We propose a ridge-penalized adaptive Mantel test (AdaMant) for evaluating the association of two high-dimensional sets of features. By introducing a ridge penalty, AdaMant tests the association across many metrics simultaneously. We demonstrate how ridge penalization bridges Euclidean and Mahalanobis distances and their corresponding linear models from the perspective of association measurement and testing. This result is not only theoretically interesting but also has important implications in penalized hypothesis testing, especially in high-dimensional settings such as imaging genetics. Applying the proposed method to an imaging genetic study of visual working memory in healthy adults, we identified interesting associations of brain connectivity (measured by electroencephalogram coherence) with selected genetic features.

**KEYWORDS**
association testing, distance, GWAS, high-dimensional statistics, neuroimaging, similarity

1 **INTRODUCTION**

The primary goal of imaging genetics is to identify genes that might influence variation in neurophysiological function and neuroanatomical structure. Thus, imaging genetic studies typically collect neuroimaging and genetic data with hundreds of thousands of features per modality, while sample sizes are usually on the order of several dozen to a few thousand. In this scenario, both explanatory and response variables of interest are high-dimensional, such as thousands of gene expressions, millions of DNA variants, dozens to hundreds of electroencephalogram (EEG) channels, and hundreds to millions of blood oxygen level dependent (BOLD) signals. Given how little is known about the relationship of neurological phenotypes with genetics, it is of interest to determine whether particular sets of features are significantly associated in aggregate. However, developing association tests that are sufficiently powered remains an important statistical challenge for the analysis of imaging genetics data.

Mantel’s test\(^1\) is one of the earliest formulations of a (ostensibly nonparametric) distance-based test for association of features from two observational modalities. In the Mantel test (MT), one computes a distance matrix between pairs of subjects in each modality, and tests for association of the two distance matrices. This simple approach is particularly appealing for two sets of high-dimensional variables as it is flexible to choose any distance metric. Improvements have been made by the use of a Procrustean superimposition approach,\(^2\) which is closely related to the double-centering transformation\(^3\) used in several articles, such as Wessel and Schork,\(^4\) Salem et al,\(^5\) and Schork and Zapala.\(^6\) As described in Section 2, in some special scenarios, after double centering, distances are converted to similarities. Closely related tests have been developed from Escoufier’s RV coefficient and its application to neuroimaging,\(^7,8\) the distance covariance (dCov),\(^9\) and...
the sum of powered score tests (aSPC). Excellent reviews and discussions of the connections of MT, the RV coefficient, and dCov are given in Omelka and Hudecová and Josse and Holme. Similar methods have frequently been used for genetic association analysis, for example, Haseman and Elston, Beckmann et al., Wessel and Schork, Salem et al., Schaid, Mukhopadhyay et al., Wei et al., Goeman et al., Schork and Zapala, Larson et al., and Zhao and Zhu.

In modern statistics, shrinkage methods play critical roles in achieving various statistical inference goals. In particular, in high-dimensional settings, shrinkage methods are necessary for numerical stability, interpretation, preventing overfitting, and constructing powerful tests. Ridge regression, which performs $L^2$ regularization, was originally proposed to deal with ill-conditioned or close to ill-conditioned problems, and has since proven extremely useful in prediction problems in various scientific areas, such as predicting human complex traits and agricultural and breeding outcomes. Ridge penalty is also widely used in learning such as neural network to reduce overfitting. This suggests that ridge penalization may yield useful association tests, but despite the existing rich literature on applications of $L^2$ penalized tests, for example, Liu et al., Lin et al., Cule et al., the role that ridge penalty plays in hypothesis testing has been rarely studied. For prediction in deep learning and estimation, the results are mixed. For example, the optimal ridge penalty is found to be positive when the coefficients are generated from a distribution whereas the optimal ridge penalty might be negative when the coefficients are fixed. However, in practice, the true model is never known. Motivated by these challenges, in this article, we propose the ridge-penalized adaptive Mantel test (AdaMant) for evaluating the association of two high-dimensional sets of features. By introducing a ridge penalty, AdaMant tests the association across many metrics simultaneously.

The structure of this article is as follows. Section 2 introduces a group of MTs and describes how ridge penalization, which induces a ridge kernel and distance, connects Euclidean distances to Mahalanobis distances and their corresponding linear models. In Section 3, we propose the AdaMant and describe its implementation. Section 4 evaluates the performance of the test using simulations. Section 5 presents results showing associations between brain connectivity (measured by EEG coherence) and genetic features from a working memory study of 350 college students. Section 6 concludes with a discussion of generalizations and directions for future work.

## 2 | MT AND LINEAR MODELS

The framework of MT is extremely general, and encompasses many well-known association tests, such as the RV coefficient, the distance covariance, and the sum of powered score tests. Consider data $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}^q$, $i = 1, \ldots, n$, where $n$ is the number of observations (participants), $p$ is the number of independent variables such as 500 single nucleotide polymorphisms (SNPs), $q$ is the number of dependent variables, such as 190 pairwise theta band coherence from 20 EEG channels. The corresponding data matrices $X_{n \times p}$ and $Y_{n \times q}$ have been column-centered. To state the general form of the MT, let $K^X(\cdot, \cdot)$ be a positive semi-definite (p.s.d.) kernel function on $X \times X$, with corresponding Gram matrix $H$ defined by $H_{ij} = K^X(X_i, X_j)$, and define $K^Y$ and $K$ similarly using $Y$. The MT statistic for these kernels is equivalent to $\text{tr}(HK)$. The reference distribution under the null hypothesis of no association between similarities measured by $K^X$ and $K^Y$, can be obtained from the observed features $X$ and $Y$ by permuting the observation labels for one set of features and calculating the empirical reference distribution. Equivalently, one can hold one matrix fixed, say $H$, and simultaneously permute the rows and columns of $K$.

Let $X_i$ and $X_j$ be the vector of independent variables for subjects $i$ and $j$, respectively. One simple kernel (similarity) metric is their inner product $(X_i, X_j) = X_i^T X_j$, which is commonly known as the linear kernel in kernel regression or the dot product similarity. It is natural to generalize it by applying a p.s.d. weight matrix $W$: $K_W(X_i, X_j) = (X_i, X_j)_W = X_i^T W X_j$. Particular choices of $W$ yield Mantel-type tests that are corresponding to different distance measures. Choosing $W = I_p$ (the $p \times p$ identity matrix) gives the standard Euclidean inner product, with Gram matrix $XX^T$, that is, the linear kernel. Setting $W = (X^T X)^{-1}$ where the superscript “−1” stands for a generalized inverse, gives Gram matrix $X(X^T X)^{-1} X^T$, which is recognizable as the projection matrix into the column space of $X$. Preconditioning the Mahalanobis weight matrix as $W = (X^T X + \lambda I_p)^{-1}$ gives Gram matrix $X(X^T X + \lambda I_p)^{-1} X^T$. As shown in Appendix A.1, the above similarity matrices are identical to their double-centered squared distances. For example, let $C = I_n - 11^T/n$ (the centering matrix, where $I_n$ is the identity matrix of size $n$ and $1_n$ is the vector of ones of size $n$) and let $D^2$ be the squared distance matrix whose $(i,j)$th element is the squared distance between subjects $i$ and $j$:

$$D^2_{i,j(i,j)} = (X_i - X_j)^T (X^T X + \lambda I_p)^{-1} (X_i - X_j).$$
We have $X(X^TX + \lambda I_p)^{-1}X^T = -\frac{1}{2}CD^2C$. Thus, it is sensible to use Euclidean and Mahalanobis similarity, respectively, to refer to $XX^T$ and $X(X^TX)^{-1}X^T$.

### 2.1 Score statistics in multivariate linear models

We next show that MTs using the weighted matrices introduced above are closely related to score statistics in linear models. Rao’s score test\(^35\) is based on the gradient of the log-likelihood evaluated when the null hypothesis is true. It is locally most powerful and asymptotically equivalent to the Wald and likelihood ratio tests. It is particularly attractive for high-dimensional inference since it does not require the estimation of the parameters to be tested, greatly reducing the computation required compared to alternative approaches. This also makes it convenient to calculate the null distribution for the score test statistic via a permutation procedure, which forms the basis of the AdaMant algorithm proposed in Section 3.1.

