 DOI: 10.1111/biom.12314

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

Wen-Yu Hua[1,*] and Debashis Ghosh[2]

[1]Department of Child and Adolescent Psychiatry, New York University School of Medicine, New York, New York, U.S.A.

[2]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, U.S.A.

[*]*email:* littlehanag@gmail.com

Summary. Associating genetic markers with a multidimensional phenotype is an important yet challenging problem. In this work, we establish the equivalence between two popular methods: kernel-machine regression (KMR), and kernel distance covariance (KDC). KMR is a semiparametric regression framework that models covariate effects parametrically and genetic markers non-parametrically, while KDC represents a class of methods that include distance covariance (DC) and Hilbert–Schmidt independence criterion (HSIC), which are nonparametric tests of independence. We show that the equivalence between the score test of KMR and the KDC statistic under certain conditions can lead to a novel generalization of the KDC test that incorporates covariates. Our contributions are 3-fold: (1) establishing the equivalence between KMR and KDC; (2) showing that the principles of KMR can be applied to the interpretation of KDC; (3) the development of a broader class of KDC statistics, where the class members are statistics corresponding to different kernel combinations. Finally, we perform simulation studies and an analysis of real data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The ADNI study suggest that SNPs of *FLJ16124* exhibit pairwise interaction effects that are strongly correlated to the changes of brain region volumes.

Key words: Confounding; Distance covariance; Hilbert–Schmidt independence criterion; Neuroimaging genomics; Permutation test.

1. Introduction

To better understand and utilize genomic data, researchers often study the associations between genetic variants and disease phenotypes to decode the hidden information. To this end, intermediate phenotypes have been attracting much attention compared to the final disease diagnosis, since intermediate phenotypes have the potential to contain stronger connections to the genetic variants. In the case of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, Ge et al. (2012), Stein et al. (2010a), and Stein et al. (2010b) used the brain structural magnetic resonance imaging (MRI) scans as the intermediate multiple phenotypes. They looked for associations between genetic variants and phenotypes; the detailed descriptions of the ADNI dataset are in Section 5. In this work, we primarily focus on the problem of correlating the genetic variants with the imaging data such as those from ADNI, while adjusting for environmental covariates.

In earlier work on the ADNI study, Stein et al. (2010b) applied linear regression using one genetic marker versus one phenotype in a massively univariate manner across all markers and all phenotypes. Such a method is feasible if the number of genetic variants and phenotypes is small. When both the dimension of genotypes and phenotypes is very large, the resulting test has limited power due to the issue of multiple comparisons. To address this issue, two popular modeling frameworks that could potentially be applied to the motivating example are the multivariate kernel machine regression model (MV-KMR) (Maity, Sullivan, and Tzeng, 2012) and the kernel distance covariance method (KDC) (Székely, Rizzo, and Bakirov, 2007; Gretton et al., 2008; Székely and Rizzo, 2009).

MV-KMR is a multivariate outcome regression framework based on kernel machine regression (KMR). KMR is a semiparametric regression approach that models the effect of the covariates parametrically, and the effect of the genetic markers non-parametrically (Liu, Lin, and Ghosh, 2007; Kwee et al., 2008). Specifically, the non-parametric effect of multiple markers is modeled by a kernel. The Gaussian Radial Basis Function (RBF) kernel is frequently used for quantitative measurements, while polynomial kernels can be considered for the qualitative variables. One advantage of this approach is that it is able to greatly simplify the specification of a nonparametric model for the effect of multiple markers (Liu et al., 2007). Since we focus our discussion on multivariate phenotypes in this work, we simply use KMR to denote the generic approach for both univariate and multiple phenotypes.

KDC is a term that we define as a class of tests for independence, and it includes distance covariance (DC) and the Hilbert–Schmidt independence criterion (HSIC). DC was established by Székely et al. (2007) to provide a test of independence in high-dimensional settings that is consistent against all alternatives. One advantage of DC is the compact representation of the corresponding statistic in terms of the product of expectations of pairwise Euclidean distances (L_2).

The statistic can be estimated empirically in a straightforward manner. On the other hand, Gretton et al. (2008) formulated the two-variable independence test (HSIC) using Reproducing Kernel Hilbert Spaces (RKHS), where the HSIC statistic is able to test for dependence in multivariate spaces, and it is consistent when a characteristic kernel (Sriperumbudur et al., 2010) is used. Sejdinovic et al. (2013) demonstrated that DC and HSIC are the same when the distance-induced kernel in HSIC is chosen. The results in Sejdinovic et al. (2013) showed that the HSIC test is more sensitive when the test statistic is derived from kernels, and that the HSIC tests can be readily extended to more structured and non-Euclidean spaces.

In this work, we establish the equivalence between KMR and KDC. We first provide some review on distance properties of the sum of squares, and the outer product formulation of linear models. This is then followed by an algebraic representation of the multiple phenotypes version of KMR. We show that the KDC statistic is equivalent to the KMR in the absence of a parametric component and when a linear kernel is used for the phenotype spaces. Furthermore, we propose a new covariate-adjusted KDC test in the presence of the covariates, and show that KDC is equivalent to the KMR of Maity et al. (2012). Three major implications of the equivalence are established in this work. First, the equivalence shows that the principles of KMR can be applied to the interpretation of KDC. Second, the new proposed covariate-adjusted KDC test shows an increase in power relative to the original KDC test in our simulation studies. Third, the KMR statistic is a member of the KDC family, in that the members correspond to distinct kernels. To conclude, our experiments suggest that KDC may yield a more powerful result when tailored to the application at hand.

