Adaptive Mantel Test for Penalized Inference, with Applications to Imaging Genetics

Dustin Pluta[1], Tong Shen[1], Hernando Ombao[2], and Zhaoxia Yu[1]

[1] *University of California, Irvine*

[2] *King Abdullah University of Science and Technology*

Abstract

Mantel's test (MT) for association is conducted by testing the linear relationship of similarity of all pairs of subjects between two observational domains. Motivated by applications to neuroimaging and genetics data, and following the succes of shrinkage and kernel methods for prediction with high-dimensional data, we here introduce the adaptive Mantel test as an extension of the MT. By utilizing kernels and penalized similarity measures, the adaptive Mantel test is able to achieve higher statistical power relative to the classical MT in many settings. Furthermore, the adaptive Mantel test is designed to simultaneously test over multiple similarity measures such that the correct type I error rate under the null hypothesis is maintained without the need to directly adjust the significance threshold for multiple testing. The performance of the adaptive Mantel test is evaluated on simulated data, and is used to investigate associations between genetics markers related to Alzheimer's Disease and heatlhy brain physiology with data from a working memory study of 350 college students from Beijing Normal University.

Keywords: High-dimensional inference; distance-based methods; kernel regression; neuroimaging; genetics.

1 Introduction

A common goal of modern scientific studies is determining statistically significant relationships between multiple sets of features. In a collaborative work with behavioral scientists, the key interest of our research is to determine the extent to which genetic factors explain the variability in brain function (activation and connectivity) and how both genetics and brain function may influence human behavior. In most biomedical studies, these sets of features are often complexly structured, with each set having dimension orders of magnitude larger than the sample size. As a consequence of the "curse of dimensionality" in high-dimensional settings, only factors with large effect sizes can be detected after adjusting for multiple comparisons. The deficiencies of massive univariate analyses have motivated methods that aggregate effects from many individual variables. The goal of such methods is to assess the joint effects of a set of variables, rather than separately estimating many univariate effects on the outcome of interest. Examining the overall effects is particularly relevant in the presence of two sets of high-dimensional data from distinct measurement modalities. For instance, in our collaborative work, we examine data from different imaging modalities, such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) which measure different aspects of brain function. Thus, our goal is to develop a method to determine whether these two modalities are significantly related and how that association might vary across different experimental conditions. Several parametric and nonparametric approaches to multi-modal association testing have been developed and revisited in the scientific and statistics literature.

Mantel's test Mantel (1967) is one of the earliest formulations of a distance-based, ostensibly nonparametric, test for association of features between two observational modalities. To conduct the Mantel test, one computes a similarity or distance between each pair of subjects in each modality, and tests for significance of the correlation of similarities via an asymptotic

distribution or permutation testing procedure. Formally, let X and Y be measurements on n subjects from two observational modalities $\mathbf{X}$ and $\mathbf{Y}$, with metrics δ_X and δ_Y giving rise to (dis)similarity matrices K and H respectively, where $K_{ij} = \delta_X(X_i, X_j)$ represents the dis(similarity) between subjects i and j with respect to the modality $\mathbf{X}$, and similarly for H with respect to the modality $\mathbf{Y}$. The original form of the MT statistic can then be stated as

$$T = \sum_{i=1}^{n} \sum_{j>i} K_{ij} H_{ij},$$

from which significance can be assessed using an asymptotic distribution or permutation procedure.

Alternative, but closely related tests, have been developed from the RV coefficient Robert and Escoufier (1976) and the distance covariance Székely et al. (2007). The RV coefficient is defined as the correlation of pairwise similarities across the two modalities, and thus is a scaled version of a similarity MT. Distance covariance Székely et al. (2007) is a more recent approach, which is defined as the covariance of distances between random vectors $X, X' \in \mathbf{X}$ and $Y, Y' \in \mathbf{Y}$. A detailed discussion of the connections of MT, the RV coefficient, and distance covariance is given in Omelka and Hudecová (2013), as well as simulation results indicating that tests with distance covariance have higher power than the standard Mantel test (using Euclidean distance matrices) in some situations. However, if we choose K and H in the Mantel test to be the Gower-centered distance matrices, then the Mantel test and the dCov tests are equivalent Omelka and Hudecová (2013).

The primary contribution of this article is the introduction of the adaptive Mantel test as a method for high-dimensional inference, which improves upon the classical Mantel test by simultaneously testing across a set of similarity measures without the need to explicitly adjust for multiple comparisons. Section 2 reviews the original MT and establishes the relationship of MT and linear model score statistics, from which a unified formulation of

the fixed effects, random effects, and ridge regression score tests is developed. Section 3 describes the implementation and use of the adaptive Mantel test (AMT). Section 4 evaluates the performance of the test through simulations, and illustrates the use of AMT with an investigation of EEG and genetic data from a working memory study of 350 college students. Section 5 concludes with a discussion of practical implications and directions for future work.

2 Mantel's Test for Linear Association

2.1 Forms of Mantel's Test

If two statistics T_1 and T_2 produce equivalent test results when calculated on the same size n sample, we refer to these statistics as *testing equivalent*, and write $T_1 \asymp T_2$. When the tests from T_1 and T_2 are asymptotically equivalent as $n \to \infty$, we say T_1 is *asymptotically testing equivalent* to T_2, and write $T_1 \doteq T_2$. To develop the theoretical foundation of the adaptive Mantel test, we first discuss the original form of Mantel's test and introduce a modified form that is asymptotically testing equivalent. Consider n independent observations each measured on $\mathbf{X} \times \mathbf{Y}$, where X is the $n \times p$ design matrix from the p-dimensional feature space $\mathbf{X}$ and Y the $n \times 1$ response vector from a univariate feature space $\mathbf{Y}$, and further suppose X is column-centered, and Y is centered. Let K be an $n \times n$ similarity matrix that measures the similarity of observations in $\mathbf{X}$. For $\delta^{\mathbf{X}}(\cdot, \cdot)$ a similarity or dissimilarity metric on $\mathbf{X} \times \mathbf{X}$, let K be the corresponding Gram matrix with $K_{ij} = \delta(X_i, X_j)$. Similarily, for some metric $\delta^{\mathbf{Y}}$ defined on $\mathbf{Y} \times \mathbf{Y}$, let H be the corresponding Gram matrix of Y. The original Mantel test statistic is defined as

