it allows estimating the effect of each individual gene-set separately, which could substantially improve both the estimation and computational efficiency. The ANBKM model is described in Section 2 and the detailed procedures for model estimations are given in Sections 3 and 4. In Section 5, we first provide results from simulation studies illustrating the performance of our proposed procedures and some of the existing procedures. Then, applying our methods to a GWAS of type I diabetes (T1D) collected by Wellcome Trust Case Control Consortium (WTCCC), we obtain a genetic risk score classifying T1D and evaluate its accuracy in classifying the T1D disease status. Some closing remarks are given in Section 6.

2. NAIVE-BAYES KERNEL MACHINE (NBKM) MODEL

Let Y denote the binary outcome of interest with $Y = 1$ being diseased and $Y = 0$ being non-diseased. Suppose there are M distinct gene-sets available for predicting Y and we let $\mathbf{Z}^{(m)}$ denote the vector of genetic markers in the m th set. The gene-sets can be created via biological criteria such as genes, pathways, or linkage disequilibrium (LD) blocks. Let $\mathbf{Z}^{(*)} = (\mathbf{Z}^{(1)\top}, \dots, \mathbf{Z}^{(M)\top})^\top$ denote the entire vector of genetic markers from all M sets. Assume that

data for analysis consist of n independent and identically distributed random vectors, $\{(Y_i, \mathbf{Z}_i^{(\bullet)}, l = 1, \dots, M), i \in \mathcal{D}\}$, where $\mathcal{D} = \{1, \dots, n\}$ indexes all subjects of the entire dataset. Throughout, we use the notation $\|\cdot\|_1$ and $\|\cdot\|_2$ to denote the L_1 and L_2 vector norm respectively.

To construct a prediction model for Y based on $\mathbf{Z}^{(\bullet)}$, we start by imposing a working *Naive Bayes* (NB) assumption that $\{\mathbf{Z}^{(m)}, m = 1, \dots, M\}$ are independent of each other conditional on Y . Under this assumption, it is straightforward to see that

$$\text{logit}P(Y=1|\mathbf{Z}^{(\bullet)})=a+\sum_{m=1}^M \text{logit}P(Y=1|\mathbf{Z}^{(m)}), \quad (1)$$

and thus $P(Y = 1 | \mathbf{Z}^{(\bullet)})$ can be approximated by first approximating $P(Y = 1 | \mathbf{Z}^{(m)})$ using data from the m th gene-set only. The NB working assumption allows genetic markers to interact within gene-sets, but not across gene-sets given the disease status. Additionally, this assumption allows us to estimate the joint effects of the gene-sets based on the marginal effects of each set. This greatly reduces both the computational and model complexity. Although this assumption seems restrictive, the prediction performance of the resulting model is quite robust in the face of deviations in independence. The reduction in model complexity in turn could result in better prediction performance due to bias and variance trade-off. Moreover, several authors have previously illustrated and discussed the strong performance of the NB classifier over more complex models and its optimality in regards to the 0–1 loss even when the conditional independence assumption is violated (Domingos and Pazzani 1997; Zhang 2005; Hastie et al. 2009).

To estimate $P(Y = 1 | \mathbf{Z}^{(m)})$, we assume a logistic KM model

$$\text{logit}P(Y_i=1|\mathbf{Z}_i^{(m)})=a^{(m)}+h^{(m)}(\mathbf{Z}_i^{(m)}) \quad (2)$$

where $h^{(m)}(\cdot)$ is an unknown smooth function belonging to the Reproducible Kernel Hilbert Space (RKHS) $\mathcal{H}_k^{(m)}$ implicitly specified by a positive definite kernel function $k(\cdot, \cdot)$. While the choice of k can vary across gene-sets, we suppress its dependence on m for the ease of presentation. More discussions on choosing an appropriate k for each gene-set are given in the discussion section.

For any pair of genetic marker vectors $(\mathbf{z}_1, \mathbf{z}_2)$, $k(\mathbf{z}_1, \mathbf{z}_2)$ measures the similarity between $\mathbf{z}_1$ and $\mathbf{z}_2$. The choice of k directly impacts the complexity and predictive performance of the model and should be selected based on the biological knowledge and empirical evidence of the relationship between $\mathbf{Z}$ and Y . The linear kernel, $k_{\text{LIN}}(\mathbf{z}_1, \mathbf{z}_2)=\mathbf{z}_1^\top \mathbf{z}_2$, models additive effects of the markers, while examples of kernel functions advocated as effective in capturing non-linear and/or interactive effects (Scholkopf and Smola 2002; Kwee et al. 2008) include

- i. polynomial kernel: $k_{\text{POLY}}(\mathbf{z}_1, \mathbf{z}_2; d) = (1 + \mathbf{z}_1^\top \mathbf{z}_2)^d$ corresponding to d -way multiplicative interactive effects;
- ii. IBS kernel for genetic markers: $k_{\text{IBS}}(\mathbf{z}_1, \mathbf{z}_2) = \sum_{l=1}^p \text{IBS}(z_{1l}, z_{2l})$, where $\text{IBS}(z_{1l}, z_{2l})$ represents the number of alleles shared identity by state;
- iii. Gaussian Kernel: $k_{\text{GAU}}(\mathbf{z}_1, \mathbf{z}_2) = \exp\{-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\rho\}$ which allows for complex non-linear smooth effects, where ρ is a tuning parameter.

Further examples of kernels useful for genomic data can be found in Schaid (2010).

Under the NB kernel machine model assumptions given in (1) and (2), the conditional like-

lihood of Y given $\mathbf{Z}^{(\bullet)}$ is a monotone function of $\sum_{m=1}^M h^{(m)}(\mathbf{Z}^{(m)})$. Therefore,

$\sum_{m=1}^M h^{(m)}(\mathbf{Z}^{(m)})$ is the optimal risk score of $\mathbf{Z}^{(\bullet)}$ for classifying Y in the sense that

$\sum_{m=1}^M h^{(m)}(\mathbf{Z}^{(m)})$ achieves the highest ROC curve among all risk scores determined by $\mathbf{Z}^{(\bullet)}$ (McIntosh and Pepe 2002). It follows that the optimal risk score can be estimated by

separately fitting the m th KM model (2) to data from the m th gene-set: $\{(Y_i, \mathbf{Z}_i^{(m)}), i = 1, \dots, n\}$.

3. KERNEL PCA ESTIMATION FOR MODELING THE M TH GENE-SET

To estimate $h^{(m)}$, we note that by Mercer's Theorem (Cristianini and Shawe-Taylor 2000),

any $h^{(m)}(\mathbf{z}^{(m)}) \in \mathcal{H}_k^{(m)}$ has a *primal representation* with respect to the eigensystem of k .

More specifically, let a nonnegative and non-increasing sequence $\{\lambda_j^{(m)}\}$ be the eigenvalues

of k under the probability measure $\mathcal{P}_{\mathbf{Z}^{(m)}}$, and $\{\varphi_j^{(m)}\}$ their corresponding eigenfunctions,

where $\mathcal{P}_{\mathbf{Z}^{(m)}}$ is the distribution of $\mathbf{Z}^{(m)}$. Since k is a Mercer kernel, $\lambda_j^{(m)}$ are square-summable

(Braun 2005). Write $\psi_j^{(m)}(\mathbf{z}) = \sqrt{\lambda_j^{(m)}} \varphi_j^{(m)}(\mathbf{z})$. Then $h^{(m)}(\mathbf{z}^{(m)}) = \sum_{j=1}^{\infty} \beta_j^{(m)} \psi_j^{(m)}(\mathbf{z})$, where

$\{\beta_j^{(m)}\}$ are the square-summable unknown coefficients. For finite samples, a suitable approach to incorporate the potentially large number of parameters associated with $h^{(m)}$ is to maximize a penalized likelihood with the penalty accounting for the smoothness of $h^{(m)}$.

However, the basis functions for $h^{(m)}$, $\{\psi_j^{(m)}(\mathbf{z})\}$, which involve the true distribution of $\mathbf{z}$, are generally unknown. It is thus not feasible to directly use the primal representation to estimate $h^{(m)}$. On the other hand, one may estimate the bases corresponding to the leading eigenvalues via the spectral decomposition of the Gram matrix

$\mathbf{K}_{n(m)} = n^{-1} [k(\mathbf{Z}_i^{(m)}, \mathbf{Z}_j^{(m)})]_{1 \leq i, j \leq n}$ (Koltchinskii and Giné 2000; Braun 2005). To this end, we apply a singular value decomposition to $\mathbf{K}_{n(m)}$ and denote the non-decreasing

eigenvalues by $(l_1^{(m)}, \dots, l_n^{(m)})$ and the corresponding eigenvectors by $(\mathbf{u}_1^{(m)}, \dots, \mathbf{u}_n^{(m)})$.

Therefore, $\mathbf{K}_{n(m)} = \mathbf{U}_{(m)} \mathbf{D}_{(m)} \mathbf{U}_{(m)}^\top$, where $\mathbf{U}_{(m)} = [\mathbf{u}_1^{(m)}, \dots, \mathbf{u}_n^{(m)}]$ and

$\mathbf{D}_{(m)} = \text{diag}\{l_1^{(m)}, \dots, l_n^{(m)}\}$. The first n basis functions evaluated at the sample points,

$\Psi_{(m)} = \{\psi_j^{(m)}(\mathbf{Z}_i)\}_{1 \leq j \leq n, 1 \leq i \leq n}$ may be estimated with $\tilde{\Psi}_{(m)} = \mathbf{D}_{(m)}^{1/2} \mathbf{U}_{(m)}$ and an estimator of $\beta^{(m)}$ may be obtained as the maximizer of

$$\mathcal{L}^{(P)}(a, \beta; \tilde{\Psi}_{(m)}) = \mathbf{Y}^\top \log g(a + \tilde{\Psi}_{(m)} \beta) + (1 - \mathbf{Y})^\top \log \{1 - g(a + \tilde{\Psi}_{(m)} \beta)\} - \tau \|\beta\|_2^2, \quad (3)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, $g(\cdot) = \text{logit}^{-1}(\cdot)$, and $\tau \geq 0$ is a tuning parameter controlling the amount of regularization. Hence, we estimate $[h^{(m)}(\mathbf{Z}_1^{(m)}), \dots, h^{(m)}(\mathbf{Z}_n^{(m)})]^\top$ with $\tilde{\Psi}_{(m)} \hat{\beta}^{(m)}$.