2. Preliminaries

2.1. *Distance Properties of the Sum of Squares*

For each $i, j \in 1, \ldots, n$, $D = (d_{ij})$ is an L_2 of $\mathbf{Y} \in \mathbb{R}^p$, where each component of d_{ij} is the square root of $d_{ij}^2 = \sum_{r=1}^{p}(y_{ir} - y_{jr})^2$. Consider $A \equiv (a_{ij})$ as the sum of squares of $\mathbf{Y}$ with (i, j)th entry, such that $a_{ij} = \sum_{r=1}^{p}(y_{ir} - \bar{y})(y_{jr} - \bar{y})$, where $\bar{y}$ is the total average of $\mathbf{Y}$. We can also express the sum of squares A as the centred inner product matrix,

$$A = \left(I - \frac{1_n 1_n'}{n}\right)\mathbf{Y}\mathbf{Y}'\left(I - \frac{1_n 1_n'}{n}\right). \tag{1}$$

where 1_n is a $n \times 1$ vector consisting of all ones, and I is an $n \times n$ identity matrix. Gower (1966) showed that if $\mathbf{Y}\mathbf{Y}'$ in (1) is replaced by $-\frac{1}{2}D^2 = -\frac{1}{2}(d_{ij}^2)$ (denoted as a Gower distance), then the sum of squares A can be interpreted as a centred distance matrix. In addition, Ch. 14 in Mardia et al. (1980) proved that D is Euclidean if and only if A is a positive semi-definite matrix.

2.2. *Linear Model*

Suppose for n subjects we observe the response $\mathbf{Y} \in \mathbb{R}^p$ and the predictor $\mathbf{Z} \in \mathbb{R}^q$. A typical approach to model the relationship between $\mathbf{Y}$ and $\mathbf{Z}$ is to apply multivariate analysis of variance (MANOVA). Traditional multivariate analysis proceeds through partitioning of the total sum of squares based on the trace of $\mathbf{Y}'\mathbf{Y}$, and the analysis can be done using the linear model $\mathbf{Y} = \mathbf{Z}\beta + \epsilon$. For a test of the association between $\mathbf{Y}$ and $\mathbf{Z}$, we formulate the null hypothesis as $H_0 : \beta = 0$, and the least-squares estimate of β is $\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. Therefore, the fitted values of $\mathbf{Y}$ are $\hat{\mathbf{Y}} = \mathbf{Z}\hat{\beta} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$, and the matrix of residuals is $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$, and $tr(\mathbf{Y}'\mathbf{Y}) = tr(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + tr(\mathbf{R}'\mathbf{R})$, where $tr(\cdot)$ is the trace operator. An appropriate statistic to test the null hypothesis of no association between $\mathbf{Y}$ and $\mathbf{Z}$ is a pseudo F-statistic (McArdle and Anderson, 2001):

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

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

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

If $\mathbf{D}$ is a Gower distance matrix, then (3) is the same as (2); if $\mathbf{D}$ is some other distance matrix, then the significance of (3) can be tested using the permutation technique.

3. Methods

McArdle and Anderson (2001) applied the outer product formulation to extend MANOVA to more general settings using distance matrices. The advantage is that when the outer product space is considered, one can apply other distances for measuring the difference between two observations. Therefore, the estimate is more flexible in capturing the nature of both the response and the predictor variables. This inspires us to apply the same method to the KMR model, and the resulting representation shows that KMR is equivalent to the KDC when a common kernel is chosen. To understand this equivalence between KMR and KDC, we first review the KMR model without the covariates. We then consider the situation when covariates are included, which leads to our proposal of a new covariate-adjusted KDC test.

3.1. *Without Covariates*

To test dependence between two random vectors, that is, the association between the phenotypes $\mathbf{Y} = (Y_1, \ldots, Y_n)^t \in \mathbb{R}^p$ and the genotypes $\mathbf{Z} = (Z_1, \ldots, Z_n)^t \in \mathbb{R}^q$, KDC (i.e., DC or HSIC) can be used. Here, we use the same algebraic formulation as in Gretton et al. (2008), and denote the kernel function $k_{ij} = k(Z_i, Z_j)$ as the element inxw 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 for the association between $\mathbf{Y}$ and $\mathbf{Z}$ is defined as

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

where $H = (I - 1_n 1_n'/n)$ is a centering matrix, I is an identity matrix of size n, and 1_n is a $n \times 1$ vector with each element equal to 1. If both k and l are L_2 distance kernels, then (4) is the DC statistic (Székely et al., 2007); if other reproducing kernels are applied (4) is the HSIC statistic (Gretton et al., 2008). In summary, the KDC statistic is used for testing the dependence between $\mathbf{Y}$ and $\mathbf{Z}$, where we do not have to assume a certain distribution of $\mathbf{Y}$ or $\mathbf{Z}$.