$$T(X, Y) = \sum_{i=1}^{n} \sum_{i<j} H_{ij} K_{ij}.$$

The reference distribution under the null hypothesis of no association between the $\delta^{\mathbf{X}}$

measures and $\delta^{\mathbf{Y}}$ measures, 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 for T. Equivalently, one can hold one matrix fixed, say H, and simultaneously permute the rows and columns of K. In Mantel's original formulation, T uses only the *upper* triangle of each matrix, excluding the diagonal. In this paper, we define a modified Mantel test statistic, denoted T^*, as

$$T^*(X, Y) = \sum_i \sum_j H_{ij} K_{ij} = tr(HK).$$

Testing T^* will not be equivalent to testing T when the diagonals of H and K are both nonconstant, but it can be shown that $T^* \doteq T$ Martins-Filho and Yao (2006). For the remainder of this article, we will focus on tests with similarity measures that can be written as weighted Euclidean inner products, as these tests have a close connection with linear models. Extending these results to a broader class of kernel similarity measures is a direction for future work.

We now define three Gram matrices calculated from weighted inner products, which we denote K_R, K_F, and K_λ since these will be shown in the following section to correspond to the score tests for the random effects, fixed effects, and ridge regression models respectively. For a positive semi-definite (p.s.d.) weight matrix $\mathcal{W}$, the corresponding weighted inner product is calculated as $\langle X_i, X_j \rangle_{\mathcal{W}} = X_i^T \mathcal{W} X_j /$. Choosing $\mathcal{W} = \mathbf{I}$ gives the standard Euclidean inner product, with Gram matrix

$$K_R := XX^T$$

where $K_{R,ij} = \langle X_i, X_j \rangle$, $K_{R,ii} = \|X_i\|^2$. Another natural choice for weight matrix is $\mathcal{W} = (X^T X)^{-1}$ (assuming the inverse exists), which is the projection matrix into $\mathcal{C}(X)$, the column space of X. The Gram matrix is

$$K_F := X(X^TX)^{-1}X^T.$$

K_F is also recognizable as the "hat matrix" from the fixed effects model, and is related to the Mahalanobis distance. When X^TX is not full rank, such as when $n > p$, we can replace $(X^TX)^{-1}$ with a generalized inverse $(X^TX)^-$, since the similarity matrix is invariant to the choice of inverse. Alternatively, we can pre-condition the weight matrix by adding a positive constant λ to the diagonal, i.e. $\mathcal{W} = (X^TX + \lambda I_p)^{-1}$. This gives similarity matrix

$$K_\lambda := X(X^TX + \lambda I_p)^{-1}X^T.$$

2.2 Linear Model Score Tests

In general, the score test is defined as follows. For a stochastic model with parameters $\theta = (\beta, \alpha)$ and observations (X, Y), and with likelihood $\mathcal{L}(\theta)$, score vector $\mathcal{U}(\theta) = \frac{\partial \log \mathcal{L}}{\partial \theta}$, and Fisher information $\mathcal{I}(\theta) = -\mathbb{E}\frac{\partial}{\partial \theta}\mathcal{U}(\theta)$, the score test statistic for testing $H_0 : \beta = 0$ is

$$S = [\mathcal{U}(\theta)^T(\mathcal{I}(\theta))^{-1}\mathcal{U}(\theta)]|_{\beta=0}.$$

The classical fixed effects model is robust and broadly applicable for simple association testing, but requires $n > p$, and so is not feasible for high-dimensional settings. It does however serve as a useful theoretical comparison to the random effects and ridge regression models to be discussed next. The fixed effects model can be written

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n). \tag{1}$$

The global score test statistic is straightforward to calculate, and can be written as

$$S_F = \frac{1}{\sigma^2}\text{tr}([X(X^TX)^{-1}X^T][YY^T]).$$

In practice, the nuisance parameter σ^2 can be replaced with $\hat{\sigma}^2 = Y^TY/(n-1)$, which is the REML estimator when there are no adjustment covariates. This is a scalar that is fixed under permutations of K_F. Therefore,

$$S_F \asymp Y^TX(X^TX)^{-1}X^TY = \text{tr}(HK_F),$$

where $H = YY^T$.

Now consider the random effects model

$$Y = Xb + \epsilon,$$

where $b \sim N(\mathbf{0}, \sigma_b^2 I_p)$, $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. Under the null hypothesis of no association between X and Y, we have $\sigma_b^2 = 0$. Similar to Liu et al. (2007), we first calculate the score and use the term that involves both X and Y as our score test statistic. Thus a testing equivalent form for the random effect score test statistic is

$$S_R \asymp \text{tr}(HK_R)/\sigma^2 \asymp \text{tr}(HK_R).$$

The relationship between similarity-based tests and score tests of random-effect or kernel machine regression has been previously discussed in detail (Kwee et al. (2008); Tzeng et al. (2009); Pan (2011)).

