frontiers in Neuroscience

HYPOTHESIS AND THEORY
published: 24 March 2022
doi: 10.3389/fnins.2022.836100

# Ridge Penalization in High-Dimensional Testing With Applications to Imaging Genetics

*Iris Ivy Gauran[1], Gui Xue[2], Chuansheng Chen[3], Hernando Ombao[1] and Zhaoxia Yu[4]**

[1] Biostatistics Group, Computer, Electrical, Mathematical Sciences, and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [2] Center for Brain and Learning Science, Beijing Normal University, Beijing, China, [3] Department of Psychological Science, University of California, Irvine, Irvine, CA, United States, [4] Department of Statistics, University of California, Irvine, Irvine, CA, United States

High-dimensionality is ubiquitous in various scientific fields such as imaging genetics, where a deluge of functional and structural data on brain-relevant genetic polymorphisms are investigated. It is crucial to identify which genetic variations are consequential in identifying neurological features of brain connectivity compared to merely random noise. Statistical inference in high-dimensional settings poses multiple challenges involving analytical and computational complexity. A widely implemented strategy in addressing inference goals is penalized inference. In particular, the role of the ridge penalty in high-dimensional prediction and estimation has been actively studied in the past several years. This study focuses on ridge-penalized tests in high-dimensional hypothesis testing problems by proposing and examining a class of methods for choosing the optimal ridge penalty. We present our findings on strategies to improve the statistical power of ridge-penalized tests and what determines the optimal ridge penalty for hypothesis testing. The application of our work to an imaging genetics study and biological research will be presented.

**Keywords: high-dimensional testing, genome-wide association studies, neuroimaging, ridge penalization, imaging genetics**

OPEN ACCESS

***Edited by:***
*Rongjie Liu,*
*Florida State University, United States*

***Reviewed by:***
*Ziliang Zhu,*
*University of North Carolina at Chapel Hill, United States*
*Jaroslaw Harezlak,*
*Indiana University, United States*
*Bingxin Zhao,*
*Purdue University, United States*

****Correspondence:***
*Zhaoxia Yu*
*zhaoxia@ics.uci.edu*

***Specialty section:***
*This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience*

***Received:*** *15 December 2021*
***Accepted:*** *24 February 2022*
***Published:*** *24 March 2022*

***Citation:***
*Gauran II, Xue G, Chen C, Ombao H and Yu Z (2022) Ridge Penalization in High-Dimensional Testing With Applications to Imaging Genetics. Front. Neurosci. 16:836100. doi: 10.3389/fnins.2022.836100*

## 1. INTRODUCTION

Even with the advancements of genome-wide association studies over the past two decades, unraveling the genetic basis of many complex neurological conditions remains to be a challenge. Often, each individual's genetic information has a small contribution to disease risk and can be highly heterogeneous (Peper et al., 2007; Marenco and Radulescu, 2010; Tost et al., 2012; Batmanghelich et al., 2013). Imaging genetics offers an approach to understanding the genetic basis of neurological disorders by investigating the integrated multi-scale genomic data, multimodal brain imaging information, and environmental risk factors (Thompson et al., 2013; Nathoo et al., 2019). The rationale for imaging genetics is that by examining single nucleotide polymorphisms (SNPs), we may discover essential insights into the brain-relevant genetic polymorphisms to understand better the neural architecture through which psychopathology may emerge. Typically, these studies involve a small number of subjects relative to the amount of information available per subject, such as millions of SNPs, thousands of genetic variants or differentially methylated probes, hundreds of thousands of voxels, and dozens to hundreds of electroencephalogram (EEG) channels. Hence, both explanatory and response variables in imaging genetics studies can be high-dimensional in nature.

However, the joint analysis of both high-dimensional imaging and genetic data presents major computational and theoretical challenges for existing analytical methods (Nathoo et al., 2019) as well as the proliferation of false discoveries (Meyer-Lindenberg et al., 2008). Widely-implemented methods to fit high-dimensional statistical models include penalized regression where some form of regularization is imposed. The penalized regression literature generally adopts the perspective of maximum likelihood theory. In the context of linear regression, the negative log likelihood or loss function has the form $\mathcal{L} = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ where $\mathbf{X}$ is an $n \times p$ design matrix of explanatory variables and $\mathbf{Y}$ is an $n \times q$ matrix of responses. The classic, unique solution minimizing the loss function $\mathcal{L}$ is $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ when $n > p$ and $\mathbf{X}$ has full column rank.