The above estimator of $h^{(m)}$ may not be efficient due to the high dimensionality in the parameter space and could be numerically challenging to obtain when the sample size n and hence the dimension of β is not small, as in many GWAS settings. To improve the computation and estimation efficiency, we propose the use of the kernel PCA (Scholkopf and Smola 2002; Bengio et al. 2004; Braun 2005) where only the principal components with large eigenvalues are included for estimation. When the eigenvalues $\{\lambda_j^{(m)}\}$ decay quickly, the feature space $\mathcal{H}_k^{(m)}$ may be approximated well with the space spanned by the leading

eigenfunctions and $\{\beta_j^{(m)} \sqrt{\lambda_j^{(m)}}\}$ may also decay quickly. Due to the bias and variance trade-off, the estimation of $h^{(m)}$ may be improved by employing the approximated feature space. At the same time, computational efficiency will be improved due to the decreased dimensionality. We may also understand the gain in stability and efficiency when we let

$b_j^{(m)} = \beta_j^{(m)} \sqrt{\lambda_j^{(m)}}$ and parameterize $h^{(m)}$ as $h^{(m)}(\mathbf{z}) = \sum_{j=1}^{\infty} b_j^{(m)} \varphi_j^{(m)}$. Here $b_j^{(m)}$ is an inner product of square-summable sequences and so is itself square-summable. Furthermore, $b_j^{(m)}$ decays as j increases and becomes difficult to estimate for large j ; and hence in finite sample, due to the bias and variance trade-off, eigenfunctions with small eigenvalues may not be useful for classification (Williams and Seeger 2000).

To select the number of eigenvalues to include in estimation, let $r_n^{(m)}$ be the smallest r such that $\sum_{i=1}^r l_i^{(m)} / \sum_{i=1}^n l_i^{(m)} \geq \varphi$, where $\varphi \in (0, 1)$ is a pre-specified proportion tending to 1 as $n \rightarrow \infty$. The kernel PCA approximation to $\mathbf{K}_{n(m)}$ corresponding to these $r_{n(m)}$ eigenvalues is then $\mathbf{K}_{n(m)}^{[r_{n(m)}]} = \mathbf{U}_{(m)} \mathbf{D}_{(m)} \mathbf{U}_{(m)}^\top$ where $\mathbf{U}_{(m)} = [\mathbf{u}_1^{(m)}, \dots, \mathbf{u}_{r_{n(m)}}^{(m)}]$ and $\mathbf{D}_{(m)} = \text{diag}\{l_1^{(m)}, \dots, l_{r_{n(m)}}^{(m)}\}$ are the truncated versions of $\mathbf{U}_{(m)}$ and $\mathbf{D}_{(m)}$. We now estimate $\beta^{(m)}$ as the maximizer of

$$\mathcal{L}^{(P)}(a, \beta; \hat{\Psi}_{(m)}) = \mathbf{Y}^\top \log g(a + \hat{\Psi}_{(m)} \beta) + (1 - \mathbf{Y})^\top \log \{1 - g(a + \hat{\Psi}_{(m)} \beta)\} - \tau \|\beta\|_2^2, \quad (4)$$

where $\hat{\Psi}_{(m)} = n^{-\frac{1}{2}} \mathbf{D}_{(m)}^{1/2} \mathbf{U}_{(m)}^\top$.

In summary, with the training samples, we essentially transform the original covariate matrix $(\mathbf{Z}_1^{(m)}, \dots, \mathbf{Z}_n^{(m)})^\top$ to $\tilde{\Psi}_{(m)}$ and estimate $\{h^{(m)}(\mathbf{Z}_1^{(m)}), \dots, h^{(m)}(\mathbf{Z}_n^{(m)})\}^\top$ as $\tilde{\Psi}_{(m)} \hat{\beta}^{(m)}$,

where $\{\hat{a}^{(m)}, \hat{\beta}^{(m)}\} = \arg\max_{a, \beta} \{ \mathcal{L}^{(m)}(a, \beta; \tilde{\Psi}_{(m)}) \}$. Note that, when $\varphi = 1$, this estimate is equivalent, with reparameterization, to the Liu et al. (2008) estimator obtained via the dual representation $h^{(m)}(\mathbf{z}) = \sum_{j=1}^n \alpha_j^{(m)} k(\mathbf{z}, \mathbf{Z}_j^{(m)})$, where $\{\alpha_j^{(m)}\}$ are the unknown regression parameters. To estimate $h^{(m)}(\mathbf{z}^{(m)})$ for a future subject with marker value $\mathbf{Z}^{(m)} = \mathbf{z}^{(m)}$, one may find the transformed covariate in the induced feature space via the Nyström method (Rasmussen and Williams 2006) as

$$\hat{\Psi}_{(m)}(\mathbf{z}^{(m)})^\top = n^{-\frac{1}{2}} \text{diag} \left(\frac{1}{\sqrt{l_1^{(m)}}}, \dots, \frac{1}{\sqrt{l_{r_n}^{(m)}}} \right) \mathbb{U}_{(m)}^\top [k(\mathbf{z}^{(m)}, \mathbf{Z}_1^{(m)}), \dots, k(\mathbf{z}^{(m)}, \mathbf{Z}_n^{(m)})]^\top. \quad (5)$$

Subsequently, we estimate $h^{(m)}(\mathbf{z}^{(m)})$ as $\hat{h}^{(m)}(\mathbf{z}^{(m)}) = \hat{\Psi}_{(m)}(\mathbf{z}^{(m)}) \hat{\beta}^{(m)}$. In Appendix A, we show that our estimator is root-n consistent for $h^{(m)}(\cdot)$ under the assumption that the reproducible kernel Hilbert space $\mathcal{H}_k^{(m)}$ is finite dimensional. This is often a reasonable assumption for GWAS settings since each gene-set has a finite set of single-nucleotide polymorphism (SNP) markers, which can only span a finite dimensional space regardless the choice of kernel.

4. COMBINING MULTIPLE GENE-SETS FOR RISK PREDICTION

4.1 Adaptive Naive Bayes (ANB) Kernel Machine Method

With the estimated $\hat{h}^{(m)}$, one may simply classify a future subject with $\mathbf{Z}^{(\bullet)} = \{\mathbf{z}^{(m)}, m = 1, \dots, M\}$ based on $\sum_{m=1}^M \hat{h}^{(m)}(\mathbf{z}^{(m)})$ under the NB assumption. However, since some of the gene-sets may not be associated with disease risk, including $\hat{h}^{(m)}$ from these gene-sets in the model may lead to a decrease in the precision of prediction and risk score estimation. To further improve the precision, we propose to employ a LASSO regularization procedure (Tibshirani 1996) in the second step to estimate the optimal weight for each individual gene-set. The regularized estimation would assign a weight zero to non-informative regions while simultaneously providing stable weight estimates for the informative regions. Specifically, based on the synthetic data $\{\mathbf{Y}, \hat{\mathbf{H}}\}$ constructed from the first step, we re-weight the gene-sets in the second step by fitting the logistic model

$$\text{logit} P(Y=1 | \mathbf{Z}^{(\bullet)}) = b_0 + \boldsymbol{\gamma}^\top \hat{\mathbf{H}}(\mathbf{Z}^{(\bullet)})$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)^\top$, $\hat{\mathbf{H}}(\mathbf{Z}^{(\bullet)}) = [\hat{h}^{(1)}(\mathbf{Z}^{(1)}), \dots, \hat{h}^{(M)}(\mathbf{Z}^{(M)})]^\top$ and $\hat{\mathbf{H}} = [\hat{h}^{(m)}(\mathbf{Z}_i^{(m)})]_{n \times M}$. We obtain a LASSO regularized estimate of $\{b_0, \boldsymbol{\gamma}\}$ with $\{b, \boldsymbol{\gamma}\}$, the maximizer of

$$\mathcal{L}_{\hat{\mathbf{H}}}(b, \boldsymbol{\gamma}) = \mathbf{Y}^\top \log g(b + \hat{\mathbf{H}}\boldsymbol{\gamma}) + (1 - \mathbf{Y})^\top \log \{1 - g(b + \hat{\mathbf{H}}\boldsymbol{\gamma})\} - \tau_2 \|\boldsymbol{\gamma}\|_1, \quad (6)$$

where $\tau_2 = 0$ is a tuning parameter such that $n^{-\frac{1}{2}} \tau_2 \rightarrow 0$ and $\tau_2 \rightarrow \infty$. This ANBKM model allows additional flexibility with possible dependence between gene-sets. It is also important

to note that our estimator $\hat{\gamma}$ is essentially an adaptive LASSO (Zou 2006) type estimator since these weights are multiplied with $\hat{h}^{(m)}(\mathbf{z})$ which are consistent for $h^{(m)}$. As a result, $\hat{\gamma}$ exhibits the gene-set selection consistency property such that $P(\hat{\mathcal{A}} = \mathcal{A}) \rightarrow 1$ as $n \rightarrow \infty$, where $\mathcal{A} = \{m : h^{(m)}(\mathbf{z}) > 0\}$ and $\hat{\mathcal{A}} = \{m : \hat{\gamma}_m > 0\}$. Therefore, this method of estimation consistently includes only informative regions in the prediction model. We show in Appendix A.2 that the proposed adaptively reweighting procedure is consistent in group selection, i.e. $P(\hat{\mathcal{A}} = \mathcal{A}) \rightarrow 1$ in probability as $n \rightarrow \infty$.

4.2 Improved Estimation of Gene-set Weights via Cross-validation

Based on the estimation procedures described in Section 4.1, we may estimate the probability of disease for a future subject with $\mathbf{Z}^{(\bullet)}$ under the NBKM as $P(\mathbf{Z}^{(\bullet)}) = g\{b + \hat{\gamma}^\top \hat{\mathbf{H}}(\mathbf{Z}^{(\bullet)})\}$. However, training of the KM model for each specific gene-set involves complex models with a potentially large number of effective model parameters, the estimation of γ in the second stage may also suffer from instability due to overfitting if we estimate γ on the same dataset that we use to estimate β for $h^{(m)}(\mathbf{z})$.

To overcome overfitting issues, we propose a K-fold cross-validation procedure to partition the training data $\mathcal{D}_t$ of size n_t into K parts of approximately equal sizes, denoted by $\{\mathcal{D}_{t(\kappa)}, \kappa = 1, \dots, K\}$. For each κ we use data *not* in $\mathcal{D}_{t(\kappa)}$ to obtain an estimate for $h^{(m)}$ as $\hat{h}_{t(-\kappa)}^{(m)}$ based on procedures described in Section 3; and then use those estimates to predict subjects in $\mathcal{D}_{t(\kappa)}$ to obtain $\hat{\mathbf{H}}_{t(\kappa)} = [\hat{h}_{t(-\kappa)}^{(m)}(\mathbf{Z}_{t(\kappa)i}^{(m)})]_{\frac{n}{K} \times M}$. Subsequently, we maximize

$$\sum_{\kappa=1}^K \left[\mathbf{Y}_{t(\kappa)}^\top \log g(b + \hat{\mathbf{H}}_{t(\kappa)} \gamma) + (1 - \mathbf{Y}_{t(\kappa)})^\top \log \{1 - g(b + \hat{\mathbf{H}}_{t(\kappa)} \gamma)\} \right] = \tau_2 \|\gamma\|_1, \quad (7)$$

with respect to $\{b, \gamma\}$ to obtain $\{\hat{b}_{cv}, \hat{\gamma}_{cv}\}$. This procedure enables us to reduce the overfitting bias without losing information from the training set. As shown in the simulation section, this method provides a more accurate estimate of γ than using the entire $\mathcal{D}_t$ without cross-validation which leads to overfitting. The consistency of $\hat{\gamma}_{cv}$ can be established using similar arguments as those given in Appendix A.2 for $\hat{\gamma}$.