The ridge regression model is a penalized form of 1, for which the estimator of β is defined as the minimizer of the penalized residual sum of squares:

$$\hat{\beta}_\lambda = \arg\min_b \left\{||Y - Xb||^2 + \lambda||b||^2\right\}.$$

Although the above penalized function does not correspond to a likelihood conditional on (X, Y), it can instead be formulated from the likelihood for the augmented data model

$$Y_\lambda \;=\; X_\lambda \beta + \epsilon \text{ where } Y_\lambda = \begin{pmatrix} Y \\ \mathbf{0}_{p\times 1} \end{pmatrix}, X_\lambda = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix}$$

From this augmented likelihood, we can derive the score statistic as

$$S_\lambda = \frac{1}{\sigma^2} Y_\lambda^T X_\lambda (X_\lambda^T X_\lambda)^{-1} X_\lambda^T Y_\lambda = \frac{1}{\sigma^2} \mathrm{tr}(HK_\lambda) \asymp \mathrm{tr}(HK_\lambda).$$

2.3 (B)ridging the Fixed-effects and Random-effects Models

In this section we demonstrate that T^* is closely related to well-known correlations in several linear models. The sample correlation of the similarity measures of all subjects is defined as

$$r(H, K) = \frac{tr(HK)}{\sqrt{tr(HH)tr(KK)}}.$$

Since $r(H, K)$ is a correlation, we have $-1 \le r(H, K) \le 1$ in general, and for H and K p.s.d., this is further restricted to $0 \le r(H, K) \le 1$. When permutations are used to assess the strength of association, using MT is equivalent to testing the significance of $r(H, K)$, since the denominator of $r(H, K)$ is fixed when simultaneously permuting the rows and columns of either H or K. The sample correlation of similarities for the three choices of K above can be conveniently related through the singular value decomposition (SVD) of X.

Theorem. Let X be an $n \times p$ column centered matrix of covariates with $\mathrm{rank}(X) = r$ and singular value decomposition $X = U_{n\times r} D_{r\times r} V_{p\times r}^T$, with squared singular values $\eta_i, i = 1, \cdots, r$. Let Y be an $n \times 1$ centered vector of scalar responses, and let $H = YY^T$ and $Z = U^T Y$.

1. Fixed Effects

$$r(H, K_F) = \frac{\sum_{i=1}^{r} z_i^2}{\sqrt{p}\sum_{i=1}^{n} y_i^2}. \tag{2}$$

2. Random Effects

$$r(H, K_R) = \frac{\sum_{i=1}^{r} \eta_i z_i^2}{\sqrt{\sum_{i=1}^{r} \eta_i^2 \sum_{i=1}^{n} y_i^2}} \tag{3}$$

3. Ridge Regression

$$r(H, K_\lambda) = \frac{\sum_{i=1}^{r} \frac{\eta_i}{\lambda+\eta_i} z_i^2}{\sqrt{\sum_{i=1}^{r} \left(\frac{\eta_i}{\eta_i+\lambda}\right)^2 \sum_{i=1}^{n} y_i^2}}. \tag{4}$$

4. Relation to Correlation of X and Y

$$r(H, K_F) = \frac{1}{\sqrt{p}} R^2(X, Y)$$

5. Asymptotic Equivalences

$$\lim_{\lambda \to 0} r(H, K_\lambda) = r(H, K_F)$$

$$\lim_{\lambda \to \infty} r(H, K_\lambda) = r(H, K_R).$$

The preceding theorem also gives a straightforward derivation of the null distributions for the score tests. Since the matrix U from the SVD of X is orthogonal, the distribution of Z is

$$Z = U^T Y \sim N_r(0, U^T \Sigma U),$$

where $\Sigma \equiv Cov(Y)$. The asymptotic distributions of the score test statistics then follow from standard results on the distribution of quadratic forms of normal random variables.

Corollary. Connections between the Mantel test statistics

Let $V_i \overset{iid}{\sim} \chi_1^2, i = 1, \cdots, r$ and $V \sim \chi_p^2$.

1. Fixed-effects model.

$$S_F \asymp \mathrm{tr}(HK_F) = Z^T D(D^T D)^{-1} D^T Z \ \sim \ cV$$

2. Random-effects model.

$$S_R \asymp \mathrm{tr}(HK_R) = Z^T D D^T Z \ \sim \ \sum_{i=1}^{r} \eta_i V_i,$$

3. Ridge regression.

$$S_\lambda \asymp \mathrm{tr}(HK_\lambda) = \sum_{i=1}^{r} \frac{\eta_i}{\lambda + \eta_i} z_i^2 \propto \lambda \sum_{i=1}^{r} \frac{\eta_i}{\lambda + \eta_i} z_i^2$$
$$\stackrel{\lambda \to \infty}{\to} \sum_{i=1}^{r} \eta_i z_i^2 \sim \ \sum_{i=1}^{r} \eta_i \chi_1^2$$

These results provide a description of the ridge regression score test as a natural intermediary between the fixed and random effects score tests. A crucial difference of the ridge test is the presence of the tuning parameter λ. It is obvious that when $\lambda = 0$, S_λ reduces to S_F, and $r(H, K_\lambda)$ reduces to $r(H, K_0)$. On the other hand, when $\lambda \to \infty$, the test based on a ridge-penalty converges to that based on the random-effect model. Note that in permutation based methods, multiplying a constant does not change the permuted p-value. Therefore, $S_\lambda \propto \lambda \mathrm{tr}(HK_\lambda)$, which converges to $\sum_{i=1}^{r} \eta_i z_i^2$, i.e., S_R.

Remark To better understand the differences in the tests based on T_F^* and T_R^*, a geometric comparison is helpful. Recognizing K_F as the projection into $\mathcal{C}(X)$ and writing $T_F^* := \mathrm{tr}(HK_F) = \|X(X^T X)^{-1} X^T Y\|^2$, we can interpret the test of T_F^* as testing the norm of projected image of Y into $\mathcal{C}(X)$. Equivalently, T_F^* tests the significance of the norm $\|X^T Y\|_{\mathcal{W}}$, which is the Mahalanobis norm, since $\mathcal{W}$ is the Fisher information matrix of the fixed effects model. In contrast, the statistic $T_R^* := \mathrm{tr}(HK_R)$ tests the unweighted L_2 norm of

$X^TY \in \mathcal{C}(X)$. Since the jth component of the $p \times 1$ vector X^TY is the covariance of the jth feature of X with Y, the test of T_R^* can be understood as giving equal weight to each feature in X in measuring the strength of the overall relationship of X and Y, whereas the test of T_F^* weights the contribution of each feature of X according to the Mahalanobis norm. In the case of independent features, these weights are inversely proportional to the observed variance of that feature.