Another powerful test of dependence can be derived using KMR, which we now briefly discuss. 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 the non-parametric effect of genotypes on the univariate response, Y, and it is determined by a specified positive semi-definite kernel function $k(.,.)$. To test the hypothesis that $h(\cdot) = 0$, Liu et al. (2007) proposed a hierarchical Gaussian process regression for the linear model (5):

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

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

$$\begin{aligned} Q &\propto (Y - \bar{Y})' K (Y - \bar{Y}) \\ &= tr[(Y - \bar{Y})' K (Y - \bar{Y})] \\ &= tr[K(Y - \bar{Y})(Y - \bar{Y})'] \\ &= tr\left[K\left(I - \frac{1_n 1_n'}{n}\right) YY' \left(I - \frac{1_n 1_n'}{n}\right)\right] \\ &= tr[KHYY'H] \end{aligned} \tag{6}$$

By using the property of "trace of a product" (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 phenotype $\mathbf{Y}$, and therefore $tr[KH\mathbf{Y}\mathbf{Y}'H]$ is equivalent to MV-KMR in Maity et al. (2012) in the absence of covariates. Second, a common kernel is used in K for both KMR and KDC, and by replacing the outer product YY' with any distance matrix L in (6) results in the equivalence of KMR and KDC in (4).

3.2. *With Covariates*

In practice, we may want to know the relationship between the genotypes ($\mathbf{Z}$) and phenotypes ($\mathbf{Y}$) adjusting for covariates ($\mathbf{X}$), with n observed 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 is

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

where $h(\cdot)$ is an non-parametric function which describes the effect of $\mathbf{Z}$ on $\mathbf{Y}$ adjusting for $\mathbf{X}$. To test the effect of $\mathbf{Z}$, one can test $H_0 : \tau_1 = \cdots = \tau_p = 0$ under the following representation that is a multivariate extension of the hierarchical Gaussian process regression from the previous section:

$$\mathbf{Y}|(\boldsymbol{\beta}, h(\mathbf{Z})) \sim MVN\{\mathbf{X}\boldsymbol{\beta} + h(\mathbf{Z}), \Sigma\}, \qquad h(\cdot) \sim GP\{0, \boldsymbol{\tau} K\},$$

and the corresponding score test of H_0 is proportional to

$$\begin{aligned} Q &\propto (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' K (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= tr[(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' K (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})] \\ &= tr[(\tilde{\mathbf{Y}} - \bar{\tilde{\mathbf{Y}}})' K (\tilde{\mathbf{Y}} - \bar{\tilde{\mathbf{Y}}})] \\ &= tr[KH\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'H], \end{aligned} \tag{7}$$

where $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, and $\bar{\tilde{\mathbf{Y}}}$ is the average of $\tilde{\mathbf{Y}}$ in (7) with H being a centering offset (normalized constant) $(I - 1_n 1_n'/n)$. Note that $\hat{\boldsymbol{\beta}}$ is the MANOVA estimate in Section 2.2. Hence (7) is equivalent to the score test in KMR, and the outer product $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'$ can be replaced with any distance measure $\tilde{L}$ such that (7) becomes

$$tr[KH\tilde{L}H]. \tag{8}$$

The original KDC was presented as a test of independence between $\mathbf{Y}$ and $\mathbf{Z}$. Here, we extend it to the case when covariates are present. Specifically, by applying a common kernel on K for both statistics, and if $\tilde{L}$ is calculated by a linear kernel, then KDC in (8) is again equivalent to (7) for the KMR. For both cases when covariate effects are absent or present in (6) and (7), respectively, we ignore the covariance matrix structure of $\mathbf{Y}$ in order to establish the connection between KDC and KMR. Our idea for this step is similar to the work in Pan (2011) that treats the covariance term as the fixed effects, while the covariance of $\mathbf{Y}$ is modeled in this work. We demonstrate that the covariance of $\mathbf{Y}$ can be captured by choosing a suitable kernel matrix in the simulation studies.

4. Numerical Simulations

The goal of the following simulations is to evaluate the performance of KMR and KDC in terms of the empirical size and power under different kernel combinations.

4.1. *Kernel Choices*

Many kernel choices exist for characterizing the similarity of individuals with respect to the variations of genotypes and phenotypes. In this work, we consider the identity-by-state (IBS), L_2, Gaussian RBF, linear and quadratic kernels for our numerical analyses. These are their definitions:

(1) IBS kernel: $k(Z_i, Z_j) = (2q)^{-1} \sum_{r=1}^{q} (2 - |Z_{ir} - Z_{jr}|)$, where q is the number of loci considered in the calculation.
(2) L_2 kernel: $k(Z_i, Z_j) = \|Z_i - Z_j\|_q = \sqrt{\sum_{r=1}^{q} (Z_{ir} - Z_{jr})^2}$.
(3) Gaussian RBF kernel: $k(Z_i, Z_j) = \exp\{-\rho \|Z_i - Z_j\|_q^2\}$, where ρ is the weight parameter.
(4) Polynomial kernel: $k(Z_i, Z_j) = (\langle Z_i, Z_j \rangle + c)^d$, where $\langle Z_i, Z_j \rangle$ denotes the inner product of Z_i and Z_j, c and d are the constants.

	Size (%)	Power (%)			
a	0	25	50	75	100
KMR (K =linear)	4.4	22.7	75.9	97.0	99.6
KMR (K =quadratic)	3.5	19.8	70.6	96.2	99.5
KDC ($\tilde{L}$,K= linear)	4.9	21.2	72.5	96.7	99.5
KDC ($\tilde{L}$,K=L_2)	4.4	22.7	75.9	97.0	99.6
KDC ($\tilde{L}$ =linear,K =quadratic)	3.5	19.8	70.6	96.2	99.5
KDC ($\tilde{L}$ =quadratic,K =linear)	4.1	17.3	58.5	87.7	93.3
KDC ($\tilde{L}$,K =quadratic)	3.9	14.6	51.9	82.2	90.8

Figure 1. Empirical size and power of KMR and KDC, represented as a heatmap with different kernels: linear, quadratic, and L_2 distance. K represents the kernel matrix for the genotypes $\mathbf{Z}$, and $\tilde{L}$ represents the kernel matrix for the adjusted phenotype $\tilde{Y}$. The header a's and content values are displayed in percentages. The heatmap colors are encoded by each column (given an a value): two tests have the same color which indicates that their empirical values are the same.

