Published in final edited form as:

Genet Epidemiol. 2018 February ; 42(1): 64–79. doi:10.1002/gepi.22096.

Inference on Phenotype-Specific Effects of Genes using Multivariate Kernel Machine Regression

Arnab Maity,

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

Jing Zhao,

Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA

Patrick F. Sullivan, and

Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill NC, 27599, USA

Jung-Ying Tzeng

Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh NC 27695, USA; Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei 100, Taiwan; Department of Statistics, National Cheng-Kung University, Tainan City, 701, Taiwan

Abstract

We consider the problem of assessing the joint effect of a set of genetic markers on multiple, possibly correlated phenotypes of interest. We develop a kernel machine based multivariate regression framework, where the joint effect of the marker set on each of the phenotypes is modeled using pre-specified kernel functions with unknown variance components. Unlike most existing methods that mainly focus on the global association between the marker set and the phenotype set, we develop estimation and testing procedures to study phenotype-specific associations. Specifically, we develop an estimation method based on the penalized likelihood approach to estimate phenotype-specific effects and their corresponding standard errors while accounting for possible correlation among the phenotypes. We develop testing procedures for the association of the marker set with any subset of phenotypes using a score based variance components testing method. We assess the performance of our proposed methodology via a simulation study and demonstrate the utility of the proposed method using the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) data.

Keywords

Kernel machine; Multivariate regression; Mixed Models; Restricted Maximum Likelihood; Variance Components

Corresponding Author: Arnab Maity, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA, Phone: 919-515-1937, amaity@ncsu.edu.

The authors have no conflict of interest.

1 Introduction

Multivariate phenotypes are commonly collected in genetic association studies. The multiple phenotypes may include subclinical phenotypes, biochemical measurements, or relevant trait information contributing to disease risk. Understanding the relationship between genetic variants and multiple phenotypes can provide profound insights into disease etiology and underlying mechanisms. Joint analysis of the potentially correlated phenotypes can also enhance power to detect genes compared to individual phenotype analyses (Zhu and Zhang, 2009).

When the response variable is univariate, there is a vast amount of literature concerning gene-based (or equivalently, marker-set) association analysis. Among these approaches, kernel machine (KM) regression has gained popularity due to its model flexibility in capturing the true genetic effects that may be nonlinear or involve interactions, and its ability to accommodate a large number of genetic variables (see, e.g., Liu et al. (2007), Kwee et al. (2008), Wu et al. (2011b) among many others). KM based methods model the genetic effect via a pre-specified kernel function that aims to quantify the similarity between a pair of subjects based on their genetic covariates. By proper specification of the kernel functions, KM can capture nonlinear main effects and complicated interactions among multiple markers. It has been shown that KM regression is often powerful in modeling multi-dimensional genetic variables while retaining computational efficiency due to its connection to the linear mixed model framework (Liu et al., 2007).

Because of the various advantages of KM, several multivariate regressions based on KM (Maity et al. (2012), Yan et al. (2015), Xu et al. (2015), Broadaway et al. (2016), Sun et al. (2016), Zhan et al. (2017)) have been developed to carry out marker-set association analysis for multivariate phenotypes. These multivariate KM approaches have been demonstrated to have comparable or better power than the KM analysis of univariate phenotype. The power gain is more substantial when the within-subject correlation of the phenotypes becomes stronger or when there exist shared genetic effects among different phenotypes. Such a power gain occurs mainly due to two reasons. First, if one simply applies univariate testing methods separately for each available phenotype, it would require correction for multiple testing and thus resulting in power loss. Multivariate testing procedures typically perform a single test and thereby avoiding the need for multiple testing corrections. Second, the strategy of combining separate univariate tests ignore the correlation within the multiple phenotypes. In contrast, multivariate tests can borrow information across phenotypes to produce larger power (Maity et al., 2012), especially when the within-subject correlation is strong.

While being flexible and powerful in assessing marker-set effects on multiple phenotypes, the existing multivariate KM methods mainly focus on examining global association and do not provide phenotype-specific information. To maximize the utility of multivariate KM approaches, in this paper, we develop the estimation and testing procedure under the multivariate KM framework to study marker-set effects and allow the analyses of any subsets of responses to make inference on phenotype-specific effects. Specifically, we make two main contributions in this paper. First, we propose an estimation procedure for the model

components in the presence of multivariate responses via a penalized least squares approach. This method accounts for the correlation present among the response variables as well as the possibly complicated joint effect of multiple markers. We allow for the covariates to have different effect levels on various outcomes. Also, we show that the multi-response KM model has a direct connection with linear mixed effects modeling framework; such a connection allows us to construct the standard errors of the estimates of the model components and provides an effective way to choose the penalty parameters minimizing through maximizing the REML criterion. As shown in our numerical studies, the standard errors of our multivariate KM method tend to be smaller than those from univariate KM analysis, resulting in a more precise estimation. Second, we generalize the global testing procedure of Maity et al. (2012) to test for any particular subset of phenotypes. Specifically, based on the above estimation method, we extend the original testing procedure of Maity et al. (2012) to allow testing for any subset of phenotypes while still accounting for possible correlation among the phenotypes. The testing results can be used to determine the exact (set of) phenotype(s) that may be driving the global association.

The rest of this article is organized as follows. In Section 2, we introduce the multivariate kernel machine regression model and derive the estimation procedure of the model components. In section 3, we develop a score type test statistic to evaluate the genetic effects of multiple markers on any particular subset of the multiple phenotypes. In section 4, we conduct a simulation study to evaluate the performance of our methodology in finite-sample settings. In section 5, we apply the method to analyze the CATIE antibody data. Concluding remarks and discussions are presented in Section 6.

2 Multivariate Kernel Machine Regression

Suppose for each individual $i = 1, \dots, n$, we observe a $p \times 1$ vector of continuous outcomes $\mathbf{Y}_i = (Y_{1i}, \dots, Y_{pi})^T$, a vector of M genetic covariates $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{im})^T$, and a vector of other covariates/confounders $\mathbf{X}_i$. In our particular data example, $\mathbf{Y}_i$ is a vector containing responses to the three herpesviruses (HSV-1, HSV-2 and CMV), $\mathbf{X}_i$ is a covariate vector containing sex and age of the subject, and $\mathbf{Z}_i$ denotes a set of M SNPs with $Z_{im} \in \{0, 1, 2\}$, $m = 1, \dots, M$, recording the number of minor alleles. We consider a multivariate response model to relate the response vector $\mathbf{Y}_i$ to covariates $\mathbf{X}_i$ and genetic covariates $\mathbf{Z}_i$:

$$Y_{\ell i} = \mathbf{X}_i^T \boldsymbol{\beta}_\ell + h_\ell(\mathbf{Z}_i) + \varepsilon_{\ell i},$$

with $\ell = 1, \dots, p$ and $i = 1, \dots, n$, where $(\varepsilon_{1i}, \dots, \varepsilon_{pi})^T$ are random error vectors assumed to have a $\text{Normal}(0, \boldsymbol{\Sigma}_{\text{true}})$ distribution, $\boldsymbol{\beta}_\ell$ is an unknown coefficient vector denoting possible different effect of $\mathbf{X}_i$ on different responses, and $h_\ell(\cdot)$ indicates joint effect of the SNP set $\mathbf{Z}_i$ on the ℓ th response. Our goal is to estimate the unknown functions $h_\ell(\cdot)$ denoting the genetic effect on the responses.