The correlation formulas (2) – (4) give some additional insight into the relationship of the three models in terms of $Z = U^TY$, the image of Y after transformation by the left eigenvectors of X. From 2, we see that the fixed effects score test is equivalent to testing the Euclidean norm of Z. Whereas from 3, the random effects score test statistic is equivalent to testing the *weighted* norm of Z, where the jth component (corresponding to the jth eigenvector) is weighted by η_j. This has the effect of emphasizing the influence of directions in $\mathbf{X}$ for which X has large variance and reducing the influence of directions with small variance. The ridge regression score test is a compromise between the fixed and random effects, with small λ yielding a test close to the fixed effects (or identical at $\lambda = 0$), and large λ yielding a test close to the random effects score test, and identical tests for $\lambda \to \infty$. Geometrically, the ridge test weights the z_j proportional to the η_j as in the random effects test, but flattens each weight by a factor of $\frac{1}{\lambda+\eta_j}$.

3 Adaptive Mantel Test

Effectively using kernel methods requires an appropriate selection of the kernel function and tuning parameters for the particular setting. Selection methods have been extensively considered in the context of prediction problems, with cross-validation as the *de facto* standard. Cross-validation is a straight-forward and practical selection method for prediction, but may be difficult to implement for hypothesis testing, since the type I error rate needs

to be controlled. Furthermore, the tuning parameter selected by minimizing the CV MSE may not necessarily yield the highest powered test. This section introduces the adaptive Mantel test (AMT), which extends the classical MT to simultaneously test across a set of tuning parameters and kernels without the need to directly apply adjustments for multiple comparisons.

3.1 Algorithm for the Adaptive Mantel Test

The "adaptive" procedure used here is similar to the adaptive sum of powered score test algorithm described in Xu et al. (2017). The procedure receives as input a list of pairs of metrics/kernels $\{(\delta_m^{\mathbf{X}}, \delta_m^{\mathbf{Y}}) | m = 1, \cdots M\}$ from which the matrices $K_m = \delta_m^{\mathbf{X}}(X)$ and $H_m = \delta_m^{\mathbf{Y}}(Y)$ are computed for each metric pair, $m = 1, \cdots, M$. These metrics may be from a single family with varying tuning parameters, such as ridge kernels with different penalization terms, or may include kernels from different families.

For each $m = 1, \cdots, M$, P_m is calculated as the P-value of the Mantel test with metrics $\delta_m^{\mathbf{X}}$ and $\delta_m^{\mathbf{Y}}$ for X and Y respectively. The AMT test statistic is defined as the minimum of these values,

$$P^{(0)} := \min_{m=1,\cdots,M} P_m.$$

A permutation procedure can be used to calculate the reference distribution for $P^{(0)}$. For each m, and $b = 1, \cdots, B$, $H_m^{(b)}$ is generated by permuting rows and columns of H simultaneously, and the corresponding test statistic $P^{(b)}$ is calculated. The AMT P-value is then calculated as

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

General pseudocode for the adaptive Mantel test is given in Algorithm 1.

1: **for** $m = 1, \ldots, M$ **do**

2: $K_m \leftarrow \delta_m^{\mathbf{X}}(X)$

3: $H_m \leftarrow \delta_m^{\mathbf{Y}}(Y)$

4: Calculate $Z_m^{(0)} \leftarrow Z_m := \text{tr}(K_m H_m)$

5: **end for**

6: Generate B permutations of H_m, labeled $H_m^{(b)} \;\; \forall\, m = 1, \ldots, M; b = 1, \ldots, B$.

7: $Z_m^{(b)} \leftarrow \text{tr}(K_m H_m^{(b)}) \;\; \forall\, m = 1, \ldots, M; b = 1, \ldots, B$

8: $P_m^{(b)} \leftarrow \frac{1}{B+1} \sum_{b=0}^{B} I\left(Z_m^{(b)} \leq Z_m^{(b')}\right) \;\; \forall\, m = 1, \ldots, M; b = 1, \ldots, B$

9: $P^{(b)} \leftarrow \min_{m=1,\cdots,M} P_m^{(b)} \;\; \forall\, b = 1, \ldots, B$

10: $P_{AMT} \leftarrow \frac{1}{B+1} \sum_{b=0}^{B} I\left(P^{(0)} \leq P^{(b)}\right)$

Algorithm 1: Adaptive Mantel algorithm

3.2 Computational Methods

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 λ are included in AMT, there are two approaches that can be used to greatly reduce the computational cost.

The first approach utilizes the SVD $X = UDV^T$. 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 $S_\lambda = \sum_{i=1}^{r} \frac{\eta_i}{\lambda + \eta_i} z_i^2$, which has a total complexity of $O(nr)$. Note that when $p \gg n$, the rank r is often the same as n; as a result, the cost needed for calculating S_λ is $O(nr) = O(n^2)$. Calculating the test statistics for B permutations requires $O(Bn^2)$, for a total computational complexity of $O(np^2 + Bn^2)$.

Alternatively, when p is very large relative to n, we can use the identity $X(X^T X + \lambda I)^{-1} = (XX^T + \lambda I)^{-1} X$, so that the matrix inverse is applied to an $n \times n$, rather than $p \times p$, matrix.