We then use the entire training set $\mathcal{D}_t$ to obtain an estimate of $\sum_{m=1}^M h^{(m)}(\mathbf{Z}^{(m)})$ as $\hat{\gamma}_{cv}^\top \hat{\mathbf{H}}(\mathbf{Z}^{(\bullet)})$ for an out of sample subject with covariate data $\mathbf{Z}^{(\bullet)}$. The final estimated risk prediction model would thus predict the risk of disease for this new subject as

$$\hat{P}(\mathbf{Z}^{(\bullet)}) = g\{\hat{b}_{cv} + \hat{\gamma}_{cv}^\top \hat{\mathbf{H}}(\mathbf{Z}^{(\bullet)})\}.$$

4.3 Tuning Parameter Selection

There are several tuning parameters involved in our model estimation. In the first stage, we select the kernel ridge regression tuning parameter τ in equation (4) with AIC since we are most concerned with prediction accuracy of the gene-set risk scores, $\hat{h}^{(m)}(\mathbf{z}^{(m)})$. In the

second stage when we aggregate across gene-sets we select the tuning parameter for LASSO in equation (7) with BIC. In this stage, we are most concerned with removing non-informative gene-sets from the final model. Therefore, BIC shrinks the estimates more aggressively than AIC and achieves adequate variable selection in finite samples. In regards to the number of folds for cross-validation when obtaining $\hat{\mu}_{(n)}$ and $\hat{\gamma}_{CV}$, it is imperative that the size of the training set is large enough to accurately estimate the kernel ridge regression model parameters. Hence, we recommend choosing K to be at least 5, and we use $K = 5$ in our numerical studies. In our data analysis we also used $K = 10$ as recommended in Breiman and Spector (1992) and saw very similar results to $K = 5$. For computational efficiency we present results from $K = 5$ in simulation studies and the data analysis.

5. NUMERICAL ANALYSES

5.1 Type I Diabetes GWAS Dataset

Type I diabetes (T1D), also known as juvenile-onset diabetes, is a chronic autoimmune disease characterized by insulin deficiency and hyperglycemia due to the destruction of pancreatic islet beta cells. Diagnosis and onset often occurs in childhood. Since the discovery of the association of the disease with the HLA sequence polymorphisms in the late 1980s, the understanding of T1D pathogenesis has advanced with the identification of additional genetic risk factors for the disease (Van Belle et al. 2011). T1D is thought to be triggered by environmental factors in genetically susceptible individuals. However, the proportion of newly diagnosed children with known high-risk genotypes has been decreasing, suggesting that further genetic risk markers have not yet been discovered (Borchers et al. 2010).

Compiling information from a number of large scale genetic studies conducted and published in recent years, the National Human Genome Research Institute (NHGRI) provides an online catalog which lists 75 single nucleotide polymorphisms (SNPs) that have been identified as T1D risk alleles (Hindorff et al. 2009; <http://www.genome.gov/gwastudies/> Accessed December 10, 2011) and 91 genes that either contain these SNPs or flank the SNP on either side on the chromosome. Expanding the search to other documented autoimmune diseases (Rheumatoid arthritis, Celiac disease, Crohn's disease, Lupus, Inflammatory bowel disease), the NHGRI lists 375 genes that containing or flanking 365 SNPs that have been found to be associated with this class of diseases.

Included among the studies listed in the NHGRI catalog is a large-scale GWAS collected by WTCCC, a group of 50 research groups across the UK that was formed in 2005. The study, detailed in Burton et al. (2007), consists of 2000 T1D cases and 3004 controls of European descent from Great Britain. The control subjects were drawn from the 1958 British Birth Cohort and the UK Blood Services. Approximately 482,509 SNPs were genotyped on an Affymetrix GeneChip 500K Mapping Array Set. We chose to segment the genome on the 22 autosomal chromosomes into gene-sets that include a gene and a flanking region of 20KB on either side of the gene. The WTCCC data we use for analysis includes 350 gene-sets that either contain or lie up- or down-stream of the 365 SNPs that were previously found to be associated with autoimmune diseases. These genes are a subset of the 375 genes in the NCBI catalog that were obtained by removing pseudo-genes, genes with no genotyped SNPs

in the WTCCC data, and genes on the X chromosome. Most genes contain just one associated SNP but some of the HLA genes contain or lie upstream or downstream of multiple associated SNPs. The data includes 40 genotyped SNPs of the 75 SNPs that were previously found to be associated with T1D. Including the flanking region of 20KB, these 350 gene-sets cover 9,256 SNPs present in the WTCCC data. The gene-sets contain on average 26.45 SNPs (median 13.5) with the largest set containing 533 SNPs and 12 sets containing one SNP.

5.2 Assessment of Prediction Accuracy

When such a risk prediction model is formed, it is crucial to assess its ability in discriminating subjects with or without disease. For a given risk score P , the discrimination accuracy can be summarized based on various measures such as the area under the receiver operating characteristic (ROC) curve (AUC) (Swets 1988; Pepe 2003). The ROC curve is determined from plotting sensitivity against 1-specificity for all possible cut-offs of the risk score. An AUC of 1 indicates a perfect prediction and 0.5 indicates a random result. Few clinical scores achieve AUCs above 0.75, and scores with an AUC of 0.95 or greater are considered excellent. Since the number of parameters involved in the training the proposed risk score could be quite large, the AUC should be estimated empirically in an independent validation set. This validation set may be a new data set, or one could set aside a random sample of the data so that $\mathcal{D}$ is partitioned into $\mathcal{D}_t$ and $\mathcal{D}_v$ prior to building the model.

5.3 Simulation Studies

We first present results from simulation studies with data generated from the SNP data from the WTCCC study. To assess the performance of our methods, we chose settings that reflect possible types of genetic association with disease risk. For illustrative purposes, we let $\mathbf{Z}^{(\bullet)}$ be the aforementioned $M = 350$ gene-sets. We generated the disease status of 1500 subjects

from the logistic regression model, $\text{logit}P(Y=1|\mathbf{Z}^{(\bullet)}) = \sum_{m=1}^4 h^{(m)}(\mathbf{Z}^{(m)})$, where the $h^{(m)}(\mathbf{z})$ for $m = 1, \dots, 4$ are set as linear or nonlinear functions of $\mathbf{Z}^{(m)}$, with varying degrees of complexity. The remaining 346 gene-sets were included as non-informative regions. The labels used in the subsequent tables are denoted in parentheses in the following model descriptions. We present the results from three settings where $h^{(m)}(\mathbf{z})$ for $m = 1, \dots, 4$ are all linear (allL), all nonlinear (allNL), or two linear and two nonlinear functions (LNL). We relegate details about the forms of these functions to Appendix B.

We partition each dataset once into a training set of 1000 and a validation set of 500 subjects. We estimate $h^{(m)}(\cdot)$ using the training set by fitting the block specific KM model with either a linear kernel, k_{LIN} , or an IBS kernel, k_{IBS} . To evaluate the effect of PCA, we obtain estimates by maximizing (3) with the full kernel matrix (noPCA, $\varphi = 1$) and also based on the PCA approximated likelihood in (4) with $\varphi = 0.99$ or 0.999. Decreasing φ from 0.999 to 0.99 gives nearly identical results so we report only $\varphi = 0.999$ which is approximately $1 - 1/n_t$. When combining information across the M blocks, we use both $\hat{\gamma}$ and $\hat{\gamma}_{cv}$ with 5-fold cross-validation as described in Section 4 to estimate γ . We compare our adaptive weighting scheme (ANB) that adaptively estimate γ to the purely NB approach where $\gamma = 1$ (NB). Additionally, we compare our methods to models that do not incorporate

the block structure of the data by fitting three global models with all 9,256 SNPs in the 350 gene-sets: (1) a global KM model with k_{IBS} (gIBS), (2) a global ridge regression model (gRidge), as well as (3) the sure independence screening procedure (SIS) described in Fan and Lv (2008). Lastly, we compare our methods to the weighted sum of the marginal log odds ratios for each of the SNPs (WLGR). The tuning parameter was selected by maximizing the AIC for the ridge regression model in the first stage and via the BIC for the LASSO model in the second stage for combining across blocks. The results are based on 1500 Monte Carlo simulations.

First, we present results on selecting informative blocks via our second stage adaptive estimation of $h(\mathbf{z})$ based on $\hat{\gamma}_{cv}$. As shown in Table 1, all estimates $\hat{h}(\mathbf{z})$ have high Pearson's correlation with the true $h(\mathbf{z})$ and low median squared error (MSE) for all blocks with linear effects. For blocks with interactive and nonlinear effects, correlation and MSE are generally best for $\hat{h}(\mathbf{z})$ with k_{IBS} methods. For the most difficult block effects that are highly interactive, the MSE remains low, though correlation decreases. Noninformative blocks are excluded from the model with very high probability with MSE of $h(\mathbf{z})$ essentially zero, illustrating the oracle property of $\hat{\gamma}_{cv}$ proved in the appendix.

We see that in general, the ANBKM methods with k_{IBS} give the lowest or approximately the lowest MSE for all types of effects studied. In Table 1, we also see that the method with k_{LIN} selects a more gene-sets on average than the method with k_{IBS} but has a lower probability of selecting the informative gene-sets with nonlinear effects. The method without PCA selects more gene-sets on average and results in a similar MSE for $\hat{h}(\mathbf{z})$ for the informative gene-sets. Overall, the best performance in estimation and gene-set selection is seen for models with k_{IBS} .