#### 2.1.1 Adjust for covariates

To simplify the presentation of mathematical derivations, we assume that $X$ only consists of variables of direct interest and covariates such as age and gender have been adjusted for. Adjusting for covariates is critical to ensure valid inference and, when omitted, can violate the exchangeability of observations under the null hypothesis, a fundamental assumption required by permutation-based tests. Fortunately, in many challenging problems, such as imaging genetics, both the outcome and explanatory variables are high-dimensional but the number of covariates to be adjusted for is usually small. These covariates are not of primary interest. One simple method to handle covariates is to first compute the residuals by regressing the outcomes on the covariates and then replace the outcomes with the residuals. This type of covariate adjustment in mixed-effects models has been commonly used.\(^29,36\) A more accurate method is to use the restricted maximum likelihood estimates (REML)-type transformation, for example, Huh and Jhun\(^37\) and Ge et al.\(^38\) A comprehensive description and investigation on relevant methods can be found in Winkler et al.\(^39\)

#### 2.1.2 Linear fixed and random effects models

The conventional linear model is a fixed effects model where the coefficients are treated as fixed but unknown parameters. When the number of explanatory variables $p$ is large relative to the sample size $n$, the design matrix is ill-conditioned or close to ill-conditioned and regularization methods are often applied. An alternative is the random effects model which assumes that the coefficients are random variables from a normal distribution. The random effects model has been widely used to handle correlated data such as repeated measurements in longitudinal studies.\(^40\) Because it can naturally analyze a large number explanatory variables and account for correlation across observations, the variance components form of the random effects model has been used for many years in genetic studies to estimate the heritability of phenotypes from a sample of related or unrelated subjects. Consider the following two linear models.

1. **Fixed effects model:** $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, I_n, \Sigma)$, which can be equivalently expressed using vectorized $Y$:
   
   \[ \text{vec}(Y) \sim N(\text{vec}(X\beta), \Sigma \otimes I_n) \]

   where $\text{vec}(\cdot)$ denotes the vectorization operator.

2. **Random effects model:** $Y = Xb + \epsilon$, where $\epsilon \sim N(0, I_n, \Sigma)$ and $b \sim N(0, I_p, \Sigma_b)$. Equivalently, $\text{vec}(Y) \sim N(0, \Sigma_b \otimes XX^T + \Sigma \otimes I_n)$

   Their null hypotheses are $\beta = 0$ and $\Sigma_b = 0$, respectively. For the random effects model, the score equals a quadratic form minus its expectation, where the expectation is evaluated under the null hypothesis. Because the expectation equals the product of a function of $X$ and a function of $Y$, it is not informative for association testing. As a result, score tests are often conducted based on the quadratic form, for example, Liu et al,\(^29\) Huang and Lin,\(^41\) Goeman et al,\(^19\) and Goeman et al.\(^36\) Using the same strategy, we have the following result:
**Proposition 1.** Let $H_0 = X(X^TX)^{-1}X^T$, $K_0 = Y(Y^TY)^{-1}Y^T$, $H_∞ = XX^T$, and $K_∞ = YY^T$. The score statistic from the fixed effects model is proportional to $\text{tr}(H_0 K_0)$. When $Σ = σ^2 I_q$ and $Σ_β = σ^2 β I_p$, which is always true when $q = 1$, the score statistic from the random effects model is proportional to $\text{tr}(H_∞ K_∞)$. Asymptotically, their null distributions are chi-squared and mixture of chi-squared, respectively.

The proof is in Appendix A.2. Remarks. (1) When $Σ_β$ in the random effects model is unstructured, we have a score matrix because there is a score for each element of $Σ_β$. If we use its trace, that is, the sum of its diagonal elements, the corresponding test statistic is still $\text{tr}(H_∞ K_∞)$. The assumption of $Σ_β = σ^2 I_p$ implicitly assumes that the coefficients in $b$ are exchangeable, that is, the density of all permutations of the coefficients are the same. (2) If $Σ$ in the random effects model is unstructured, the score statistic involves $YΣ^{-2}YT$, which is proportional to $Y(Y^TY)^{-2}YT$ when $Σ$ is replaced by a consistent estimator. This result is consistent with those presented in earlier work such as Maity and Sullivan and Huang and Lin. Although $Σ = σ^2 I_q$ is an oversimplification, when the dimension $q$ is large, it is costly to estimate $Σ$. The ridge penalized similarity introduced next might be considered as a compromise.

### 2.1.3 Ridge penalization and a data-augmentation approach

The ridge regression is a $L_2$-penalized form of the fixed effects model with the following penalized log-likelihood:

$$
\ell_\lambda = -\frac{n}{2} \log|Σ| - \frac{1}{2} \text{tr}[(Y - Xβ)Σ^{-1}(Y - Xβ)^T] - \frac{λ X}{2} \text{tr}(ββ^T),
$$

where $λ X ≥ 0$ is the ridge parameter.

The ridge estimator of $β$ is $β_\lambda = \arg\min_β \left\{ \text{tr}(Y - Xβ)Σ^{-1}(Y - Xβ)^T + λ X \text{tr}(ββ^T) \right\}$. Recall that ridge regression can be formulated by augmenting both $X$ and $Y$ (“12.3.1 Physical interpretations of ridge regression” in Sen and Srivastava). For example, the log-likelihood and ordinary least squares estimate (OLSE) based on the following augmented data is mathematically identical to the log-likelihood in (1) and the ridge-penalized estimate $β_\lambda$, respectively:

$$
\bar{Y} = [Y^T: \ 0_{q×p}]^T, \bar{X} = [X^T: \sqrt{λ_p} I_p]^T.
$$

Because ridge regression does not involve $λ Y$, it is sensible to consider a symmetric treatment of $X$ and $Y$ when both of them are high-dimensional. Similar data augmentation strategies have been proposed in Hodges for the purpose of model diagnostics in hierarchical models. Consider two tuning parameters $λ X > 0, λ Y > 0$. We propose to use the following augmented data

$$
\bar{Y} = \begin{pmatrix} Y \\ 0 \\ \sqrt{λ_p} I_q \end{pmatrix}, \quad \bar{X} = \begin{pmatrix} X \\ \sqrt{λ_p} I_p \\ 0 \end{pmatrix}.
$$

If we model $\bar{X}$ and $\bar{Y}$ with the same distributional assumptions as used to the fixed-effects model, we can define an augmented-data likelihood. Although it is not a real likelihood, nevertheless, one can still define score statistics. By Proposition 1, the corresponding score statistic for $β = 0$ is proportional to $\text{tr}(H\bar{K})$, where $H = \bar{X}(X^TX + λ X I_n)^{-1}X^T$ and $K = \bar{Y}(\bar{Y}^T\bar{Y})^{-1}\bar{Y}^T$. Let $H_λ = X(X^TX + λ X I_n)^{-1}X^T$, and $K_λ = Y(Y^TY + λ Y I_n)^{-1}Y^T$. It is not difficult to show that $H = H_λ$ and $K = K_λ$. This observation provides an augmented-data interpretation of the MT based on the ridge-penalized similarity matrices.

Note that the OLSE based on the augmented data depends only on $λ X$, but not on $λ Y$. To examine the role played by $λ Y$, we consider the augmented-data log-likelihood,

$$
\ell_{λ X, λ Y} = C - (n + p + q) \log|Σ| - \text{tr}[(Y - Xβ)Σ^{-1}(Y - Xβ)^T] - λ X \text{tr}[βΣ^{-1}β^T] - λ Y \text{tr}[Σ^{-1}],
$$

where $C$ is a constant that does not depend on any parameters. One interesting observation is that the likelihood corresponding to Equation (3) is mathematically identical to the posterior distribution of $β$ and $Σ$ where the joint prior is a normal-inverse-Wishart distribution:

$$
p(β, Σ) ∝ |Σ|^{-\frac{p+q}{2}} \exp\left\{ -\frac{λ X}{2} \text{tr}[βΣ^{-1}β^T] - \frac{λ Y}{2} \text{tr}[Σ^{-1}] \right\}.
$$
2.2 The ridge as a bridge between Euclidean and Mahalanobis metrics

The Escoufier’s RV coefficient\(^7\) based on \(H\) and \(K\) is defined as

\[
R(H, K) = \frac{\text{tr}(HK)}{\sqrt{\text{tr}(H^2)\text{tr}(K^2)}},
\]

where \(H\) and \(K\) are similarity matrices, which are often p.s.d., and \(H^2\) and \(K^2\) represent the squares of \(H\) and \(K\), respectively, under the definition of matrix multiplication. Note that \(R(H, K)\) is the sample Pearson’s correlation of the vectorized similarity matrices \(H\) and \(K\). By examining the RV coefficients with different ridge penalty, we provide a novel unification of the fixed effects, ridge, and random effects model as a two-parameter class of correlation measures, parameterized by the ridge penalty terms \((\lambda_X, \lambda_Y) \in [0, \infty]^2\), where \(\lambda_X = 0\) corresponds to Euclidean similarity, and \(\lambda_X = \infty\) corresponds to the Mahalanobis similarity.