From this identity, K_λ can be rewritten as

$$\begin{aligned} K_\lambda &= X(X^TX + \lambda I_p)^{-1}X^T \\ &= (XX^T + \lambda I_n)^{-1}XX^T. \end{aligned}$$

Note that calculating K_λ 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

$$S_\lambda = \text{tr}(YY^TK_\lambda) = Y^TK_\lambda Y,$$

which has a complexity of $O(n^2)$. With B permutations, the total computational complexity is $O(n^2p + Bn^2)$, which is less than the required computational complexity using SVD. Thus, switching from the feature space to the subject space (i.e., from a $p \times p$ similarity matrix of the features to an $n \times n$ similarity matrix of the subjects), has a computational advantage. Additionally, in some situations the SVD computation may be unstable, thus the matrix identity method may be recommended as the more robust approach.

3.3 Variance Explained and the Ridge Penalty

By allowing for simultaneous testing over a set of tuning parameter values, AMT lessens the challenge of parameter selection, but does not completely resolve it. Although it does not require an overly conservative adjustment for multiple testing, the power of AMT does decrease as the number of metrics considered increases. Conversely, the test results are highly sensitive to the choice of parameters to test. Consequently, one must still take some care in the selection of the included parameters, balancing the desire to use a wide range of parameter values, with the gains of using a small set of parameters. When only ridge kernels are included in AMT, previous results on the role of the ridge penalty term in predictive

modeling can help with the identification of a reasonable set of values to test. Specifically, it has been shown that when the ridge penalty λ is chosen to be the noise to signal ratio, the resulting shrinkage estimator $\hat{\beta}_\lambda$ is identical to the best linear unbiased predictor (BLUP) $\hat{b}$ for the random effects model, and moreover, for a new observation with unknown response value the predictions using the ridge and random effects models are the same de los Campos et al. (2013). Thus, when using ridge regression for prediction, it is recommended that the penalty should reflect the relevant level of noise versus signal, i.e., $\lambda = \sigma_e^2/\sigma^2$.

To apply this result to practical settings, if one can determine *a priori* a likely range for the noise to signal ratio or a related quantity, this will determine a reasonable range of penalty terms. For instance, in assessing the genetic influence on observed phenotypes, the noise to signal ratio is related to what is known as the *heritability* of the phenotype, which can be understood as the proportion of variance in the observed trait explained by the genetic data. Formally, for (standardized) genetic data matrix X consisting of alleles for p single nucleotide polymorphisms (SNPs), and observed phenotype vector Y, the random effects model given in Eq. 3 is commonly used to estimate the genetic heritability of the trait Yang et al. (2011); Liu et al. (2007). From this model, the *narrow-sense heritability* h^2 of the phenotype is defined as

$$h^2 := \frac{p\sigma^2}{p\sigma^2 + \sigma_\varepsilon^2}.$$

Plugging in $Var(Y) = \sigma^2 + \sigma_\varepsilon^2 XX^T$ gives

$$h^2 = \frac{\sigma^2 tr(XX^T/n)}{\sigma^2 tr(XX^T/n) + \sigma_\varepsilon^2 tr(I_n/n)} = \frac{tr(XX^T/n)}{tr(XX^T/n) + \sigma_\varepsilon^2/\sigma^2}.$$

We see then that if h^2 is known, the optimal penalty $\tilde{\lambda}$ (for prediction) can be found by solving for the noise to signal ratio. In practice, the scientific interpretation and range of plausible values of h^2 will depend on the specific modalities of $\mathbf{X}$ and $\mathbf{Y}$. For instance, in the

genetics literature, $h^2 > 0.5$ would generally indicate high heritability, while a heritibility of $h^2 < 0.1$ is probably not scientifically interesting. As a point of reference, most estimated heritability in the UK Biobank data is between 0.1 to 0.4 Ge et al. (2017).

4 Simulations and Applications to Imaging Genetics Data

To verify the theoretical connection between the fixed, random, and ridge regression models and to illustrate the differences in testing results, we now compare the power of these models via a simulation study and application to a real-world imaging genetics data set.

4.1 Simulations

For this simulation study of the adaptive Mantel test, the simulated data was generated from the random effects model (Eq. 3), with $n = 500$ observations, number of covariates p ranging from 250 to 2000, and $\sigma_\varepsilon^2 = 1$ fixed. For each setting, 500 simulations were run. The design matrix X was generated from n draws from a p-variate normal distribution $N(0, \Sigma_X)$, where Σ_X is chosen to have an $AR(1)$ structure with $\rho = 0.1$ and jth diagonal element equal to $\log(j + 1)$. The hypothesis of interest for heritability analyses is $H_0 : h^2 = 0$, which is equivalent to $H_0 : \sigma^2 = 0$. AMT was applied with $H = YY^T$ and K_λ for $\lambda \in \{1, 5, 10, 50, 100, 1000, \infty\}$, where ∞ corresponds to the L_2 similarity measure, with 1000 permutations.

Overall, the simulation results show that the power of the adaptive Mantel test is not substantially lower than the P-values for the individual Mantel tests. Figure 1 shows the simulation results for the heritability fixed at $h^2 = 0.035$, requiring the effect size σ^2 decrease as p increases. Comparing the power of AMT and the simple MT for each of the λ, we observe

that the power of AMT is competitive with the best of the simple Mantel tests for the λ considered. For this setting, $\lambda = 1000$ and $\lambda = \infty$ exhibited the highest power among the simple Mantel tests; smaller values had lower power relative to the larger values for all values of p. The power of all the tests decreased as p increased, leveling off with approximately 45% power for $p > 1500$. In Figure 1, $\sigma = 0.008$ is fixed, resulting in h^2 increasing as p increases. In this setting, the power of AMT and MT increase as p increases since each added feature increases h^2. AMT is again competitive with the best simple Mantel tests, and again the smaller penalty terms perform relatively poorly across the range of P-values. All of the tests converge to roughly the same power when $p > 1500$, for which the power is near 90%.