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 $d = 2$.

4.2. *Simulation 1*

The first simulation examined the association between a single phenotype Y and a multivariate genotype $\mathbf{Z}$ adjusted by a single covariate X, and the design of the simulation was based on Liu et al. (2007) where the model and parameter settings were chosen according to the data from the Michigan prostate cancer study (Dhanasekaran et al., 2001). We used the same simulation study to compare the performance of KMR and KDC. The true linear model was $Y = \beta_0 + \beta_1 X + h(Z_1, \ldots, Z_q) + \epsilon$, 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 $Z_j's (j = 1, \ldots, 5)$ were generated from uniform(0,1) while u and ϵ were generated from the standard normal distribution.

To estimate the coefficients, we adopted the same procedure from Liu et al. (2007) which assumed the effect h was zero, and used the *lm* function from the stat package in R to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, and set $\tilde{Y} = Y - \hat{\beta}_0 - \hat{\beta}_1 X$. The empirical size and power of the association tests between $\tilde{Y}$ and $\mathbf{Z}$ were evaluated by generating data under $a = 0$ and $a = 25\%, 50\%, 75\%, 100\%$ at the significance level of 0.05. The sample size was 60; the p-value of the statistic was 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 $\tilde{Y}$.

Figure 1 shows the results of the empirical size and power of the KMR and KDC tests, where linear and quadratic kernels were used for K in KMR, and the L_2 distance, linear and quadratic kernels were used in both K and $\tilde{L}$ in KDC. As seen in the heatmap, these two tests had the same color which indicates that their empirical values are the same, therefore providing a numerical validation of our claim of equivalence between KMR and KDC in the linear case. The performance of the quadratic kernel resulted in lower power relative to the other kernel combinations. This suggests that for this simulation scenario, a linear kernel or L_2 distance is sufficient.

4.3. *Simulation 2*

For the second simulation, the design was based on Maity et al. (2012), where the authors used estimates from the CATIE study (Lieberman et al., 2005). The part of the CATIE study relevant to our purposes focused on determining the association between schizophrenia-associated SNPs with antibody responses of three neurotrophic herpesviruses. Here, we wished to study the performance of KMR and KDC with correlated responses. For $k = 1, 2, 3$, the data were generated from the model

$$Y_k = \mathbf{X}\boldsymbol{\beta}_k + h_k(\mathbf{Z}) + \epsilon_k, \tag{9}$$

where $\mathbf{X} = (X_1, X_2)^T$ were generated from bivariate normal distribution with mean vector $(0.2, 0.4)^T$ and identity variance-covariance matrix. The ϵ_k were generated from a multivariate normal distribution with mean zero vector and variance-covariance matrix Σ, which will be defined shortly. The q-dimensional SNP genotype data $\mathbf{Z} \equiv (Z_1, \ldots, Z_q)$, with $q = 9$, were simulated using the linkage disequilibrium structure of the gene *SLC17A1*. Two choices for the effects of h_k were considered. The first was the sparse effect, where $h_1 = a(z_1 + z_2 + z_3 + z_1z_4z_5 - z_6/3 - z_7z_8/2 + (1 - z_9))$, $h_2 = h_3 = 0$, and $a = 0, 10\%, 20\%$. The second was the common effect, where $h_1^* = h_1 + az_3$, and $h_2 = h_3 = az_3$ with $a = 0, 10\%, 20\%$. In addition, we also investigated the performance of KMR and KDC by setting the variance–covariance matrix Σ to have an independent structure ($\Sigma = \Sigma_1$) and a more dependent structure ($\Sigma = \Sigma_2$), where

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

The empirical size ($a = 0$) and power ($a = 10\%, 20\%$) were examined at a significance level of 5%. The sample size, n, was 100, the dimension of genotypes, q, was 9, and the dimension of phenotypes, p, was 3. To adjust for $\mathbf{X}$, we again used the *lm* function from the stat package in R to obtain $\hat{\boldsymbol{\beta}}$, so that $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\mathbf{Y} = (Y_1, \ldots, Y_p)$, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$. The linear, quadratic, IBS and L_2 distance kernels were used in this simulation.

The empirical size of all tests for both sparse and common effects were all close to 0.05 (Web Figure 1), which suggests that the tests were able to control type I error. Figure 2 displays the empirical power of case 1 when covariates were adjusted for (test of independence between $\mathbf{Z}$ and $\tilde{\mathbf{Y}}$). In Case 2, the covariates were excluded (test of independence between $\mathbf{Z}$ and $\mathbf{Y}$). The results showed that case 1 had greater power than case 2. This suggests that our proposed covariate-adjusted KDC test results in an increase in power. However, the power gain depends on the choice of Σ.