As $p$ increases for a fixed $n$, the direct application of regression model and likelihood-based methods are encumbered by several issues. For example, an overfitted model may lead to large variance and low performance in testing data, i.e., low generalization. When the number of explanatory variables exceeds the sample size, the least squares estimate is not unique because the computation involves inverting the singular $\mathbf{X}^\top\mathbf{X}$. Regularization methods are often adopted to overcome these problems, in which case the objective function is modified to be $Q(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}) = \mathcal{L}(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}) + P_\lambda(\boldsymbol{\beta})$ where $P_\lambda$ is a penalty function and $\lambda$ is a tuning parameter. For $\gamma \geq 0$, the $L_\gamma$ norm of $\boldsymbol{\beta}$ is formally defined as

$$\|\boldsymbol{\beta}\|_\gamma = \left(\sum_{j=1}^{p} |\beta_j|^\gamma\right)^{1/\gamma}. \tag{1}$$

This class of well-known penalization functions and criteria aim to balance the trade-off between bias and variance or between complexity and generalization. For example, both AIC and BIC, two well-known criteria, belong to the $L_0$ norm, as $\|\boldsymbol{\beta}\|_0$ is the number of non-zero elements in $\boldsymbol{\beta}$. The $L_1$ norm is often considered as a convex relaxation of the $L_0$ norm, and it achieves both sparsity and computational efficiency. The $L_2$ norm, the penalty of which is often known as the ridge penalty or Tikhonov regularization, was motivated for ill-conditioned or close to ill-conditioned problems (Tikhonov, 1943; Hoerl, 1962; Hoerl and Kennard, 1970). Ridge penalty has also been used alone or combined with $L_1$ in high-dimensional inference problems and deep neural networks. For example, for genetic predictive problems or association studies, ridge penalty has been widely used (Hayes et al., 2001; Liu et al., 2007; Cule et al., 2011; de los Campos et al., 2013; Lin et al., 2013, 2016; Zhao and Zhu, 2019).

More recently, ridge regression has been intensively studied as a way to try to understand why overfitted models can have satisfactory predictive performance in testing data. For example, it has been observed that models trained using deep neural networks not only have an almost perfect fit to the training data, but also generalize well to testing data (Zhang et al., 2016). Recall that a ridge regression applies an $L_2$ penalty, i.e., the corresponding objective function is

$$Q(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\,\|\boldsymbol{\beta}\|_2^2. \tag{2}$$

The ridge penalty is particularly attractive to work with because the maximum penalized likelihood estimator has a simple closed form. This objective function is differentiable and it is straightforward to show that its minimum occurs at

$$\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}. \tag{3}$$

Thus, the ridge solution includes the ordinary least squares solutions as a special case when $\lambda = 0$.

The ridge regression methodology yields a class of biased estimators, and massive literature is driven toward identifying an optimal ridge penalty parameter to be used in practice. The primary objective is to ensure that the ridge estimator has lowest mean squared error (Hoerl and Kennard, 1970). This translates to the pragmatic goal of developing methods which produce ridge estimates that are more useful than the least squares estimates. Despite the numerous available methods for choosing an optimal value, the ultimate choice of $\lambda$ for a specific application still remains unsolved because the optimal level of regularization usually depends on the unknown characteristics of the data generating distribution (Patil et al., 2021).

As ridge regression is mathematically neat and relatively easy to study, it has been recently widely used as the first attempt to understand under what circumstances overfitting is harmless or benign, especially in high-dimensional settings. A variety of combinations have been examined, such as asymptotic or fixed sample sizes, random coefficients or fixed coefficients, the ratio of $p$ to $n$, and conditional on or marginalize the covariate matrix. A representative but by no means complete list of studies include (Randolph et al., 2012; Dobriban and Wager, 2018; Hastie et al., 2019; Bartlett et al., 2020; Kobak et al., 2020; Patil et al., 2021) and some of them are credited for introducing eye-catching phrases such as "benign overfitting" and "double descent". These studies involving ridge regression are devoted to either performance of prediction or regression coefficient estimation. To the best of our knowledge, no systematic work has been conducted to investigate the role that ridge penalty plays in high-dimensional hypothesis testing.