A classical way to model the functions $h_\ell(\cdot)$ is to assume parametric models. Examples of such models include the main effects only model, where the effect of $\mathbf{Z}$ is modeled parametrically using a linear term, or the two-way interaction model where all possible two-

way interactions are modeled linearly in addition to the linear and quadratic main effects. However, such a parametric modeling strategy can often be very simplistic and may not capture the possible nonlinear nature of the genetic effect and the interactions among the genetic covariates. Also, accommodating more complex interaction terms in the model may result in the curse of dimensionality, even if the number of genetic covariates M is moderate. In this paper, we use the statistical kernel machine framework to model $h_\ell(\cdot)$ instead of a parametric specification. Such kernel based modeling allows one to specify $h_\ell(\cdot)$ parametrically or nonparametrically depending on the choice of kernels. Such an approach has been shown to be convenient and powerful especially for multi-dimensional genetic data. In particular, we assume that $h_\ell(\cdot)$ belongs to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\ell$ with positive definite kernel function $K_\ell(\cdot, \cdot)$. Thus, by Representer Theorem, $h_\ell(\cdot)$ can be written in a dual form $h_\ell(\mathbf{Z}) = \sum_{i=1}^n K_\ell(\mathbf{Z}_i, \mathbf{Z}) \alpha_{\ell i}$ for some constants $\alpha_{\ell 1}, \dots, \alpha_{\ell n}$.

Typically the true kernel function is unknown. In practice, there are many choices of kernel functions available. Two most commonly used kernels for SNP data are the d th order polynomial kernel, $K_\ell(\mathbf{z}_1, \mathbf{z}_2) = (1 + \mathbf{z}_1^T \mathbf{z}_2)^d$, and the identical by state (IBS) kernel, $K(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{m=1}^M (2 - |Z_{1m} - Z_{2m}|) / (2M)$. In many practical situations, the IBS kernel performs well and can capture possible nonlinear effects of $\mathbf{Z}$ (Wu et al. (2010)). We use the IBS kernel in our data analysis.

2.1 Estimation of Model Components

We adopt a penalized approach to estimate the model components. For a fixed error covariance matrix Σ , we propose to minimize the penalized least squares criterion

$$\mathcal{L} = \sum_{i=1}^n \sum_{\ell=1}^p \sum_{r=1}^p \sigma^{\ell r} \{Y_{\ell i} - \mathbf{X}_i^T \beta_\ell - h_\ell(\mathbf{Z}_i)\} \{Y_{ri} - \mathbf{X}_i^T \beta_r - h_r(\mathbf{Z}_i)\} + \sum_{\ell=1}^p \lambda_\ell \|h_\ell(\cdot)\|_{\mathcal{H}_\ell}^2,$$

where $\sigma^{\ell r}$ is the (ℓ, r) -th element of Σ^{-1} , $\|h_\ell(\cdot)\|_{\mathcal{H}_\ell}^2$ denotes the function norm of $h_\ell(\cdot)$ and λ_ℓ is the roughness penalty parameter associated with $h_\ell(\cdot)$. Such a penalization is required in the estimation procedure of the model parameters to avoid overfitting the data. The value of λ_ℓ controls the estimators of $h_\ell(\cdot)$: larger values of λ_ℓ produces flatter and near zero estimates of $h_\ell(\cdot)$ while smaller values of λ_ℓ results in estimator that interpolate the data closely (Liu et al., 2007). Thus the choice of the penalty parameters plays a crucial role in estimation. We shall discuss an effective way to choose these penalty parameters in section 2.2.

Now we adopt the dual representation of h_ℓ that is, $h_\ell(\mathbf{Z}) = \sum_{i=1}^n K_\ell(\mathbf{Z}_i, \mathbf{Z}) \alpha_{\ell i}$. With a slight abuse of notation, define the $n \times n$ kernel matrices $\mathbf{K}_\ell$ such that the (i, j) -th element of $\mathbf{K}_\ell$ is $K_\ell(\mathbf{Z}_i, \mathbf{Z}_j)$. It can be shown that the penalty terms reduces to

$$\|h_\ell(\cdot)\|_{\mathcal{H}_\ell}^2 = \alpha_\ell^T \mathbf{K}_\ell \alpha_\ell,$$

where $\mathbf{a}_l = (\mathbf{a}_{1l}, \dots, \mathbf{a}_{nl})^T$ (Liu et al., 2007). We emphasize again that $K(\cdot, \cdot)$ is the kernel function whereas $\mathbf{K}_l$ is the resulting kernel matrix.

Define the combined response vector $\mathbf{Y} = (Y_{11}, \dots, Y_{1n}, \dots, Y_{p1}, \dots, Y_{pn})^T$, and the combined covariance matrix $\mathbf{X} = \mathbf{I}_{p \times p} \otimes \mathbf{X}$, where $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$. Accordingly, define the combined parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)^T$, the vector of genetic effects $\mathbf{h} = [h_1(\mathbf{Z}_1), \dots, h_1(\mathbf{Z}_n), \dots, h_p(\mathbf{Z}_1), \dots, h_p(\mathbf{Z}_n)]^T$ and the error vector $\boldsymbol{\epsilon} = (\epsilon_{11}, \dots, \epsilon_{1n}, \dots, \epsilon_{p1}, \dots, \epsilon_{pn})^T$. Then we have $\mathbf{h} = \mathbb{K}\mathbf{a}$, where $\mathbb{K} = \text{blockdiag}(\mathbf{K}_1, \dots, \mathbf{K}_p)$ and $\mathbf{a} = (\mathbf{a}_1^T, \dots, \mathbf{a}_p^T)^T$. Substituting the dual form of $h(\cdot)$ and the corresponding penalty terms in the penalized least square criteria $\mathcal{L}$, we can write

$$\mathcal{L} = (-\boldsymbol{\beta} - \boldsymbol{\alpha})^T \tilde{\Sigma}^{-1} (-\boldsymbol{\beta} - \boldsymbol{\alpha}) + \boldsymbol{\alpha}^T \boldsymbol{\Lambda} \boldsymbol{\alpha},$$

where $\tilde{\Sigma} = \tilde{\Sigma} \otimes \mathbf{I}_n$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p) \otimes \mathbf{I}_n$. Minimizing $\mathcal{L}$ with respect to $\mathbf{a}$ and $\boldsymbol{\beta}$, we obtain estimates of $\boldsymbol{\beta}$ and $\mathbf{h}$ as

$$\hat{\boldsymbol{\beta}} = \{T(\boldsymbol{\Lambda}^{-1} + \tilde{\Sigma})^{-1}\}^{-1} T(\boldsymbol{\Lambda}^{-1} + \tilde{\Sigma})^{-1}; \quad \hat{\mathbf{h}} = (\tilde{\Sigma}^{-1} + \boldsymbol{\Lambda})^{-1} \tilde{\Sigma}^{-1} (-\hat{\boldsymbol{\beta}}). \quad (1)$$

If the kernel matrix $\mathbb{K}$ is nonsingular, then further simplification can be done to the expression of $\hat{\mathbf{h}}$; specifically,

$$\hat{\mathbf{h}} = \boldsymbol{\Lambda}^{-1} (\boldsymbol{\Lambda}^{-1} + \tilde{\Sigma})^{-1} (-\hat{\boldsymbol{\beta}}). \quad (2)$$

Estimates of the model components depend on the penalty parameters as well as the covariance matrix. Typically the true covariance matrix is unknown, and one often has to specify a working covariance matrix. We propose to fit a working independence model first, assuming a diagonal covariance matrix for errors and compute the residuals. Then we form an estimate of by using the sample covariance matrix of the residuals and use the resulting covariance estimate to re-fit the model for final estimation. Such a strategy of estimating true from a working independence model fit has been previously applied in the context of global hypothesis testing in multivariate kernel machine regression by other authors such as Zhan et al. (2017), Wu and Pankow (2016) and Maity et al. (2012) using the residuals under the null model. Note that the point estimators of the parameters are still valid while using such working covariance matrix in the second stage of estimation, however, no investigation has been done on the impact of such two-stage estimation on the variance-covariance matrices of the resulting estimators of the model components. In Section 2.2, we will propose a model based and a sandwich type estimator for the variance-covariance matrices of the estimators of $\boldsymbol{\beta}$ and $\mathbf{h}$.