Overall, when the independent covariance matrix Σ_1 was used, the KMR and KDC tests computed by a linear kernel consistently resulted in good performance in terms of power in both sparse and common effect. When the common effect and the dependent structure Σ_2 were considered, the KDC with an L_2 kernel on both L and K resulted in the largest empirical power. This suggests that the L_2 kernel was able to

	Sparse effect (%)				Common effect (%)			
	Σ_1		Σ_2		Σ_1		Σ_2	
a	10	20	10	20	10	20	10	20
Case 1: Covariate effects included								
KMR (K =linear)	41.0	98.4	32.7	97.3	33.6	94.7	25.5	96.0
KMR (K =quadratic)	40.5	98.0	32.2	97.4	33.0	94.9	24.2	95.4
KMR (K =IBS)	40.0	97.7	30.8	96.6	33.0	94.6	23.7	94.6
KDC ($\tilde{L}$,K= L_2)	36.4	97.0	35.5	98.2	30.6	93.1	30.4	97.7
KDC ($\tilde{L}$,K= linear)	41.0	98.4	32.7	97.3	33.6	94.7	25.5	96.0
KDC ($\tilde{L}$ =linear,K =quadratic)	40.0	98.0	32.2	97.4	33.0	94.9	24.2	95.4
KDC ($\tilde{L}$ =linear,K =IBS)	40.0	97.7	30.8	96.6	33.0	94.6	23.7	94.6
KDC ($\tilde{L}$ = quadratic,K =linear)	25.2	92.2	17.1	75.6	21.6	82.1	13.6	66.9
KDC ($\tilde{L}$,K= quadratic)	25.3	92.6	17.9	74.8	21.8	81.8	13.3	66.0
KDC ($\tilde{L}$ =quadratic,K =IBS)	25.8	90.8	17.7	75.7	21.6	80.7	13.7	65.1
Case 2: Covariate effects excluded								
KMR (K =linear)	33.6	94.7	24.2	93.0	27.8	90.2	18.7	88.4
KMR (K =quadratic)	32.6	95.4	23.7	92.7	26.2	89.6	17.6	87.3
KMR (K =IBS)	32.1	94.2	22.2	91.1	26.1	88.6	17.2	86.3
KDC ($\tilde{L}$,K= L_2)	31.4	93.0	26.6	96.4	25.3	88.6	22.7	94.1
KDC ($\tilde{L}$,K= linear)	33.6	94.7	24.2	93.0	27.8	90.2	18.7	88.4
KDC ($\tilde{L}$ =linear,K =quadratic)	32.6	95.4	23.7	92.7	26.2	89.6	17.6	87.3
KDC ($\tilde{L}$ =linear,K =IBS)	32.1	94.2	22.2	91.1	26.1	88.6	17.2	86.3
KDC ($\tilde{L}$ = quadratic,K =linear)	20.1	87.6	14.4	77.9	17.2	76.4	7.5	58.7
KDC ($\tilde{L}$,K= quadratic)	20.6	88.4	14.2	78.6	17.0	75.9	7.8	57.8
KDC ($\tilde{L}$ =quadratic,K =IBS)	21.3	87.5	13.8	77.4	15.8	76.2	8.4	57.4

Figure 2. Power (a=10%, 20%) of KMR and KDC represented as a heatmap, including case 1: covariates included; and case 2: covariate effect excluded. The different choice of kernels are linear, quadratic, IBS, and L_2 distance for both cases. 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. The header a's and content values are displayed in percentages. The heatmap colors are encoded by each column (given an a value): two tests have the same color which indicates that their empirical values are the same.

incorporate the information from the dependent covariance structure, that is, Σ_2, and identify the association between $\mathbf{Y}$ and $\mathbf{Z}$ at the same time.

5. Experiments with the Alzheimer Disease Neuroimaging Initiative (ADNI) Study

We now evaluate our approaches using simulated and real data from the ADNI study. Data used in the preparation of this article were obtained from the ADNI database. One of the goals of the ADNI study is to perform genome-wide association tests, and identify the genetic variants that influence the voxel-level differences in brain MRI. We used parts of data from the ADNI study for the association test, where the multiple phenotypes are the T_1 weighted brain structural MRI scans (31,662 brain voxels). Rather than using the original MRI scans, where a voxel's intensity value represents the anatomical structure of the scanned subject, here we used a tensor-based-morphometry (TBM) to compute a 3D map. This map gives the difference in brain volume between each scanned individual and an average brain template based on healthy elderly subjects. Thus, a voxel's intensity value represents the volumetric difference from that of a healthy brain. The genotypes are encoded by 448,244 SNPs across the entire genome, and the demographic variables include gender and age.

In this work, the phenotypes were grouped from 31,662 total voxels into 119 regions-of-interest (ROIs), where the mapping of the ROIs was based on the GSK CIC Atlas (Tziortzi et al., 2011). The average voxel value of each region was used to represent each of the 119 ROIs. With these 119 regions, Hua, Nichols, and Ghosh (2015)used the DC test and discovered that the difference in brain volumes were highly associated with a common variant *rs*11891634 in the intron region of gene *FLJ16124*, with a total of 141 SNPs within gene *FLJ16124* that were identified by the SNP-gene mapping from Hibar et al. (2011). The subject pool consists of 741 subjects from the ADNI study that have passed the quality control filtering according to Stein et al. (2010b), which we retained for the simulation and real data analysis. Figure 3a describes the data preparation.

