2015

Equivalence of Kernel Machine Regression and Kernel Distance Covariance for Multidimensional Trait Association Studies

Wen-Yu Hua
Debashis Ghosh, *university of colorado denver*

Equivalence of Kernel Machine Regression and Kernel Distance Covariance for Multidimensional Trait Association Studies

Wen-Yu Hua, Debashis Ghosh and the Alzheimers Disease Neuroimaging Initiative[1]

Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

Correspondence author : Wen-Yu Hua

Telephone : (814) 865-1348

Fax : (814) 863-7114

Email : wxh182@psu.edu

Address : Department of Statistics, 323 Thomas Building, University Park, PA 16802, USA

---

[1] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

Abstract

Finding associations between genetic markers and disease traits in high dimensional samples is a challenging problem in statistics and science. Two classes of methods are used to examine the interactions between genetic markers and disease phenotypes, i.e., kernel-machine regression (KMR), and kernel distance covariance (KDC). In the field of statistics, KMR is a semiparametric regression model that modeled the covariate effects parametrically, while the genetic markers are considered non-parametrically. KDC is a term that we have defined as a class of methods that includes distance covariance (DC) and Hilbert-Schmidt Independence Criterion (HSIC), which is a non-parametric statistic popularly used in the machine learning community for the test of independence hypothesis, given the particular kernels. In this work, we show that the score test of KMR is equivalent to the KDC statistic under certain kernel conditions. We also propose a novel KDC test that incorporates covariate effects and show that these two tests are the same in the presence of the covariates. Our contributions are three-fold: (1) establishing the equivalence between KMR and KDC; (2) the principles of kernel machine regression can be applied to the interpretation of KDC; (3) the KMR statistic is a member of a broader class of KDC statistics, that the members are the quantities of different kernels. We present the theoretical representation of the KMR and KDC equivalence, while the empirical justifications are demonstrated in our simulation studies for both single and multi-traits. Finally, the ADNI study is used to explore the association between the genetic variants on gene *FLJ16124* and phenotypes represented in 3D structural brain MR images adjusting for age and gender. The results suggest that SNPs of *FLJ16124* exhibit strong pairwise interaction effects that are correlated to the changes of brain region volumes.

Key words: Distance covariance; Hilbert-Schmidt independence criterion; Ker-

nel machine regression; Kernel distance covariance; Genetic association study; Score test; Permutation test.

Introduction

Association studies allow for inferences between the disease phenotypes and genetic variants. Compared to the final disease diagnosis, intermediate traits have been studied in the recent years, since such phenotypes have stronger connections to the genetic variants and provide comprehensive information for a given disease. For example, a structural magnetic resonance imaging (MRI) scan contains the structural details of a brain that are useful for detecting complex diseases, such as schizophrenia (Potkin et al., 2009), and Alzheimer's disease (Furney et al., 2010). Therefore, studying the multiple phenotypes on the genetic analysis should be added in the identification of genetic variants.

The traditional method in these studies is to regress data from one voxel on one variant at a time over all voxels and all variants (Stein et al., 2010b). Such a method is feasible in the cases of small numbers of genetic variants and traits. However, when the dimension of genotypes and phenotypes is both very high, the results has limited power because of the multiple testing problem. Therefore, several alternative strategies have been proposed to address this issue of multivariate studies, and they can be grouped into two categories: multiple markers versus univariate phenotype, and multiple marker versus multiple phenotypes.

The first type of the analysis is to model multiple markers with a univariate phenotype, and to then combine results across all phenotypes. This procedure is more powerful than the traditional method due to the reduction in the number of tests. Many flexibly parametric and nonparametric methods could be applied here, including generalized additive model (GAMs) (Hastie and Tibshirani, 1986), thin-plate splines (Wahba (1990) and Wood (2003)), and penalized regression splines (Ruppert et al., 2003), kernel-machine regression (KMR) (Liu et al. (2007) and Kwee et al. (2008)). These is able to provide more informative association between multiple bio-markers and single phenotype. However, the multiple testing issues are still need to be address over all phenotypes.

In this work, the discussion is focused on modelling multiple genotypes versus multiple phenotypes. Two classes of methods have been used here are the multivariate kernel machine regression model (MV-KMR) (Maity et al., 2012), and the kernel distance covariance method (KDC) (Szekely et al. (2007), Szekely and Rizzo (2009), Gretton et al. (2005), and Smola et al. (2007)).

MV-KMR is a multiple traits extension of KMR that was proposed by Maity et al. (2012). It is a dimension reduction approach to model linear or nonlinear multi-marker effects. In this work, we focus on the multivariate phenotypes, therefore we will refer to MV-KMR as KMR, which means the KMR model is feasible for both single and multiple phenotypes. Notice that this approach of modeling the genotypes is similar to spline-based approach theoretically, but the implications behind the variable space fitting are different. The kernel machine determines the kernel based on the nature of variable values, where we generally use Gaussian RBF or linear kernel for the quantitative variables, and IBS kernel or polynomial kernel for the qualitative samples. In general, kernel machine methods can greatly simplify specification of a non-parametric model, especially for high dimensional data (Liu et al., 2007).

KDC was presented for the test of independence i.e., distance covariance (DC) and Hilbert-Schmidt Independence Criterion (HSIC). DC were established by Szekely and Bakirov (2003), Szekely et al. (2007) for testing independence in Euclidean spaces. Such a test is applicable in high dimensional spaces and it is consistent against all alternatives as long as the first moment for both samples are bounded. One advantage of distance covariance is the compact representation of the statistic which is the product of expectations of pairwise Euclidean distance, and this give the straightforward empirical estimates. On the other hand, Gretton et al. (2005), and Smola et al. (2007) formulated the two-sample independence test (HSIC) in reproducing kernel Hilbert spaces (RKHS). Again, the HSIC statistic is able to find the

test of dependence in multivariate spaces, and it is consistent when a characteristic RKHS is used, (Sriperumbudur et al., 2010). Sejdinovic et al. (2013) demonstrated that the above two statistics are the same when the distance-induced kernels in HSIC are chosen. The results showed that the HSIC test was more sensitive when the quantity was derived from other kernels, and the HSIC tests can be readily extended to more structured and non-Euclidean spaces.

The application of multivariate traits association study is focused in this work, and we first provide a compact representation of the multiple phenotypes version of KMR, and show that the KMR and KDC are equivalent under the condition of no covariate and Gower distance (Gower, 1966) is used for the phenotype spaces. Then, we extend to the setting when the covariates are present, propose a new KDC test with the presence of the covariates and show that the KDC is equivalence to the KMR of Maity et al. (2012). Two major implications are found by the equivalence established in this work: first, the equivalence shows that the principles of kernel machine regression can be applied to the interpretation of KDC. Second, the KMR statistic is a member of a family of KDC statistics, that the members are the quantities of different kernels. Finally, our numerical results suggest that the KDC may yield a more powerful results with the data-oriented kernels.