Remark 1 To implement the estimation procedure in practice, we use the following reparameterization for computational stability. Assuming the kernel functions are positive semi-definite, that is $\mathbf{K}_\ell$, $\ell = 1, \dots, p$, are positive semi-definite matrices, we can always find a symmetric positive semi-definite matrix $\mathbf{K}_\ell^{1/2}$ such that $\mathbf{K}_\ell = \mathbf{K}_\ell^{1/2} \mathbf{K}_\ell^{1/2}$. Consequently, we can define $\mathbf{K}^{1/2} = \text{blockdiag}(\mathbf{K}_1^{1/2}, \dots, \mathbf{K}_p^{1/2})$ so that $\mathbf{K} = \mathbf{K}^{1/2} \mathbf{K}^{1/2}$. Thus we can reparameterize $\mathbf{h} = \mathbf{K} \mathbf{a} = \mathbf{K}^{1/2} \mathbf{a}_*$, where $\mathbf{a}_* = \mathbf{K}^{1/2} \mathbf{a}$. Similarly, the penalty term can be rewritten as $\mathbf{a}^T \mathbf{K} \mathbf{a} = \mathbf{a}_*^T \mathbf{A} \mathbf{a}_*$. Thus the penalized least squares criterion can be written as

$$\mathcal{L} = (-\beta - \mathbf{1}^T \alpha_*)^T \tilde{\Sigma}^{-1} (-\beta - \mathbf{1}^T \alpha_*) + \alpha_*^T \mathbf{A} \alpha_*.$$

Straightforward calculations show that the solutions of β and $\mathbf{h}$ from the reparameterized model are identical to (1). In practice, We compute $\mathbf{K}_\ell^{1/2}$ using the eigen decomposition. Specifically, suppose $\mathbf{K}_\ell = \mathbf{U}_\ell [\text{diag}(d_1, \dots, d_r, 0, \dots, 0)] \mathbf{U}_\ell^T$, where r is the rank of $\mathbf{K}_\ell$ and $d_1, \dots, d_r$ are the positive eigenvalues of $\mathbf{K}_\ell$. Then we define $\mathbf{K}_\ell^{1/2} = \mathbf{U}_\ell [\text{diag}(d_1^{1/2}, \dots, d_r^{1/2}, 0, \dots, 0)] \mathbf{U}_\ell^T$. In our numerical implementation, we only take the eigenvalues that are greater than a pre-set threshold (we use 10^{-10}) for numerical stability.

2.2 Standard Error Estimation and Penalty Parameter Selection

If we assume that the kernel matrices $\mathbf{K}_\ell$ are nonsingular, then the solution given in (2) can be viewed as the solution of the minimization problem

$$\min_{\beta, \mathbf{h}} \left[\{-\beta - \mathbf{h}\}^T \tilde{\Sigma}^{-1} \{-\beta - \mathbf{h}\} + \mathbf{h}^T \mathbf{A} \mathbf{h} \right].$$

In other words, the solution of $(\beta, \mathbf{h})$ solves the normal equations

$$\begin{bmatrix} \mathbf{1}^T \tilde{\Sigma}^{-1} & \mathbf{1}^T \tilde{\Sigma}^{-1} \\ \tilde{\Sigma}^{-1} & \tilde{\Sigma}^{-1} + \mathbf{A} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{h} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \tilde{\Sigma}^{-1} \\ \tilde{\Sigma}^{-1} \end{bmatrix}.$$

From a computational point of view, these normal equations are identical to those of the mixed effects model

$$\mathbf{Y} = \beta \mathbf{1} + \mathbf{h} + \varepsilon, \quad (3)$$

where $\mathbf{h} \sim \text{Normal}(0, \mathbf{K} \mathbf{A}^{-1})$ and $\varepsilon \sim \text{Normal}(0, \tilde{\Sigma})$. Hence one can think of the parameters λ_ℓ^{-1} , $\ell = 1, \dots, p$ as the variance components. We can view $\hat{\beta}$ and $\hat{\mathbf{h}}$ as the best linear

unbiased estimator (BLUE) and best linear unbiased predictor (BLUP) of β and h , respectively. Letting $V = K\Lambda^{-1} + \tilde{\Sigma}$, the solution of $\hat{\beta}$ and $\hat{h}$ in (2) can be rewritten as

$$\hat{\beta} = (X^T V^{-1})^{-1} X^T V^{-1} y, \quad \hat{h} = \Lambda^{-1} P y, \quad (4)$$

where $P = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}$.

To estimate the variance-covariance matrices of the model components, we use the linear mixed model formulation in (3) treating the true h as a random effect, and opt to use the so-called Bayesian covariances (Liu et al., 2007). The covariance of $\hat{\beta}$ can be computed as $\text{cov}(\hat{\beta}) = (X^T V^{-1} X)^{-1}$. However, such an estimator assumes that the first stage estimator of β is indeed the true error covariance, and does not take into account the variability due to the second stage estimation. Alternatively, one can compute a sandwich type covariance for β as

$$\text{cov}(\hat{\beta}) = (X^T V^{-1})^{-1} X^T V^{-1} \widehat{\text{cov}}(y) V^{-1} (X^T V^{-1})^{-1},$$

where $\widehat{\text{cov}}(y)$ can be constructed as $S \otimes I_n + K\Lambda^{-1}$ with S denoting the sample covariance of the residuals from the final model fit. Similarly, a model based estimator of the variance of the BLUP of h can be formed as $\text{cov}(\hat{h} - h)$ can be formed as $K\Lambda^{-1} - K\Lambda^{-1} P \Lambda^{-1} K$, and a sandwich type estimator is given by

$$\text{cov}(\hat{h} - h) = \Lambda^{-1} + \Lambda^{-1} P (S \otimes I_n + \Lambda^{-1}) P \Lambda^{-1} - 2\Lambda^{-1} P \Lambda^{-1}.$$

Note that if one takes S to be the estimated residual covariance from the initial model fit, the sandwich estimators reduce to the model based estimators due to the fact that $(S \otimes I_n + K\Lambda^{-1})$ becomes V and $PVP = P$. In practice, we opt to use the sandwich estimators for robustness.

Remark 2 In establishing the connection between the penalized least squares model and the mixed model, we have assumed that K is nonsingular. It is interesting to note that the corresponding mixed model in (3) is still valid even if K is singular but positive semi-definite. One can still posit that $h \sim \text{Normal}(0, K\Lambda^{-1})$ with the understanding that h now follows a singular normal distribution. The mixed model normal equation for such cases (where the variance-covariance matrix of the random effects may be singular) are discussed in Christensen (2011), see their p. 230. Specific to our case, the normal equations are

$$\begin{bmatrix} T\tilde{\Sigma}^{-1} & T\tilde{\Sigma}^{-1}\Lambda^{-1} \\ \tilde{\Sigma}^{-1} & I + \tilde{\Sigma}^{-1}\Lambda^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \zeta \end{bmatrix} = \begin{bmatrix} T\tilde{\Sigma}^{-1}y \\ \tilde{\Sigma}^{-1}y \end{bmatrix},$$

where a BLUP of $\mathbf{h}$ is given by $\hat{\mathbf{h}} = \mathbb{K}\mathbf{\Lambda}^{-1}\hat{\boldsymbol{\zeta}}$. Simple algebra shows that $\hat{\boldsymbol{\zeta}} = \mathbf{V}^{-1}(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}) = \mathbf{P}\mathbf{Y}$, and thus $\hat{\mathbf{h}}$ in this case is identical to (4). Thus the expression of $\hat{\mathbf{h}}$ in (4) is still valid regardless of whether $\mathbb{K}$ is nonsingular or not.