Moreover, to explore the relationship of neurological and genetic information in imaging genetic studies, we are interested in determining whether given sets of features are significantly associated in aggregate. In this study, we will utilize one of the extensions (Pluta et al., 2021) of the classical Mantel test (Mantel, 1967) to characterize the association between two potentially high-dimensional distance matrices. The Mantel test (Mantel, 1967) is an easy-to-implement and flexible procedure, which was originally motivated by assessing the association between the temporal and spatial relationship of leukemia cases and similar diseases. As presented in Mantel (1967), the temporal-spatial association can be examined by using the correlation of the temporal and spatial distance matrices of the observed leukemia cases. Similar or modified approaches have been commonly applied, such as identifying the spatial pattern of genetic variation by correlating genetic and geographic distances (Diniz-Filho et al., 2013). In the extension presented by Pluta et al. (2021), a Mantel-type of test with ridge regularization was presented as a compromise between the score tests from fixed-effects and

random-effects model. The overarching goal of this study is to examine a class of methods for choosing the optimal ridge penalty parameter and incorporate these in the Adaptive Mantel test (Pluta et al., 2021) for hypothesis testing problems with high-dimensional data set up.

Our contribution to the initial work by Pluta et al. (2021) is three-fold. First, we propose a thresholding procedure aligned to the philosophical considerations of ridge regression in high-dimensional settings. In this study, we allow the set of candidate values of $\lambda$ to include negative values and investigate how these negative penalty parameters can affect the corresponding Type I error and empirical power of the Adaptive Mantel Test. Second, we extend the AdaMant algorithm to include the selection of the optimal ridge penalty parameter *via* generalized cross-validation. To illustrate the almost sure convergence results in Patil et al. (2021) using imaging genetics data, we also implement the selection of the ridge penalty parameter using leave-one-out cross-validation. The resulting optimal choice between the two cross-validation procedures will be compared. Third, we also investigate the Type I error rate and empirical power of the test using the parametric asymptotic null distribution.

This article is outlined as follows. The general frameworks of Mantel Test and score test in linear models are presented in Sections 2.1 and 2.2. Some existing procedures for selecting the ridge penalty parameter are discussed in Section 2.3. The rationale and contributions of our work are illustrated in Section 2.4. The class of methods for choosing the optimal ridge penalty are presented in Section 3. Finally, the numerical studies involving the proposed methods and the application to an imaging genetics data set are available in Section 4.

# 2. RELATED WORK

## 2.1. Mantel Test

Suppose we have $(\mathbf{X}_i, \mathbf{Y}_i) \in \mathbb{R}^p \times \mathbb{R}^q$ for all subjects $i = 1, 2, \ldots, n$ where $p$ is the number of explanatory variables and $q$ is the number of response variables. In imaging genetics studies, the value of $p$ usually correspond to the total number of genetic variations, such as single nucleotide polymorphisms (SNPs) in genomics or differentially methylated probes in epigenetics. Meanwhile, the response variables correspond to the brain imaging information, such as pairwise alpha-band coherence measures obtained from several EEG channels.

Suppose $\mathbf{X}_i$ and $\mathbf{X}_j$ correspond to the vector of explanatory variables for subjects $i$ and $j$, respectively. As described in Pluta et al. (2021), let $\mathcal{K}^{\mathbf{X}}(\cdot,\cdot)$ and $\mathcal{K}^{\mathbf{Y}}(\cdot,\cdot)$ be positive semi-definite kernel functions on $\mathbf{X} \times \mathbf{X}$ and $\mathbf{Y} \times \mathbf{Y}$, respectively where the data matrices $\mathbf{X}$ and $\mathbf{Y}$ are column-centered. Specifically, we are interested in investigating the kernel function $\mathcal{K}^{\mathbf{X}}(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{X}_i^\top \mathbf{W}_{\lambda_X} \mathbf{X}_j$ where $\mathbf{W}_{\lambda_X} = (\mathbf{X}^\top\mathbf{X} + \lambda_X \mathbf{I}_p)^{-1}$ is the ridge-penalized weight matrix. The corresponding Gram matrix for this kernel is denoted by

$$\mathbf{H}_{\lambda_X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda_X \mathbf{I}_p)^{-1}\mathbf{X}^\top. \tag{4}$$