**Theorem 1** (Limits of multivariate ridge correlations). Let \(X_{n \times p}\) and \(Y_{n \times q}\) be column-centered data matrices of two sets of features, respectively. Let \(H_{\lambda_X} = X(X^TX + \lambda_X I_p)^{-1}X^T\), \(K_{\lambda_Y} = Y(Y^TY + \lambda_Y I_q)^{-1}Y^T\). Then

\[
\lim_{\lambda_X \to \lambda_X^*} R(H_{\lambda_X}, K_{\lambda_Y}) = R(H_{\lambda_X^*}, K_{\lambda_Y^*}),
\]

where \(H_0 = X(X^TX)^{-1}X^T\), \(K_0 = Y(Y^TY)^{-1}Y^T\), \(H_\infty = XX^T\), and \(K_\infty = YY^T\).

The proof is detailed in Appendix A.3 and outlined here. We first prove that

\[
\lim_{\lambda_X \to \lambda_X^*} R(H_{\lambda_X}, K_{\lambda_Y}) = R(H_{\lambda_X^*}, K_{\lambda_Y^*}).
\]

The result is obvious when \(\lambda_X^* = 0\). The proof for \(\lambda_X^* = \infty\) is accomplished by performing singular value decomposition (SVD) of \(X\). Performing SVD of \(Y\) and using a similar method will lead to the desired conclusion, that is, \(\lim_{\lambda_X \to \infty} R(H_{\lambda_X}, K_{\lambda_Y}) = R(H_\infty, K_\infty)\).

This result indicates that, as we increase the ridge penalty, we are moving from Mahalanobis distances/similarities toward Euclidean distances/similarities. The quantities \(R(H, K)\) and \(tr(HK)\) are related to known statistics for many choices of \(K\) and \(H\). For example, \(tr(H_0K_0)\) equals the sum of the squared canonical correlations and is the Pillai’s trace statistic; in particular, it is multiple \(R^2\) when \(q = 1\); it is also proportional to Hooper’s trace correlation \(R^2\)\(^{4,5}\). Because \(tr(H_\infty K_\infty)\) was derived from random effects model, after some derivations, it is not surprising to see that it is linked to method of moment estimates of variance components in the random effects model.

To better understand the underlying geometric interpretation of different choices, we focus on the case of univariate outcomes, that is, \(q = 1\) and assume that \(Y\) has been standardized with mean zero and variance one. Note that the fixed and random effects models have the same null distribution of \(Y\). Their corresponding score test statistics can be compactly written by considering the SVD of \(X\): \(X = UDV^T\), where \(U\) is \(n \times r\) with \(U^TU = I_r\), \(D\) is an \(r \times r\) diagonal matrix of singular values, \(r = \text{rank}(X)\), \(d_j, j = 1, \ldots, r\) are the nonzero singular values, and \(V\) is \(p \times r\), with \(V^TV = I_r\). We refer to the columns of \(U\) as the principal directions of \(X\), and to \(Z = U^TV\) as the principal correlation vector, which is an \(r \times 1\) vector with \(j\)th component \(Z_j = R(Y, U_j)\), the Pearson correlation of \(Y\) with \(X\) along the \(j\)th principal direction. With these notations, \(tr(H_0 Y Y^T) = \sum_{j=1}^{r} Z_j^2\) and \(tr(H_\infty Y Y^T) = \sum d_j^2 Z_j^2\). Thus, we can interpret the fix effects score statistic as testing the Euclidean norm of the correlation vector; in contrast, the random effects (ie, variance components) score statistic is equivalent to testing the weighted norm of \(Z\), where the \(j\)th principal direction of \(X\) is weighted by \(d_j^2\), the variance of \(X\) along the \(j\)th principal direction. This has the effect of emphasizing the influence of directions in \(X\) for which \(X\) has large variance and reducing the influence of directions with small variance. The ridge regression score statistic equals

\[
tr(H_{\lambda_X} Y Y^T) = \sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \lambda_X^*} Z_j^2,
\]

which is a compromise between the fixed effects and variance components tests, with small \(\lambda_X\) yielding a test close to the fixed effects (and identical at \(\lambda_X = 0\)), and large \(\lambda_X\) yielding a test close to the random effects score statistic, converging to identical tests as \(\lambda_X \to \infty\). In other words, the ridge statistic weights \(Z_j^2\) proportional to \(d_j^2\) as in the random effects statistic, but flattens each weight by a factor of \(\frac{1}{d_j^2 + \lambda_X^*}\). This geometric interpretation motivates us to consider a ridge-penalized AdaMant that considers a set of ridge tuning parameters.
3 | ADAPTIVE MANTEL TEST

In this section, we develop the ridge-penalized AdaMant, a novel extension of the classical MT to utilize the ridge penalty in association testing. By using a permutation procedure on the minimum P-value from a set of test statistics, AdaMant is able to simultaneously test across a set of tuning parameters and kernels without increasing the type I error rate. In comparison to the RV-test, which relies on a single metric, AdaMant tests across a set of metrics, and can therefore achieve higher power, as shown by the simulations in Section 4.

The idea of AdaMant can be understood through an example varying the relationship of the principal components (PCs) of $X$ with $Y$ generated from a variance components model. For the ease of presentation of the idea, we assume $Y$ is univariate, which implies that $\lambda_Y$ is not needed. Suppose that $X$ is a $100 \times 2$ matrix of covariates drawn iid from a standard normal distribution, and $Y \sim N(0, \sigma^2 K(\theta) + \sigma^2 I_{100})$, with $K(\theta) = U \Theta \Delta^2 \Theta^T U^T$ for $\Theta$ the $2 \times 2$ rotation matrix for angle $\theta$. Figure 1 shows the results of applying AdaMant with weights for the Euclidean, Mahalanobis, and ridge kernels ($\lambda_X = 10$) for data generated with $\theta = 0, 0.12\pi, 0.3\pi$. In each case, AdaMant selects the weighting that most emphasizes the empirical direction of association of $X$ and $Y$. This clearly demonstrates the potential advantage of AdaMant if the direction of association of $X$ and $Y$ is different from the first PC of $X$. In this case, AdaMant will likely be a more powerful test than methods such as the RV test, which corresponds to using only the Euclidean weighting.

3.1 | Algorithm for the AdaMant

The “adaptive” procedure receives as input a list of pairs of metrics/kernels $\{(\mathcal{K}_m^X, \mathcal{K}_m^Y) | m = 1, \ldots, M\}$ from which the matrices $K_m = \mathcal{K}_m^X(X)$ and $H_m = \mathcal{K}_m^Y(Y)$ are computed for each metric pair, $m = 1, \ldots, M$. These metrics may be from a single family with varying tuning parameters, such as the ridge penalty tuning parameter, or may include kernels from different families. For each $m = 1, \ldots, M$, $P_m$ is calculated as the P-value of the MT with metrics $\mathcal{K}_m^X$ and $\mathcal{K}_m^Y$ for $X$ and $Y$, respectively. The AdaMant test statistic is defined as the minimum of these values, $P^{(0)} := \min_{m \in \{1, \ldots, M\}} P_m$. A permutation procedure is practical for calculating the reference distribution for $P^{(0)}$. For each $m$, and $b = 1, \ldots, B$, $H^{(b)}_m$ is generated by permuting rows and columns of $H$ simultaneously, and the corresponding test statistic $P^{(b)}$ is calculated. The AdaMant P-value is then calculated as the proportion of P-values no greater than $P^{(0)}$, that is,

$$P_{AdaMant} = \frac{1}{B+1} \sum_{b=0}^{B} I(P^{(b)} \leq P^{(0)}) .$$

![Figure 1](https://wileyonlinelibrary.com)

**Figure 1** Illustration of AdaMant for $Y$ generated from covariates $X$ rotated by angle $\theta$. Gray shaded ellipses show the true covariance of $X$; pink shaded ellipses show the rotated covariances used to generate $Y$; magenta line segments show the direction of association estimated from a fixed effects model. The solid lined ellipses correspond to the weights chosen by AdaMant; dashed ellipses are other weightings considered. In each case, AdaMant selects the weighting that most emphasizes the empirical direction of association of $X$ and $Y$. A, $\theta = 0$; B, $\theta = 0.12\pi$; C, $\theta = 0.3\pi$ [Colour figure can be viewed at wileyonlinelibrary.com]
General pseudocode for AdaMant is given in Algorithm 1. In this article, we focus on ridge kernels, which include Euclidean and Mahalanobis similarities as special cases. The R code is available at https://sites.google.com/view/zhaoyixiu-homepage. A discussion of the computational strategies and complexity for ridge kernels is given in Appendix A.4.