The above-mentioned connection to the linear mixed effects model also provides us with an effective solution for choosing the parameters λ_ℓ . Treating λ_ℓ^{-1} as variance components in mixed effects model from (3), the restricted maximum likelihood (REML) criteria under (3) can be written as

$$\mathcal{L}_{\text{REML}} = -\log|\mathbf{V}|/2 - \log|\mathbf{V}^{-1}|/2 - (-\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1}(-\hat{\boldsymbol{\beta}})/2.$$

One can now choose the penalty parameters by maximizing the REML criteria. In our simulation study and data analysis, we found that the REML estimation of penalty parameters produced good results in all the cases we considered. We have implemented this procedure using the gam function in the R package mgcv, which automates the optimization process using Newton's method (Wood, 2011) by default. We note that computation time is mainly taken by the REML optimization procedure due to the optimization problem being multi-dimensional, and the computation time increases with sample size. For larger sample sizes, we recommend the bam function within mgcv which is designed to fit such models for very large datasets.

Remark 3 Note that for the special case $p = 1$, with $\tilde{\tau} = \sigma^2 \mathbf{I}_n$ and only one variance parameter $\tau = \lambda^{-1}$, our results translate to

$$\begin{aligned}\hat{\mathbf{h}} &= \tau \mathbf{K}(\tau \mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \hat{\boldsymbol{\beta}} &= \{\mathbf{X}(\tau \mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{X}^T\}^{-1} \mathbf{X}(\tau \mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Y}.\end{aligned}$$

These expressions correspond to univariate kernel machine regression (Liu et al. (2007)). In the special case with $\tilde{\tau}$ being a diagonal matrix, our procedure is equivalent to performing p separate regressions.

2.3 Prediction

Suppose now we have a new set of SNPs that is not included in the given sample, $\mathbf{Z}_{\text{new}}$, and we want to predict the joint genetic effect, $\mathbf{h}(\mathbf{Z}_{\text{new}})$, based on the estimation. Let

$\mathbf{K}_{\text{new}} = \text{diag}(\mathbf{K}_{\text{new},1}^T, \dots, \mathbf{K}_{\text{new},p}^T)$, where $\mathbf{K}_{\text{new},\ell} = (\mathbf{K}(\mathbf{Z}_{\text{new}}, \mathbf{Z}_1), \dots, \mathbf{K}(\mathbf{Z}_{\text{new}}, \mathbf{Z}_n))^T$. Then we have

$$\begin{aligned}\hat{\mathbf{h}}(\mathbf{Z}_{\text{new}}) &= \mathbf{K}_{\text{new}} \mathbf{\Lambda}^{-1} \mathbf{P}, \\ \text{COV}\{\hat{\mathbf{h}}(\mathbf{Z}_{\text{new}}) - \mathbf{h}(\mathbf{Z}_{\text{new}})\} &= \mathbf{K}_{\text{new}} \mathbf{\Lambda}^{-1} \mathbf{P}(\mathbf{S} \otimes \mathbf{I}_n + \mathbf{\Lambda}^{-1}) \mathbf{P} \mathbf{\Lambda}^{-1} \mathbf{K}_{\text{new}}^T + \mathbf{D} - 2\mathbf{K}_{\text{new}} \mathbf{\Lambda}^{-1} \mathbf{P} \mathbf{\Lambda}^{-1} \mathbf{K}_{\text{new}}^T,\end{aligned}$$

where $D = \text{diag}(\lambda_1^{-1} K_1(\mathbf{Z}_{\text{new}}, \mathbf{Z}_{\text{new}}), \dots, \lambda_p^{-1} K_p(\mathbf{Z}_{\text{new}}, \mathbf{Z}_{\text{new}}))$. In the formula above, $\mathbf{\Lambda}$ and $\mathbf{P}$ are computed from the sample following the procedure in Section 2.2 and $\mathbf{K}_{\text{new}}$ is the kernel matrix constructed based on the genotype of the new set of SNPs and genotypes of the SNP sets in the sample.

3 Hypothesis Testing

In this section, we describe a testing procedure in the multivariate kernel machine regression model. Recently Maity et al. (2012) developed a global testing procedure to test the null hypothesis $H_0 : h_1(\cdot) = \dots = h_p(\cdot) = 0$. While such a test provides information on whether the genetic covariates are associated to *any* of the response variables of interest, this test cannot identify the exact set of responses that drive the association. In this section we derive a testing procedure where one can test a particular set of functions, e.g., $H_0 : h_1(\cdot) = \dots = h_{p_0}(\cdot) = 0$ where $p_0 \leq p$ corresponds to the number of functions that are being tested for.

For simplicity, we start with the case $p_0 = 1$, that is, testing only a single function. We then generalize this procedure for any arbitrary value of p_0 . In the case of $p_0 = 1$, we make use of the connection to the linear mixed effects models as described in Section 2.2. In this framework, we only need to test

$$H_0^{(1)} : \tau_1 = 0.$$

We use the restricted maximum likelihood (REML) estimation procedure (see for example, Liu et al. (2007)) to derive a score test. The score function of τ_1 derived from $\mathcal{L}_{REML}$, evaluated at H_0 is

$$S_{\tau_1, n} = (-\hat{\beta})^T \mathbf{V}_0^{-1} \mathbf{K}_1 \mathbf{V}_0^{-1} (-\hat{\beta}) / 2 - \text{trace}(\mathbf{K}_1 \mathbf{P}_0),$$

where $\hat{\beta}$ is estimated under null, $\mathbf{V}_0$ denotes $\mathbf{V}$ evaluated under null, and $\mathbf{P}_0$ has the same definition of $\mathbf{P}$ with $\mathbf{V}$ replaced with $\mathbf{V}_0$. To test $H_0 : \tau_1 = 0$, we propose to use the score type test statistic

$$T_{\tau_1, n} = (-\hat{\beta})^T \mathbf{V}_0^{-1*} \mathbf{V}_0^{-1} (-\hat{\beta}),$$

where $\mathbf{V}_0^{-1*}$ is a block diagonal matrix with $\mathbf{K}_1$ as the first block and zero otherwise.

To derive the null distribution of the test statistic, we first observe that $T_{\tau_1, n}$ is a quadratic form in $\mathbf{r} = \mathbf{V}_0^{-1} (\mathbf{Y} - \mathbf{X}^T \hat{\beta})$. Note that $\mathbf{r}$, under $H_0^{(1)}$, follows a Gaussian distribution with mean zero and covariance matrix $\mathbf{P}_0[\text{cov}(\mathbf{Y})]\mathbf{P}_0$. If we assume $\text{cov}(\mathbf{Y}) = \mathbf{V}_0$, that is, the initial estimator of $\mathbf{V}$ from the first stage is indeed the truth, the covariance reduces to $\mathbf{P}_0$. In practice however, this may not be the case as such an assumption does not take into account

the variability due to the second stage estimation. Instead, we propose to form $\text{cov}(\mathbf{Y})$ from the second stage residuals as, $\widehat{\text{cov}}() = \mathbf{S} \otimes \mathbf{I}_n + \mathbf{\Lambda}^{-1}$, where $\mathbf{S}$ is the empirical covariance matrix of the residuals from the final model fit under H_0 .

It is readily seen that the null distribution can be written as a weighted linear combination of independent χ_1^2 random variables, where the weights are the eigenvalues of $[\mathbf{P}_0 \widehat{\text{cov}}() \mathbf{P}_0]^{1/2} \mathbf{I}_1 [\mathbf{P}_0 \widehat{\text{cov}}() \mathbf{P}_0]^{1/2}$. The p-value of the test can be computed using the Davies approach (Davies, 1980). We use the R package CompQuadForm for this purpose.

Now we take up the problem of testing p_0 components, that is, $H_0 : h_1(\cdot) = \dots = h_{p_0}(\cdot) \equiv 0$, where p_0 corresponds to the number of functions that are being tested for. Using similar arguments as in Maity et al. (2012), we note that it is sufficient to test for

$$H_0^{(2)} : \tau_1 + \dots + \tau_{p_0} = 0,$$

as each τ_ℓ is a non-negative parameter. Then we propose to use the score type test statistic

$$T_{2,n} = \sum_{\ell=1}^{p_0} T_{\tau_\ell, n} = (-\hat{\beta})^T \mathbf{V}_0^{-1} \left(\sum_{\ell=1}^{p_0} \mathbf{I}_\ell^* \right) \mathbf{V}_0^{-1} (-\hat{\beta}),$$

where $\mathbf{V}_0$ denotes $\mathbf{V}$ evaluated under null and $\mathbf{I}_\ell^*$ is a block diagonal matrix with $\mathbf{K}_\ell$ as the ℓ^{th} block and zero otherwise. Note that for $p_0 = 1$, $T_{2,n}$ reduces to the test statistic for a single function. We can compute the null distribution of $T_{2,n}$ easily following the same procedure as described for testing for single function where we replace $\mathbf{I}_1^*$ with $\sum_{\ell=1}^{p_0} \mathbf{I}_\ell^*$.