We define $\mathcal{K}^{\mathbf{Y}}$ and the associated Gram matrix $\mathbf{K}_{\lambda_Y}$ similarly using $\mathbf{Y}$. The Mantel test statistic is equivalent to $\mathrm{tr}(\mathbf{H}_{\lambda_X}\mathbf{K}_{\lambda_Y})$. Under the null hypothesis, there is no association between the similarities measured by $\mathcal{K}^{\mathbf{X}}$ and $\mathcal{K}^{\mathbf{Y}}$. In practice, the reference distribution can be obtained *via* a permutation procedure (Nichols and Holmes, 2002; Shaw and Proschan, 2013; Zhou et al., 2014). For instance, we can simultaneously permute the rows and columns of $\mathbf{Y}$, while keeping $\mathbf{X}$ fixed. Equivalently, for a fixed matrix $\mathbf{H}_{\lambda_X}$, we can permute the observation labels for $\mathbf{K}_{\lambda_Y}$ and calculate the empirical null distribution.

## 2.2. Score Test in Linear Models

The general framework of Mantel Test presented in Section 2.1 encompasses several association tests (for examples, see Robert and Escoufier, 1976; Székely et al., 2007; Xu et al., 2017) and various kernel functions can be investigated to reflect model complexity and detect underlying linear or non-linear associations. Moreover, Pluta et al. (2021) developed a unified framework of linear models that links the Mantel test and Rao's score test (Rao, 1948) in a class of tests indexed by the ridge penalty. Following the discussion of Pluta et al. (2021), we consider the following linear models:

1. Fixed Effects Model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma})$ or alternatively using the vectorized response variables, $\mathrm{vec}(\mathbf{Y}) \sim N(\mathrm{vec}(\mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$ where $\mathrm{vec}(\cdot)$ is the vectorization operator and $\otimes$ refers to the Kronecker product operator on two matrices.
2. Random Effects Model: $\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}_q)$ and $\mathbf{b} \sim N(\mathbf{0}, \mathbf{I}_p, \boldsymbol{\Sigma}_b)$ or equivalently, $\mathrm{vec}(\mathbf{Y}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_b \otimes \mathbf{X}\mathbf{X}^\top + \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$.

To describe the score statistic compactly, we consider the Singular Value Decomposition (SVD) of the matrix $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and $\mathbf{D}$ is a diagonal matrix with the (non-negative) singular values. To perform the global test $H_0 : \boldsymbol{\beta} = \mathbf{0}$ under the fixed effects model, the score test statistic is given by

$$S_{FE} \asymp \mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-(\mathbf{X}^\top\mathbf{Y}) = \mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top) = \sum_{j=1}^{r} Z_j^2 \overset{H_0}{\sim} \sum_{j=1}^{r} \chi^2_{1,j} \tag{5}$$

where $\mathbf{Z} = \mathbf{U}^\top\mathbf{Y}$ and $r = \mathrm{rank}(\mathbf{X})$. The notation $\mathbf{A}^-$ denotes the Moore-Penrose pseudoinverse of the matrix $\mathbf{A}$. It is well-known that the Moore-Penrose psedoinverse leads to the minimum norm solution to the least-squares problem. On the other hand, to test $H_0 : \boldsymbol{\Sigma}_b = \mathbf{0}$ under certain conditions, the score test statistic for the random effects (variance components) model is

$$S_{RE} \asymp \mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{Y}) = \mathrm{tr}(\mathbf{Z}^\top\mathbf{D}\mathbf{D}^\top\mathbf{Z}) = \sum_{j=1}^{r} d_j^2 Z_j^2 \overset{H_0}{\sim} \sum_{j=1}^{r} d_j^2 \chi^2_{1,j}. \tag{6}$$