Preliminaries

Distance properties of the sum of squares

Before we introduce the equivalence between KMR and KDC, the review of the distance properties of the sums of squares (Gower, 1966) is provided here. The multivariate sample has a set of n points $\mathbf{Y} = (Y_1, ..., Y_n)^t$ in p space; where Y_i has coordinates $(y_{i1}, ..., y_{ip})$. Thus, the distance d_{ij} between Y_i and Y_j is $d_{ij}^2 = \sum_{r=1}^{p}(y_{ir} - y_{jr})^2$. Suppose $\mathbf{A}$ is a $n \times n$ symmetric sum of squares matrix of $\mathbf{Y}$ with eigenvalues $\lambda_1, ..., \lambda_n$ and eigenvectors $\mathbf{c}_1, ..., \mathbf{c}_n$. If $\mathbf{c}_1, ..., \mathbf{c}_n$

are normalized, then $\mathbf{A} = \mathbf{c}_1\mathbf{c}_1' + \mathbf{c}_2\mathbf{c}_2' + ..., \mathbf{c}_n\mathbf{c}_n'.$; where $a_{ii} = \sum_{r=1}^{n} c_{ir}^2$, and $a_{ij} = \sum_{r=1}^{n} c_{ir}c_{jr}$. Hence, the distance between Y_i and Y_j can be derived as follows:

$$\begin{aligned} d_{ij}^2 &= \sum_{r=1}^{p}(y_{ir} - y_{jr})^2 \\ &= \sum_{r=1}^{n}(c_{ir} - c_{jr})^2 \\ &= a_{ii} + a_{jj} - 2a_{ij} \end{aligned} \tag{1}$$

Now, A is centered by subtracting the means, then $a_{ii} = 0$, and $a_{ij} = -\frac{1}{2}d_{ij}^2$ (denoted as Gower distance (Gower, 1966)), then the sum of square of $\mathbf{Y}$ became

$$\begin{aligned} \mathbf{YY}' &= \left(I - \frac{11'}{n}\right)\mathbf{A}\left(I - \frac{11'}{n}\right) \\ &= -\frac{1}{2}\left(I - \frac{11'}{n}\right)d_{ij}^2\left(I - \frac{11'}{n}\right), \end{aligned}$$

here $\mathbf{Y}$ is also centered by column means, such that $\sum_{i=1}^{n} y_{ir} = 0$.

Linear model

The linear model can be formulated to examine the effects between two variables. Suppose we observe the n subjects with index $i = 1, ..., n$, the response $\mathbf{Y} = (Y_1, .., Y_n)^t$ in $p-$dimensional space and the predictor $\mathbf{Z} = (Z_1, ..., Z_n)$ in $q-$dimensional space. To understand the relationships between $\mathbf{Y}$ and $\mathbf{Z}$, a typical way is to apply the multivariate linear model (e.g., MANOVA). Traditional multivariate analysis proceeds through partitioning of the total sum of squares $tr(\mathbf{Y}'\mathbf{Y})$. The analysis can be done by the linear model $\mathbf{Y} = \mathbf{Z}\beta + \epsilon$, and to test of effects between $\mathbf{Y}$ and $\mathbf{Z}$, we do the hypothesis for testing $H_0 : \beta = 0$, and the least square estimates for β is $\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. Therefore, the fitted values of $\mathbf{Y}$ is $\hat{\mathbf{Y}} = \mathbf{Z}\hat{\beta} = \mathbf{HY}$, where $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$. The sum of residuals is $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$, and $tr(\mathbf{Y}'\mathbf{Y}) = tr(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + tr(\mathbf{R}'\mathbf{R})$. Hence an appropriate statistic to test the null hypothesis of the model having no effects is a pseudo F statistic (McArdle and Anderson, 2001):

$$F = \frac{tr(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})/(q-1)}{tr(\mathbf{R}'\mathbf{R})/(n-q)} = \frac{tr(\mathbf{HYY}'\mathbf{H})/(q-1)}{tr((\mathbf{I}-\mathbf{H})\mathbf{YY}'(\mathbf{I}-\mathbf{H}))/(n-q)}. \tag{2}$$

McArdle and Anderson (2001) suggested that the above partitioning procedure can be done using the outer product matrix, i.e., $tr(\mathbf{YY}')$, since $tr(\mathbf{Y}'\mathbf{Y}) = tr(\mathbf{YY}')$. Therefore, we can replace $\mathbf{YY}'$ with any $n \times n$ distance matrix D.

$$\frac{tr(\mathbf{HDH})/(q-1)}{tr((\mathbf{I}-\mathbf{H})\mathbf{D}(\mathbf{I}-\mathbf{H}))/(n-q)}. \tag{3}$$

If D is a Gower distance matrix, then Eq. (3) is the same as Eq. (2), and Eq. (2) is a pseudo F statistic. If D is other distance (kernel) matrix, then the significance of Eq. (3) can be tested using the permutation technique. The estimate in Eq. (3) is flexible in capturing the nature of the response, and measures the effect between the predictors and the responses at the same time.

Methods

McArdle and Anderson (2001) applied the outer product tool to extend the MANOVA to the general MANOVA. This inspired us to apply the same argument to the response variable in the KMR model. It turns out that it is equivalent to the KDC when a common kernel is chosen. To understand this link between KMR and KDC, we first assume no covariate effects. Later, we extend the situation that the covariate effects are included, proposed a new KDC test with the presence of the covariates and show that the statistics of KMR and KDC are equivalent.

Without the covariates

In order to test of dependence of two random variables, i.e., the association between the phenotypes $\mathbf{Y} = (Y_1, ..., Y_n)^t \in \mathbb{R}^p$ and the genotypes $\mathbf{Z} = (Z_1, ..., Z_n)^t \in \mathbb{R}^m$, one can use KDC (i.e., DC or HSIC) for testing the independence. We denote the kernel function $k_{ij} = k(Z_i, Z_j)$ as an element of row i and column j of the kernel matrix K in $\mathbf{Z}$, and $l_{ij} = l(Y_i, Y_j)$ is the kernel function of Y_i and Y_j in the kernel matrix L in $\mathbf{Y}$ space. Therefore,

the KDC statistic is defined as follows:

$$\text{KDC}_n = \frac{1}{n^2} tr(KHLH) \propto tr(KHLH), \tag{4}$$

where $tr(\mathbf{X})$ is the trace of $\mathbf{X}$ and $H = (I - \frac{11'}{n})$. If both k and l are distance kernels, then Eq. (4) is the DC statistic (Szekely and Bakirov, 2003), otherwise Eq. (4) is the HSIC statistic (Gretton et al., 2005). In summary, the KDC statistic is used to test the dependence between $\mathbf{Y}$ and $\mathbf{Z}$ without fitting the any model.