4 Simulation Study

We conduct a simulation study to investigate the finite sample performance of our proposed estimation and testing procedures. Recall that we denote the number of individuals by n , the number of SNPs by M , the number of phenotypes observed for each subject by p , and the number of $h(\cdot)$ functions to test by p_0 . We generate data from the model, for $i = 1, \dots, n$ and $\ell = 1, \dots, p$,

$$Y_{\ell i} = \mathbf{X}_i^T \boldsymbol{\beta}_\ell + h_\ell(\mathbf{Z}_i) + \varepsilon_{\ell i},$$

with $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iM})^T$, $\mathbf{X}_i = (X_{i1}, X_{i2})^T$, $\boldsymbol{\beta}_\ell = (\beta_{\ell 1}, \beta_{\ell 2})^T$ and $(\varepsilon_{1i}, \dots, \varepsilon_{pi})^T = \text{Normal}(0, \text{true})$. In accordance to the real data, we set the dimension of responses to $p = 3$. We allow $\mathbf{X}_i$ to be the same for all the items in the outcome and generate $(X_{i1}, X_{i2})^T$ from $\text{Normal}(0, \mathbf{I}_2)$ distribution. We set the true values of $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \dots, \beta_{31}, \beta_{32})^T = (0.7, 1, 0.7, 1, 0.7, 1)^T$.

We consider the following three simulation scenarios: (1) $n = 100$, $M = 9$, (2) $n = 200$, $M = 9$ and (3) $n = 200$, $M = 30$. For scenarios (1) and (2), we generate the genetic variable $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{iM})$ based on the SLC17A1 gene (containing 9 SNPs) in the CATIE antibody study, the frequency distribution is shown in Table 1. We only take genotypes with 7 (1% of sample size in data example, $n = 690$) occurrences in the real data for this simulation study. In scenario (3), we add 21 nuisance SNPs, each taking value 0, 1 or 2 with probability 0.3, 0.5 and 0.2, respectively.

For both estimation and testing simulations, we will use three choices of Σ_{true} :

$$\Sigma_{\text{true},1} = \begin{pmatrix} 0.95 & 0.07 & 0.23 \\ 0.07 & 0.86 & 0.24 \\ 0.23 & 0.24 & 0.89 \end{pmatrix}, \quad \Sigma_{\text{true},2} = \begin{pmatrix} 0.95 & 0 & 0 \\ 0 & 0.86 & 0 \\ 0 & 0 & 0.89 \end{pmatrix}, \quad \Sigma_{\text{true},3} = \begin{pmatrix} 0.95 & 0.57 & 0.43 \\ 0.57 & 0.86 & 0.24 \\ 0.43 & 0.24 & 0.89 \end{pmatrix}$$

The first covariance matrix $\Sigma_{\text{true},1}$ is the estimated covariance matrix from the real data under the assumption that there is no SNP effect. The second choice $\Sigma_{\text{true},2}$ is a matrix that is expanded by the diagonal elements of $\Sigma_{\text{true},1}$ and will be used to evaluate the performance of our procedure where the outcomes are independent. The third choice $\Sigma_{\text{true},3}$ introduces high correlation between Y_1 and (Y_2, Y_3) and is used to demonstrate the effect of correlation on the performance of our procedure.

In the numerical study presented in the following sections, we used the IBS kernel. Similar results for the quadratic kernel are presented in the Supplementary materials.

4.1 Estimation

In the simulation to evaluate our estimation procedure, we use $h_1(\cdot) = z_1 + z_2 + z_3 + z_1 z_4 z_5 - z_6/3 - z_7 z_8/2 + (1 - z_9)$, $h_2(\cdot) = z_1 z_2 + z_5(z_7 + z_3) + z_2 z_9$,

$h_3(\cdot) = z_2 + z_3 + (1 - z_6)^2/3 - z_5^2/2 + z_8^2 - z_1^3/3$ for h for scenarios (1)-(3). These specific choices for the functions are taken so that we can investigate the accuracy of the estimation procedure for different types of effect patterns. For example, $h_1(\cdot)$ includes linear main effects, and two-way and three-way interaction terms; $h_2(\cdot)$ includes purely two-way interaction terms with no main effects, and $h_3(\cdot)$ includes linear as well quadratic terms. Thus this setting allows us to investigate the performance of the estimation procedure beyond simple linearity. For each of the scenarios, we generate 1000 data sets.

First, we evaluate the accuracy of the parameter estimates of β . Specifically, we compute the empirical bias (Bias), empirical standard deviation (SD), estimated standard errors averaged over 1000 simulations using sandwich (SE.S) and model based (SE.M) estimator, and corresponding coverage probabilities (C.S and CM, respectively) of 95% confidence intervals. The results are reported in Table 2. It is evident that the parameter estimates are unbiased. Also, the model based and sandwich estimators of the standard errors are close in all the settings we consider, and both provide near nominal coverage.

To measure the accuracy of the estimation of $h_\ell(\cdot)$, we report the following evaluating criteria ($\ell = 1, 2, 3$):

$$M_{1\ell} = n^{-1} \sum_{i=1}^n \{\hat{h}_\ell(Z_i) - h_\ell(Z_i)\}^2$$

$$M_{2\ell} = n^{-1} \sum_{i=1}^n |\hat{h}_\ell(Z_i) - h_\ell(Z_i)|.$$

To visualize the goodness of fit, we note that in our simulation settings, there are only 19 unique values (corresponding to 19 different genotype combinations in Table 1) for each $h_k(\cdot)$. Thus for each true value of $h_k(\cdot)$, we collect the estimated values across the 1000 simulations and create side-by-side boxplots. We overlay a fitted linear regression line with median of estimated $h(\cdot)$ values of each group as response and the corresponding true $h(\cdot)$ values as predictors. The results are shown Figures 1 - 3. It is evident that the estimated values are close to the true values on average for all the three scenarios.

To demonstrate the advantage of applying the proposed multivariate regression, we generate 100 data sets for a combination of $(n = 100, m = 9)$ under different settings of correlation between responses. For each generated data set, we compute the ratio average prediction error of $h(\cdot)$ for fitting three separate univariate KM regression models versus fitting the joint multivariate regression model. We consider the cases when the true error covariance is $\Sigma = I_3$ or $\Sigma_{\text{true},3}$, or

$$\Sigma = \Sigma_{\text{true},4} = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.8 \\ 0.6 & 0.8 & 1 \end{pmatrix}.$$

For the case with $\Sigma = I_3$, recall that we expect the multivariate procedure to be essentially equivalent to fitting separate univariate KM regression models. In this case, the ratio between the average prediction errors for $h_k(\cdot)$ were 0.999, 0.997 and 0.996, for $\ell = 1, 2$ and 3, respectively. Thus when the errors are indeed independent and thus an univariate KM model is sufficient, fitting the multivariate model does not result in a loss of estimation accuracy. For the case with $\Sigma_{\text{true},3}$, the ratio between the average prediction errors were 1.059, 1.013 and 1.030; for $\Sigma = \Sigma_{\text{true},4}$, they were 1.093, 1.105 and 1.099, for $\ell = 1, 2$ and 3, respectively. We can see that for stronger within-subject correlation, the estimates from univariate models produce 5% - 10% larger prediction errors compared to those from multivariate model.