To compare the methods with respect to predictive performance, we project the model estimates into a validation set of 500 subjects and report the AUC estimates and their standard errors from all models in Table 2. We first focus on our risk estimates based on the recommended $\hat{\gamma}_{cv}$ for block reweighting. The global methods (gRidge, gIBS, SIS, WLGR) generally have substantially worse predictive performances compared to our proposed ANBKM procedures, suggesting the benefit of taking advantage of blocking combined with adaptive weighting. The benefit of blocking is also highlighted when we compare results between the ANBKM procedures and the SIS procedures. The SIS procedures outperform global ridge and WLGR procedures with higher AUC values, but also have larger standard errors than any other method. Even when all effects are linear and SIS performs fairly well with higher AUC than other global methods as well as NBKM methods, we still see substantial improvement in prediction when applying an ANBKM method with either the linear or IBS kernel. Although both procedures allow for marker selection, the ANBKM procedures can more effectively estimate the effect of informative blocks and remove the non-informative blocks. When comparing ANBKM and NBKM procedures, we see that similar to the global methods, the purely NBKM methods tend to result in a substantially lower AUC with a higher standard error compared to our ANBKM methods due to the inclusion of non-informative gene-sets. The IBS kernel generally performs well, resulting in similar performances as the linear kernel when the effects are linear; and better performances than the linear kernel when the effects are non-linear. In particular, for non-

linear effects settings, the IBS kernel leads to higher AUCs for our ANBKM procedure with smaller standard errors than the linear kernel. Our methods with PCA perform very similarly to methods without PCA in terms of prediction with very slight improvement in prediction accuracy, but the computational efficiency is much greater when using PCA. Overall, we observe the strengths of the PCA and adaptively weighted blocking models, and note that we obtain the best prediction accuracy with k_{IBS} . The average number of PCs included in the first stage for $\varphi = 0.999$ (mean 26, median 12) is typically quite larger than those for $\varphi = 0.99$ (mean 12, median 7). It is important to note that both procedures select substantially fewer PCs than the total number of nonzero eigenvalues, which is the number used in the noPCA methods. Furthermore, most computational algorithms to estimate eigenvalues have difficulty exactly estimating true zero eigenvalues and so selecting all of the PCs corresponding to estimated nonzero eigenvalues can lead to much instability and can increase the computational burden, especially with large n .

To illustrate the advantage of the proposed cross-validation based estimator for γ , we also contract results for the predictive performance of the resulting risk estimates based on $\hat{\gamma}$ and $\hat{\gamma}_{\text{cv}}$ as shown in Table 2. Interestingly, the simulation results suggest that $\hat{\gamma}$ suffered from over-fitting and led to less accurate estimates. The average AUCs from $\hat{\gamma}$ were consistently about 10% lower than the AUCs from those corresponding to $\hat{\gamma}_{\text{cv}}$. Additionally, the standard errors of the AUCs from k_{IBS} with ANB were high for $\hat{\gamma}$, suggesting instability in the estimates of γ . This demonstrates the substantial advantage of employing the cross-validation to improve the estimation of block weights.

5.4 Data Example

Using the methods described above, we also constructed T1D risk prediction models based on the WTCCC GWAS dataset. To build our prediction model, we randomly selected 1600 cases and 1500 controls as a training set to implement the first stage and the cross-validation procedures for the second stage, and left the remaining 400 cases and 1500 controls as a validation set. To avoid dependency on the selection of the training set, we randomly selected 15 partitions of the data into training and validation sets of these sizes and report the median accuracy measures across the validation sets.

Although our dataset includes 40 SNPs that are known to be associated with T1D disease status, they do not explain all of the genetic variability and there may be many other SNPs that are associated with disease status through complex effects. Furthermore, many autoimmune diseases may be caused by the same SNPs or genes and therefore investigating SNPs or genes associated with other autoimmune diseases might improve prediction of T1D disease status. We hope to gain predictive power by allowing other SNPs to be included in the model via the gene-sets constructed with the NHGRI catalog.

We compare our methods to the same methods described in the simulation section. The AUC estimates in the validation set for selected procedures are shown in Table 3. In addition, we compare our methods to univariate SNP based methods that include only the 40 SNPs found to be associated with T1D disease risk (reported by the NHGRI) that were genotyped in our data. These methods reflect the traditional procedure of testing for association and subsequently building a fairly low dimensional prediction model with $p=40$

SNP predictors. We combine these 40 SNPs into a risk score through either ridge regression, a kernel machine model with k_{IBS} , or a weighted log odds ratio risk score (univariate WLGR) with log odds ratio SNP weights. In general, our proposed ANB KM estimators have much higher AUC than the global methods and purely NB methods. With k_{IBS} and $\rho = 0.999$, our ANB KM PCA method obtains a high AUC. Across the 15 random partitions of the data into training and validation sets of sizes described above, the median AUC was about 0.94. This method obtains similar prediction accuracy to the same method that does not use PCA, but it required much less computational time (in this example, running the gene-set analyses in parallel across 350 CPUs and aggregating across gene-sets on one CPU ran in total less than 40 minutes on a multi-core cluster with computers with 48GB of memory, whereas the no PCA analysis ran for approximately twice as long). These results also improve upon previously published results on the WTCCC T1D data, including an AUC of 0.79 from a polygenic analysis (Evans et al. 2009), and are comparable to an AUC of 0.91 from a gene pathway analysis (Eleftherohorinou et al. 2009). The pathway results described in Eleftherohorinou et al. (2009) also provides evidence that signal remains outside of known associated SNPs, and our methods further increase predictive power by allowing for complex effects within gene-sets. Our procedure estimates 23 of the 350 gene-sets to have nonzero effects in the second stage. It includes five of the 92 genes that have been associated with T1D in the final model. The other 18 genes that were included in the model were not reported as being associated with T1D specifically, but have been shown to be linked to other autoimmune disease risk. The k_{LIN} ANB blockwise methods select more gene-sets to be included in the final model and have much lower AUC.

6. DISCUSSION

The successful incorporation of genetic markers in risk prediction models has important implications in personalized medicine and disease prevention. However, standard methods for building such models are hindered by large datasets and nonlinear genetic associations with the outcome of interest. To overcome these challenges, we propose a multi-stage prediction model that includes genetic markers partitioned into gene-sets based on prior knowledge about the LD structure or pathway information. To achieve a balance between allowing a flexible model that captures complex nonlinear effects and efficient estimation in the model parameters, we utilize an ANBKM regression framework that builds non-linear risk models separately for each gene-set and then aggregates information from multiple gene-sets efficiently via an adaptive block-wise weighting scheme. The complexity and flexibility of machine learning complements the simplicity of a combining gene-set specific risk scores with regularized regression. These methods are used together to create a powerful method for risk modeling with genome-wide data. Through simulation studies and a real data example, we show that our ANBKM model performs well and maintains high prediction accuracy even when the underlying association of covariates and disease risk is complex. In addition, we justify the theoretical properties of our model, including the consistency of the kernel machine estimator, and contribute to the statistical learning field by providing insight into the behavior of the sample eigenspace of kernel functions within the regression framework.

The ANBKM model requires *a priori* selection of the gene-set structure and the kernel function. We observe in our numerical studies that the IBS kernel performs well in the presence of both nonlinear and linear effects. This is in part due to the fact that IBS kernel effectively captures non-linear additive effects across the SNPs. For any kernel function that can be written as $K = K_1 + K_2$, the eigenspace corresponding to the RKHS $\mathcal{H}_K$ is equal to the span of $\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$. Thus, the IBS kernel implies that the effects from different SNPs are additive. It is related to the triangular kernel (Fleuret and Sahbi 2003) and models SNP effects additively but non-linearly, so the space spanned by the IBS kernel will be identical to the space spanned by $\{I(Z_1 = 1), I(Z_1 = 2), \dots, I(Z_p = 1), I(Z_p = 2)\}$. Wu et al. (2011) found that the IBS kernel increases power over the linear kernel when the number of interactions is modest. This increase of power occurs when the SNPs are correlated and the effects due to the interaction between SNPs are somewhat captured by the inclusion of non-linear effects. More flexible kernels such as the Gaussian kernel can be used to capture more complex non-linear and interactive effects. In general, the performance of the procedure would depend on the kernel function and it would be desirable to select an optimal kernel for each gene-set to maximize the prediction accuracy of the resulting $\hat{h}^{(m)}(\mathbf{z}^{(m)})$. Hence, it may be useful to employ Multiple Kernel Learning (MKL) or composite kernel strategies (Bach et al. 2004; Lanckriet et al. 2004) when estimating the gene-set specific risk score $\hat{h}^{(m)}(\mathbf{z}^{(m)})$. One simple approach would be to compare the AIC for each individual gene-set risk score $\hat{h}^{(m)}(\mathbf{z}^{(m)})$ from the fitted models with different kernels and use the $\hat{h}^{(m)}(\mathbf{z}^{(m)})$ from the model with the highest AIC for the subsequent analysis. When we used this approach on our data, 275 gene-sets had lower individual AICs from the IBS kernel and 75 had 20 lower AICs from the linear kernel. Combining the $\hat{h}^{(m)}(\mathbf{z}^{(m)})$ across gene-sets with LASSO in the second stage, 32 gene-sets remained in the model with nonzero coefficients (as opposed to 23 and 47 from the purely IBS and linear kernel models, respectively), where 30 of those gene-sets were modeled with the IBS kernel and two were modeled with the linear kernel. When we included the stronger kernel for each block in the training set and calculated the AUC in the validation set after estimating γ with this mixture of $\hat{h}$ from the winning kernels, the AUC in the validation set was almost identical to the AUC from the method that only used the IBS kernel (AUC=0.94246 for the kernel choosing method and 0.94252 for the method with only IBS kernel). This gives further evidence that the IBS kernel is robust even for linear effects and this method of choosing the “winner” based on AIC in the training set for each gene-set appears to be effective.

Another factor that may affect the performance of our proposed procedure is the selection of φ for the kernel PCA. In our numerical studies, we see that kernel PCA approximation improves over the noPCA methods mainly in the computational efficiency, but also slightly in model selection and prediction accuracy. Hence in practice, we would recommend applying the kernel PCA with a relatively stringent threshold such as $1 - n^{-1}$ to estimate the eigenspace well while still substantially reducing dimensionality, although the optimal selection of threshold warrants further investigation.

Incorporating the block structure of the gene-sets in our model could potentially improve prediction accuracy over global methods that attempt to build one-stage models with a large number of unstructured genetic markers. Of course, one would expect that their relative

performance may also depend on how well the gene-sets are grouped together. The NB working assumption implies that the markers may interact within the gene-set but not across gene-sets. Thus it may be preferable to group genes that tend to interact with each other into a gene set. We recommend creating gene-sets based on biological knowledge related to how genes might interact with each other. In our numerical studies, we partitioned the genome based on the gene structure. Other examples of such knowledge base include recombination hotspots, protein-protein interaction networks and pathways. We note that when partitioning the entire genome into gene-sets, one may first screen these blocks using a testing procedure such as the logistic kernel machine score test proposed by Liu et al. (2008) to reduce the number of blocks in the model which may improve efficiency and prediction accuracy. It would also be interesting to explore the best procedures for this initial screening stage. We have found the KM score test for associations within gene blocks to perform well in other numerical examples. However, further research is needed to explore how the proposed procedure is affected by the screening procedure and the criteria used for forming the gene-sets.