4.2 Association of EEG Coherence and Selected SNPs

We next consider data from 350 healthy college students from Beijing Normal University who participated in a visual working memory task, during which 64-channel EEG was recorded at 1 kHz. The total duration of the experiment was approximately 10 minutes for each subject. Approximately 5×10^5 SNPs were also measured for each subject. Standard pre-processing and quality control steps were applied to both the EEG and genetic data. We are interested in testing the association of alpha and theta band coherence with a group of 11 SNPs that have been identified as potentially related to Alzheimer's Disease.

The coherence between two EEG channels at a particular frequency ω is a measure of the oscillatory concordance of the the two signals at ω. 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 $n \times \binom{q}{2}$ matrix X. This results in 2080 distinct features when using all 64 channels, and 300 distinct features for the 25 selected frontal channels. The adaptive Mantel test was performed with $\lambda \in \{0.5, 1, 5, 10, 100, 1000, \infty\}$ and using 1000 permutations. Genetic similarity of subjects was calculated as the L_2 inner product of the centered standardized

Test	P-value (Best λ)
α, All Channels	0.065 (5)
α, Frontal Channels	0.381 (1)
θ, All Channels	0.416 (0.5)
θ, Frontal Channels	0.085 (0.5)

Table 1: Testing results for associations of EEG coherence and AD SNPs.

SNP data for all tests. For the alpha band, all 64 channels gave $P = 0.065$, and the selected frontal channels gave $P = 0.381$. Results for the theta band were $P = 0.416$ and $P = 0.085$ for all 64 channels and frontal channels respectively. Since the adaptive Mantel test was used, these P-values already take into account testing across multiple λ. In the case of the alpha band, the test results suggest that coherence involving channels outside of the selected frontal channels may be associated with genetic similarity determined by the 11 AD SNPs, whereas for the theta band, the SNP association appears stronger when considering only the frontal channels. For a better sense of the significance of these 11 SNPs, the AMT P-values for 200 sets of 11 randomly selected SNPs were calculated for each of the four tests considered here. The boxplot of these values are given in Figure 2. For the alpha – all channels test and the theta – frontal channels test, the P-values from the AD SNPs (0.065 and 0.085 respectively) are outside the ranges of P-values from the randomly selected SNPs.

To assess possible contributions of individual channel pairs, each channel pair was separately tested for significance with the AD SNPs with the adaptive Mantel test. Figure 2 shows the most significant channel pairs across all channel pairs for the alpha band. The TP8 – CP2 connection is the most significant at $P < 0.0001$. Other top pairs are C4 – PO4, C4 – AF7, and T8 – FT7. The most significant connections are mostly between the right temporal regions with the left frontal regions. For the theta band, the most significant channel pairs were P9 – Cz, P9 – CP3, CP3 – AF3, and P8 – F3. The P9 channel also had

relatively significant connections with many other channels in the frontal left hemisphere. These results are supported by a number of previous studies that have established links between working memory performance and features measured by EEG. For instance, Onton et al. (2005) found increases in frontal midline theta power with increasing memory load during a verbal-working memory task; Sauseng et al. (2005) also found that alpha coherence plays a significant role in "top-down" control during working memory tasks; and Simons and Spiers (2003) identified important interactions between the prefrontal and medial temporal lobes for the processing of long-term memory.

Similarly, the association of each individual SNP with EEG coherence was assessed with AMT. For alpha coherence with all channels, the most significant SNP ($P = 0.02$) is rs2227564, a functional polymorphism within plasminogen activator urokinase (PLAU) gene. An allele of this SNP has been linked to significantly high plaque counts in AD, although its role is not well-established. The second most significant SNP from the individual tests is rs3851179, a SNP upstream of the PICALM gene. This SNP has been repeatedly implicated as a factor in AD, as well as Parkinson's Disease and schizophrenia, although there are dissenting results regarding its significance in particular Chinese populations. In the genetics literature on cognitive function in healthy subjects, polymorphisms in neurotransmitter genes, such as those in the dopamine pathway, have been shown to be significantly associated with increased neuronal activity in the prefrontal cortex during working memory tasks, as measured by fMRI Bertolino et al. (2006). Vogler et al. analyzed data from the n-back memory task for 2298 subjects, and estimated genome-wide heritability of working memory accuracy to be 41% (95% CI: 0.13, 0.69) Vogler et al. (2014). Taken together, these results make it plausible genetic factors have an important influence on brain function related to working memory, but determining the specific nature of the role of genetics in brain function remains a challenging problem that will require repeated validation through a variety of different studies and experiments.

While the results of the present analysis are preliminary, they do suggest that the selected SNPs may influence characteristics of brain connectivity during a working memory task in the alpha and theta bands. If these selected SNPs truly are significant factors of memory-related brain function in healthy individuals, and given that many of these SNPs are known to discriminate healthy individuals from AD cases, further studying these SNPs in healthy subjects may lead to identification of new targets for treatment, or provide insight into the genetic mechanisms of AD.

5 Discussion

Stated as a test of the correlation of similarities or distances, the Mantel test is a geometrically motivated method for association testing. We have here shown that the Mantel test with ridge kernel similarity measures is in fact equivalent to the score tests for the fixed effects, ridge regression, and random effects models for particular choices of similarity on the covariate space $\mathbf{X}$. In high-dimensional settings, the random effects score test has been shown to have reasonable performance when a large number of covariates contribute small independent effects, but with real genetic data the validity of this assumption is questionable, as it is known that there exists correlation and more complicated relationships between SNPs, and the proportion of unrelated SNPs for a given trait is difficult to know *a priori*. Ridge regression is designed to both address collinearity between covariates and reduce the influence of noisy covariates in prediction settings. In predictive modeling, tuning parameters are often selected via cross-validation to minimize squared error loss. This is sensible and practical when the end goal is prediction, but has two major drawbacks for inference. Firstly, one must compute the post-selection null distribution of the test statistic; secondly, the best predictive model is not necessarily the highest powered for null hypothesis testing. The topic of post-selection inference has received increasing attention in recent years, lead-

ing to important developments for general post-selection inference Lee et al. (2016), and for post-selection estimation of heritability Gorfine et al. (2017), although these methods still rely on CV and minimizing squared error loss to select the tuning parameters.