Algorithm 1. Adaptive Mantel test

1: procedure ADA Mantel \((X, Y, \{K^X_m, K^Y_m\}_{m=1}^M, B)\)
2: \[ H_m \leftarrow K^X_m(X), m = 1, \ldots, M \]
3: \[ K_m \leftarrow K^Y_m(Y), m = 1, \ldots, M \]
4: Calculate \( T_m^{(0)} \leftarrow T_m := \text{tr}(H_m K_m) \)
5: Generate \( B \) permutations of \( H_m \), labeled \( H_m^{(b)}, m = 1, \ldots, M; b = 1, \ldots, B \).
6: \[ T_m^{(b)} \leftarrow \text{tr}(H_m^{(b)} K_m), m = 1, \ldots, M; b = 1, \ldots, B \]
7: \[ p_m^{(b)} \leftarrow \frac{1}{B+1} \sum_{b'=0}^B I (T_m^{(b')} \leq T_m^{(b)}), m = 1, \ldots, M; b = 1, \ldots, B \]
8: \[ p^{(b)} \leftarrow \min_{m=1,\ldots,M} p_m^{(b)}, b = 1, \ldots, B \]
9: \[ P_{\text{AdaMant}} \leftarrow \frac{1}{B+1} \sum_{b=0}^B I (p^{(b)} \leq p^{(0)}) \]
10: end procedure

Remark 1. Algorithm 1 computes the AdaMant \(P\)-value via a permutation procedure. This method can be computationally expensive, but is likely feasible in practice when some computational tricks are used (see Appendix A.4). It is also possible to compute the asymptotic null distribution of the AdaMant test statistic using a moment-matching normal approximation and results on the order statistics of normal random variables. However, this normal approximation for a mixture of chi-squared random variables is not accurate for small sample sizes, and so the permutation procedure is generally recommended.

3.2 Choosing the ridge penalty

A main advantage of the ridge-penalized AdaMant is to allow for simultaneous testing over a set of ridge tuning parameter values while incurring substantially less loss of power compared to multiple testing adjustment methods such as Bonferroni. When applying AdaMant, the set of candidate penalty terms should be chosen as small as possible, since the test power will decrease as the number of penalty terms increases. When only ridge kernels of form \(X^T X \lambda X I_p^{-1} X^T\) are included in AdaMant, previous results on the role of the ridge penalty term in predictive modeling can help with the identification of a reasonable set of candidate values.

Specifically, it has been shown that, for column-standardized \(X\), when the ridge penalty \(\lambda X\) is chosen to be proportional to the noise-to-signal ratio \(\sigma^2 / \sigma_b^2\), the resulting shrinkage estimator \(\hat{\beta}_{\lambda X}\) is identical to the best linear unbiased predictor (BLUP) for the random effects model. Recall that the genetic heritability is defined as \(h^2 = \frac{p \sigma^2}{\sigma^2 + p \sigma_b^2}\), which implies that \(\sigma^2 / \sigma_b^2 = p \frac{1-h^2}{h^2}\). Moreover, for a new observation with unknown response the predictions using the ridge and random effects models are the same when replacing \((H_\infty / p)^{-1}\) with \((H_\infty / [p \frac{1-h^2}{h^2} + I_p])^{-1}\), which is equivalent to choose \(\lambda X = p \frac{1-h^2}{h^2}\). Note that, with fixed heritability, the choice of \(\lambda X\) should increase with \(p\), which is not surprising as shrinkage/penalization is usually advantageous in high-dimensional settings. Thus, a reasonable set of penalty terms can be determined in practice by positing a priori a likely range for \(h^2\), the scientific interpretation and range of plausible values of which depend on the specific modalities of \(X\) and \(Y\). For instance, for a common disease, \(h^2 > 0.5\) is considered high (and would likely have already been discovered by previous studies), whereas \(h^2 < 0.1\) is probably not scientifically interesting. As a point of reference, most estimated heritability in the UK Biobank data is between 0.1 and 0.4.
4 | SIMULATIONS

4.1 Univariate simulations \((q = 1)\)

Because AdaMant is a permutation-based test, it naturally controls its type I error rate at any desired level. Using simulated data generated from varying numbers of covariates and several covariance structures (homoskedastic, heteroskedastic, and compound symmetric), we confirmed that the type I error rate was controlled appropriately. To assess the power of AdaMant, we first conducted simulations with \(q = 1\), \(n = 200\), \(p\) ranging from 50 to 500, \(\sigma^2 = 1\), \(\sigma^2_b = 0.035^2\) (random effects), \(\beta_j = (-1)0.05, j = 1, \ldots, p\) (fixed effects), and 500 permutations. The design matrix \(X\) was generated from the matrix normal \(N(0, I_n, \Sigma_X)\), where \(\Sigma_X\) is chosen to have a compound symmetric structure with \(\rho = 0.1\). For each setting, 500 simulations were run to estimate power. AdaMant was conducted using our \textit{adamant} R-package with \(K = YY^T\) and \(H_{\lambda_X}\) for \(\lambda_X \in \{100, 1000, 2500, 5000, 7500, 25000, \infty\}\). Three other testing methods (aSPC,\(^1\) dCOV,\(^9\) and RV-coefficient) were included for comparison.

Figure 2 shows the estimated power for data generated from the variance components model (the left panel) and the fixed effects model (the right panel). The simulation results indicate that the power of AdaMant is competitive with the best of the single-parameter MT for the values considered. For the variance components setting, \(\lambda_X = 1000\) attains the highest power for the ridge MT. As anticipated, the power of the ridge MT exhibits unimodal behavior, such that the power increases to its peak for \(\lambda_X = 1000\), and decreases as \(\lambda_X \to \infty\). We note that even though the variance components model is the true data generating mechanism, the test power is substantially increased over the variance components score test through the use of the ridge kernels. The optimal ridge parameter selected by AdaMant (\(\lambda_X = 1000\)) confirms the relationship of the optimal parameter and the signal-to-noise ratio discussed in Section 3.2, which predicts the optimal parameter is \(\lambda^* = 1/0.035^2 \approx 800\).

When the response data are generated from the fixed effects model, the simulation results show that using a penalized ridge kernel produces much higher power than competing tests, with \(\lambda_X = 100\) giving the highest power across all values of \(p\). These results suggest that the variance components test is severely underpowered when the relationship of \(X\) and \(Y\) is strongest along PCs of \(X\) with low variance. Moreover, the RV coefficient, dCov, and aSPC tests perform equally poorly in this setting.

![Figure 2](https://wileyonlinelibrary.com)
TABLE 1  Estimated power for association testing with AdaMant compared to other testing methods for simulated data generated from the multivariate random effects model with $n = 200$, $p = 40$, $q = 20$

| $\sigma^2_A$ | AdaMant | Mantel | RV | dCov | aSPC |
|---------------|---------|--------|----|------|------|
| 0             | 0.059   | 0.021  | 0.045 | 0.043 | 0.052 |
| 0.01          | 0.251   | 0.070  | 0.244 | 0.253 | 0.077 |
| 0.02          | 0.759   | 0.115  | 0.756 | 0.752 | 0.177 |
| 0.03          | 0.924   | 0.191  | 0.935 | 0.919 | 0.378 |
| 0.04          | 0.981   | 0.313  | 0.980 | 0.981 | 0.462 |

4.2  Multivariate simulations

4.2.1  Simulations with the multivariate variance components model

To compare the performance of AdaMant with existing methods for testing association of two multivariate feature sets, we generate data from the multivariate variance components model. In the variance components simulation, $Y = G + \epsilon$, with $Y \sim N(0, \Sigma_A \otimes XX^T/p + I_q \otimes I_n)$, where $\Sigma_A$ is a compound symmetric covariance matrix with diagonal entries equal to $\sigma^2_A$ and off-diagonal entries $0.1\sigma^2_A$. We present simulation results for $n = 200$, $p = 40$, $q = 20$, and $\sigma^2_A \in \{0, 0.01, 0.02, 0.03, 0.04\}$. We use ridge penalization for both $X$ and $Y$ features with $\lambda_X \in \{10, 100, \infty\}$ and $\lambda_Y \in \{10, 100, 500, 1000, \infty\}$. The results (Table 1) show that AdaMant’s power is comparable to the RV-test and dCov tests, all of which are substantially more powerful than the classical MT and aSPC test.

4.2.2  Simulated EEG data ($q = 20$)

Since we are primarily interested in applying AdaMant to imaging genetics data, we also consider the test performance on simulated data with structure similar to realistic EEG and SNP data. Simulated EEG data were produced as a mixture of AR(2) components using the method proposed in Gao et al\(^49\) to ensure the spectral properties, characterized by spectral density, of the generated data are similar to real EEG data (Figure 3). Association of the EEG data with the simulated genetic data was induced by generating the mixture weights with a variance components model with the genetic features as covariates. Similar to our real data analysis, the outcome of interest in this simulated EEG study is EEG coherence, where the coherence of two signals is a measure of the spectral similarity of two signals within a specified frequency band.\(^50,51\) Because the coherence matrix is symmetric, we use its vectorized upper triangles as the outcome variables. Further details on the simulation method and calculating of EEG coherence are given in Appendices A.5 and A.6.