Finally, the ridge regression score test statistic for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is

$$\begin{aligned} S_{RR} &\asymp \mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda_X\mathbf{I}_p)^{-1}(\mathbf{X}^\top\mathbf{Y}) \\ &= \mathrm{tr}(\mathbf{Z}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda_X\mathbf{I}_p)^{-1}\mathbf{D}^\top\mathbf{Z}). \end{aligned} \tag{7}$$

Hence,

$$\sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \lambda_X} Z_j^2 \overset{H_0}{\sim} \sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \lambda_X} \chi_{1,j}^2. \tag{8}$$

As summarized in Pluta et al. (2021), the score test statistics described in (5) – (7) can be formulated equivalently as $\text{tr}(\mathbf{H}_{\lambda_X}\mathbf{K}_{\lambda_Y})$ which is the expression for the Mantel test statistic described in Section 2.1. In particular, the fixed effects score test statistic is equivalent to $\text{tr}(\mathbf{H}_0\mathbf{Y}\mathbf{Y}^\top)$ where $\mathbf{H}_0 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top$. Meanwhile, when $\mathbf{\Sigma} = \sigma^2\mathbf{I}_q$ and $\mathbf{\Sigma}_b = \sigma_b^2\mathbf{I}_q$, then the score statistic corresponding to the random effects model is proportional to $\text{tr}(\mathbf{H}_\infty\mathbf{K}_\infty)$ where $\mathbf{H}_\infty = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{K}_\infty = \mathbf{Y}\mathbf{Y}^\top$ (Pluta et al., 2021). Lastly, the ridge regression score test statistic can be written as $\text{tr}(\mathbf{H}_{\lambda_X}\mathbf{Y}\mathbf{Y}^\top)$ using $\mathbf{H}_{\lambda_X}$ provided in (4). Furthermore, Pluta et al. (2021) highlights that the ridge regression score statistic is a compromise between the fixed effects and variance components tests. For small values of the ridge penalty $\lambda_X$, the test statistic in (7) approaches the fixed effects score test statistic, and is identical at $\lambda_X = 0$. Also, Pluta et al. (2021) remarked that a large choice of $\lambda_X$ yields a test close to the random effects score statistic, converging to identical tests as $\lambda_X \to \infty$.

## 2.3. Examining the Choice of Ridge Penalty Parameter

Motivated by the framework introduced by Pluta et al. (2021), which categorizes the association test and score tests into a single class of tests characterized by the ridge penalty, we examine the choice of this parameter in the high-dimensional hypothesis testing set-up. In practice, the optimal choice of ridge penalty parameter is based on the observed data and proper data-dependent tuning is among the central tasks in statistical learning (Patil et al., 2021).

### 2.3.1. Ridge Predictive Performance

The role of the ridge penalty in high-dimensional prediction and estimation has been an active area of research in the past several years. For both asymptotic and non-asymptotic settings, the predictive performance of ridge regression has been studied extensively (see Hsu et al., 2012; Cule and De Iorio, 2013; Karoui, 2013; Dobriban and Wager, 2018; Hastie et al., 2019; Wu and Xu, 2020; Richards et al., 2021 for examples). Furthermore, Kobak et al. (2020) demonstrated that an explicit positive ridge penalty can fail to provide any improvement over the minimum-norm least squares estimator using simulations and real-life high-dimensional data sets. In particular, they showed that the optimal value of ridge penalty in this situation could be negative when $n \ll p$. Similar to these work, in this article, we focus on the role of ridge penalty in hypothesis testing for a univariate response, i.e., $q = 1$. The extension of to multivariate responses will be considered in future research. In Sections 2.1 and 2.2, $\lambda_X$ corresponds to the tuning parameter in the Gram matrix of the ridge kernel associated with $\mathbf{X}$ which is not necessarily the same as $\lambda_Y$, the tuning parameter in the ridge kernel corresponding to $\mathbf{Y}$. However, under the univariate response $\mathbf{y}$ setting, we only have to specify the ridge penalty parameter $\lambda_X$. For brevity, we will refer to $\lambda_X$ as $\lambda$ in the next sections.