Lastly, the proposed procedures can be easily extended to adjust for covariates. For example, if there are existing clinical variables or population structure principal components $\mathbf{X}$ available for risk prediction, one may impose a conditional NBKM model by extending (1) and (2) to

$$\begin{aligned} \text{logit } P(Y_i=1|\mathbf{Z}_i^{(\bullet)}, \mathbf{X}_i) &= a_0 + \mathbf{X}_i^\top \mathbf{b}_0 + \sum_{m=1}^M \text{logit } P(Y_i=1|\mathbf{Z}_i^{(m)}, \mathbf{X}_i) \\ \text{and } \text{logit } P(Y_i=1|\mathbf{Z}_i^{(m)}, \mathbf{X}_i) &= a_0^{(m)} + \mathbf{X}_i^\top \mathbf{b}_0^{(m)} + h^{(m)}(\mathbf{Z}_i^{(m)}), \quad \text{respectively.} \end{aligned}$$

The proposed procedures can be carried out by first fitting M separate models with $(\mathbf{X}_i, \mathbf{Z}_i^{(m)})$ and then adaptively weighting to obtain a sparse combination of $h^{(m)}$ across M gene-sets.

Acknowledgments

This research was supported by National Institutes of Health grants T32 AI007358, R01 GM079330, U54 LM008748, and National Science Foundation grants 0846234, 0854970 and DMS 1007762. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. The authors thank the editor, the associate editor, and two referees for their insightful and constructive comments that greatly improved the article.

References

- Bach, FR.; Lanckriet, GR.; Jordan, MI. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. Proceedings of the Twenty-first International Conference on Machine Learning; ACM; 2004. p. 6
- Bengio Y, Delalleau O, Le Roux N, Paiement J, Vincent P, Ouimet M. Learning Eigenfunctions Links Spectral Embedding and Kernel PCA. Neural Computation. 2004; 16:2197–2219. [PubMed: 15333211]
- Borchers A, Uibo R, Gershwin M. The Geoepidemiology of Type 1 Diabetes. Autoimmunity Reviews. 2010; 9:A355–A365. [PubMed: 19969107]

- Braun, M. PhD thesis. University of Bonn; 2005. Spectral Properties of the Kernel Matrix and their Application to Kernel Methods in Machine Learning.
- Breiman L, Spector P. Submodel Selection and Evaluation in Regression: the x-Random Case. *International Statistical Review/Revue Internationale de Statistique*. 1992; 60:291–319.
- Burton P, Clayton D, Cardon L, Craddock N, Deloukas P, Duncanson A, Kwiatkowski D, McCarthy M, Ouwehand W, Samani N, et al. Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature*. 2007; 447:661–678. [PubMed: 17554300]
- Casanova R, Hsu F-C, Sink KM, Rapp SR, Williamson JD, Resnick SM, Espeland MA, Initiative ADN, et al. Alzheimer's Disease Risk Assessment Using Large-Scale Machine Learning Methods. *PloS One*. 2013; 8:e77949. [PubMed: 24250789]
- Cassidy A, Myles J, van Tongeren M, Page R, Liloglou T, Duffy S, Field J. The LLP Risk Model: an Individual Risk Prediction Model for Lung Cancer. *British Journal of Cancer*. 2008; 98:270–276. [PubMed: 18087271]
- Chatterjee N, Carroll R. Semiparametric Maximum Likelihood Estimation Exploiting Gene-environment Independence in Case-control Studies. *Biometrika*. 2005; 92:399–418.
- Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail M. Projecting Absolute Invasive Breast Cancer Risk in White Women with a Model that Includes Mammographic Density. *Journal of the National Cancer Institute*. 2006; 98:1215–1226. [PubMed: 16954474]
- Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*. Cambridge University Press; 2000.
- D'Agostino R, Wolf P, Belanger A, Kannel W. Stroke Risk Profile: Adjustment for Antihypertensive Medication. The Framingham Study. *Stroke*. 1994; 25:40–43. [PubMed: 8266381]
- Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier Under Zero-one Loss. *Machine Learning*. 1997; 29:103–130.
- Eleftherohorinou H, Wright V, Hoggart C, Hartikainen A, Jarvelin M, Balding D, Coin L, Levin M. Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases. *PLoS One*. 2009; 4:e8068. [PubMed: 19956648]
- Evans D, Visscher P, Wray N. Harnessing the Information Contained Within Genome-wide Association Studies to Improve Individual Prediction of Complex Disease Risk. *Human Molecular Genetics*. 2009; 18:3525–3531. [PubMed: 19553258]
- Fan J, Lv J. Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society: Series B*. 2008; 70:849–911.
- Fleuret, F.; Sahbi, H. Scale-invariance of Support Vector Machines Based on the Triangular Kernel. 3rd International Workshop on Statistical and Computational Theories of Vision; 2003.
- Gail M. Discriminatory Accuracy from Single-nucleotide Polymorphisms in Models to Predict Breast Cancer Risk. *Journal of the National Cancer Institute*. 2008; 100:1037–1041. [PubMed: 18612136]
- Gail M, Brinton L, Byar D, Corle D, Green S, Schairer C, Mulvihill J. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who are Being Examined Annually. *Journal of the National Cancer Institute*. 1989; 81:1879–1886. [PubMed: 2593165]
- Gail M, Costantino J. Validating and Improving Models for Projecting the Absolute Risk of Breast Cancer. *Journal of the National Cancer Institute*. 2001; 93:334–335. [PubMed: 11238688]
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*. Vol. 2. Springer; New York; 2009.
- Hindorf L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, Manolio T. Potential Etiologic and Functional Implications of Genome-wide Association Loci for Human Diseases and Traits. *Proceedings of the National Academy of Sciences*. 2009; 106:9362–9367.
- Janssens A, van Duijn C. Genome-based Prediction of Common Diseases: Advances and Prospects. *Human molecular genetics*. 2008; 17:R166–R173. [PubMed: 18852206]
- Johansen C, Hegele R. Predictive Genetic Testing for Coronary Artery Disease. *Critical Reviews in Clinical Laboratory Sciences*. 2009; 46:343–360. [PubMed: 19958218]
- Knight K, Fu W. Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*. 2000; 28:1356–1378.

- Koltchinskii V, Giné E. Random Matrix Approximation of Spectra of Integral Operators. *Bernoulli*. 2000; 6:113–167.
- Kwee L, Liu D, Lin X, Ghosh D, Epstein M. A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *The American Journal of Human Genetics*. 2008; 82:386–397. [PubMed: 18252219]
- Lanczkiet G, Cristianini N, Bartlett P, Ghaoui L, Jordan M. Learning the Kernel Matrix with Semidefinite Programming. *The Journal of Machine Learning Research*. 2004; 5:27–72.
- Lee S, DeCandia T, Ripke S, Yang J, Sullivan P, Goddard M, Keller M, Visscher P, Wray N, et al. Estimating the Proportion of Variation in Susceptibility to Schizophrenia Captured by Common SNPs. *Nature Genetics*. 2012; 44:247–250. [PubMed: 22344220]
- Li, H.; Luan, Y. Kernel Cox Regression Models for Linking Gene Expression Profiles to Censored Survival Data. *Pacific Symposium on Biocomputing*; World Scientific Pub Co Inc; 2003. p. 65-76.
- Liu D, Ghosh D, Lin X. Estimation and Testing for the Effect of a Genetic Pathway on a Disease Outcome Using Logistic Kernel Machine Regression via Logistic Mixed Models. *BMC Bioinformatics*. 2008; 9:292. [PubMed: 18577223]
- Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*. 2007; 63:1079–1088. [PubMed: 18078480]
- Machiela M, Chen C, Chen C, Chanock S, Hunter D, Kraft P. Evaluation of Polygenic Risk Scores for Predicting Breast and Prostate Cancer Risk. *Genetic Epidemiology*. 2011; 35:506–514. [PubMed: 21618606]
- Makowsky R, Pajewski N, Klimentidis Y, Vazquez A, Duarte C, Allison D, de los Campos G. Beyond Missing Heritability: Prediction of Complex Traits. *PLoS genetics*. 2011; 7:e1002051. [PubMed: 21552331]
- Marchini J, Donnelly P, Cardon L. Genome-wide Strategies for Detecting Multiple Loci that Influence Complex Diseases. *Nature Genetics*. 2005; 37:413–417. [PubMed: 15793588]
- Mardis E. The Impact of Next-generation Sequencing Technology on Genetics. *Trends in Genetics*. 2008; 24:133–141. [PubMed: 18262675]
- McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, Ioannidis J, Hirschhorn J. Genome-wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges. *Nature Reviews Genetics*. 2008; 9:356–369.
- McIntosh MW, Pepe MS. Combining Several Screening Tests: Optimality of the Risk Score. *Biometrics*. 2002; 58:657–664. [PubMed: 12230001]
- McKinney B, Reif D, Ritchie M, Moore J. Machine Learning for Detecting Gene-gene Interactions: a Review. *Applied Bioinformatics*. 2006; 5:77–88. [PubMed: 16722772]
- Meigs J, Shrader P, Sullivan L, McAteer J, Fox C, Dupuis J, Manning A, Florez J, Wilson P, D'Agostino R Sr, et al. Genotype Score in Addition to Common Risk Factors for Prediction of Type 2 Diabetes. *The New England Journal of Medicine*. 2008; 359:2208–2219. [PubMed: 19020323]
- Murcay C, Lewinger J, Gauderman W. Gene-Environment Interaction in Genome-Wide Association Studies. *American Journal of Epidemiology*. 2009; 169:219–226. [PubMed: 19022827]
- Paynter N, Chasman D, Paré G, Buring J, Cook N, Miletich J, Ridker P. Association Between a Literature-based Genetic Risk Score and Cardiovascular Events in Women. *The Journal of the American Medical Association*. 2010; 303:631–637. [PubMed: 20159871]
- Pearson T, Manolio T. How to Interpret a Genome-wide Association Study. *Journal of the American Medical Association*. 2008; 299:1335. [PubMed: 18349094]
- Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. United Kingdom: Oxford University Press; 2003.
- Pepe MS, Cai T, Longton G. Combining Predictors for Classification Using the Area Under the Receiver Operating Characteristic Curve. *Biometrics*. 2006; 62:221–229. [PubMed: 16542249]
- Purcell S, Wray N, Stone J, Visscher P, O'Donovan M, Sullivan P, Sklar P, Ruderfer D, McQuillan A, Morris D, et al. Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature*. 2009; 460:748–752. [PubMed: 19571811]
- Rasmussen, C.; Williams, C. *Gaussian Processes for Machine Learning*. Springer; 2006.