Another powerful test for test of association is kernel machine regression, which we now briefly discuss and link to the KDC. The linear model in Liu et al. (2007) and Kwee et al. (2008) is given by

$$Y = \beta_0 + h(\mathbf{Z}) + \epsilon, \tag{5}$$

where $h(\cdot)$ is an unknown function to be estimated by the effects of the SNPs on the univariate response Y, and it is determined by a specified positive semi-definite kernel function $k(.,.)$. To test the effects $h(\cdot)$, Liu et al. (2007) proposed a hierarchical Gaussian process regression for the linear model (5):

$$Y|h(\mathbf{Z}) \sim N\{\beta_0 + h(\mathbf{Z}), \sigma^2\}, \quad h(\cdot) \sim GP\{0, \tau K\}.$$

Therefore, the null hypothesis is that there is no association between phenotype Y and the SNPs $\mathbf{Z}$. One can test $H_0 : \tau = 0$ since h can be treated as the subject-specific random effects with mean 0 and covariance matrix τK. Thus, the corresponding variance component score test is proportional to:

$$\begin{aligned} Q &\propto (Y - \bar{Y})' K (Y - \bar{Y}) \\ &= tr[(Y - \bar{Y})' K (Y - \bar{Y})] \\ &= tr[K(Y - \bar{Y})(Y - \bar{Y})'] \\ &= tr[K(I - \frac{11'}{n})YY'(I - \frac{11'}{n})] \\ &= tr[KHYY'H] \end{aligned} \tag{6}$$

By using the trace trick, Eq. (6) can be extended into two directions. First, the previous work in Liu et al. (2007) and Kwee et al. (2008) focused on a single phenotype Y, but we can also replace Y with a multivariate phenotypes $\mathbf{Y}$ and it turns out that $tr[KH\mathbf{Y}\mathbf{Y}'H]$ is equivalent to MV-KMR in Maity et al. (2012) in absence of covariates. Second, we can replace the outer product YY' with any distance matrix L, and it is equivalent to KDC in Eq. (4).

With covariates

In practice, we may want to know the relationship between the genotypes ($\mathbf{Z}$) and phenotypes ($\mathbf{Y}$) where the phenotype is adjusted by the covariates ($\mathbf{X}$), and we observe n samples from $\mathbf{X} \in \mathbb{R}^m$, $\mathbf{Y} \in \mathbb{R}^p$, and $\mathbf{Z} \in \mathbb{R}^q$. Under this setting, the multivariate traits KMR model (Maity et al., 2012) is

$$\mathbf{Y} = \mathbf{X}\beta + h(\mathbf{Z}) + \epsilon, \tag{7}$$

where $h(\cdot)$ is an non-parametric function which describes the effect of Z. To test the effect of $h(\cdot)$, one can test $H_0 : \tau = 0$ under the following representation that is a multivariate extension of the hierarchical Gaussian process regression from the previous section:

$$\mathbf{Y}|(\beta, h(\mathbf{Z})) \sim N\{\mathbf{X}\beta + h(\mathbf{Z}), \Sigma\}, \quad h(\cdot) \sim GP\{0, \tau K\}.$$

, and the corresponding score test of H_0 is proportional to

$$\begin{aligned} Q &\propto (\mathbf{Y} - \mathbf{X}\hat{\beta})'K(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= tr[(\mathbf{Y} - \mathbf{X}\hat{\beta})'K(\mathbf{Y} - \mathbf{X}\hat{\beta})] \\ &= tr[(\tilde{\mathbf{Y}} - \bar{\tilde{\mathbf{Y}}})'K(\tilde{\mathbf{Y}} - \bar{\tilde{\mathbf{Y}}})] \\ &= tr[KH\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'H], \end{aligned} \tag{8}$$

Notice that the $\hat{\beta}$ is the MANOVA estimates in the linear model section. So, we can set $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ in Eq. (8), here H is an normalized constant matrix $(I - \frac{11'}{n})$. Hence, E.q (8)

is equivalent to the score test in KMR from Maity et al. (2012), and the outer product $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'$ can be replaced with any distance measure $\tilde{L}$, such that Eq. (8) becomes

$$tr[KH\tilde{L}H]. \tag{9}$$

The original idea of KDC (i.e., DC in Szekely and Bakirov (2003) and HSIC in Gretton et al. (2005)) was presented to test for independence. Here, we can extend it to the case when the covariates are present. Also, if $\tilde{L} = -\frac{1}{2}d_{ij}^2$, where $d_{ij}^2 = \sum_r^p (\tilde{y}_{ir} - \tilde{y}_{jr})^2$, then KDC in E.q (9) is again equivalent to the KMR in Eq (8).

For both the cases of absent or present the covariate effects, the equivalence was built by ignoring the variance factor of the score test in Eq. (6) and Eq. (8), and we treat the statistic Q as non-random and fixed. Therefore, the variance factor can be absorbed to Q by standardized the distribution of Q, similar to the assumption presented by Pan (2011).

Simulation studies

Possible kernel choices

There are a number of kernels for characterizing the similarity of individuals with respect to the variations of genotypes and phenotypes. We consider the following kernels for our numerical data analyses:

(1) Identity-by-state (IBS) kernel

Since humans have two copies of an allele at each position on the genome, one can count how many alleles (0,1, or 2) a pair of individuals shares. The IBS-sharing similarity, K can be calculated for individuals i and j as

$$K(x_i^l, x_j^l) = \frac{\sum_{l=1}^S k_{i,j}^l(x_i^l, x_j^l)}{2S},$$

where S is the number of loci considered in the calculation; x_i^l and x_j^l are genotypes of individuals i and j, respectively at the lth locus ($l = 1, ..., S$); and $k_{i,j}^l(x_i^l, x_j^l)$ is a

function mapping the genotype information for individuals i and j at locus l. $k^l_{i,j}(x^l_i, x^l_j)$ has a value of zero if individual i and j are homozygous for different SNP alleles, a value of one if they share one allele, and a value of two if they share both alleles (Wessel and Schork, 2006).

(2) Euclidean distance (ED) kernel

$$K(x_i, x_j) = \sqrt{\sum_{r=1}^{m}(x_{ir} - x_{jr})^2},$$

where m is the dimensionality of sample x.

(3) Gaussian RBF kernel

$$K(x_i, x_j) = \exp\{-\rho\|x_i - x_j\|^2\},$$

where ρ is the weight parameter.