### 2.3.2. Ridge Cross-Validation

The performance of the fitted model is affected by the calibration of the regularization parameter. One of the most widely used methods for regularization tuning is cross-validation (for examples, see Allen, 1971; Stone, 1974; Delaney and Chatterjee, 1986; Arlot and Celisse, 2010). In ridge regression, two commonly used cross-validation procedures are generalized cross-validation (GCV) (Golub et al., 1979) and leave-one-out cross-validation (LOOCV), a variant of the $k$-fold cross-validation (Hastie et al., 2009). GCV, a rotation-invariant version of the predicted residual error sum of squares (PRESS), is a popular choice in practice because it does not require model refitting. Similarly, approximation methods to LOOCV (e.g., Kumar et al., 2013; Meijer and Goeman, 2013) to circumvent the problem of computational complexity brought by multiple model refitting.

The LOOCV estimate for a response vector $\mathbf{y}$ containing $n$ observations is defined as

$$\text{loocv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_{-i,\lambda}\right)^2 \tag{9}$$

where $\widehat{\boldsymbol{\beta}}_{-i,\lambda}$ is the ridge estimate when the $i$th observation is not included in the training set. As cited in Patil et al. (2021), an alternative formula for the LOOCV (Hastie et al., 2009) is given by

$$\text{loocv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_\lambda}{1 - [\mathbf{H}_\lambda]_{ii}}\right)^2 \tag{10}$$

where $[\mathbf{H}_\lambda]_{ii}$ corresponds to the $i$th diagonal entry of the matrix $\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-}\mathbf{X}^\top$. Closely similar to (10), the GCV estimate formulation provided by Patil et al. (2021) is

$$\text{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_\lambda}{1 - \text{tr}(\mathbf{H}_\lambda)/n}\right)^2 \tag{11}$$

where the average of the trace elements is used instead of the $i$th diagonal entry. When $\lambda = 0$ and $\text{rank}(\mathbf{X}) = n$, the diagonal elements of $\mathbf{H}_0$ is equal to 1 and $\text{tr}(\mathbf{H}_0)$ reduces to $n$. In this case, the ridge regression is an interpolator of $\mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{y}$ (Patil et al., 2021). Since both numerator and denominator of the expressions in (10) and (11) are 0, Hastie et al. (2019) defined the LOOCV and GCV estimates based on the limits $\lambda \to 0$, respectively.

Moreover, the asymptotic optimality of LOOCV and GCV tuning for ridge regression in high-dimensional setting is presented by Hastie et al. (2019). Patil et al. (2021) generalized the scope discussed in Hastie et al. (2019) by showing that the GCV converges almost surely to the expected out-of-sample prediction error, uniformly over a set of candidate ridge regularization parameters. The discussion provided by Patil et al. (2021) is aligned with Kobak et al. (2020) wherein the optimal ridge penalty parameter can be positive, negative, or zero.

## 2.4. Rationale and Illustration of Contributions of Our Work

The Adaptive Mantel test (AdaMant) coined by Pluta et al. (2021) is an extension of the classical Mantel Test by incorporating the ridge penalty parameter to association testing. The adaptive procedure involves the calculation of similarity matrices $\mathbf{H}_m = \mathcal{K}_m^{\mathbf{X}}(\mathbf{X})$ and $\mathbf{K}_m = \mathcal{K}_m^{\mathbf{Y}}(\mathbf{Y})$ for every pair of input metrics or kernels $(\mathcal{K}^{\mathbf{X}}, \mathcal{K}^{\mathbf{Y}}), m = 1, 2, \ldots, M$. Under the null hypothesis of no association between the similarities measured by $\mathcal{K}^{\mathbf{X}}$ and $\mathcal{K}^{\mathbf{Y}}$, Pluta et al. (2021) proposed a permutation procedure where they generate $B$ permutations of the observation labels for $\mathbf{H}_m$ for a fixed matrix $\mathbf{K}_m$. The $p$-value $P_m^{(b)}$ is computed as a function of the test statistic $\mathrm{tr}(\mathbf{H}_m^{(b)}\mathbf{K}_m)$ for each $m = 1, 2, \ldots, M$ and permutations $b = 1, 2, \ldots, B$. Finally, Pluta et al. (2021) defined the AdaMant test statistic as $P^{(0)} = \min_{m\in\{1,2,\ldots,M\}} P_m^{(0)}$ where $b = 0$ refers to the original data set. Using the permutation procedure to obtain the empirical null distribution of $P^{(0)}$, the corresponding AdaMant $p$-value is the proportion of $P^{(b)}$ less than or equal to $P^{(0)}$, that is,

$$P_{\text{AdaMant}} = \frac{1}{B+1}\sum_{b=0}^{B} \mathrm{I}(P^{(b)} \leq P^{(0)}) \tag{12}$$

where $P^{(b)} = \min_{m\in\{1,2,\ldots,M\}} P_m^{(b)}$.