- Schaid D. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. *Human Heredity*. 2010; 70:132–140. [PubMed: 20606458]
- Scholkopf, B.; Smola, A. *Learning with Kernels*. MIT Press Cambridge; Mass: 2002.
- Spiegelman D, Colditz G, Hunter D, Hertzmark E. Validation of the Gail et al. Model for Predicting Individual Breast Cancer Risk. *Journal of the National Cancer Institute*. 1994; 86:600–607. [PubMed: 8145275]
- Su JQ, Liu JS. Linear Combinations of Multiple Diagnostic Markers. *Journal of the American Statistical Association*. 1993; 88:1350–1355.
- Swets J. Measuring the Accuracy of Diagnostic Systems. *Science*. 1988; 240:1285–1293. [PubMed: 3287615]
- Thompson I, Ankerst D, Chi C, Goodman P, Tangen C, Lucia M, Feng Z, Parnes H, Coltman C Jr. Assessing Prostate Cancer Risk: Results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute*. 2006; 98:529–534. [PubMed: 16622122]
- Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Umbach D, Weinberg C. Designing and Analysing Case-control Studies to Exploit Independence of Genotype and Exposure. *Statistics in Medicine*. 1997; 16:1731–1743. [PubMed: 9265696]
- Van Belle T, Coppieters K, Von Herrath M. Type 1 diabetes: Etiology, Immunology, and Therapeutic Strategies. *Physiological Reviews*. 2011; 91:79–118. [PubMed: 21248163]
- Vasan R. Biomarkers of Cardiovascular Disease: Molecular Basis and Practical Considerations. *Circulation*. 2006; 113:2335–2362. [PubMed: 16702488]
- Visscher P, Hill W, Wray N. Heritability in the Genomics Era: Concepts and Misconceptions. *Nature Reviews Genetics*. 2008; 9:255–266.
- Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson H, Diver W, Thun M, Cox D, Hankinson S, Kraft P, et al. Performance of Common Genetic Variants in Breast-cancer Risk Models. *New England Journal of Medicine*. 2010; 362:986–993. [PubMed: 20237344]
- Wei Z, Wang K, Qu H, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner J, Chiavacci R, et al. From Disease Association to Risk Assessment: an Optimistic View from Genome-wide Association Studies on Type 1 Diabetes. *PLoS Genetics*. 2009; 5:e1000678. [PubMed: 19816555]
- Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackleton E, Kim C, Mentch F, Van Steen K, Visscher PM, et al. Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease. *The American Journal of Human Genetics*. 2013; 92:1008–1012. [PubMed: 23731541]
- Williams, C.; Seeger, M. The Effect of the Input Density Distribution on Kernel-based Classifiers. *Proceedings of the 17th International Conference on Machine Learning*; Morgan Kaufmann; 2000. p. 1159–1166.
- Wolf P, D'agostino R, Belanger A, Kannel W. Probability of Stroke: a Risk Profile from the Framingham Study. *Stroke*. 1991; 22:312–318. [PubMed: 2003301]
- Wray N, Goddard M, Visscher P. Prediction of Individual Genetic Risk of Complex Disease. *Current Opinion in Genetics & Development*. 2008; 18:257–263. [PubMed: 18682292]
- Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics*. 2011; 89:82–93. [PubMed: 21737059]
- Yang Q, Khoury M. Evolving Methods in Genetic Epidemiology. III. Gene-Environment Interaction in Epidemiologic Research. *Epidemiologic Reviews*. 1997; 19:33–43. [PubMed: 9360900]
- Yang Q, Khoury M, Botto L, Friedman J, Flanders W. Improving the Prediction of Complex Diseases by Testing for Multiple Disease-susceptibility Genes. *The American Journal of Human Genetics*. 2003; 72:636–649. [PubMed: 12592605]
- Zhang H. Exploring Conditions for the Optimality of Naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*. 2005; 19:183–198.
- Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.

A. APPENDIX A: THEORETICAL JUSTIFICATION

A.1 Convergence of the kernel PCA estimator

We now provide some theoretical justification to the proposed adaptive NBKM. To this end, we first show that the kernel PCA estimator $\hat{h}^{(m)}(\cdot)$ introduced earlier is a root-n consistent estimator of the true $h^{(m)}(\cdot)$. Recall that $\hat{h}^{(m)}(\cdot)$ is constructed based on the rank $r_n^{(m)}$ approximation $\mathbf{K}_{n(m)}^{[r_n^{(m)}]}$ to the Gram matrix $\mathbf{K}_{n(m)}$ where

$$r_n^{(m)} = \arg \min \left\{ r : \sum_{i=1}^r l_i^{(m)} \geq \rho \sum_{i=1}^n l_i^{(m)} \right\}$$

and $l_1^{(m)} \geq \dots \geq l_n^{(m)}$ are the eigenvalues of $\mathbf{K}_{n(m)}$. More specifically, let

$\mathbf{K}_{n(m)}^{[r_n^{(m)}]} = \mathbf{U}_{(m)} \mathbf{D}_{(m)} \mathbf{U}_{(m)}^\top$ be its eigenvalue decomposition, then $\hat{h}^{(m)}(\cdot) = (\hat{\boldsymbol{\beta}}^{(m)})^\top \hat{\boldsymbol{\Psi}}_{(m)}(\cdot)$ where

$$\hat{\boldsymbol{\Psi}}_{(m)}(\cdot) = n^{-1/2} \mathbf{D}^{-1} \mathbf{U}^\top (k(\cdot, \mathbf{Z}_i))_{1 \leq i \leq n}^\top,$$

and $\hat{\boldsymbol{\beta}}^{(m)}$ maximizes

$$\mathcal{L}^{(P)}(a, \boldsymbol{\beta}; \tilde{\boldsymbol{\Psi}}_{(m)}) = \mathbf{Y}^\top \log g(a + \tilde{\boldsymbol{\Psi}}_{(m)} \boldsymbol{\beta}) + (1 - \mathbf{Y})^\top \log \{1 - g(a + \tilde{\boldsymbol{\Psi}}_{(m)} \boldsymbol{\beta})\} - \tau \|\boldsymbol{\beta}\|_2^2.$$

Theorem 1

Assume that

$$\text{logit} P(Y=1 | \mathbf{Z}^{(m)}) = a^{(m)} + h^{(m)}(\mathbf{Z}^{(m)})$$

where $h^{(m)}(\cdot)$ belongs to the reproducing kernel Hilbert space $\mathcal{H}_k^{(m)}$ identified with a kernel $k(\cdot, \cdot)$ of finite rank. If $\rho \rightarrow 1$ and $\tau = o(\sqrt{n})$, then

$$\|\hat{h}^{(m)} - h^{(m)}\| = \left(\int (\hat{h}^{(m)} - h^{(m)})^2 dP_{\mathbf{Z}^{(m)}} \right)^{1/2} = O_p(n^{-1/2}), \quad (\text{A.1})$$

as n tends to infinity.

Proof—For brevity, we shall abbreviate the superscript throughout the proof. In addition, since it is somewhat lengthy we break the argument into several steps.

We first show that with probability tending to one, $r_n = r$. Observe that $\text{rank}(\mathbf{K})$ is no greater than the rank of k . Therefore,

$$P(r_n < r) = P(r_n \leq r-1) = P\left(\sum_{j=1}^{r-1} l_j \geq \rho \sum_{j=1}^r l_j\right) = P\left(l_r \leq \frac{1}{1-\rho} \sum_{j=1}^{r-1} l_j\right).$$

We now show that the rightmost hand side goes to 0 as $n \rightarrow \infty$. It is well known (see, e.g., Koltchinskii and Giné 2000) that $\max_{1 \leq j \leq r} |l_j - \lambda_j| = o_p(1)$. In particular, under the event that $\lambda_j/2 < l_j < 2\lambda_j$ for all $1 \leq j \leq r$ which holds with probability tending to one,

$$\sum_{j=1}^{r-1} l_j \leq 2 \sum_{j=1}^{r-1} \lambda_j = \left(4\lambda_r^{-1} \sum_{j=1}^{r-1} \lambda_j\right) \lambda_r/2 < \left(4\lambda_r^{-1} \sum_{j=1}^{r-1} \lambda_j\right) l_r.$$

Because $\rho \rightarrow 1$, for large enough n , we get $\rho > 1 - (4\lambda_r^{-1} \sum_{j=1}^{r-1} \lambda_j)$. As a result,

$$P\left(l_r \leq \frac{1}{1-\rho} \sum_{j=1}^{r-1} l_j\right) \rightarrow 0.$$

In the light of this observation, we shall consider the case conditional on the event that $r_n = r$ in what follows.

Next, we argue that $\hat{\Psi} = \sqrt{n} \mathbb{D}^{1/2} \mathbb{U}^\top$ approximate $\Psi = (\psi_j(\mathbf{Z}_i))_{1 \leq i \leq n, 1 \leq j \leq r}$ well in that

$$\|\hat{\Psi} - \Psi\|_{\text{F}}^2 := \sum_{i,j=1}^n \left(\hat{\psi}_j(\mathbf{Z}_i) - \psi_j(\mathbf{Z}_i)\right)^2 = O_p(1).$$

where $\|\cdot\|_{\text{F}}$ represents the usual matrix Frobenius norm. Recall that

$k(s, t) = \sum_{j=1}^r \lambda_j \varphi_j(s) \varphi_j(t)$, where $\lambda_1, \dots, \lambda_r$ and $\mathbb{E} \phi_j(X) \phi_l(X) = \delta_{jl}$, where δ_{jl} is the Kronecker's delta. Therefore, $\mathbf{K}_n = n^{-1} \Phi \Lambda \Phi^\top$ where $\Phi = \{\phi_j(\mathbf{Z}_i)\}_{1 \leq i \leq n, 1 \leq j \leq r}$, and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$. It is not hard to see that

$$\mathbb{E} \|n^{-1} \Phi^\top \Phi - I\|_{\text{F}}^2 = \mathbb{E} \sum_{j_1, j_2=1}^r (\mathbb{E}_n \varphi_{j_1}(X) \varphi_{j_2}(X) - \mathbb{E} \varphi_{j_1}(X) \varphi_{j_2}(X))^2$$

where $\mathbb{E}_n$ stands for the empirical expectation, i.e., $\mathbb{E}_n f(X) = n^{-1} \sum_{i=1}^n f(x_i)$. Letting $\Phi = \mathcal{U}_{n \times r} \mathcal{D}_{r \times r} \mathcal{V}_{r \times r}^\top$ be its singular value decomposition, we have

$$\sum_{j=1}^r (n^{-1}d_j^2 - 1)^2 = \|n^{-1}\mathbf{\Phi}^\top \mathbf{\Phi} - I\|_F^2 = O_p(n^{-1}),$$

which implies that $\max_j |n^{-\frac{1}{2}}d_j - 1| = O_p(n^{-\frac{1}{2}})$. Write $\tilde{\mathbf{G}} = \mathbf{U}\mathbf{V}^\top \mathbf{\Lambda} \mathbf{V}\mathbf{U}^\top$. Then

$$\begin{aligned} \|\tilde{\mathbf{G}} - \mathbf{K}_n\|_F^2 &= \|\mathcal{U}\mathcal{V}^\top \mathbf{\Lambda} \mathcal{V}\mathcal{U}^\top - n^{-1}\mathcal{U}\mathcal{D}\mathcal{V}^\top \mathbf{\Lambda} \mathcal{V}\mathcal{D}\mathcal{U}^\top\|_F^2 = \|\mathcal{V}^\top \mathbf{\Lambda} \mathcal{V} - n^{-1}\mathcal{D}\mathcal{V}^\top \mathbf{\Lambda} \mathcal{V}\mathcal{D}\|_F^2 \\ &\leq \|\mathcal{V}^\top \mathbf{\Lambda} \mathcal{V}\|_{\max}^2 \sum_{j_1, j_2=1}^r (1 - n^{-1}d_{j_1}d_{j_2})^2 = O_p(n^{-1}). \end{aligned}$$