In applying AdaMant, similarity of observations in the EEG domain was computed from the ridge kernel applied to theta coherence matrices (resulting in 190 features); similarity in the genetics domain was computed from the ridge kernel applied to the observations’ SNP vectors. Candidate penalty terms were $\lambda_X = \lambda_Y = \{10, 100, \infty\}$. Five hundred simulations were run for effect sizes equal to $\sigma^2 = \{0, 2.5 \times 10^{-6}, 5 \times 10^{-6}, 7.5 \times 10^{-6}, 1 \times 10^{-5}\}$. The simulation results show (Figure 4) that AdaMant has power competitive with other association testing methods in the two ways used for generating multi-channel EEG data.

5  Applications

We next consider data from 350 healthy college students from Beijing Normal University (BNU) who participated in a visual working memory task conducted by our co-author, during which 64-channel EEG was recorded at 1 kHz. These data are part of a broader set of imaging, genetics, and behavioral data collected with the goal of identifying neurological features of brain connectivity that are significantly associated with both genetic features and cognitive performance. In the present study, we focus on testing the association of EEG coherence with a set of genome-wide SNPs, and a set of candidate SNPs identified in a meta-analysis of educational attainment (EA). EEG coherence has been implicated in many cognitive processes and neurological diseases, for instance, in associative learning,\(^52,53\) as a predictor of ability in certain linguistic and visuospatial tasks,\(^55,56\) and in Alzheimer’s disease\(^57,58\) and schizophrenia.\(^59\)
**FIGURE 3** A comparison of the simulated and real EEG time series shows the simulated series exhibit similar spectral characteristics as the real data after applying the same Butterworth bandpass filter (5-50 Hz). A, Filtered simulated EEG time series and corresponding spectral density. B, Filtered EEG for subject 2002, channel FC4, and corresponding spectral density.

**FIGURE 4** Power plots of association testing methods for simulated EEG data with $n = 200$ observations, $p = 300$ genetic features, $q = 20$ EEG channels. Left, Simulation with strength of association controlled by the within-group genetic similarity, for $\sigma_g^2 = 5 \times 10^{-5}$ fixed. Right, Simulation with strength of association controlled by $\sigma_q^2$. [Colour figure can be viewed at wileyonlinelibrary.com]
The experimental task is depicted in Figure 5. The task for the participants was to recall the positioning of the red bars (the targets) on the left or right side of the image as indicated by the arrow in the center of the image. Six experimental conditions were used for 2, 3, 4, 6, and 8 targets, and for 2 targets + 2 blue distractors. Further details on experimental design and preprocessing of the EEG data are given in Appendix A.7. The pairwise coherence for $q$ EEG channels is a $q \times q$ symmetric matrix, from which we extract the upper triangle and vectorize to form the $n \times \left( \frac{q^2}{2} \right)$ matrix $Y$. For our analysis, we consider coherence restricted to the 25 most anterior channels (“frontal”), and fronto-parietal (FP) coherence from 30 channels covering the frontal and parietal lobes. This results in 300 distinct features for the 25 frontal channels, and 435 features for the FP channels. Channel topography and coordinates are available from the BioSemi website.

Genotype data from approximately $5 \times 10^5$ SNPs were measured, of which 484 496 autosomal SNPs meeting quality control thresholds (minor allele frequency (MAF) > 1%, Hardy-Weinberg equilibrium (HWE) P-value $\geq 0.001$) were included for association testing. Data processing was conducted using PLINK v1.90b4.460 and GCTA 1.91.5 beta2.61 Let $X$ be the column-standardized SNP data matrix. Note that the purpose of association testing is to determine whether the heritability of EEG coherence in the theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-50 Hz) bands is significantly greater than zero. AdaMant was performed with $H_\infty = XX^T$ (which is proportional to the usual genetic relationship matrix), $K_{\lambda_Y} = Y(Y^TY + \lambda_Y I_q)^{-1}Y^T$ for $\lambda_Y \in \{10, 100, 1000, \infty\}$, using 5000 permutations.

### 5.1 Association of EEG coherence and all SNPs

As mentioned above, we are interested in the coherence in the four bands (theta, alpha, beta, and gamma) at the frontal region (300 coherence features from 25 channels) and the FP region (435 coherence features from 30 channels). We tested the association between each of the eight region-band combinations with all SNPs. Among them, the most significantly heritable coherence was found to be the theta band for the frontal channels, with $P$-value = 0.0458. Note that the $P$-value only provides significance but not magnitude or strength of the total genetic effect. We therefore estimated the multivariate genetic heritability using the method of moment estimator (see Ge et al38). The estimated mean heritability of theta
coherence for the frontal channels is $h^2 = 0.165$. These evidences are at best marginal or weak, which is probably due to the small sample size, a potentially large number SNPs that do not contribute to EEG coherence, and low signal-to-noise ratio in EEG. Next, we focus on a subset of SNPs that are known to be associated with EA.

5.2 Association of EEG coherence and SNPs related to EA

SNPs were selected from a meta-analysis of EA, which identified over $3 \times 10^5$ SNPs with $P$-value less than 0.01, from which the top 100 EA SNPs were selected for testing. AdaMant test was performed with $\lambda_Y \in \{0.5, 1, 5, 10, 100, 1000, \infty\}$ for ridge kernel similarity of the coherence data, and using 5000 permutations. Genetic similarity of subjects was calculated as the $L_2$ inner product of the standardized SNP data. Testing the association of EEG coherence of the frontal and FP channels with the selected SNPs, we found that similarities from beta and gamma (30-50 Hz) band coherence were significantly associated with the genetic features. The beta band was the most significantly associated, with $P$-value $= 0.0104$ for the frontal channels and $P = 0.013$ for the FP channels. These results are consistent with existing evidence on the role that the beta and gamma bands play in working memory.

6 DISCUSSION

We have developed a unified framework of linear models that links the association measurements and score tests of the random effects (variance components), fixed effects, and ridge regression models as a single class of tests indexed by ridge penalty. The unified framework we discovered motivated us to develop the ridge-penalized AdaMant as a metric-based testing procedure, which reduces the difficulty of kernel parameter selection by simultaneously testing across a set of candidate parameters and automatically selecting the parameter yielding the most significant test from the candidate set. As a tool for high-dimensional inference, AdaMant improves over competing association tests by data-adaptively selecting the kernel that yields the most significant association of $X$ and $Y$, without the need to explicitly adjust the calculated $P$-value. AdaMant is simple to implement, yet can be usefully applied for any data modality that admits a meaningful measure of similarity. As part of the present contributions, and to facilitate the use of AdaMant by other researchers, an R-package providing a basic implementation of AdaMant is available on the author's website.

AdaMant assesses the overall significance of association between two sets of variables. For example, in our application of AdaMant to the imaging genetics problem in Section 5.2, we examined the association between a set of brain signal variables defined by EEG coherence and a set of genetic variables based on SNPs. It is a global test thus exploratory. Once such overall significance is established, it is perhaps useful to further examine the contribution of individual variables. In Choi et al., they developed tools to visualize the contribution of individual genetic variants to multiple structural changes in the brain measured from patients with Alzheimer's disease. The effect they aimed to quantify is related to pleiotropy, the situation when one gene is associated with multiple phenotypes. In many other studies, it might be interesting to estimate the overall genetic effects, that is, genetic heritability, for one specific or multiple targeting phenotypes. Whether one should examine the individual effects of $X$ on all/subsets of $Y$ or the individual effects of $Y$ on all/subsets of $X$ depends on the study goal and the desired direction of interpretation. Another relevant work is model selection with ridge penalization. Fan and Lv proposed an iteratively thresholded ridge regression screener that balances the computational speed and accuracy in variable selection for a univariate outcome variable. Future research that adopts shrinking strategies, such as the ridge shrinkage considered here, to better understand the detailed relationship between two sets of variables might be helpful.