(4) Polynomial kernel

$$K(x_i, x_j) = (\langle x_i, x_j\rangle + c)^d,$$

where $\langle x_i, x_j\rangle$ denotes the inner product of x_i and x_j, and c is a constant.

Notice that the polynomial kernel can be simplified into a linear kernel when c=0 and $d = 1$ or into a quadratic kernel when $c = 1$ and $d = 2$.

Simulation studies

The design of the simulation studies is based on Liu et al. (2007), and Maity et al. (2012), and the goal of the following three simulations is to compare the performances of KMR and KDC in terms of the empirical error rates and power analysis under different kernel combinations.

Simulation 1

The first simulation examines the association between the effect of a single phenotype Y and the multivariate $\mathbf{Z}$ adjusted by a single covariate X, and the design of the simulation is

based on Liu et al. (2007). Consider the true linear model:

$$Y = \beta X + h(Z_1, ..., Z_q) + \epsilon,$$

where $h(\mathbf{Z}) = ah_1(\mathbf{Z})$, $h_1(\mathbf{Z}) = 2\cos(Z_1) - 3Z_2^2 + 2\exp(-Z_3)Z_4 - 1.6\sin(Z_5)\cos(Z_3) + 4Z_1Z_5$, and $X = 3\cos(Z_1) + u$. The $Z_j's (j = 1, ..., 5)$ were generated from uniform(0,1) while u and ϵ were generated from independent $N(0, 1)$.

To adjust for X thoughout the paper, we first used the *lm* function from `stat` package in R to solve $\hat{\beta}$, and then $\tilde{Y} = Y - \hat{\beta}X$. The empirical error rates and power of the tests were evaluated by generating data under $a = 0$ and $a = 0.25, 0.5, 0.75, 1.0$ at significance level of 0.05. The sample size was 60; the p-values of the statistic were computed based on 10^4 permutations, and this experiment was repeated 1000 times. In the following, we used K to represent the kernel matrix for the genotypes $\mathbf{Z}$, and $\tilde{L}$ to represent the kernel matrix for the adjusted phenotype $\tilde{Y}$. Table 1 shows the equivalence between KMR and KDC when $\tilde{L}$ is a linear kernel, and the performances of the quadratic kernel are shown to have less power than its counterparts. Table 2 displays the size and power estimates of KMR when Gaussian RBF is used, and the power is high when the scale parameter ρ is small(less than or equal 1 in our experiment setup), while the power is low when ρ is large. The KDC test results when using Gaussian RBF kernel with combinations of a linear or a quadratic kernel are shown in Table 3. Case 1 in Table 3 shows the KDC results using a linear kernel for $\tilde{L}$ and Gaussian RBF for K, and they are the same as KMR estimates in Table 2. This is because the KMR implicity models the phenotype effect linearly. Furthermore, the results of all four cases in Table 3 are similar, due to the design of the simulation which was based on single phenotype to all other genotypes. Therefor, all kernels captured the simple effects between the phenotype and genotypes. Overall, the results of Table 1, 2, and 3 show that the empirical error rates ($a = 0$) are all very close the nominal level 0.05, and the power increases as a increased. Specifically, the best power was shown by KMR with Gaussian RBF kernel

($\rho = 0.5$) for K, and KDC with Gaussian RBF kernel ($\rho = 0.5$) for K and linear kernel for $\tilde{L}$. This suggests our proposed KDC approach is able to achieve the same performance as KMR to capture the association between the genotypes and phenotype when the covariate is present.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

Simulation 2

For our second simulation, we examine the effects of KMR and KDC in a high-dimensional setting, and the design of the simulation was based on Maity et al. (2012). For $k = 1, ..., p$, the data was generated through the model

$$Y_k = \mathbf{X}'\beta_k + h_k(\mathbf{Z}) + \epsilon_k, \tag{10}$$

where $\mathbf{X} = (X_1, X_2)^T$ were generated from bivariate normal $BN((0.2, 0.4)^T, I)$, and $\epsilon_k's$ were generate $MVN(0, \Sigma_{true})$. The q-SNP genotype data $\mathbf{Z} \equiv (Z_1, ..., Z_q)$ were simulated based on the first gene $SLC17A1$ in the CATIE antibody study that contained nine ($q = 9$) SNPs.

The empirical error rates ($a = 0$) and power ($a = 0.1, 0.2$) were examined at significance level of 0.05. The sample size n was 100, the dimension of genotypes q was 9, and the dimension of phenotypes p was 3. Two choices for the effects of $h_k; k = 1, 2, 3$ were also considered: first is the sparse effect, i.e., $h_1 = a(z_1 + z_2 + z_3 + z_1z_4z_5 - \frac{z_6}{3} - \frac{z_7z_8}{2} + (1 - z_9))$, $h_2 = h_3 = 0$, where $a = 0, 0.1, 0.2$; and then the common effect, i.e., $h_1^* = h_1 + az_3$, and $h_2 = h_3 = az_3$ with $a = 0, 0.1, 0.2$. In addition, we also investigated the performances of KMR and KDC by varying the variance-covariance of Σ_{true} from an independent structure

(Σ_1), and a highly dependent structure (Σ_2).

$$\Sigma_1 = \begin{pmatrix} 0.95 & 0 & 0 \\ 0 & 0.86 & 0 \\ 0 & 0 & 0.89 \end{pmatrix}; \quad \Sigma_2 = \begin{pmatrix} 0.95 & 0.57 & 0.43 \\ 0.57 & 0.86 & 0.24 \\ 0.43 & 0.24 & 0.89 \end{pmatrix}.$$

Table 4 shows the empirical error rates of KMR and KDC tests when the linear, quadratic, and Euclidean distance kernels are used, and the values are all close to $\alpha = 0.05$. This suggests that the tests satisfactory control the error rates. Table 5 displays the power of case 1 when the covariate effects are included, where the phenotypes $\tilde{Y}$ are adjusted by the covariates (same as simulation 1); and case 2 when the covariate effects are excluded, where the phenotypes Y are the raw samples. It shows that case 1 had more powerful results than case 2, and it suggests that our proposed KDC test is able to incorporate the covariates and improve the power; and although the highly dependent structure (Σ_2) weakens the power estimates but the values are still very high, which means that the KMR and KDC tests are both able to identify the associations even when the variables are correlated. Table 6 shows the empirical size and power estimates; again the empirical error rates are close to $\alpha = 0.05$, and the power increases when a increased. The results of Table 6 are similar to the previous simulation, that the power performances are poor when the quadratic and Gaussian RBF kernel with large scale ρ are used.