Recall that $\mathbf{U}\mathbf{V}^\top$ is the eigenvector of $\tilde{\mathbf{G}}$, and $\mathbf{K}_n = \mathbf{U}_{n \times r} \mathbf{D}_{r \times r} \mathbf{U}_{n \times r}^\top$ is its singular value decomposition. A standard perturbation analysis yields

$$\|n^{-1/2}\hat{\mathbf{\Psi}} - \mathbf{\Lambda}^{1/2}\mathcal{V}\mathcal{U}^\top\|_F^2 = O(\|\tilde{\mathbf{G}} - \mathbf{K}_n\|_F^2) = O_p(n^{-1}).$$

Together with the fact that

$$\|\mathbf{\Lambda}^{1/2}\mathcal{V}\mathcal{U}^\top - n^{-1/2}\mathbf{\Psi}\|_F \leq \lambda_1^{1/2} \|\mathcal{V}\mathcal{U}^\top - n^{-1/2}\mathbf{\Phi}\|_F = O_p(n^{-1/2}),$$

we get

$$\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}\|_F \leq \|\hat{\mathbf{\Psi}} - \sqrt{n}\mathbf{\Lambda}^{1/2}\mathcal{V}\mathcal{U}^\top\|_F + \|\sqrt{n}\mathbf{\Lambda}^{1/2}\mathcal{V}\mathcal{U}^\top - \mathbf{\Psi}\|_F = O_p(1),$$

by the triangular inequality.

We now prove that $\hat{\boldsymbol{\beta}}$ is root-n consistent. It suffices to show that for any $\varepsilon > 0$, there exists a constant $C > 0$ such that

$$P \left\{ \sup_{\|\mathbf{n}^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| \geq C} [\mathcal{L}^{(P)}(a, \boldsymbol{\beta}; \hat{\mathbf{\Psi}}) - \mathcal{L}^{(P)}(a, \boldsymbol{\beta}_0; \hat{\mathbf{\Psi}})] < 0 \right\} \geq 1 - \varepsilon$$

To this end, write $\mathbf{q} = \mathbf{n}^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Then, by Taylor expansion,

$$\begin{aligned}
D_n(\mathbf{q}) &:= \mathcal{L}^{(P)}(a, \beta_0 + n^{-\frac{1}{2}}\mathbf{q}; \hat{\Psi}) - \mathcal{L}^{(P)}(a, \beta_0; \hat{\Psi}) \\
&= n^{-\frac{1}{2}} \left[\mathbf{Y} - g(a + \hat{\Psi}\beta_0) \right]^\top \hat{\Psi}\mathbf{q} - \frac{1}{2}\mathbf{q}^\top \left[n^{-1} \sum_{i=1}^n \xi(a + \hat{\Psi}_i^\top \beta_0) \hat{\Psi}_i \hat{\Psi}_i^\top \right] \mathbf{q} - 2n^{-\frac{1}{2}} \tau \sum_{j=1}^r \beta_{0j} \mathbf{q} + n^{-1} \tau r \mathbf{q}^\top \mathbf{q} + o_p \left(n^{-1} \|\mathbf{q}\|^2 \right) \quad (\text{A.}) \\
&\leq n^{-\frac{1}{2}} \{ \mathbf{Y} - g(a + \Psi\beta_0) \}^\top \Psi\mathbf{q} - \frac{1}{2}\mathbf{q}^\top \{ n^{-1} \Psi^\top \text{diag}\{\xi(a + \Psi\beta_0)\} \Psi \} \mathbf{q} - 2n^{-\frac{1}{2}} \tau \sum_{j=1}^r \beta_{0j} \mathbf{q} + n^{-1} \tau r \mathbf{q}^\top \mathbf{q} + c(n^{-\frac{1}{2}}(\|\mathbf{q}\| + \|\mathbf{q}\|^2)) \quad (2) \\
&\leq c\|\mathbf{q}\| - \mathbf{q}^\top \Lambda \mathbf{q} + cn^{-\frac{1}{2}}\|\mathbf{q}\|^2,
\end{aligned}$$

where $c > 0$ is a generic constant that may take different values at each appearance. It is now clear that by taking $C > 0$ large enough, $D_n(\mathbf{q}) < 0$ because the second term on the rightmost hand side dominates the remaining terms. This yields the root- n consistency of $\hat{\beta}$.

Finally, we establish the convergence rates for the estimated basis function via Nyström projection, that is

$$\|\hat{\Psi}(\cdot) - \Psi(\cdot)\|^2 := \sum_{j=1}^r \|\hat{\Psi}_j(\cdot) - \Psi_j(\cdot)\|^2 = O_p(n^{-1}).$$

where $\hat{\Psi}_j$ is the j th component function of $\hat{\Psi}(\cdot)$. Write $\mathbf{K}_Z = [k(\mathbf{z}, \mathbf{Z}_1), \dots, k(\mathbf{z}, \mathbf{Z}_n)]^\top = \Phi \Lambda$ $[\phi_1(\mathbf{z}), \dots, \phi_r(\mathbf{z})]^\top =: \Phi \Lambda \Phi_Z$. Then

$$\begin{aligned}
\|\hat{\Psi}(\cdot) - \Psi(\cdot)\|^2 &= E_Z \|\hat{\Psi}(\mathbf{z}) - \Psi(\mathbf{z})\|^2 = E_Z \|n^{-1/2} \mathbb{D}^{-1/2} \mathbb{U}^\top \Phi \Lambda \Phi_Z - \Lambda^{1/2} \Phi_Z\|^2 \\
&= \|n^{-1/2} \mathbb{D}^{-1/2} \mathbb{U}^\top \Phi \Lambda - \Lambda^{1/2}\|_F^2 \leq \lambda_1 \|n^{-1/2} \mathbb{D}^{-1/2} \mathbb{U}^\top \Phi \Lambda^{1/2} - I\|_F^2,
\end{aligned}$$

where E_Z is the expectation taken over $\mathbf{z}$ which follows distribution P_Z , and the last equality follows from the fact that $\{\phi_j\}$ is an orthonormal basis in $L_2(P_Z)$. Now since $\Phi \Lambda^{1/2} = \Psi$,

$$\|n^{-1/2} \mathbb{D}^{-1/2} \mathbb{U}^\top \Phi \Lambda^{1/2} - I\|_F^2 = \|\hat{\Psi}^{-1} \Psi - I\|_F^2.$$

The desired statement then follows from the fact that $\|\hat{\Psi} - \Psi\|_F^2 = O_p(1/n)$ as shown before.

To summarize, we conclude that

$$\begin{aligned}
\|\hat{h} - h\| &= \|\hat{\beta}^\top \hat{\Psi}(\cdot) - \beta_0^\top \Psi(\cdot)\| = \|(\hat{\beta} - \beta_0)^\top \hat{\Psi}(\cdot) + \beta_0^\top (\hat{\Psi}(\cdot) - \Psi(\cdot))\| \\
&\leq \|\hat{\beta} - \beta_0\| \|\hat{\Psi}(\cdot)\| + \|\beta_0\| \|\hat{\Psi}(\cdot) - \Psi(\cdot)\| \leq O_p(n^{-1/2}).
\end{aligned}$$

A.2 Model selection consistency

We now provide justification for the selection consistency for the gene sets. Recall that

$$\{\hat{b}, \hat{\gamma}\} = \underset{b, \gamma}{\operatorname{argmax}} \left\{ \mathbf{Y}^\top \log g(b + \hat{\mathbf{H}}\gamma) + (1 - \mathbf{Y})^\top \log \{1 - g(b + \hat{\mathbf{H}}\gamma) - \tau_2 \|\gamma\|_1\} \right\}.$$

Denote by $\mathcal{A} = \{m : h^{(m)}(\mathbf{z}) > 0\}$ and $\mathcal{A}^c = \{m : h^{(m)}(\mathbf{z}) = 0\}$. Then we have

Theorem 2

Assume that $\tau_2 \rightarrow \infty$ in such a fashion that $n^{-\frac{1}{2}}\tau_2 \rightarrow 0$. Then

$$\lim_{n \rightarrow \infty} \mathcal{P}(\hat{\gamma}_{\mathcal{A}^c} = \mathbf{0}) = 1. \quad (\text{A.3})$$

Proof—The proof follows from very similar arguments as those provided for the LASSO and adaptive LASSO (Knight and Fu 2000; Zou 2006) and hence we only provide a sketched outline below. First, we may reparameterize the objective function from (6) by

defining $\theta_m = \gamma_m \|\hat{h}^{(m)}\|$, $\|\hat{h}^{(m)}\| = \sqrt{n^{-1} \sum_{i=1}^n (\hat{h}^{(m)}(\mathbf{Z}_i^{(m)}))^2}$, $\tilde{H}_{im} = \hat{h}^{(m)}(\mathbf{Z}_i^{(m)}) / \|\hat{h}^{(m)}\|$, $\tilde{\mathbf{H}}_m = [H_{1m}, \dots, H_{nm}]^\top$, and $\tilde{\mathbf{H}} = [\tilde{\mathbf{H}}_1, \dots, \tilde{\mathbf{H}}_M]$. The reparameterized estimator can then be represented as

$$\{\hat{b}, \hat{\theta}\} = \underset{b, \theta}{\operatorname{argmax}} \left\{ \mathbf{Y}^\top \log g(b + \tilde{\mathbf{H}}\theta) + (1 - \mathbf{Y})^\top \log \{1 - g(b + \tilde{\mathbf{H}}\theta) - \tau_2 \sum_{m=1}^M \frac{|\theta_m|}{\|\hat{h}^{(m)}\|} \} \right\}.$$