In summary, fitting the multivariate model even when the errors are independent does not result in any loss of estimation accuracy. However, when the errors have stronger correlation, the multivariate model can procure more accurate predictions compared to fitting separate univariate models.

4.2 Testing

To evaluate our testing procedure, we consider the following choices of the functions:

- A.** $h_1 = \delta(z_1 + z_2 + z_3 + z_1 z_4 z_5 - z_6/3 - z_7 z_8/2 + (1 - z_9))$, $h_2 = h_3 = 0$; we only test for $H_0 : h_1(\cdot) = 0$ while $h_2(\cdot)$ and $h_3(\cdot)$ need to be estimated as nuisance parameters.
- B.** $h_1 = \delta(z_1 + z_2 + z_3 + z_1 z_4 z_5 - z_6/3 - z_7 z_8/2 + (1 - z_9))$, $h_2 = \delta(z_1 z_2 + z_5(z_7 + z_3) + z_2 z_9)$, h_3 is as in the estimation section. Here we test for $H_0 : h(\cdot) = h_2(\cdot) = 0$ while $h_3(\cdot)$ needs to be estimated as nuisance parameter.
- C.** The functions are as in scenario (A) but we test $H_0 : h_1(\cdot) = h_2(\cdot) = 0$.
- D.** $h_1(\cdot)$ as in scenario (B), $h_2(\cdot) = -h_1(\cdot)$, and $h_3(\cdot)$ as in (B). Here we test for $H_0 : h_1(\cdot) = h_2(\cdot) = 0$.

In each of these scenarios, we take $\delta = 0, 0.04, 0.08, 0.12$. Here, $\delta = 0$ corresponds to the Type I error of our procedure, and nonzero values of δ correspond to power. For each combination of (n, m) described in scenarios (1)-(3), we investigate the performance of our test with compare our test with the standard univariate kernel machine test as implemented in the R package SKAT (Wu et al., 2011a). When using SKAT, we perform separate tests for each phenotype to be tested and then correct for multiple testing using a Bonferroni correction. Notice that Scenario (A) favors SKAT as only one test is being performed here while scenarios (B) – (D) is suited for the multivariate testing procedure. Significance level $\alpha = 0.05$ is used in all the testing procedures.

The proportions of rejections using the IBS kernel function are shown in Figure 4-7 for scenarios (A) - (D), respectively. From the figures, it is evident that the Type I error rates of both our proposed method as well as the Bonferroni corrected univariate SKAT method are controlled near the nominal level for all scenarios.

For Scenario (A), where only $h_1(\cdot)$ is associated with the corresponding phenotype, both the multivariate test and SKAT test have comparable power for small within-subject correlations, although in presence of a stronger correlation ($\rho_{\text{true},3}$) the multivariate method produces larger power than SKAT. This is because SKAT only uses the data on one phenotype to test for the corresponding genetic effect. However, the multivariate testing method can borrow information across phenotypes due to accounting for the within-subject covariance. Thus, our method, despite estimating two nuisance components, does not show any loss of power. This conclusion remains valid for all the three choices of (n, M) .

Similar trend are observed in scenarios (B) – (D) as well. In Scenario (B), where we test for two components simultaneously, our proposed method is either similar or more powerful than SKAT with Bonferroni correction. In Scenario (C), where we test for two components simultaneously when in fact only one is significant, our proposed method shows similar power to SKAT when the phenotypes are independent or have small correlation among themselves ($\rho_{\text{true},2}$ and $\rho_{\text{true},1}$). However, in the presence of stronger correlation ($\rho_{\text{true},3}$), the multivariate test shows a significant power gain over SKAT. This gain is because while $h_2(\cdot) = 0$ in this case, Y_2 still retains some information about Y_1 due to the high correlation among them. Thus the multivariate test uses this information to produce better power where SKAT cannot. Furthermore, SKAT also suffers from the need of multiple testing correction. Such a phenomenon has been previously observed by Maity et al. (2012) in the context of global testing, that is, testing for $H_0 : h_1(\cdot) = h_2(\cdot) = h_3(\cdot) = 0$ when only $h_1(\cdot)$ has nonzero

effect (see their scenario (a): sparse effect). Finally, in Scenario (D), where $h_2 = -h_1$, similar conclusions can be drawn as in the previous scenarios.

We have repeated these simulations for the quadratic kernel function as well, and arrived at similar conclusions as those present above. The results are presented in the Supplementary materials.

5 Data Analysis

We applied the proposed methodology to the CATIE antibody study (Lieberman et al. (2005); Sullivan et al. (2008); Yolken et al. (2011)) data as analyzed in Maity et al. (2012). The data set consists of 690 CATIE-GWAS participants with schizophrenia (Stroup et al. (2003); Keefe et al. (2006); Sullivan et al. (2008)). Each sample has the information of SNPs in the major histocompatibility complex (MHC) region and the IgG class antibodies to Herpes Simplex Virus type 1 (HSV-1), Herpes Simplex Virus type 2 (HSV-2), and to Cytomegalovirus (CMV) (Yolken et al. (2011); Sullivan et al. (2008)). Here we focus our analysis on *SLC17A1*, a gene consisting of 9 SNPs with a global association (p-value 0.006) with the antibodies of the three herpesviruses. The goal of our analysis is to understand antibody-specific effects for *SLC17A1* using the proposed procedure.

For $i = 1, \dots, 690$ and $\ell = 1, 2, 3$, we consider the following model,

$$Y_{\ell i} = \mathbf{X}_i^T \beta_{\ell} + h_{\ell}(\mathbf{Z}_i) + \varepsilon_{\ell i},$$

with $Y_{1i} = \text{HSV1}_i$, $Y_{2i} = \text{HSV2}_i$, $Y_{3i} = \text{CMV}_i$, $\mathbf{Z}_i = (\text{SNP}_{i1}, \dots, \text{SNP}_{i9})^T$, $\mathbf{X}_i = (\text{sex}_i, \text{age}_i)^T$, $\beta_{\ell} = (\beta_{\ell 1}, \beta_{\ell 2})^T$ and $(\varepsilon_{i1}, \dots, \varepsilon_{i3})^T = \text{Normal}(0, \text{true})$. We fit this model and estimate $\beta_1, \beta_2, h_1, h_2$ and h_3 using the methodology developed in Section 2, with IBS kernel applied when estimating the SNP-set effect. We construct the point-wise confidence interval for h_1, h_2 and h_3 using the sandwich estimator.

The estimated values of h_1, h_2 and h_3 along with the simultaneous 95% confidence intervals, are displayed in Figure 8 (left panel). We do not observe any significant genetic effect on HSV-1 or CMV since the estimated 95% confidence intervals contain 0 for the genotype combinations in the dataset. For two genotype combinations, estimated intervals of h_2 are below 0, which indicates a possible negative effect of the SNP set on HSV-2. We also report the point estimates of h_1, h_2 and h_3 as well as their 95% point-wise confidence intervals when separate univariate models are fitted (right panel of Figure 8). Compared to the multivariate method, univariate models produce similar points estimates but slightly wider confidence intervals. This is not surprising due to low correlations between the responses ($\text{true}_{1,1}$ in section 4).

We next test for the genetic effect of these SNPs on the antibody levels to the three herpesviruses, first testing for single functions and, then testing for two functions simultaneously. The results are shown in Table 4. Based on the multivariate testing results (first row of 4), it is clear that the SNP set does not have any significant association with

HSV1 but produces much smaller p-values for HSV2 and CMV. After Bonferroni correction (since three functions are being tested), only HSV2 shows a significant association. The same conclusion can be reached using the univariate SKAT results as well. This finding matches with our findings in the estimation section where we observed significant negative effects for two genotype combinations for HSV2 (left column, middle panel in Figure 8).

Examining the results from simultaneous tests for two functions, we see that, after Bonferroni correction for three tests, both (HSV1, HSV2) and (HSV2, CMV) tests show significant association with the SNP set. Together with the single function test results, it seems that HSV2 is indeed the driving factor behind the association.