However, the main limitation in the ridge-penalized AdaMant procedure by Pluta et al. (2021) is the optimal selection of the ridge penalty parameter. When kernels of the form $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top$ are considered in AdaMant, $\lambda$ is chosen to be proportional to the signal-to-noise ratio re-expressed as a function of genetic heritability $h^2$ and number of explanatory variables $p$ (Pluta et al., 2021). This implies that the value of the chosen ridge penalty is restricted to be non-negative. In their examples, the ridge penalty is chosen from a set with only a few values such as $\lambda \in \{100, 1,000, 2,500, 5,000, 7,500, 25,000, \infty\}$. With a limited number of ridge penalty parameters to choose from, the $\lambda$ which yields the highest empirical power may not be captured by the initial study of Pluta et al. (2021).

As highlighted in Section 1, the primary objective of this article is to examine the optimal choice of ridge penalty in the high-dimensional hypothesis testing scenario. To illustrate the utility of addressing this goal and our subsequent contributions, we study liver.toxicity data set in Bushel et al. (2007). This data contains microarray expression levels of $p = 3,116$ genes and 10 clinical chemistry measurements in liver tissue of $n = 64$ rats. First, we replicate the results presented in Kobak et al. (2020) using 10-fold cross-validation for varying ridge penalty parameter $\lambda$ using one dependent variable at a time. The cross-validated MSE plotted for each dependent variable is displayed in **Figure 1**. In **Figure 1A** where $n > p$, Kobak et al. (2020) showed that this result is in agreement with the seminal article by Hoerl and Kennard (1970) wherein the optimal penalty is always larger than zero under the low-dimensional setting. However, in **Figure 1B**, five out of ten dependent variables yielded a minimum cross-validated MSE corresponding to the smallest value of $\lambda$ considered when $n \ll p$ (Kobak et al., 2020).

Motivated by the aforementioned results, we investigated the empirical power and average of the -$\log_{10}$ $p$-values of the Adaptive Mantel test for several values of $\lambda$. We employ the liver toxicity data as our motivating example because it has been widely used recently to better understand overfitting. It was found that the clinical variables may not facilitate in the detection of paracetamol toxicity in the liver, but gene expression could be an alternative solution (Heinloth et al., 2004; Bushel et al., 2007). In this illustration, we compute the empirical power for a fixed $\lambda$, using one dependent variable at a time. For each replication, we add a vector of random noise to the vector of response, that is, $\mathbf{y}_s = \mathbf{y} + \boldsymbol{\varepsilon}_s$ for $s = 1, 2, \ldots, S$. Under the null hypothesis when $\boldsymbol{\beta} = 0$, the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ reduces to $\mathbf{y} = \boldsymbol{\varepsilon}$. Hence, we can view the recursive expression as $\mathbf{y}_s = (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \boldsymbol{\varepsilon}_s \neq \boldsymbol{\varepsilon}$ in favor of the case that the alternative hypothesis is true. For $s = 1, 2, \ldots, S$, we compute the AdaMant $p$-value at each $\lambda$ using $\mathbf{y}_s$ and the entire matrix of gene expression $\mathbf{X}$ as inputs to the ridge kernel described in (4). After repeating this procedure for a total of $S$ replications, the empirical power is computed as the proportion of replications where the Adamant $p$-value is less than the nominal level of significance $\alpha$.

$$\text{Power}_\lambda = \frac{1}{S}\sum_{s=1}^{S} \mathrm{I}(P_{\text{AdaMant},\lambda,s} \leq \alpha) \tag{13}$$

The results are presented in **Figure 2**.