The present article has primarily focused on the properties of AdaMant when only ridge kernel similarities are used, but this class of weights may not be optimal. For instance, in Figure 1B,C, weights that are concentrated along the true direction of association will produce a more powerful test than the ridge kernel weights. Considering a wider class of similarity metrics, such as rotations of the Mahalanobis metric or nonlinear kernel functions, is an important direction of future work for further improving the power and efficiency of high-dimensional testing methods. Many currently used association testing methods are powered for only a specific functional association of the two feature sets, for example, linear or quadratic, or for general dependence testing (as with dCov). The idea of using a ridge penalty in defining distance is flexible and can be incorporated into other powerful tests. For example, the dCov test with the Euclidean distance is able to capture both linear and nonlinear relationship as a zero dCov implies statistical dependence of the two feature sets. One way might improve the dCov test is to replace the Euclidean distance in dCov with a ridge distance indexed
by a penalty parameter $\lambda$. AdaMant using the Gaussian radial basis kernel with a suitable selection of tuning parameters may be able to detect a wide variety of associations (although not all). Building on kernel-based methods for estimating heritability, future work will consider extensions of AdaMant to a more general class of similarity measures, and investigate methods for improving testing power while selecting across a large number of tuning parameters.

The primary goal of this article is to understand the role that ridge penalization plays in similarity or distance measurements and to develop a ridge-penalized AdaMant. We notice that in the past few years, with the counterintuitive discovery that overfitting often leads to accurate predictive models in deep learning, there has been increased interest in investigating overfitting for estimation, variable selection, and prediction when $p \gg n$. The work of Fan and Lv focused on variable selection and the asymptotic behavior of their novel iterative sure independence screening under varying high-dimensional or ultrahigh-dimensional settings. During the preparation of the current version (the first draft was uploaded to arXiv in 2017), it has been found that the optimal ridge penalty is positive when the coefficients are generated from a distribution whereas the optimal ridge penalty could be negative when the coefficients are fixed. These studies rely on random matrix theorems and assumptions of the covariance structure of the explanatory variable. Although our goal is different, our simulation results nevertheless are consistent with theirs. As shown in Figure 2, ridge penalization is more helpful when data were generated using a random effects model than when data were generated from a fixed effects model. In practice, the true mechanism is never known; thus it is perhaps reasonable to be data driven, that is, let the data determine the optimal amount of penalization. We will continue examining whether and when mitigating overfitting by ridge penalty is beneficial for hypothesis testing.

ACKNOWLEDGEMENTS
We thank Professor Daniel L. Gillen, University of California, Irvine for the helpful discussions. We greatly appreciate the reviewers’ insightful, careful, and constructive comments on our manuscript. These valuable comments have helped us improve the quality of our work.

DATA AVAILABILITY STATEMENT
The simulated data are available from the corresponding author upon reasonable request.

ORCID
Zhaoxia Yu https://orcid.org/0000-0001-9700-1795

REFERENCES
1. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967;27(2 Part 1):209-220.
2. Peres-Neto P, Jackson D. How well do multivariate data sets match? the advantages of a procrustean superimposition approach over the Mantel test. Oecologia. 2001;129:169-178.
3. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3-4):325-338.
4. Wessel J, Schork NJ. Generalized genomic distance–based regression methodology for multilocus association analysis. Am J Hum Genet. 2006;79(5):792-806.
5. Salem RM, O’Connor DT, Schork NJ. Curve-based multivariate distance matrix regression analysis: application to genetic association analyses involving repeated measures. Physiol Genomics. 2010;42(2):236-247.
6. Schork NJ, Zapala MA. Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. Front Genet. 2012;3:190.
7. Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient. Appl Stat. 1976;25(3):257-265.
8. Shinkareva SV, Ombao HC, Sutton BP, Mohanty A, Miller GA. Classification of functional brain images with a spatio-temporal dissimilarity map. NeuroImage. 2006;33(1):63-71.
9. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35(6):2769-2794.
10. Xu Z, Xu G, Pan W. Adaptive testing for association between two random vectors in moderate to high dimensions. Genetic Epidemiol. 2017;41(7):599-609.
11. Omelka M, Hudcová Š. A comparison of the Mantel test with a generalised distance covariance test. Environmetrics. 2013;24(7):449-460.
12. Josse J, Holmes S. Measuring multivariate association and beyond. Stat Surv. 2016;10:132.
13. Haseman J, Elston R. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet. 1972;2(1):3-19.
14. Schaid DJ, McDonnell SK, Hebing SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet. 2005;76(5):780-793.
15. Beckmann L, Thomas D, Fischer C, Chang-Claude J. Haplotype sharing analysis using Mantel statistics. Hum Hered. 2005;59(2):67-78.
16. Schaid DJ. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered. 2010;70(2):109-131.
APPENDIX

A.1 Relationship between squared distance and similarity

As assumed in this article, both $X_{n \times p}$ and $Y_{n \times q}$ have been column centered. The squared weighted distance between subjects $i$ and $j$ is

$$D_{w,ij}^2 = (x_i - x_j)^T W (x_i - x_j) = x_i^T W x_i - 2 x_i^T W x_j + x_j^T W x_j,$$

where $x_i^T \in \mathbb{R}^p$ is the $i$th row of the data matrix $X$.

Let $C$ be the “centering” matrix, that is, $C = I_n - \frac{1}{n} 1_n 1_n^T$, where $I_n$ is the identity matrix of size $n$ and $1_n$ is the vector of ones of size $n$. Let $e_i$ denote the vector whose $i$th element is 1 and 1 and the other elements are 0. Then the $(i,j)$th element of $-\frac{1}{2} CD_{w} C$ is $-\frac{1}{2} (e_i - 1_n/n)^T D_{w}^2 (e_j - 1_n/n)$. Note that

$$(e_i - 1_n/n)^T D_{w}^2 (e_j - 1_n/n) = e_i^T D_{w}^2 e_j - e_i^T D_{w}^2 1_n/n - 1_n^T D_{w}^2 e_j/n + 1_n^T D_{w}^2 1_n/n^2$$

$$= D_{w,ij}^2 - \frac{1}{n} \sum_j D_{w,jk}^2 - \frac{1}{n} \sum_j D_{w,jk}^2 + \frac{1}{n^2} \sum_{ij} D_{w,ij}^2$$

$$= x_i^T W x_j - 2 x_i^T W x_j + x_j^T W x_j$$

How to cite this article: Pluta D, Shen T, Xue G, Chen C, Ombao H, Yu Z. Ridge-penalized adaptive Mantel test and its application in imaging genetics. Statistics in Medicine. 2021;40:5313-5332. https://doi.org/10.1002/sim.9127
\[-\frac{2}{n} \sum_j \left( x_j^T W x_j - 2 x_j^T W y_j + x_j^T W x_j \right) + \frac{1}{n^2} \sum_{ij} \left( x_i^T W x_i - 2 x_i^T W y_j + x_i^T W x_j \right) = -2x_i^T W x_j - \frac{2}{n} x_j^T W \sum_j x_j + \frac{1}{n^2} \sum_i \left( x_i^T W \sum_j x_j \right) = -2x_i^T W x_j.\]

The last step is true because the columns of \( X \) have been centered. As a result, the \((i,j)\)th element of \(-\frac{1}{2}CD_w C\) is \(x_i^T W x_j\), which is the \((i,j)\)th element of \(XWx^T\).

Note that

- **Mahalanobis**: \( W = (X^T X)^{-1} \Rightarrow XWX^T = H_0, \)
- **Ridge penalized**: \( W = (X^T X + \lambda_X I_p)^{-1} \Rightarrow XWX^T = H_\lambda, \)
- **Euclidean**: \( W = I_p \Rightarrow XWX^T = H_\infty. \)

This established the relationship between squared distance and similarity (weighted inner product).

### A.2 Proof of Proposition 1

**Proposition 1.** Let \( H_0 = X(X^T X)^{-1} X^T, K_0 = Y(Y^T Y)^{-1} Y, H_\infty = XX^T, \) and \( K_\infty = YY^T. \) The score statistic from the fixed effects model is proportional to \( tr(H_0 K_0). \) When \( \Sigma = \sigma^2 I_q \) and \( \Sigma = \sigma^2 I_q, \) which is always true when \( q = 1, \) the score statistic from the random effects model is proportional to \( tr(H_\infty K_\infty). \) Asymptotically, their null distributions are chi-squared and mixture of chi-squared, respectively.

**Proof.** The derivations are mainly based on gradients with respect to vectors and matrices. See the Matrix Cook Book\(^70\) for a collection of useful derivative results.

**Fixed effects model.** Suppose the random matrix \( Y \) follows a matrix normal distribution: \( Y \sim N(X\beta, \Sigma, \Sigma). \) Equivalently, \( vec(Y) \sim N(vec(X\beta), \Sigma \otimes I_n), \) where \( vec(\cdot) \) denotes the vectorization operator. The log-likelihood is

\[ l = C - \frac{1}{2} tr[\Sigma^{-1}(Y - X\beta)^T(Y - X\beta)] = \hat{C} - \frac{1}{2} tr[\Sigma^{-1}\beta^TXX^T\beta] + tr[\Sigma^{-1}Y^TX\beta], \]

where \( C \) and \( \hat{C} \) are constant terms that do not involve parameters of interest.