We have here proposed the adaptive Mantel test as a procedure to simultaneously test across a range of tuning parameters as an alternative to other selection methods for testing. The adaptive Mantel method is also comparatively simple to describe and implement, and can naturally accomodate selecting across different families of kernels (or models). As a tool for high-dimensional inference, the AMT is a straightforward and flexible method for quickly testing the strength of association between different features of the data, and can be used as a sanity check before proceeding with more complicated modeling.

There are a number of other generalizations and extensions that can be implemented within the framework of the Mantel test. A necessary extension for application purposes is the inclusion of adjustment covariates. Specifically, suppose W is a matrix of covariates on n subjects, and we wish to test the association of X and Y adjusting for W using the Mantel test. A straightforward solution is to apply a restricted maximum likelihood approach, for which X and Y are each separately regressed on W, and then the Mantel test is performed with X and Y replaced by their corresponding residuals. Future work is also needed to characterize the class of metrics that admit a likelihood model, and describe the mapping of a metric in this class to its associated model, which could have important implications for kernel selection and geometric interpretations of model-based tests.

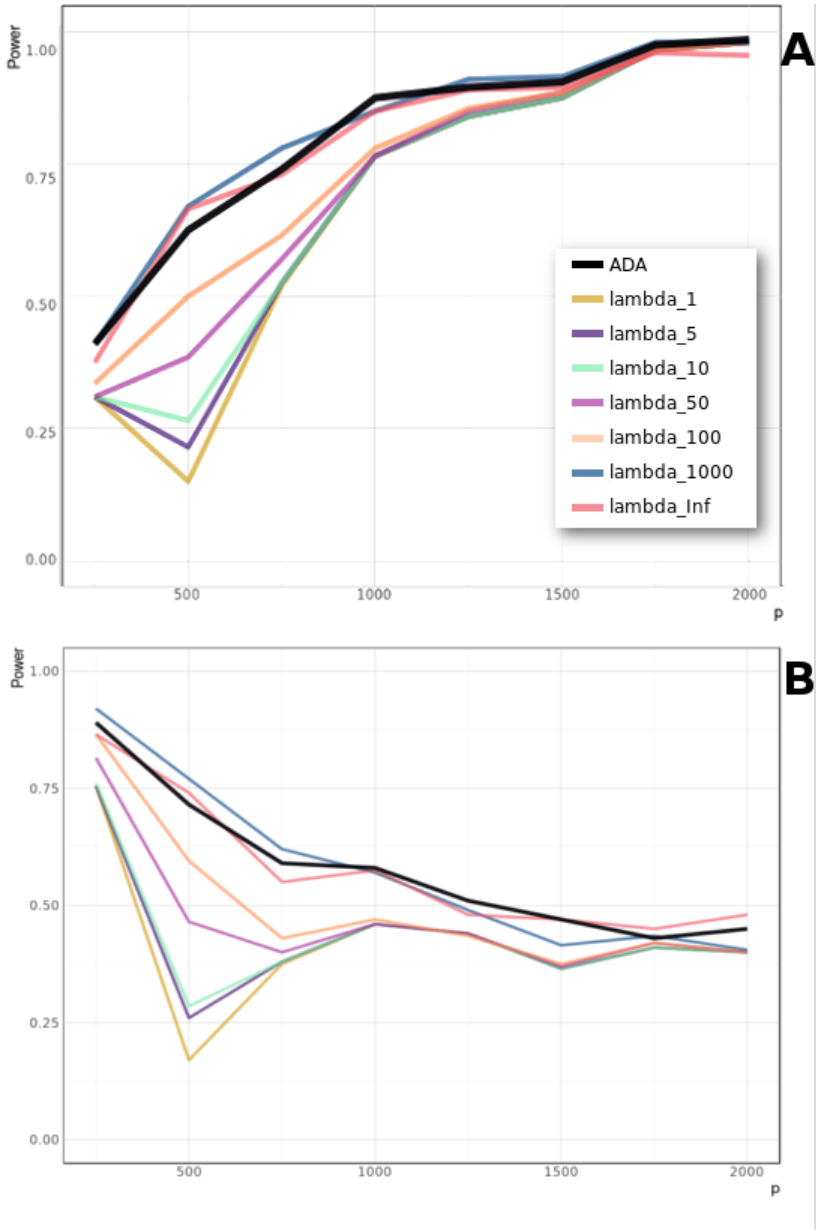

Figure 1: Simulation study of the adaptive Mantel test with $n = 500$ observations simulated from the random effects model Eq. 3. The black curve is the adaptive Mantel power; the other curves are the power for the simple Mantel test with the ridge kernel with indicated penalty term.**A** Power for data generated with constant effect size $\sigma = 0.008$ for each included feature. **B** Power for data generated from a random effects model with fixed heritability across values of p.