5.1. *Simulation Based on ADNI*

For this simulation study, a partial linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + h(\mathbf{Z}) + \boldsymbol{\epsilon}$ was used. To mimic the ADNI samples, a correlated structure among the phenotypes was considered, and this design was based on Vounou, Nichols, and Montana

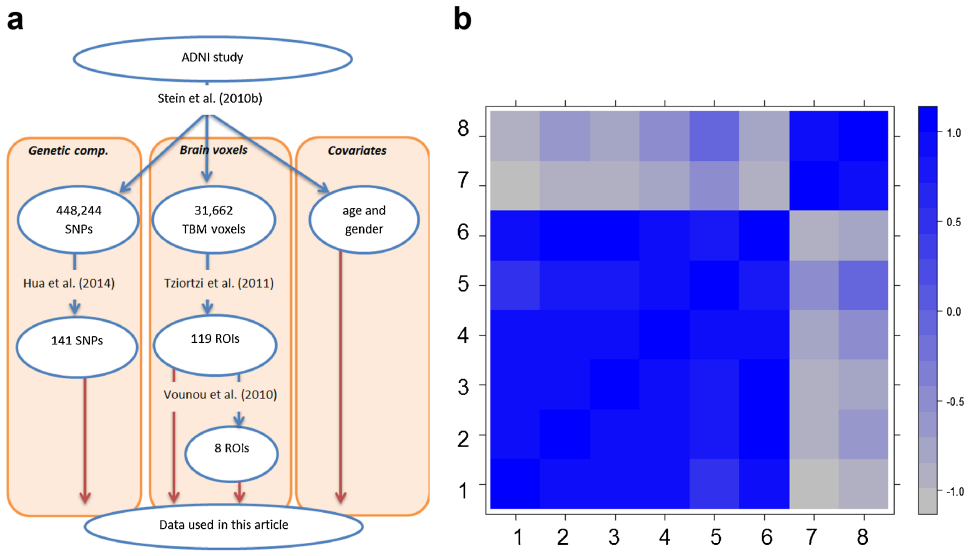

Figure 3. (a) Flow chart of the data preparation for our real data analysis; (b) all-pairwise correlations of the 358 MCI subjects using eight ($q = 8$) prefrontal cortex regions (1–2: left and right (l/r) anterior dorsolateral, 3–4: l/r posterior dorsolateral, 5–6: l/r anterior medial, 7–8: l/r posterior medial) of the 119 region of interests (ROIs) for the corrected structure in ϵ.

(2010), where the authors suggest to use the frontal cortex regions according to the GSK CIC atlas (Tziortzi et al., 2011), and estimate the pairwise correlations from those regions using subjects from the ADNI dataset with mild cognitive impairment (MCI). We followed the same procedure as in (Vounou et al., 2010) and used eight ($p = 8$) frontal cortex regions of the 119 ROIs. Figure 3b shows all-pairwise correlations of the eight frontal cortex regions that were based on the 358 MCI subjects.

We then estimated the eight ROIs' covariance matrix Σ using data from the 358 subjects with MCI. They were selected for the simulation design due to their relatively uniform MRI voxel values, and ϵ was then simulated from $MVN(0, \Sigma)$. For the genotypes elements, all 141 SNPs on gene *FLJ16124* were used, that is, $\mathbf{Z} = (Z_1, \ldots, Z_{141})$. The effect of h is defined as $h(Z_1, \ldots, Z_{141}) = a \times h_1$, with only the first 5 SNPs, $(Z_1, \ldots, Z_5)$ of 141 Z's were the causative SNPs, such that $h_1(Z_1, \ldots, Z_5) = 2\cos(Z_1) - 3Z_2^2 + 2\exp(-Z_3)Z_4 - 1.6\sin(Z_5)\cos(Z_3) + 4Z_1Z_5$. 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 distribution. A total of 100 samples were generated, and the empirical size $(a = 0)$ and power $(a = 5\%, 10\%)$ were computed based on 10^4 permutations. This simulation was repeated 1000 times and the significance level was set at 0.05.

Figure 4 shows the results of empirical size and power with the linear, quadratic, IBS and L_2 distance kernel of KMR and KDC. The KDC test with the L_2 distance measure resulted in the highest power among all the evaluated kernels. We also implemented the KDC test with a Gaussian RBF kernel for $\tilde{L}$ ($\rho \in 0.1, 0.5, 1, 5, 10$), and linear, quadratic, and IBS kernel for K, and the results can be found in Web Table 1. The highest power was observed when a Gaussian RBF kernel (ρ=0.1) was used for $\tilde{L}$ and a linear kernel for K of the KDC test (5%, 42.3%, and 97.6% when $a = 0, 5\%$, and 10%, respectively). It was close to the results of KDC with the L_2 distance kernel in Figure 4. This suggests that when the dimensions of phenotype and genotype are both very high, both the Gaussian RBF kernel (with optimal ρ) and the L_2 distance kernel are able to model the high-dimensional interactions which results in more powerful performance.