Overall, when the sparse effect and the independent covariance matrix Σ_1 was used, the best power performance could be achieved when KMR and KDC is linear on $\tilde{L}$ in Table 5. This is because the linear kernel only identified the single connection between the Y_1 and h_1 without considering pairwise interactions. In Table 6, the best power performances occurred with KDC using the Gaussian RBF kernel (ρ=0.1) for $\tilde{L}$, when the common effects for a highly dependent covariance matrix Σ_2 was introduced; and this is due to the design of this simulation using the common effects and dependent covariance structure, while a Gaussian RBF kernel is able to identify the nonlinear interaction effects.

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

Experiments with the ADNI study

Data used in the preparation for this project was obtained from the ADNI study (ADNI, 2003). One of the goals of the ADNI study is to perform genome-wide association tests on the entire genome, and identify the genetic variants that influence the voxel-level differences. Several work have been published to investigate this goal, i.e., Stein et al. (2010b), Stein et al. (2010a), Shen et al. (2010), Vounou et al. (2010), and Hibar et al. (2011). In the ADNI study, the phenotype is presented by brain structural magnetic resonance imaging (MRI) scans (31,662 brain voxels), each containing a value that represents the volumetric difference of such voxel from a healthy reference brain - a tensor based morphometry is used to compute the 3D map of regional brain volume differences compared to an average template image based on healthy elderly subjects. The genotypes are encoded by 448,244 SNPs across the entire genome, and the demographic variables include gender and age of all the participants.

In this work, the phenotypes were summarized from 31,662 total voxels into 119 region-of-interests (ROIs), and the mapping was based on the GSK CIC Atlas (Tziortzi et al., 2011). The average voxel volumetric differences within each region was used to represent the 119 ROIs. Hua et al. (2013) used the DC test on the same 119 brain MRI regions, and discovered that the difference in brain volumes were highly associated with a common variant $rs11891634$ in the intron region of gene *FLJ16124*. A total of 183 SNPs within gene *FLJ16124* were identified by SNP-gene mapping from Hibar et al. (2011). Furthermore, 741 of all subjects from the ADNI study passed the quality control filtering according to Stein

et al. (2010b) (206 normal older controls, 358 mild cognitive impairment (MCI) subjects, and 177 Alzheimer's Disease (AD) patients), which we used in the following simulation and real data analysis.

Simulation based on ADNI

For this simulation study, a linear model (same as simulation 2) $\mathbf{Y}_k = \beta\mathbf{X} + h(\mathbf{Z}) + \epsilon_k$; where $k = 1, ..., p$ was used to associate between the phenotypes and genotypes. (11) shows the all-pairwise correlations (r) that were based on the 358 MCI subjects, where eight ($p = 8$) frontal cortex regions of the 119 ROIs were used for the corrected structure in ϵ, i.e., the left and right anterior dorsolateral prefrontal cortex (corresponds to row 1 and 2), posterior dorsolateral prefrontal cortex (row 3,4), anterior medial prefrontal cortex (row 5,6), and posterior medial prefrontal cortex (row 7,8).

$$r = \begin{pmatrix} 1.00 & 0.95 & 0.97 & 0.87 & 0.53 & 0.97 & -0.99 & -0.87 \\ 0.95 & 1.00 & 1.00 & 0.98 & 0.77 & 1.00 & -0.90 & -0.66 \\ 0.97 & 1.00 & 1.00 & 0.96 & 0.72 & 1.00 & -0.94 & -0.72 \\ 0.87 & 0.98 & 0.96 & 1.00 & 0.88 & 0.97 & -0.81 & -0.51 \\ 0.53 & 0.77 & 0.72 & 0.88 & 1.00 & 0.73 & -0.43 & -0.04 \\ 0.97 & 1.00 & 1.00 & 0.97 & 0.73 & 1.00 & -0.93 & -0.71 \\ -0.99 & -0.90 & -0.94 & -0.81 & -0.43 & -0.93 & 1.00 & 0.92 \\ -0.87 & -0.66 & -0.72 & -0.51 & -0.04 & -0.71 & 0.92 & 1.00 \end{pmatrix} \quad (11)$$

We then estimated the eight ROIs' covariance matrix Σ using 358 MCI participants. The 358 MCI samples were selected for the simulation design due to the relatively uniform MRI outputs in the MCI group (Vounou et al., 2010), and hence all $\epsilon_k s$ were generated from $MVN(0, \Sigma)$. All SNPs on gene *FLJ16124* were used for the genotype element ($\mathbf{Z}$), i.e., $h(Z_1, ..., Z_{183}) = a \times h_1$, while only the first 5 SNPs, $\mathbf{Z} = (Z_1, ..., Z_5)$ of the 183 were the

causative SNPs, such that $h_1(\mathbf{Z}) = 2\cos(Z_1) - 3Z_2^2 + 2\exp(-Z_3)Z_4 - 1.6\sin(Z_5)\cos(Z_3) + 4Z_1Z_5$. For the covariate effects, we considered gender and standardized age based on the same 358 MCI subjects, where gender (X_1) was generated from a Bernoulli distribution with $p = 0.36$, and standardized age (X_2) was generated from a standard normal. A total of 100 samples were generated, and the empirical error rate ($a = 0$) and powers ($a = 0.05, 0.1$) were computed based on 10^4 permutation. This simulation was repeated 1000 times and the significance level was set as 0.05.

Table 7 shows the KMR and KDC results with the linear, quadratic, IBS and Euclidean distance kernel, and the KDC test with the Euclidean distance measure has the best performance among all the kernels. We also implemented the KDC test with a Gaussian kernel for $\tilde{L}$ ($\rho \in 0.1, 0.5, 1, 5, 10$), and the linear, quadratic, and IBS kernel for K in case 1 of Table 8. The highest power estimate was observed when we used a Gaussian kernel for $\tilde{L}$ (ρ=0.1) and a linear kernel for K. This result is similar to the power performance of KDC test with the Euclidean distance kernel in Table 7. This suggests that when the dimensions of phenotype and genotype are both very high, both the Gaussian RBF kernel (with optimal parameter ρ) and the Euclidean distance kernel are able to describe the high dimensional interactions, and also results in powerful performances.

In addition, Table 8 displays the empirical error rates and power estimates in two cases: the tests that include the effect of covariates (case 1), and exclude the effect of covariates (case 2). The table shows that the power values of case 1 were all greater than the power values of case 2, this implies that the phenotype and genotypes were confounded by the covariates, which suggests our proposed KDC test is able to greatly improve the power performances in high dimensional settings when the covaraite effects are included.

[Table 7 about here.]

[Table 8 about here.]