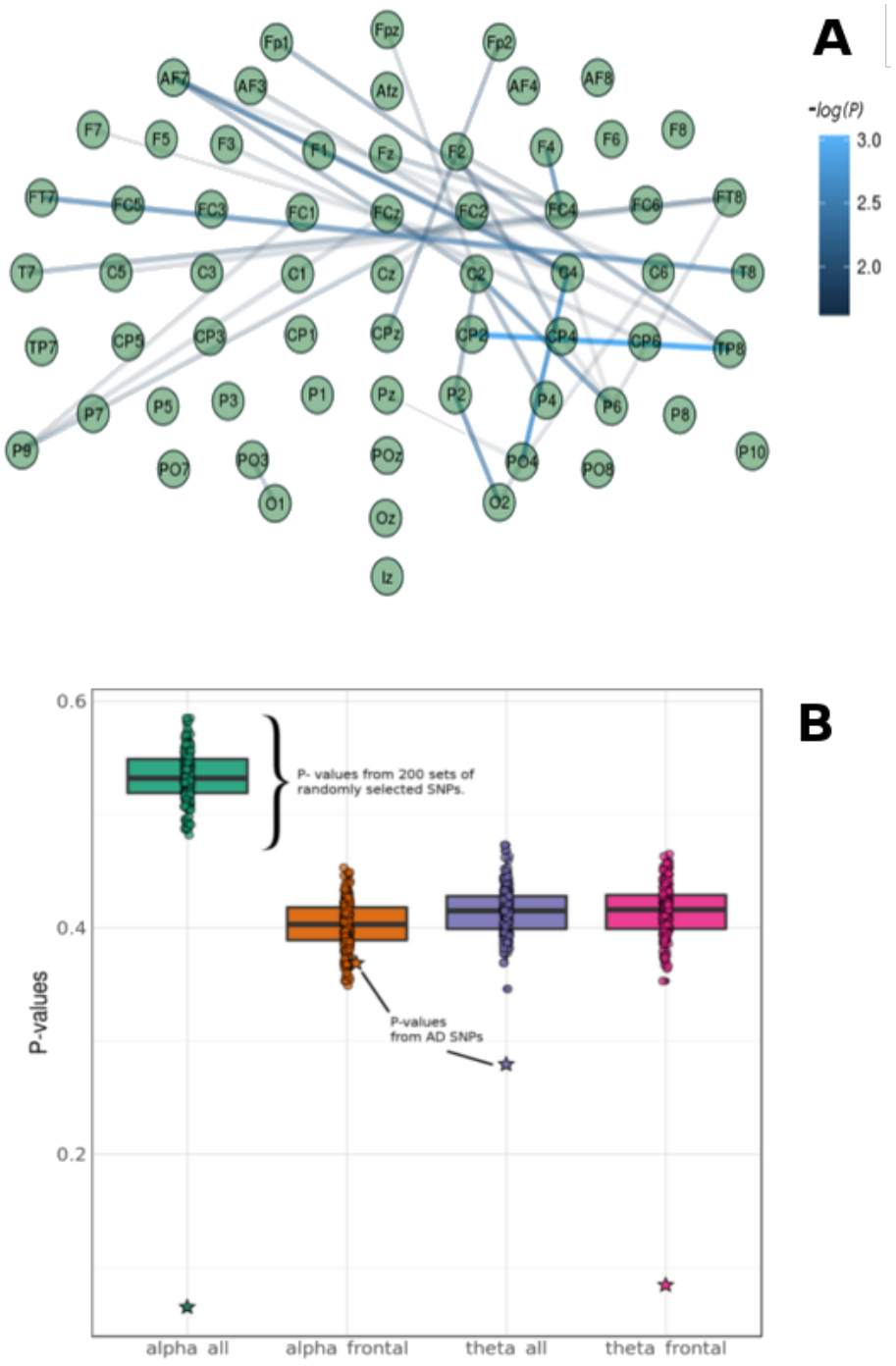

Figure 2: **A** Most significant channel pairs for α band coherence during working memory task. Edges are colored by $-log_{10}(P)$ from the univariate adaptive Mantel test for the α coherence of that channel pair with the AD SNPs. **B** Range of P-values for 200 sets of 11 randomly selected SNPs tested for significance with EEG coherence. Stars indicate the P-value from the 11 AD SNPs of interest.

References

Bertolino, A., Blasi, G., Latorre, V., Rubino, V., Rampino, A., Sinibaldi, L., Caforio, G., Petruzzella, V., Pizzuti, A., Scarabino, T., Nardini, M., Weinberger, D. R., and Dallapiccola, B. (2006). Additive effects of genetic variation in dopamine regulating genes on working memory cortical activity in human brain. *Journal of Neuroscience*, 26(15):3918–3922.

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS genetics*, 9(7):e1003608.

Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R., and Smoller, J. W. (2017). Phenome-wide heritability analysis of the uk biobank. *PLoS genetics*, 13(4):e1006711.

Gorfine, M., Berndt, S. I., Chang-Claude, J., Hoffmeister, M., Le Marchand, L., Potter, J., Slattery, M. L., Keret, N., Peters, U., and Hsu, L. (2017). Heritability estimation using a regularized regression approach (herra): Applicable to continuous, dichotomous or age-at-onset outcome. *PloS one*, 12(8):e0181269.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

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(4):1079–1088.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220.

Martins-Filho, C. and Yao, F. (2006). A note on the use of v and u statistics in non-parametric models of regression. *Annals of the Institute of Statistical Mathematics*, 58(2):389–406.

Omelka, M. and Hudecová, Š. (2013). A comparison of the mantel test with a generalised distance covariance test. *Environmetrics*, 24(7):449–460.

Onton, J., Delorme, A., and Makeig, S. (2005). Frontal midline eeg dynamics during working memory. *Neuroimage*, 27(2):341–356.

Pan, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic epidemiology*, 35(4):211–216.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Applied statistics*, pages 257–265.

Sauseng, P., Klimesch, W., Schabus, M., and Doppelmayr, M. (2005). Fronto-parietal eeg coherence in theta and upper alpha reflect central executive functions of working memory. *International Journal of Psychophysiology*, 57(2):97–103.

Simons, J. S. and Spiers, H. J. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nature reviews neuroscience*, 4(8):637–648.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

Tzeng, J.-Y., Zhang, D., Chang, S.-M., Thomas, D. C., and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, 65(3):822–832.

Vogler, C., Gschwind, L., Coynel, D., Freytag, V., Milnik, A., Egli, T., Heck, A., De Quervain, D. J., and Papassotiropoulos, A. (2014). Substantial snp-based heritability estimates for working memory performance. *Translational psychiatry*, 4(9):e438.

Xu, Z., Xu, G., and Pan, W. (2017). Adaptive testing for association between two random vectors in moderate to high dimensions. *Genetic Epidemiology*.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82.