The score is \( U_\beta = \frac{\partial l}{\partial \beta}|_{\beta=0} = X^T Y \Sigma^{-1}. \) Under the null hypothesis \( \beta = 0, \) it follows the matrix normal distribution \( N(0, X^T X, \Sigma^{-1}). \) Equivalently, we have \( vec(U_\beta) \sim N(0, \Sigma^{-1} \otimes (X^T X)). \) The corresponding score statistic is

\[ vec(U_\beta)^T[\Sigma^{-1} \otimes (X^T X)]^-1 vec(U_\beta) = tr[X(X^T X)^{-1}X^T Y \Sigma^{-1} Y^T] = tr[X(X^T X)^{-1}X^T Y (Y^T Y)^{-1} Y^T]/n. \]

In the last step, we replaced \( \Sigma \) with a consistent estimator \( Y^T Y/n. \) Thus, the score statistic for the fixed effects model is proportional to \( tr(H_0 K_0), \) where \( H_0 = X(X^T X)^{-1} X^T \) and \( K_0 = Y(Y^T Y)^{-1} Y. \) Note that \( tr(H_0 K_0) \) is Pillai’s trace statistic, which also equals the sum of the squared canonical correlations.

**Random effects model.** Under the random effects model, \( vec(Y) \sim N(0, V), \) and the log-likelihood is

\[ l = C - \frac{1}{2} \log|V| - \frac{1}{2} vec(Y)^T V^{-1} vec(Y), \]

where \( V = \Sigma_b \otimes XX^T + \Sigma \otimes I_n. \) When \( \Sigma_b = \sigma^2_b I_q, \) the score is

\[ U_{\sigma^2} = -\frac{1}{2} tr[\Sigma^{-1}] tr[XX^T] + \frac{1}{2} vec(Y)^T (\Sigma^{-2} \otimes XX^T) vec(Y). \]
Note that the first term in $U_{\sigma^2}$ is the expected value of the second term under the null hypothesis. Similar to others, we use the quadratic part, which can be rewritten as $tr[XX^TY\Sigma^{-2}Y^T]$. When $\Sigma = \sigma^2I_p$, up to a scaling factor, the score test statistic is equivalent to $tr[XX^TYY^T] = tr(H_{\infty}K_{\infty})$.

**Large-sample distributions.** It is known that when the sample size $n$ is large and the null hypothesis is true, the distribution of a quadratic form can be approximated by a mixture of chi-squared distributions.

For $tr(H_0K_0)$, the weights are equal, as the nonzero eigenvalues of $H_0 = X(X^TX)^{-1}X^T$ are all 1. Thus, after rescaling $tr(H_0K_0)$ appropriately, its null distribution is approximately a chi-squared distribution when the null hypothesis is true. As a comparison, for $tr(H_{\infty}K_{\infty})$, the weights depend on the singular values of $X$, which is noted in earlier work.

**A.3 Proof of Theorem 1**

**Theorem 1** (Limits of multivariate ridge correlations). Let $X_{nxp}$ and $Y_{nqx}$ be column-centered data matrices of $p$ features and $q$ features, respectively. Let

$$H_{\lambda_x} = X(X^TX + \lambda_xI_p)^{-1}X^T, \quad K_{\lambda_y} = Y(Y^TY + \lambda_yI_q)^{-1}Y^T.$$ 

Then

$$\lim_{\lambda_x \to \lambda_x^*} R(H_{\lambda_x}, K_{\lambda_y}) = R(H_{\lambda_x^*}, K_{\lambda_y^*}), \quad (A1)$$

where $H_0 = X(X^TX)^{-1}X^T$, $K_0 = Y(Y^TY)^{-1}Y^T$, $H_{\infty} = XX^T$, $K_{\infty} = YY^T$.

**Proof.** Consider the SVD of $X$. Let $\text{rank}(X) = r$ where $r \leq \min(p, n)$, the SVD is $X = UDVT$, where $U$ is $n \times r$ with $U^TU = I_r$, $D$ is an $r \times r$ diagonal matrix of singular values $d_j, j = 1, \ldots, r$, and $V$ is $p \times r$, with $V^TV = I_r$. Let $Z = U^TY(Y^TY + \lambda_yI_q)^{-1/2}$. We have

$$\lim_{\lambda_x \to \infty} R(H_{\lambda_x}, K_{\lambda_y}) = \frac{1}{\sqrt{\text{tr}(K_{\lambda_y}^2)}} \lim_{\lambda_x \to \infty} \frac{\text{tr}(H_{\lambda_x}K_{\lambda_y})}{\text{tr}(H_{\lambda_x}^2)}$$

$$= \frac{1}{\sqrt{\text{tr}(K_{\lambda_y}^2)}} \lim_{\lambda_x \to \infty} \frac{\text{tr}(Z^TDVT(X^TX + \lambda_xI_p)^{-1}V^TDZ)}{\text{tr}(U^DV^T(X^TX + \lambda_xI_p)^{-1}VD^2V^T(X^TX + \lambda_xI_p)^{-1}VD^2VT)}$$

$$= \frac{1}{\sqrt{\text{tr}(K_{\lambda_y}^2)}} \lim_{\lambda_x \to \infty} \frac{\sum_{j=1}^r d_j^2Z_j^2/(d_j^2 + \lambda_x)}{\sum_{k=1}^r \left(\frac{d_k^2}{d_k^2 + \lambda_x}\right)^2}$$

$$= \frac{1}{\sqrt{\text{tr}(K_{\lambda_y}^2)}} \sum_{j=1}^r d_j^2Z_j^2 \left[\sum_{k=1}^r \frac{d_k^2}{d_k^2 + \lambda_x^*} \left(\frac{d_j^2 + \lambda_x^*}{d_k^2 + \lambda_x^*}\right)^2\right]^{-1/2}$$

$$= \frac{1}{\sqrt{\text{tr}(K_{\lambda_y}^2)}} \sum_{j=1}^r d_j^2Z_j^2 \left[\sum_{k=1}^r \frac{d_k^2}{d_k^2 + \lambda_x^*} \left(\frac{d_j^2 + \lambda_x^*}{d_k^2 + \lambda_x^*}\right)^2\right]^{-1/2}$$

$$= \frac{\sum_{j=1}^r d_j^2Z_j^2}{\sqrt{\text{tr}(K_{\lambda_y}^2)} \sqrt{\sum_{k=1}^r d_k^2}}$$

$$= \frac{\text{tr}(H_{\infty}K_{\lambda_y})}{\sqrt{\text{tr}(H_{\infty}^2)} \sqrt{\text{tr}(K_{\lambda_y}^2)}}$$

$$= R(H_{\infty}, K_{\lambda_y}).$$
Similarly, we can show that
\[
\lim_{\lambda_i \to \infty} R(H_{\infty}, K_{\lambda_i}) = R(H_{\infty}, K_{\infty}).
\]
and it follows that
\[
\lim_{\lambda_i \to \infty} \lim_{\lambda_j \to \infty} R(H_{\lambda_i, \lambda_j}) = R(H_{\infty, \infty}).
\]

**A.4 Computational complexity**

If the feature space is very high-dimensional or if \( n \) is large, a straightforward implementation of Algorithm 1 may be computationally impractical. However, when only ridge kernels with varying values of \( \lambda \) are included in AdaMant, two approaches can be considered to greatly reduce the computational cost. For the ease of presentation, we illustrate our computational strategy for univariate \( Y \), that is, \( q = 1 \).

The first approach utilizes the SVD \( X = UDD^T \). The computational complexity needed for finding the SVD for \( X \) is \( O(np^2) \). Once the SVD is computed, we can compute \( Z = U^T Y \) and \( S_i = \sum_{i=1}^r \frac{n}{\lambda_i + \eta_i} Z_i^2 \), which has a total complexity of \( O(nr) \). Note that when \( n \gg p \), the rank \( r \) is no greater than \( p \); as a result, the cost needed for calculating \( S_i \) is \( O(np) \approx O(np) \). Calculating the test statistics for \( B \) permutations requires \( O(Bn^2) \), for a total computational complexity of \( O(np^2 + Bnp) \).

On the other hand, when \( p \) is very large relative to \( n \), SVD of \( X \) is not the best option. In this situation, we can use the identity derived by Henderson and Searle:71

\[
X(X^TX + \lambda I_p)^{-1} = (XX^T + \lambda I_n)^{-1}X.
\]

Thus, the dimension of the matrix to be inverted is lowered to \( n \times n \), rather than \( p \times p \). From this identity, \( K_i \) can be rewritten as \( K_i = (XX^T + \lambda I_n)^{-1}XX^T \). Note that calculating \( K_i \) involves multiplying the \( n \times p \) matrix \( X \) and the \( p \times n \) matrix \( X^T \), multiplying two \( n \times n \) matrices, and inverting an \( n \times n \) matrix. When \( p \gg n \), the computation cost is dominated by calculating \( XX^T \), which has a complexity of \( O(n^2p) \). The Mantel test statistic can be calculated as \( tr(YY^TK_i) = Y^TK_iY \), which has a complexity of \( O(n^2) \). With \( B \) permutations, the total computational complexity is \( O(n^2p + n^2B) \), which is less than the required computational complexity using SVD.