Real data analysis

KMR and KDC tests were conducted using the ADNI study to find the associations between the genetic variants and the multivariate brain MRI voxels with the demographic effects of gender and age. In contrast to the previous simulation setup, we utilize all 741 subjects of the ADNI study, 183 SNPs within gene *FLJ16124*, 119 ROIs and two covariates, i.e., gender and age. Table 9 displays the p-values of KMR and KDC with different choice of kernels (linear, quadratic, IBS, and Euclidean distance kernel), where p-values are based on 10^4 permutations. KMR (K=quadratic) and KDC ($\tilde{L}$=linear,K=quadratic) identified the smallest p-value, which means quadratic kernel on the 183 SNPs has the most powerful performance in the KDC family among all the different choices of kernels. This suggests that there are strong pairwise interaction effects among the SNPs in *FLJ16124* that associate with brain region volumes, adjusting for age and gender.

[Table 9 about here.]

Conclusion and discussion

We have established the link between KMR and KDC. The advantage of such equivalence allows the use of KMR interpretation to explain the KDC test. Liu et al. (2007) and Maity et al. (2012) provided the REML score test of KMR, and it suggests that the null distribution of the KMR test is able to fit on the KDC test for the hypothesis testing, which greatly reduces computational costs compared to the permutation approach.

In addition, the KMR tests can be treated as the members of a larger family of KDC, and more powerful tests could be designed by looking at the optimal kernel among the family members. Although there is no best kernel from our experiments, the linear kernel for KMR or KDC have better performances than other kernels when the single phenotype was introduced; when the multiple phenotypes with multiple correlations or dependent covariance

were presented, the Gaussian RBF or the Euclidean kernel achieves better performances than other kernels. This result can be extended into designing a strategy to select for the the optimal kernel from the members of the large KDC family (Gretton et al. (2012) and Wu et al. (2013)).

Finally, several work have utilized the KDC/KMR family members in the applications including Genetic pathway analysis using KMR (Liu et al., 2007), voxel-wise genome-wide association studies using least square KMR (Ge et al., 2012), Neuroimaging genome-wide association using DC (Hua et al., 2013), and multiple change point analysis using DC by Matteson and James (2013). Two recent work have presented and discussed the equivalence between the family members of our work, such as the relationships between Genomic Distance-Based Regression (GDBR) and KMR from Pan (2011) and the equivalence between DC and HSIC from Sejdinovic et al. (2013). Therefore, our establishment of the equivalence between KMR and KDC in this work is an important unification of all the above applications.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda

Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

ADNI (2003). Alzheimer's disease neuroimaging initiative. http://www.loni.ucla.edu/ADNI/.

Furney, S., Simmons, A., Breen, G., Pedroso, I., Lunnon, K., Proitsi, P., Hodges, A., Powell, J., Wahlund, L., Kloszewska, I., et al. (2010). Genome-wide association with mri atrophy measures as a quantitative trait locus for alzheimer's disease. *Molecular psychiatry* **16,** 1130–1138.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage* .

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325–338.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005). Kernel methods for measuring independence. *The Journal of Machine Learning Research* **6,** 2075–2129.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1214–1222.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science* pages 297–310.

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M. J., et al. (2011). Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage* **56,** 1875–1891.

Hua, W.-Y., Nichols, T. E., and Ghosh, D. (2013). Multiple comparison procedures for neuroimaging genomewide association studies. *Submitted to Biostatistics* .

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics* **82,** 386–397.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63,** 1079–1088.

Maity, A., Sullivan, P. F., and Tzeng, J.-i. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology* **36,** 686–695.

Matteson, D. S. and James, N. A. (2013). A nonparametric approach for multiple change point analysis of multivariate data. *Submitted to Annals of Statistics* .

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82,** 290–297.

Pan, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic epidemiology* **35,** 211–216.

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Fallon, J. H., Nguyen, D. D., Mathalon, D., Ford, J., Lauriello, J., Macciardi, F., et al. (2009). A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophrenia bulletin*

35, 96–108.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12. Cambridge University Press.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Submitted to Annals of Statistics* .

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., Foroud, T., Pankratz, N., Moore, J. H., Sloan, C. D., et al. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: a study of the adni cohort. *Neuroimage* **53,** 1051–1063.

Smola, A., Gretton, A., Song, L., and Scholkopf, B. (2007). A hilbert space embedding for distributions. In *In Algorithmic Learning Theory: 18th International Conference*, pages 13–31. Springer-Verlag.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* **99,** 1517–1561.

Stein, J. L. et al. (2010a). Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in alzheimer's disease. *Neuroimage* **51,** 542–554.

Stein, J. L. et al. (2010b). Voxelwise genome-wide association study (vgwas). *Neuroimage* **53,** 1160–1174.

Szekely, G. J. and Bakirov, N. K. (2003). Extremal probabilities for gaussian quadratic forms. *Probability Theory and Related Fields* **126,** 184–202.

Szekely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics* **3,** 1236–1265.

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35,** 2769–2794.

Tziortzi, A. C. et al. (2011). Imaging dopamine receptors in humans with [< sup> 11</sup> c]-(+)-phno: Dissection of d3 signal and anatomy. *Neuroimage* **54,** 264–277.

Vounou, M., Nichols, T. E., and Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *Neuroimage* **53,** 1147–1159.

Wahba, G. (1990). *Spline models for observational data*. Number 59. Siam.

Wessel, J. and Schork, N. J. (2006). Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics* **79,** 792–806.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 95–114.

Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., Engel, S. M., Molldrem, J. J., and Armistead, P. M. (2013). Kernel machine snp-set testing under multiple candidate kernels. *Genetic epidemiology* .

Table 1

Empirical error rates and the power of KMR and KDC with the choices of kernel: linear, quadratic, and Euclidean distance; where K represents the kernel matrix for the genotypes $\mathbf{Z}$, and $\tilde{L}$ represents the kernel matrix for the adjusted phenotype $\tilde{Y}$

	Size	Power			
	a=0	a=0.25	a=0.5	a=0.75	a=1
KMR (K=linear)	0.044	0.227	0.759	0.970	0.996
KMR (K=quadratic)	0.035	0.198	0.706	0.962	0.995
KDC ($\tilde{L}$,K=ED)	0.049	0.212	0.725	0.967	0.995
KDC ($\tilde{L}$,K=linear)	0.044	0.227	0.759	0.970	0.996
KDC ($\tilde{L}$=linear,K=quadratic)	0.035	0.198	0.706	0.962	0.995
KDC ($\tilde{L}$=quadratic,K=linear)	0.041	0.173	0.585	0.877	0.933
KDC ($\tilde{L}$,K=quadratic)	0.039	0.146	0.519	0.822	0.908

Table 2

Empirical error rates and the power of KMR with Gaussian RBF kernel for K

	Size	Power			
Scale (ρ)	a=0	a=0.25	a=0.5	a=0.75	a=1
0.1	0.051	0.213	0.772	0.985	1.000
0.5	0.040	0.217	0.766	1.000	1.000
1	0.045	0.211	0.752	1.000	1.000
5	0.037	0.156	0.653	0.939	0.997
10	0.037	0.116	0.453	0.792	0.982

Table 3

Empirical error rates and the power of simulation 1 under 4 cases. Case 1: KDC with Gaussian RBF for K and a linear kernel for $\tilde{L}$; Case 2: KDC with Gaussian RBF for K and a quadratic kernel for $\tilde{L}$; Case 3: KDC with Gaussian RBF for $\tilde{L}$ and a linear kernel for K; Case 4: KDC with Gaussian RBF for $\tilde{L}$ and a quadratic kernel for K.