It then follows with an identical argument as the epi-convergence approach of Knight and

Fu (2000) and Zou (2006) that $n^{\frac{1}{2}}(\hat{\theta}_{\mathcal{A}^c} - \theta_{0, \mathcal{A}^c}) = O_p(1)$ and $n^{\frac{1}{2}}\hat{\theta}_{\mathcal{A}^c} \rightarrow_d 0$. This, together with the convergence of $\hat{h}$, also implies that for a given $\mathbf{Z}^{(\bullet)}$, $b + \gamma^\top \hat{\mathbf{H}}(\mathbf{Z}^{(\bullet)})$ is a root-n consistent

estimator of $b_0 + \sum_{m=1}^M h(\mathbf{Z}^{(m)})$. To show that for all $m \in \mathcal{A}^c$, $P(\hat{\theta}_m > 0) \rightarrow 0$, define $\phi(\tilde{\mathbf{H}}^\top \theta) = \log(1 + \exp(\tilde{\mathbf{H}}^\top \theta))$. For the event $\hat{\theta}_m > 0$, the Karush-Kuhn-Tucker (KKT) optimality conditions imply that

$$\tilde{\mathbf{H}}_m^\top \left[\mathbf{Y} - \varphi'(b + \tilde{\mathbf{H}}\theta) \right] = \frac{\tau_2}{\|\hat{h}^{(m)}\|}.$$

From the Taylor expansion along with similar arguments as in Zou (2006), we have that the

left side of the equation is $O_p(n^{\frac{1}{2}})$ while the right side tends to infinity. Therefore, $\hat{\gamma}_m = 0$ implies $\hat{\theta}_m = 0$ almost everywhere, and so $\hat{\gamma}$ exhibits the oracle property.

B. APPENDIX B: SIMULATION DETAILS

For the simulation settings, we generated disease status through various functions of the

SNPs in four regions. Specifically, $\text{logit } P(Y=1|\mathbf{Z}^{(\bullet)}) = \sum_{m=1}^4 h^{(m)}(\mathbf{z})$, where $h^{(m)}(\mathbf{z}) = h^{(NLm)}(\mathbf{z})$ for the nonlinear (allNL) model, $h^{(m)}(\mathbf{z}) = h^{(Lm)}(\mathbf{z})$ for the linear (allL) model, and $h^{(m)}(\mathbf{z}) = h^{(Lm)}(\mathbf{z}), m = 1, 2$ and $h^{(m)}(\mathbf{z}) = h^{(NL(m-1))}(\mathbf{z}), m = 3, 4$ for the partially linear and nonlinear (NLN) model. The forms of these functions are as follows:

$h^{(NL1)}$ includes many two- and three-way interactions:

$$h^{(NL1)}(\mathbf{z}) = 2 * \left(\sum_{i=1}^{10} (Z_{25} - Z_{26} + Z_{27}) * (Z_i - .7 * Z_{30} * Z_{p/2}) * (Z_3 - Z_5) \right) * \sin\left(\sum_{i=1}^p Z_i\right)$$

$h^{(NL2)}$ has exponential effects and a many-way interaction:

$$h^{(NL2)}(\mathbf{z}) = 2 * \left(\sum_{i=1}^{10} \exp(Z_i - 4) - \prod_{i=10}^p \frac{Z_i}{2} + 1 \right)$$

$h^{(NL3)}$ includes exponential effects, a many-way interaction, and a $\tan(\mathbf{z})$ function with 24 causal SNPs:

$$h^{(NL3)} = 0.1 * \sum_{i=1}^{20} \exp(Z_i - Z_{22} - 1) - 2 * \sum_{i=1}^{10} (Z_{25} - Z_{26} + Z_{27}) * Z_i - Z_{11} * Z_{i+10} - 4 * \sin\left(\sum_{i=1}^{20} Z_i\right) - \tan\left(\sum_{i=1}^{20} Z_i\right)$$

$h^{(NL4)}$ includes exponential effects and a $\sin(\mathbf{z})$ function with 10 causal SNPs:

$$h^{(NL4)} = \left[\sum_{i=1}^{10} \exp(Z_i) \right] * \sin\left(\sum_{i=1}^{15} Z_i - 1\right) - 2 * \prod_{i=40}^p \frac{Z_i}{2}$$

$h^{(L1)}$ is additive for ten SNPs with equal weight: $h^{(L1)} = \sum_{i=1}^{10} 0.4 * Z_i$

$h^{(L2)}$ is additive for all SNPs in a region with equal weight: $h^{(L2)} = \sum_{i=1}^p 0.3 * Z_i$

$h^{(L3)}$ is additive for 12 SNPs with 6 having a small weight of .1 and the others a weight of .6: $h^{(L3)} = \sum_{i=1}^6 0.1 * Z_i + \sum_{i=7}^{12} 0.6 * Z_i$

$h^{(L4)}$ is additive of one-third of the SNPs in a region with half having a small weight of .35 and the others having a weight of .75: $h^{(L4)} = \sum_{i=1}^{p/6} 0.35 * Z_i + \sum_{i=p-p/6}^p 0.75 * Z_i$

Table 1

Correlation and Median squared error (MSE) of $\hat{h}(\mathbf{z})$ with $h(\mathbf{z})$ from simulation studies for the adaptively weighted gene-set regression model with $\hat{\gamma}_{cv}$. Presented are correlation (MSE) estimates for the four informative gene-sets for three types of settings representing different types of effects within each informative gene-set (all linear, two linear and two nonlinear, and all nonlinear). The MSE for the noninformative gene-sets are always essentially zero. Shown also are the average number of gene-sets selected (average number of informative gene-sets selected).

Effect Setting		k_{LIS} $\phi = .999$	k_{LIN} $\phi = .999$	k_{LIS} $\phi = 1$	k_{LIN} $\phi = 1$
allL	L1	0.95 (0.102)	0.97 (0.076)	0.95 (0.107)	0.97 (0.092)
	L2	0.81 (0.140)	0.86 (0.113)	0.81 (0.139)	0.85 (0.137)
	L3	0.95 (0.080)	0.99 (0.059)	0.97 (0.089)	0.99 (0.074)
	L4	0.82 (0.195)	0.88 (0.105)	0.82 (0.187)	0.87 (0.110)
	Gene-sets Selected	4.1 (2.7)	5.4 (2.8)	4.2 (2.7)	6.6 (2.9)
LNL	L1	0.90 (0.309)	0.93 (0.298)	0.90 (0.309)	0.93 (0.289)
	L2	0.71 (0.144)	0.76 (0.144)	0.70 (0.144)	0.73 (0.144)
	NL2	0.86 (2.364)	0.66 (1.333)	0.86 (2.204)	0.65 (1.170)
	NL3	0.88 (2.336)	0.84 (3.213)	0.88 (2.422)	0.84 (3.375)
	Gene-sets Selected	4.0 (3.7)	5.7 (2.8)	4.1 (3.0)	7.0 (2.8)
allNL	NL1	0.57 (1.175)	0.20 (1.649)	0.57 (1.162)	0.20 (1.605)
	NL2	0.76 (0.728)	0.78 (0.727)	0.75 (0.728)	0.77 (0.735)
	NL3	0.86 (1.394)	0.83 (2.337)	0.87 (1.434)	0.83 (2.393)
	NL4	0.88 (0.699)	0.76 (0.812)	0.88 (0.696)	0.75 (0.831)
	Gene-sets Selected	4.5 (3.1)	4.4 (2.5)	4.5 (3.1)	5.3 (2.5)

Table 2

AUC $\times 100$ (empirical standard error) for the simulation studies with various approaches under 3 different settings for generating h : all linear (allL), linear and non-linear (LNL) as well as all non-linear (NL). For the proposed approach, we include results corresponding to γ estimated as $\hat{\gamma}$ and $\hat{\gamma}_{cv}$.

$\mathcal{X}$	φ	Weight	allL		LNL		NL	
			$\hat{\gamma}_{cv}$	$\hat{\gamma}$	$\hat{\gamma}_{cv}$	$\hat{\gamma}$	$\hat{\gamma}_{cv}$	$\hat{\gamma}$
k_{IBS}	.999	ANB	80.9 (2.1)	72.5 (3.2)	87.5 (1.8)	81.7 (2.8)	84.1 (2.2)	74.8 (3.6)
k_{LIN}	.999	ANB	81.2 (2.3)	78.3 (2.5)	81.3 (2.5)	78.6 (2.5)	76.2 (3.2)	72.0 (3.1)
k_{IBS}	.999	NB	71.2 (2.3)		73.0 (2.6)		70.8 (2.8)	
k_{LIN}	.999	NB	71.9 (2.4)		70.3 (2.6)		67.2 (2.9)	
k_{IBS}	1	ANB	80.7 (2.2)	71.8 (3.4)	87.4 (1.8)	81.1 (3.0)	84.0 (2.1)	74.1 (3.9)
k_{LIN}	1	ANB	80.9 (2.4)	74.7 (2.6)	81.2 (2.5)	75.4 (2.7)	75.5 (3.5)	68.6 (3.2)
k_{IBS}	1	NB	71.2 (2.3)		73.0 (2.5)		70.6 (2.8)	
k_{LIN}	1	NB	68.6 (5.9)		68.6 (2.8)		65.2 (3.1)	
Global Method			allL		LNL		NL	
gRidge			70.7 (2.4)		69.5 (2.5)		64.1 (2.9)	
gKernelIBS			73.6 (2.3)		75.6 (2.3)		68.5 (2.8)	
SIS			75.4 (3.9)		72.5 (5.0)		66.7 (4.2)	
WLGR			65.7 (2.6)		63.8 (2.8)		57.5 (3.2)	

Table 3

AUC $\times 100$ for the models used to predict type 1 diabetes risk in the WTCCC dataset using 350 gene-sets. Median AUC across 15 random partitions of the dataset. The last column reflects the number of SNPs (genes) that are included in the final prediction model based on entire data set.

$\mathcal{K}$	φ	Block Weighting	AUC	# SNPs (genes)
k_{IBS}	.999	ANB	94.3	1041 (23)
k_{LIN}	.999	ANB	84.5	2086 (47)
k_{IBS}	.999	NB	85.5	9257 (350)
k_{LIN}	.999	NB	83.6	9257 (350)
k_{IBS}	1	ANB	94.1	2409 (54)
k_{LIN}	1	ANB	85.1	3580 (68)
k_{IBS}	1	NB	84.4	9257 (350)
k_{LIN}	1	NB	82.2	9257 (350)
gRidge			80.1	9257
gKernel k_{IBS}			82.2	9257
gWLGR			82.0	9257
Ridge			76.1	40
Kernel k_{IBS}			78.1	40
WLGR			78.3	40