**A.5 Simulating realistic imaging genetics data**

We generated imaging and genetic data for \( n = 200 \) subjects, each with \( p = 300 \) SNPs and \( q = 20 \) EEG channels. To generate the genetic data, observations were placed into one of two balanced groups to form two groups of genetically similar observations. For groups \( k = 1, 2 \), a group-wide SNP row vector \( X^{(k)} \) of length 300 is generated by iid sampling from a discrete uniform distribution over \{−1, 0, 1\}; for subject \( i \), a subject-specific error row vector \( \varepsilon_i \sim N(0, I_{300}) \) is added to the group SNP vector to get the subject SNP vector \( X^{(k)} + \varepsilon_i \), where \( k \) is the group that subject \( i \) belongs to. Finally, we stack the data together to create the \( n \times p (=200 \times 300) \) genetic data matrix \( X \):

\[
X = \begin{pmatrix}
1 & 0 \\
\vdots & \vdots \\
1 & 0 \\
0 & 1 \\
\vdots & \vdots \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
X^{(1)} \\
X^{(2)}
\end{pmatrix} + \varepsilon.
\]

To generate data with spectral characteristics similar to real EEG data, for each subject 20 time series of length 1000 were generated (for 20 “channels”) using weighted mixtures of two independent latent AR(2) processes (see Gao et al.49). Let \( Z^{(1)}(t) \) and \( Z^{(2)}(t) \) denote the two AR(2) processes with the following coefficients:

\[
Z^{(1)}(t) = 1.975Z^{(1)}(t-1) - 0.99Z^{(1)}(t-2) + w_i^{(1)}, Z^{(2)}(t) = 1.996Z^{(2)}(t-1) - 0.99Z^{(2)}(t-2) + w_i^{(2)},
\]
where \( w_1^{(1)} \) and \( w_1^{(2)} \) are independent white noises with the noise level \( \sigma_w = 0.1 \). A weighted mixture of the two AR(2) processes produces a time series with spectral peaks at 5.12 Hz (from \( Z^{(1)}_t \)) and 12.8 Hz (from \( Z^{(2)}_t \)), with the magnitude of the peaks determined by the mixing weights. The mixing weights for the first 10 channels (common across all subjects) are drawn iid from N(0,1), so that no genetic effect exists for these channels. We used two methods to generate AR mixing weights linking the genetic features to the remaining channels. In each case, after generating the initial weights using the methods described below, the weights for each subject and channel were then squared and scaled so that the weights are nonnegative and sum to one. From the resulting EEG data, the connectivity matrix for each observation was calculated from the pairwise coherence for all pairs of channels.

In the first approach for generating mixing weights, a set of common weights was produced for each group of “genetically” similar subjects, where the weight for channel \( j \), group \( k \), and AR(2) component \( m \) was generated as \( W_{jkm} = W_{\text{scale}} \cdot W_{\text{genetic}}^m \), for \( W_{jkm} \sim \text{Bern}(p) \). With this method, \( W_{\text{scale}} \) controls the within-group similarity of the mixing weights, and \( p \in [0,1] \) controls the expected number of channels that have similar weights within each group; in the present simulations we set \( p = 0.1 \). The mixing weight for subject \( i \), channel \( j \), group \( k \), and component \( m \) was then generated as \( W_{jkm} + \omega_{ijkm} \), where \( \omega_{ijkm} \) is an idiosyncratic positive perturbation drawn from a log-normal distribution with parameters \( \mu_\omega = 0.5, \sigma_\omega = 0.5 \). The results for data generated with this approach are shown in the left panel of Figure 4.

In the second approach, we generated mixing weights from a multivariate variance components model so that genetically similar subjects have similar weights. Let \( Y_j(t) \) denote the \( 200 \times 1 \) vector for the \( j \)th EEG channel of the 200 subjects. This can be compactly expressed as mixing of the two latent AR(2) time series, which is specified by the following equation

\[
Y_j(t) = W_{1j}(1 \le j \le 10) + \left( W_{1j} \right) S^{(1)}(t) + \left( W_{2j} \right) S^{(2)}(t), \quad (11 \le j \le 20),
\]

where \( W_{1j} \sim \text{N}(0,1) \) and \( W_{2j} \sim \text{iid} \). For this approach, we can control the strength of association by increasing \( \sigma_w^2 \), which increases the likelihood that subjects with similar genetic covariates will have similar mixing weights (globally across both groups). The results for data generated with this approach are shown in the right panel of Figure 4.

### A.6 Calculating coherence

The coherence between two EEG channels at a particular frequency band is a measure of the oscillatory concordance of the two signals at that band. Coherence matrices for each subject were calculated as follows.

1. Suppose we have \( Q \) channels and denote the time series at the \( q \)th channel of the \( r \)th trial to be \( X_q^r(t) \) where \( t = 1, \ldots, T \), \( r = 1, \ldots, R \), and \( q = 1, \ldots, Q \).
2. Fourier transform: We compute the discrete Fourier transform (DFT) of the time series at each channel \( q \) and trial \( r \)

\[
d_q^r(\omega_j) = T^{-1/2} \sum_{t=1}^{T} X_q^r(t) \exp(-2\pi i \omega_j t)
\]

for \( j = 0, 1, \ldots, T-1 \), where the standardized frequencies \( \omega_j = j/T \in (-0.5, 0.5] \). Denote the DFT of the \( q \)th channel at trial \( r \) and frequency \( \omega_j \) to be \( \tilde{d}(q,j) \). The cross spectrum of the \((q_1,q_2)\)th channels at frequency \( \omega_j \) is \( d^r(q_1,j) \cdot d^r(q_2,j)^\ast \) (where \( \ast \) denotes conjugate transpose). This is the \((q_1,q_2)\)th element of spectral matrix at frequency \( \omega_j \) and trial \( r \). Denote the spectral matrix at frequency \( \omega_j \) and trial \( r \) as

\[
S^r(\omega_j) = \begin{pmatrix}
S_{11}^r(\omega_j) & S_{12}^r(\omega_j) & \cdots & S_{1Q}^r(\omega_j) \\
S_{21}^r(\omega_j) & S_{22}^r(\omega_j) & \cdots & S_{2Q}^r(\omega_j) \\
\vdots & \vdots & \ddots & \vdots \\
S_{Q1}^r(\omega_j) & S_{Q2}^r(\omega_j) & \cdots & S_{QQ}^r(\omega_j)
\end{pmatrix}.
\]

3. Averaging over frequencies within bands: We set the cutoff value for frequency to be 45 Hz (which is equivalent to the standardized frequency \( \omega = 45/256 \)). For each frequency band, we average the spectral matrix over frequencies within that band. Let \( S^r(f) \) denote the \( Q \times Q \) spectral matrix at trial \( r \) and band \( f \), with its \((m,m')\)th element being \( S_{mm'}^r(f) \).

4. Averaging over trials: For frequency band \( f \), we average the spectrum matrices across the \( R \) trials and denote the resulting spectral matrix by \( S(f) \), that is, \( S(f) = \frac{1}{R} \sum_{r=1}^{R} S^r(f) \).
5. Calculate coherence matrix: For each subject, we calculate the coherence matrix for each frequency band based on the corresponding spectral matrix. The coherence between channel $m$ and $m'$ at frequency band $f$ can be calculated as

$$r_{mm'}(f) = \frac{|S_{mm'}(f)|^2}{S_{mm}(f)S_{m'm'}(f)}.$$ 

A.7 Design of EEG working-memory experiment

Each trial consists of the presentation of four images in sequence, with a total duration being approximately three seconds, structured as follows: (i) fixation image for 400 ± 200 millisecond; (ii) the standard array to be memorized (100 millisecond); (iii) the blank maintenance image (900 millisecond); (iv) comparison array is shown, and subjects are prompted to respond within 2000 millisecond if the targets in the comparison array match or mismatch the standard array. The total duration of the experiment, which consists of 359 trials, was 5 to 10 minutes for each subject. EEG data were downsampled to 256 Hz, channels with corrupted signals were removed, and a band pass filter from 1 to 40 Hz was applied. Independent components analysis was then used to remove physiological artifacts such as eye blinks and other muscle movement. Lastly, all subjects were manually inspected to remove any remaining artifacts.