	Size	Power				Size	Power			
Scale (ρ)	a=0	a=0.25	a=0.5	a=0.75	a=1	a=0	a=0.25	a=0.5	a=0.75	a=1
	case 1					case 2				
0.1	0.051	0.213	0.772	0.985	1.000	0.067	0.146	0.596	0.850	0.960
0.5	0.040	0.217	0.766	1.000	1.000	0.065	0.155	0.582	0.871	0.959
1	0.045	0.211	0.752	1.000	1.000	0.061	0.146	0.601	0.856	0.969
5	0.037	0.156	0.653	0.939	0.997	0.049	0.108	0.448	0.725	0.948
10	0.037	0.116	0.453	0.792	0.982	0.053	0.082	0.288	0.523	0.813
	case 3					case 4				
0.1	0.043	0.197	0.737	0.962	1.000	0.051	0.177	0.702	0.970	1.000
0.5	0.044	0.161	0.596	0.909	0.987	0.044	0.141	0.535	0.871	0.982
1	0.045	0.124	0.511	0.826	0.961	0.053	0.104	0.446	0.758	0.922
5	0.065	0.108	0.292	0.519	0.743	0.069	0.088	0.238	0.481	0.688
10	0.078	0.068	0.204	0.408	0.618	0.090	0.063	0.179	0.354	0.571

Table 4

Empirical type I error rates (a=0)of KMR and KDC with the different choice of kernels: linear, quadratic, IBS, and Euclidean distance, where K represents the kernel matrix for the genotypes $\mathbf{Z}$*, and* $\tilde{L}$ *represents the kernel matrix for the adjusted phenotype* $\tilde{Y}$

	Spare effect		Common effect	
	Σ_1	Σ_2	Σ_1	Σ_2
KMR (K=linear)	0.047	0.055	0.050	0.055
KMR (K=quadratic)	0.047	0.056	0.047	0.059
KMR (K=IBS)	0.043	0.060	0.050	0.059
KDC ($\tilde{L}$,K=ED)	0.036	0.059	0.045	0.061
KDC ($\tilde{L}$,K=linear)	0.047	0.055	0.050	0.055
KDC ($\tilde{L}$=linear,K=quadratic)	0.047	0.056	0.047	0.059
KDC ($\tilde{L}$=linear,K=IBS)	0.043	0.060	0.050	0.059
KDC ($\tilde{L}$=quadratic,K=linear)	0.043	0.057	0.052	0.052
KDC ($\tilde{L}$,K=quadratic)	0.043	0.062	0.051	0.061
KDC ($\tilde{L}$=quadratic,K=IBS)	0.034	0.066	0.052	0.061

Table 5

Power (a=0.1, 0.2) of KMR and KDC of case 1: covariate effects included; and case 2: covariate effect excluded. The different choice of kernels are linear, quadratic, IBS, and Euclidean distance for both cases, where K represents the kernel matrix for the genotypes $\mathbf{Z}$ and $\tilde{L}$ represents the kernel matrix for the adjusted phenotype $\tilde{Y}$. Note that $\tilde{\cdot}$ represents the kernel matrices that are adjusted by the covariates.

	Spare effect				Common effect			
	Σ_1		Σ_2		Σ_1		Σ_2	
	a =0.1	a =0.2	a =0.1	a =0.2	a =0.1	a =0.2	a =0.1	a =0.2
Case 1: Covariate effects included								
KMR (K=linear)	0.410	0.984	0.327	0.973	0.336	0.947	0.255	0.960
KMR (K=quadratic)	0.405	0.980	0.322	0.974	0.330	0.949	0.242	0.954
KMR (K=IBS)	0.400	0.977	0.308	0.966	0.330	0.946	0.237	0.946
KDC ($\tilde{L}$,K=ED)	0.364	0.970	0.355	0.982	0.306	0.931	0.304	0.977
KDC ($\tilde{L}$,K=linear)	0.410	0.984	0.327	0.973	0.336	0.947	0.255	0.960
KDC ($\tilde{L}$=linear,K=quadratic)	0.400	0.980	0.322	0.974	0.330	0.949	0.242	0.954
KDC ($\tilde{L}$=linear,K=IBS)	0.400	0.977	0.308	0.966	0.330	0.946	0.237	0.946
KDC ($\tilde{L}$=quadratic,K=linear)	0.252	0.922	0.171	0.756	0.216	0.821	0.136	0.669
KDC ($\tilde{L}$,K=quadratic)	0.253	0.926	0.179	0.748	0.218	0.818	0.133	0.660
KDC ($\tilde{L}$=quadratic,K=IBS)	0.258	0.908	0.177	0.757	0.216	0.807	0.137	0.651
Case 2: Covariate effect excluded								
KMR (K=linear)	0.336	0.947	0.242	0.930	0.278	0.902	0.187	0.884
KMR (K=quadratic)	0.326	0.954	0.237	0.927	0.262	0.896	0.176	0.873
KMR (K=IBS)	0.321	0.942	0.222	0.911	0.261	0.886	0.172	0.863
KDC (L,K=ED)	0.314	0.930	0.266	0.964	0.253	0.886	0.227	0.941
KDC (L,K=linear)	0.336	0.947	0.242	0.930	0.278	0.902	0.187	0.884
KDC (L=linear,K=quadratic)	0.326	0.954	0.237	0.927	0.262	0.896	0.176	0.873
KDC (L=linear,K=IBS)	0.321	0.942	0.222	0.911	0.261	0.886	0.172	0.863
KDC (L=quadratic,K=linear)	0.201	0.876	0.144	0.779	0.172	0.764	0.075	0.587
KDC (L,K=quadratic)	0.206	0.884	0.142	0.786	0.170	0.759	0.078	0.578
KDC (L=quadratic,K=IBS)	0.213	0.875	0.138	0.774	0.158	0.762	0.084	0.574

Table 6

Empirical type I error rates (a=0) and powers (a=0.1, 0.2) of KMR and KDC with Gaussian RBF for $\tilde{L}$ and linear, quadratic, or IBS kernel for K; where K represents the kernel matrix on the genotypes $\mathbf{Z}$ *and $\tilde{L}$ represents the kernel matrix on the adjusted phenotype $\tilde{Y}$*