6 Discussion

In this article, we propose a multivariate kernel machine regression framework to study the joint effect of a set of genetic markers on multiple phenotypes while adjusting for the demographic and clinical covariates. We apply kernel machine framework to model the joint effect of genetic variables to capture the possibly nonlinear genetic main effects and their complicated interactions, and we use multivariate regression to incorporate information of correlation between phenotypes. We derive the estimation procedure to estimate the genetic effects as well as the demographic/clinic effects on each of the phenotype. Also, we derive a score-based testing method for testing the genetic effects. To evaluate the performance of our proposed method, we conduct simulations for both estimation and testing procedure. We also compare our method with univariate kernel machine regression procedure for both estimation and SKAT with Bonferroni correction for hypothesis testing. The estimation results suggest that our method accounts for the underlying correlation among the responses to produce estimators of $h(\cdot)$ functions with smaller variance than those from separate univariate KM regressions, especially when the within-subject correlation is high. Moreover, our method has better power than univariate SKAT procedure when testing, especially when the correlations between responses increase. Therefore, applying proposed multivariate framework will lose little and produce more beneficial results.

Another advantage of the multivariate kernel machine framework is its flexibility: one can choose different kernel functions to accommodate different assumptions of the underlying model. However, the choice of a kernel function can impact the results of whether a set of genetic markers has significant effects on particular phenotypes or not. Also, genetic markers can have different effects on each phenotype, which may require different choices of kernel function to model them precisely. Wu et al. (2013) has proposed strategies for kernel machine testing in the presence of multiple candidate kernels. Although their approaches guarantee improved power over poor choices of kernels and only slight differences in power compared with the optimal kernel, their methods are developed only for univariate responses. Generalizing their methods to multivariate framework remains an interesting problem and a topic for future research.

As pointed out by an anonymous reviewer, another important issue in multivariate kernel machine testing is to whether the responses should be standardized to have the same scale or not. In our implementation, we do not standardize our responses. However, we did not find a

general consensus on whether to always standardize the responses or not. Many authors use the raw phenotypes as they are (see e.g., Wu and Pankow (2016), Broadaway et al. (2016), Zhan et al. (2017), and Maity et al. (2012)), while other authors advocate standardization (e.g., Sun et al. (2016), see their Simulation studies section). In general, the impact of such standardization on estimation and testing needs further investigation.

In our proposed method, we only consider continuous responses. In practice, traits can be either quantitative or qualitative, or a mixture of both. In such cases, the multivariate normal assumption in our model would not be met, and thus the following inference procedure would fail. Limited methods are available when multivariate phenotypes contain a mixture of quantitative and qualitative, see, for example, Davenport et al. (2017) for a global association test for multivariate binary responses. Therefore, an extension to qualitative responses is a direction that needs further investigation.

Finally, our method is intended as a gene-based or set-based approach, rather than a whole genome collapsing approach. While theoretically, one could collapse a large number of SNPs across the whole genome, and the algorithm still remains operational since the kernel matrix will still be $n \times n$ regardless of the number of SNPs, doing so would result in a loss of resolution. Instead, this approach is intended to analyze one SNP-set at a time. One could take a window-based approach for a genome-wide analysis.

7 Supplementary materials and software

We have added additional simulation results using the quadratic kernel for estimation and testing in supplementary material. The proposed methods were implemented in R, and the programs are available at the first author's web page under the software section: <http://www4.stat.ncsu.edu/~maity/software/>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank two anonymous referees for their insightful comments that helped to genuinely improve the overall quality of the article, and Dr. Robert Yolken at Johns Hopkins University for providing CATIE antibody data. This work was supported by National Institutes of Health grants R00 ES017744 (to A.M.) and P01 CA142538 (to J.Y.T.).

References

- Broadaway K, Cutler D, Duncan R, Moore J, Ware E, Jhun M, Bielak L, Zhao W, Smith J, Peyser P, Kardia S, Ghosh D, Epstein M. A statistical approach for testing cross-phenotype effects of rare variants. *American Journal of Human Genetics*. 2016; 98(3):525–540. [PubMed: 26942286]
- Christensen, R. *Plane Answers to Complex Questions*. Springer-Verlag; New York: 2011.
- Davenport CA, Maity A, Sullivan PF, Tzeng JY. A powerful test for snp effects on multivariate binary outcomes using kernel machine regression. *Statistics in Biosciences*. 2017
- Davies RB. Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1980; 3(29):323–333.

- Keefe RSE, Bilder RM, Harvey PD, et al. Baseline neurocognitive deficits in the catie schizophrenia trial. *Neuropsychopharmacology*. 2006; 31:2033–2046. [PubMed: 16641947]
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. 2008
- Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keefe RS, Davis SM, Davis CE, Lebowitz BD, Severe J, Hsiao JK. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine*. 2005; 353(12):1209–1223. [PubMed: 16172203]
- Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*. 2007; 63(4):1079–1088. [PubMed: 18078480]
- Maity A, Sullivan PF, Tzeng Ji. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol*. 2012; 36:686–695. [PubMed: 22899176]
- Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, McGee MF, Simpson GM, Stevens MC, Lieberman JA. The national institute of mental health clinical antipsychotic trials of intervention effectiveness (catie) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin*. 2003; 29(1):15–31. [PubMed: 12908658]
- Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL. Genomewide association for schizophrenia in the catie study: results of stage 1. *Mol Psychiatry*. 2008; 13(6):570–584. [PubMed: 18347602]
- Sun J, Ouakacha K, V F, HF Z, BJ R, A C, CM G. UK10K Consortium. A small-sample multivariate kernel machine test for microbiome association studies. *Eur J Hum Genet*. 2016; 24(9):1344–1351. [PubMed: 26860061]
- Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2011; 73(1):3–36.
- Wu B, Pankow JS. Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology*. 2016; 40:91–100. [PubMed: 26782911]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (skat). *American Journal of Human Genetics*. 2011a; 89:82–93. [PubMed: 21737059]
- Wu MC, 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*. 2011b; 89(1):82–93. [PubMed: 21737059]
- Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Molldrem JJ, Armistead PM. Kernel machine snp-set testing under multiple candidate kernels. *Genet Epidemiol*. 2013; 37(3):267–275. [PubMed: 23471868]
- Wu, Michael C., Kraft, P., Epstein, MP., Taylor, DM., Chanock, SJ., Hunter, DJ., Lin, X. Powerful snp-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010; 86(6):929–942. [PubMed: 20560208]
- Xu Z, Pan W. Initiative”, A. D. N. Approximate score-based testing with application to multivariate trait association analysis. *Genet Epidemiol*. 2015; 39:469–79. [PubMed: 26198454]
- Yan Q, Weeks D, Celed?n J, Tiwari H, Li B, Wang X, Lin W, Lou X, Gao G, Chen W, Liu N. Associating multivariate quantitative phenotypes with genetic variants in family samples with a novel kernel machine regression method. *Genetics*. 2015; 201:1329–1339. [PubMed: 26482791]
- Yolken RH, Torrey EF, Lieberman JA, Yang S, Dickerson FB. Serological evidence of exposure to herpes simplex virus type 1 is associated with cognitive deficits in the catie schizophrenia sample. *Schizophrenia research*. 2011; 128(1):61–65. [PubMed: 21353483]
- Zhan X, Tong X, Zhao N, Maity A, Wu MC, Chen J. A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*. 2017; 41(3):210–220. [PubMed: 28019040]
- Zhu W, Zhang H. Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society*. 2009; 38(1):1–10. [PubMed: 19655045]

Figure 1.

Results for estimation of $h_k(\cdot)$ for $k = 1$ (left column), $k = 2$ (middle column) and $k = 3$ (right column) using different covariance settings for $n = 100$ and $M = 9$.

Figure 2.

Results for estimation of $h_k(\cdot)$ for $k = 1$ (left column), $k = 2$ (middle column) and $k = 3$ (right column) using different covariance settings for $n = 200$ and $M = 9$.

Figure 3.