5.2. *Real Data Analysis*

The KMR and KDC tests were conducted to find the associations between the genetic variants and the multivariate

	Size (%)	Power (%)	
a	0	5	10
KMR (K=linear)	5.1	27.0	92.8
KMR (K=quadratic)	5.2	24.3	88.7
KMR (K=IBS)	5.8	23.2	87.5
KDC ($\tilde{L}$,K=L_2)	4.6	37.8	97.4
KDC ($\tilde{L}$,K= linear)	5.1	27.0	92.8
KDC ($\tilde{L}$ =linear,K =quadratic)	5.2	24.3	88.7
KDC ($\tilde{L}$ =linear,K =IBS)	5.8	23.2	87.5
KDC ($\tilde{L}$ = quadratic,K =linear)	5.6	6.7	14.5
KDC ($\tilde{L}$,K= quadratic)	5.4	5.8	12.1
KDC ($\tilde{L}$ =quadratic,K =IBS)	5.5	6.8	16.5

Figure 4. Empirical type I error rate (a=0) and power (a=5%, 10%) of KMR and KDC with different choice of kernels: linear, quadratic, IBS and L_2 distance. K represents the kernel matrix on the genotypes $\mathbf{Z}$, and $\tilde{L}$ represents the kernel matrix on the adjusted phenotype $\tilde{Y}$. The header a's and content values are displayed in percentages. The heatmap colors are encoded by each column (given an a value): two tests have the same color which indicates that their empirical values are the same.

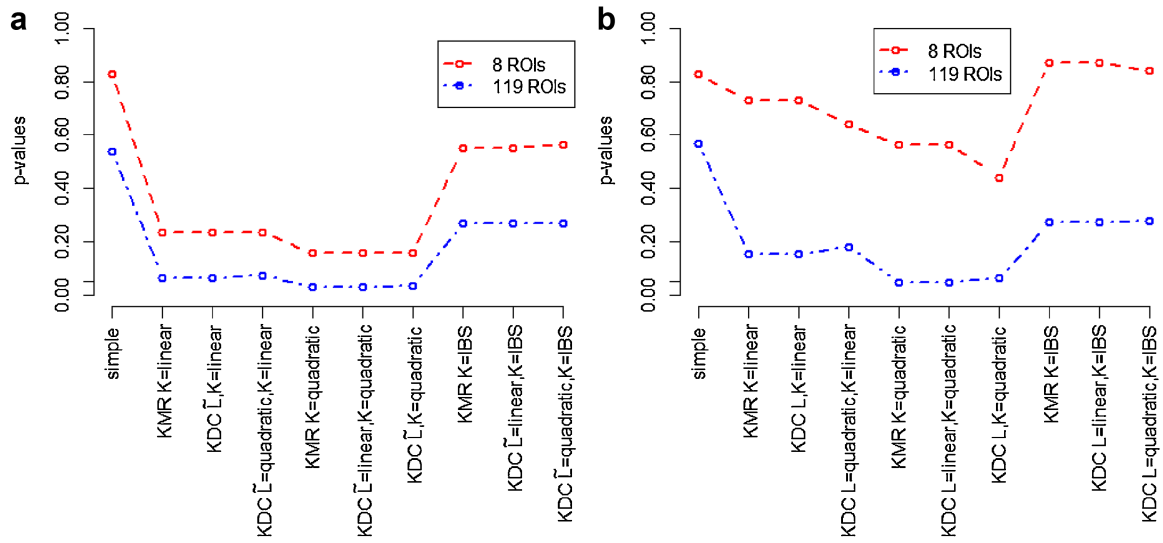

Figure 5. p-values of tests with (a) covariates included; and (b) covariate effect excluded. The dashed line shows the results based on the eight frontal cortex ROIs, and the dash-dot line represents the results based on all 119 ROIs.

brain phenotypes with the demographic variables. We first designed the analysis using all 741 subjects, 141 SNPs at gene *FLJ16124*, eight frontal cortex regions, and two covariates, that is, gender and age. In addition to using the eight ROIs, we also considered a setting in which the entire 119 ROIs were used as the multivariate response. Other than the KMR and KDC test, we also added one more method as the baseline comparison in which we extracted the first principal component (PC) from both $\mathbf{Z}$ and $\mathbf{Y}/\tilde{\mathbf{Y}}$, and applied Pearson's r to estimate their correlations. We denote this as the "simple" method.

Figure 5a and b display the p-values with covariate effects (left) and without covariate effects (right) of all the tests for both the eight frontal cortex ROI and the 119 ROI settings, where p-values were based on the 10^4 permutations.

From both plots of Figure 2, the p-values of KMR were equivalent to KDC when a linear kernel was used for the phenotype space. In addition to the equivalence, the p-values with the covariates were all less than the p-values without the covariates, and this implies that the proposed covariate-adjusted KDC method is able to better detect the associations between the 141 SNPs and the 119 brain regions, with the demographic effects being taken into consideration. Also, the dash-dot line was under the dashed line from both Figure 5a and b, meaning that the p-values from the 119 ROIs were more significant than the results from the eight frontal cortex ROI setting, which suggests that the use of the entire 119 ROIs exhibit stronger associations to the genetic variants.

KMR (with $K =$ quadratic) and KDC (with $\tilde{L} =$ linear, $K =$ quadratic) both identified the smallest p-value with the covariate effects, meaning that applying a quadratic kernel on the 141 SNPs yields the most powerful result among all other evaluated kernels. This implies that there exist strong pairwise interaction effects among the SNPs in *FLJ16124* that associate with brain region volume changes when age and gender are included. Finally, the simple kernels outperformed other kernel combinations, but the simple method did not. This shows that the KDC test with simple kernel combination ($\tilde{L}$=linear,K=quadratic) is able to detect non-linear correlations in the dataset, while the simple method (PC analysis + Pearson's r) is not.