	Sparse effect						Common effect					
	Σ_1			Σ_2			Σ_1			Σ_2		
	a=0	a=0.1	a=0.2	a=0	a=0.1	a=0.2	a=0	a=0.1	a=0.2	a=0	a=0.1	a=0.2
DC ($\tilde{L}$=RBF ρ= 0.1,K= linear)	0.049	0.404	0.974	0.045	0.367	0.989	0.052	0.340	0.924	0.054	0.342	0.982
DC ($\tilde{L}$=RBF ρ= 0.1,K=quadratic)	0.056	0.392	0.971	0.047	0.363	0.990	0.048	0.337	0.919	0.052	0.330	0.982
DC ($\tilde{L}$=RBF ρ= 0.1,K=IBS)	0.056	0.380	0.966	0.051	0.354	0.986	0.048	0.334	0.905	0.053	0.328	0.976
DC ($\tilde{L}$=RBF ρ=0.5,K=linear)	0.049	0.259	0.831	0.047	0.324	0.965	0.047	0.215	0.748	0.056	0.333	0.956
DC ($\tilde{L}$=RBF ρ=0.5,K=quadratic)	0.051	0.253	0.833	0.045	0.316	0.964	0.049	0.196	0.743	0.056	0.320	0.954
DC ($\tilde{L}$=RBF ρ=0.5,K=IBS)	0.052	0.243	0.822	0.054	0.298	0.958	0.058	0.204	0.730	0.059	0.324	0.953
DC ($\tilde{L}$=RBF ρ=1,K=linear)	0.053	0.188	0.668	0.045	0.277	0.904	0.046	0.154	0.555	0.051	0.272	0.897
DC ($\tilde{L}$=RBF ρ=1,K=quadratic)	0.058	0.188	0.655	0.052	0.262	0.892	0.053	0.154	0.538	0.052	0.270	0.898
DC ($\tilde{L}$=RBF ρ=1,K=IBS)	0.064	0.169	0.660	0.051	0.261	0.888	0.050	0.146	0.542	0.056	0.263	0.893
DC ($\tilde{L}$=RBF ρ=5,K=linear)	0.052	0.086	0.224	0.038	0.126	0.404	0.048	0.080	0.172	0.053	0.127	0.457
DC ($\tilde{L}$=RBF ρ=5,K=quadratic)	0.056	0.082	0.211	0.033	0.123	0.391	0.049	0.080	0.171	0.050	0.131	0.442
DC ($\tilde{L}$=RBF ρ=5,K=IBS)	0.061	0.079	0.215	0.041	0.126	0.420	0.051	0.083	0.171	0.050	0.119	0.474
DC ($\tilde{L}$=RBF ρ=10,K=linear)	0.049	0.067	0.132	0.031	0.094	0.245	0.048	0.070	0.100	0.048	0.098	0.276
DC ($\tilde{L}$=RBF ρ=10,K=quadratic)	0.050	0.068	0.121	0.031	0.087	0.232	0.046	0.065	0.102	0.050	0.101	0.280
DC ($\tilde{L}$=RBF ρ=10,K=IBS)	0.047	0.063	0.123	0.034	0.086	0.257	0.052	0.069	0.102	0.049	0.091	0.287

Table 7

Empirical type I error rates (a=0) and power (a=0.05, 0.1) of KMR and KDC with different choice of kernels: linear, quadratic, IBS and Euclidean distance; where K represents the kernel matrix on the genotypes $\mathbf{Z}$ and $\tilde{L}$ represents the kernel matrix on the adjusted phenotype $\tilde{Y}$

	Size	Power	
	$a = 0.0$	$a = 0.05$	$a = 0.1$
KMR (K=linear)	0.051	0.270	0.928
KMR (K=quadratic)	0.052	0.243	0.887
KMR (K=IBS)	0.058	0.232	0.875
KDC ($\tilde{L}$,K=ED)	0.046	0.378	0.974
KDC ($\tilde{L}$,K=linear)	0.051	0.270	0.928
KDC ($\tilde{L}$=linear,K=quadratic)	0.052	0.243	0.887
KDC ($\tilde{L}$=linear,K=IBS)	0.058	0.232	0.875
KDC ($\tilde{L}$=quadratic,K=linear)	0.056	0.067	0.145
KDC ($\tilde{L}$,K=quadratic)	0.054	0.058	0.121
KDC ($\tilde{L}$=quadratic,K=IBS)	0.055	0.068	0.165

Table 8

Empirical type I error rates (a=0) and powers (a=0.05, 0.1) of case (a): Covariate effects included in the KDC test with a Gaussian RBF kernel for $\tilde{L}$ and the linear, quadratic, or IBS kernel for K; and case (b) Covariate effects excluded in the KDC test with Gaussian RBF for L and the linear, quadratic or IBS kernel for K.

	Size			Power					
	$a = 0.0$			$a = 0.05$			$a = 0.1$		
	Linear	Quadratic	IBS	Linear	Quadratic	IBS	Linear	Quadratic	IBS
Case 1: Covariate effects included									
0.1	0.050	0.054	0.047	0.423	0.374	0.379	0.976	0.949	0.948
0.5	0.046	0.046	0.047	0.469	0.429	0.422	0.899	0.854	0.893
1	0.046	0.045	0.042	0.481	0.428	0.434	0.781	0.732	0.803
5	0.039	0.036	0.040	0.315	0.288	0.337	0.389	0.337	0.437
10	0.049	0.044	0.037	0.219	0.211	0.256	0.261	0.229	0.308
Case 2: Covariate effect excluded									
0.1	0.041	0.046	0.043	0.084	0.083	0.073	0.209	0.173	0.189
0.5	0.047	0.044	0.043	0.054	0.054	0.055	0.102	0.089	0.104
1	0.045	0.044	0.049	0.059	0.047	0.049	0.073	0.062	0.070
5	0.038	0.043	0.043	0.057	0.052	0.059	0.051	0.051	0.050
10	0.043	0.041	0.045	0.053	0.057	0.055	0.054	0.045	0.059

Table 9
Real data results: p-values of KMR and KDC with different choice of kernels

Test	p-values
KMR (K=linear)	0.063
KMR (K=quadratic)	0.029
KMR (K=IBS)	0.267
KDC ($\tilde{L}$,K=ED)	0.071
KDC ($\tilde{L}$,K=linear)	0.063
KDC ($\tilde{L}$=linear,K=quadratic)	0.029
KDC ($\tilde{L}$=linear,K=IBS)	0.267
KDC ($\tilde{L}$=quadratic,K=linear)	0.073
KDC ($\tilde{L}$,K=quadratic)	0.035
KDC ($\tilde{L}$=quadratic,K=IBS)	0.270