Results for estimation of $h_k(\cdot)$ for $k = 1$ (left column), $k = 2$ (middle column) and $k = 3$ (right column) using different covariance settings for $n = 200$ and $M = 30$.

Figure 4. Simulation results for testing one h function (Scenario (A)). Displayed are the power of our test (shaded bars) along with SKAT test with Bonferroni correction (Non-shaded bars) for different settings with $\alpha = 0.05$ (dashed horizontal line).

Figure 5.

Simulation results for testing two h functions (Scenario (B)). Displayed are the power of our test (shaded bars) along with SKAT test with Bonferroni correction (Non-shaded bars) for different settings with $\alpha = 0.05$ (dashed horizontal line).

Figure 6. Simulation results for testing two h functions when only one is significant (Scenario (C)). Displayed are the power of our test (shaded bars) along with SKAT test with Bonferroni correction (Non-shaded bars) for different settings with $\alpha = 0.05$ (dashed horizontal line).

Figure 7. Simulation results for testing two h functions they have different directions of effect for different phenotypes (Scenario (D)). Displayed are the power of our test (shaded bars) along with SKAT test with Bonferroni correction (Non-shaded bars) for different settings with $\alpha = 0.05$ (dashed horizontal line).

Figure 8. Estimation results of CATIE Antibody analysis. Displayed are the 95% simultaneous confidence intervals for point estimates of h_1 , h_2 and h_3 for different genotype combinations as present in the data set (sorted by their point estimate values). Estimates from the multivariate model and separate univariate models are displayed in the left and right columns, respectively.

Table 1

Frequency of genotypes as observed in our real data example. Displayed are the genotypes appearing 7 times.

Genotype	Frequency (in real data)	Relative frequency (used in simulation)
221212211	114	0.191
222202220	103	0.174
112111121	72	0.121
111121112	46	0.078
220222202	41	0.069
112112221	38	0.064
111122212	28	0.047
211122212	26	0.044
111222212	24	0.041
002021122	21	0.034
212112221	15	0.025
002020022	10	0.017
102122222	10	0.017
102022222	9	0.015
002121122	8	0.013
002122222	8	0.013
220222211	8	0.013
102021122	7	0.012
202022222	7	0.012

Table 2

Estimation results of β for simulation study in Section 4. We report the empirical bias (Bias), empirical standard deviation (SD), average estimated standard errors using sandwich (SE.S) and model based (SE.M) estimators, and coverage probabilities using sandwich (C.S) and model based (C.M) estimators of 95% confidence intervals. All the original values are multiplied by 100 and rounded to their 1st decimal places. We used the IBS kernel.

	1						2						3					
	Bias	SD	SE.S	C.S	SE.M	C.M	Bias	SD	SE.S	C.S	SE.M	C.M	Bias	SD	SE.S	C.S	SE.M	C.M
(100, 9)																		
β_{11}	0.1	11.2	10.1	93.0	10.1	93.0	0.1	11.3	10.1	93.1	10.1	93.1	0.1	11.2	10.1	92.9	10.1	92.8
β_{12}	0.5	10.8	10.1	93.6	10.1	93.6	0.5	10.8	10.1	93.6	10.1	93.6	0.2	10.9	10.1	92.8	10.1	92.8
β_{21}	0.2	10.1	9.5	92.4	9.5	92.4	0.3	10.1	9.5	93.2	9.5	93.2	0.2	10.2	9.5	93.4	9.4	93.3
β_{22}	-0.7	10.3	9.5	93.7	9.5	93.6	-0.7	10.3	9.5	92.7	9.5	92.7	-0.5	10.3	9.5	92.6	9.5	92.5
β_{31}	-0.1	9.9	9.4	93.5	9.4	93.5	-0.1	9.8	9.4	93.7	9.4	93.7	-0.1	9.9	9.4	93.1	9.4	93.1
β_{32}	-0.2	10.1	9.4	92.9	9.4	92.8	-0.2	10.1	9.4	93.0	9.4	93.0	-0.1	10.1	9.4	93.3	9.4	93.3
(200, 9)																		
β_{11}	0.0	7.5	7.2	93.7	7.2	93.7	0.0	7.5	7.2	93.6	7.2	93.6	0.0	7.5	7.2	92.8	7.2	92.8
β_{12}	0.2	7.4	7.2	93.8	7.2	93.8	0.2	7.4	7.2	94.6	7.2	94.6	0.0	7.4	7.2	94.4	7.2	94.5
β_{21}	0.2	7.0	6.7	92.7	6.7	92.7	0.3	7.0	6.7	93.0	6.7	92.9	0.2	7.0	6.7	93.2	6.7	93.1
β_{22}	-0.3	6.9	6.7	94.1	6.7	94.1	-0.3	6.9	6.7	93.7	6.7	93.7	-0.2	6.9	6.7	94.0	6.7	94.0
β_{31}	-0.5	7.1	6.7	93.2	6.7	93.2	-0.5	7.1	6.7	93.1	6.7	93.1	-0.5	7.0	6.7	94.0	6.7	94.0
β_{32}	-0.1	7.0	6.7	93.9	6.7	93.9	-0.1	7.0	6.7	93.2	6.7	93.3	-0.1	7.0	6.7	93.9	6.7	93.9
(200, 30)																		
β_{11}	0.1	7.3	6.9	93.2	6.9	93.2	0.1	7.4	6.9	93.4	6.9	93.4	0.0	7.3	6.9	93.3	6.9	93.4
β_{12}	0.4	7.1	6.9	94.2	6.9	94.2	0.4	7.1	6.9	94.3	6.9	94.3	0.3	7.1	6.9	94.4	6.9	94.4
β_{21}	-0.3	6.8	6.6	94.4	6.6	94.4	-0.3	6.8	6.6	94.1	6.6	94.1	-0.2	6.9	6.6	93.4	6.6	93.4
β_{22}	-0.2	6.9	6.5	93.0	6.5	93.0	-0.2	6.9	6.6	92.9	6.6	92.8	-0.1	6.8	6.5	93.6	6.5	93.6
β_{31}	-0.1	7.3	6.6	92.7	6.6	92.7	-0.1	7.4	6.6	92.4	6.6	92.4	-0.1	7.3	6.6	92.4	6.6	92.4
β_{32}	0.3	6.9	6.6	94.6	6.6	94.6	0.3	6.8	6.6	94.1	6.6	94.1	0.4	6.9	6.6	94.2	6.6	94.2

Table 3

Simulation results for estimation of $h_k(\cdot)$ for $k = 1, 2$ and 3 . Displayed are the values of M_1 and M_2 defined in Section 4 for different combinations of (n, M) and different covariance settings. We used the IBS kernel.

	1			2			3		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(100, 9)									
M1	0.177	0.119	0.090	0.178	0.120	0.091	0.169	0.119	0.089
M2	0.305	0.262	0.228	0.306	0.263	0.231	0.301	0.263	0.228
(200, 9)									
M1	0.136	0.078	0.050	0.136	0.079	0.051	0.133	0.078	0.050
M2	0.257	0.213	0.168	0.257	0.214	0.169	0.255	0.214	0.167
(200, 30)									
M1	0.085	0.073	0.074	0.086	0.074	0.075	0.081	0.073	0.073
M2	0.209	0.195	0.198	0.210	0.196	0.199	0.207	0.195	0.198

Table 4

Testing results of CATIE Antibody analysis. The p-values of testings for one and two function(s) cases are displayed, both for our multivariate method and univariate SKAT with Bonferroni correction (SKAT). Note that SKAT does not need to perform Bonferroni correction for testing only one function (the first three columns)

pvalue	H_0					
	$h_1 = 0$	$h_2 = 0$	$h_3 = 0$	$h_1 = h_2 = 0$	$h_1 = h_3 = 0$	$h_2 = h_3 = 0$
Multivariate	0.400	0.004	0.019	0.005	0.037	0.004
SKAT	0.324	0.002	0.009	0.004	0.018	0.004