Furthermore, it is important to know whether to use the eight frontal cortex ROIs or the entire 119 ROIs as the multi-dimensional response for the ADNI genetic association study. To investigate this, we proposed to adopt the jacknife (leave-one-out) procedure for ranking the SNPs' importance as a criteria, and focused on the cases where the covariate effects were included in the analysis, since the p-values from Figure 5a were all smaller than in Figure 5b. The steps of the jacknife procedure were as follows:

(1) Given a kernel pair, compute the KDC statistic between the 141 SNPs and the eight ROIs adjusted by age and gender, denoted as T.
(2) Remove the ith SNP and compute the KDC statistic between the 140 SNPs and the eight ROIs adjusted by age and gender, denoted as T_{-i}.
(3) Compute $T - T_{-i}$, the difference between the settings with and without the ith SNP.
(4) Repeat steps 2 and 3 for all SNPS, and rank them by the differences.
(5) Apply steps 1–4 by replacing the eight ROIs with 119 ROIs.
(6) Use Cohen's Kappa agreement test (Landis and Koch, 1977) to determine if the ranks from the eight ROIs and the 119 ROIs are similar (fmsb in R).

Table 1 shows the Cohen's Kappa agreement results, and the statistics were all above zero with all p-values being less than 5%, which indicates that the ranks between the eight ROIs and the 119 ROIs are in slight agreement. This implies that for the purpose of genetic component selection, these two

Table 1
Cohen's Kappa agreement between the ranks of eight prefrontal cortex ROIs and all 119 ROIs. Note: the values are displayed in percentages.

Test	Cohen's kappa statistic (%)	p-Value (%)	Judgement
KMR (K = linear)	2.857	3E−03	Slight agreement
KMR (K = quadratic)	2.143	1E−01	
KMR (K = IBS)	2.143	1E−01	
KDC ($\tilde{L}$,K = linear)	3.571	3E−05	
KDC ($\tilde{L}$ = linear,K = quadratic)	2.143	1E−01	
KDC ($\tilde{L}$ = linear,K = IBS)	2.143	1E−01	
KDC ($\tilde{L}$ = quadratic,K = linear)	3.571	3E−05	
KDC ($\tilde{L}$,K = quadratic)	7.857	2E−14	
KDC ($\tilde{L}$ = quadratic,K = IBS)	2.857	3E−03	

phenotypes show similar results; but for the purpose of the association study, the association tests that applied on the 119 ROIs resulted in more significant outcomes than those that used only the eight ROIs.

6. Conclusion and Discussion

In this work, we provided a formulation to show that KMR is equivalent to the KDC statistic. The advantage of this equivalence allows the use of regression modeling interpretations to explain the KDC test. For instance, Liu et al. (2007) and Maity et al. (2012) provided a restricted maximum likelihood (REML) based score test of KMR, and these results in conjunction with our findings allow for a fast computation of the null distribution for the KDC test. Exploring the parametric distribution of the KDC statistic deserves further investigation, but is beyond the scope of this article.

In addition, the KMR tests can be treated as members of a larger family of tests, and more powerful tests can theoretically be designed by looking at the "optimal" kernel among the family members. Although there was no single kernel that was chosen as the best from our simulations, the linear kernel for KMR or KDC achieved better performance than other kernels in the settings with a single phenotype. For multiple phenotypes with multiple correlations or the presence of dependent covariance, the Gaussian RBF or the L_2 kernel achieved better performance than other kernels. However, the exact optimal choice of kernel is data-dependent, and the strategy for selecting the optimal kernel from the KDC family is worthy of further study.

Finally, several works have utilized the KDC/KMR family members in applications that include genetic pathway analysis using KMR Liu et al. (2007), voxel-wise genome-wide association studies using least squares KMR (Ge et al., 2012), neuroimaging genome-wide association using DC (Hua et al., 2015), and multiple change point analysis using DC (Matteson and James, 2014). Some recent studies have presented and discussed the equivalence between these statistics, such as distance-based permutation test for between group comparisons from Reiss et al. (2010), the relationships between Genomic Distance-Based Regression and KMR from Pan (2011) and the equivalence between DC and HSIC from Sejdinovic et al. (2013). Therefore, our establishment of the KDC family (the tests of different kernel combinations) is an important unification of all the above applications.

7. Supplementary Materials

An R code package implementing the proposed methods, and Web Appendices, Tables and Figures referenced in Sections 4 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's 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`

References

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., et al. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826.

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* **63**, 858–873.

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

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, Cambridge, MA.

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., 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., for the Alzheimer's Disease Neuroimaging Initiative. (2015). Multiple comparison procedures for neuroimaging genomewide association studies. *Biostatistics* **16**, 17–30.

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

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., et al. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine* **353**, 1209–1223.

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.-Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology* **36**, 686–695.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980). *Multivariate analysis (probability and mathematical statistics)*. London: Academic Press.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* **109**, 334–345.

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.

Reiss, P. T., Stevens, M. H. H., Shehzad, Z., Petkova, E., and Milham, M. P. (2010). On distance-based permutation tests for between-group comparisons. *Biometrics* **66**, 636–643.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics* **41**, 2263–2291.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* **11**, 1517–1561.

Stein, J. L., Hua, X., Morra, J. H., Lee, S., Hibar, D. P., Ho, A. J., 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., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., et al. (2010b). Voxelwise genome-wide association study (vGWAS). *Neuroimage* **53**, 1160–1174.

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

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

Tziortzi, A. C., Searle, G. E., Tzimopoulou, S., Salinas, C., Beaver, J. D., Jenkinson, M., et al. (2011). Imaging dopamine receptors in humans with [¹¹c]-(+)-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.

Received April 2014. *Revised February* 2015.
Accepted March 2015.