To circumvent the limitations of the range of ridge penalty parameter considered in Pluta et al. (2021), we allowed the interval of $\lambda$ to include negative, zero and positive values. According to **Figure 2A**, even though there is a more distinct gradient in the values of the average of the $-\log_{10}$ $p$-values compared to the empirical power in **Figure 2B**, eight out of ten dependent variables displayed more or less similar patterns in terms of empirical power. Also, based on **Figure 2B**, some $\lambda < 0$ lead to an empirical power approaching 1 when $n \ll p$. This result is in alignment with the main result reported by Kobak et al. (2020) where the optimal ridge penalty for real-world high-dimensional data can be negative due to implicit ridge regularization. This phenomenon prompted us to further investigate real-valued ridge penalty parameters using imaging genetics data where the signals are weaker and sparsity is much more evident.

# 3. METHODOLOGY

In this section, we discuss the proposed methodology to incorporate the optimal selection of the ridge penalty parameter *via* cross-validation in the Adaptive Mantel test. The interval from which the optimal ridge penalty parameter will be selected from is discussed in Section 3.1 while the proposed methods are discussed in 3.2. The two algorithms to be compared are discussed thoroughly in Sections 3.2.1 and 3.2.3. Meanwhile, the features of the score test statistic are discussed in Section 3.2.2.

## 3.1. Range of Ridge Penalty Parameter

Before delving into the optimal choice of ridge penalty parameter to be used in hypothesis testing, it is crucial to specify the domain of sensible values first. Formally, we describe how to choose the interval $\mathcal{I} = (\lambda_{\min}, \infty)$ in this section. Following the third assumption in the main results of Patil et al. (2021), the

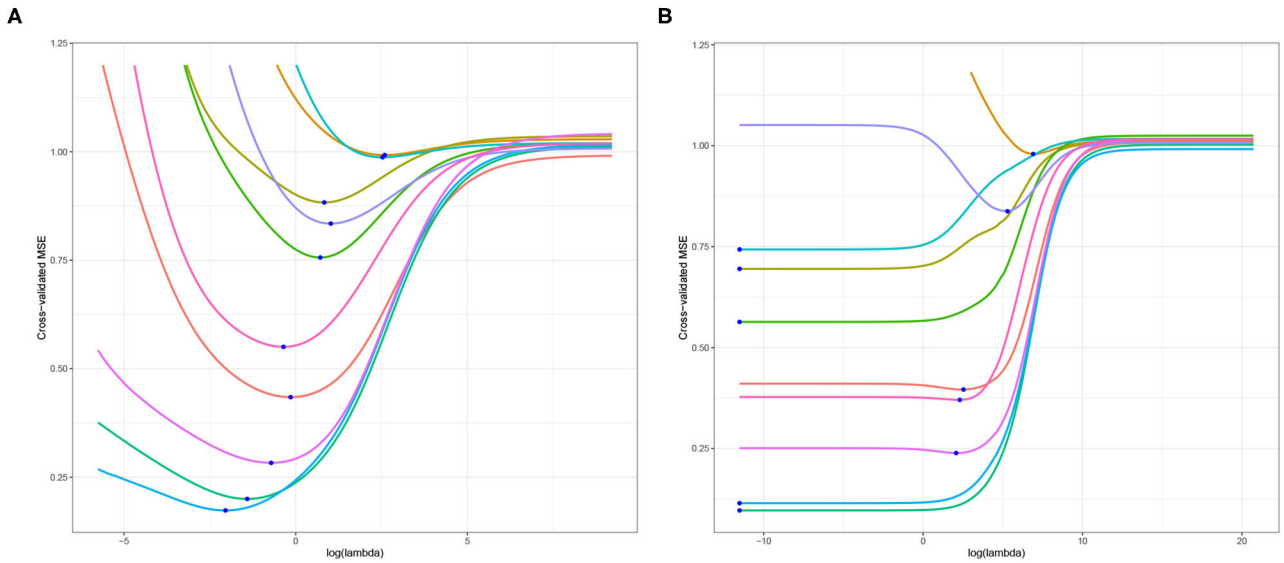


**FIGURE 1 |** Cross-validated MSE of ridge regression using **(A)** $n = 64$ and $p = 50$ randomly selected explanatory variables; **(B)** $n = 64$ and $p = 3,116$, all explanatory variables. The blue dot corresponds to the minimum cross-validated MSE for each dependent variable.

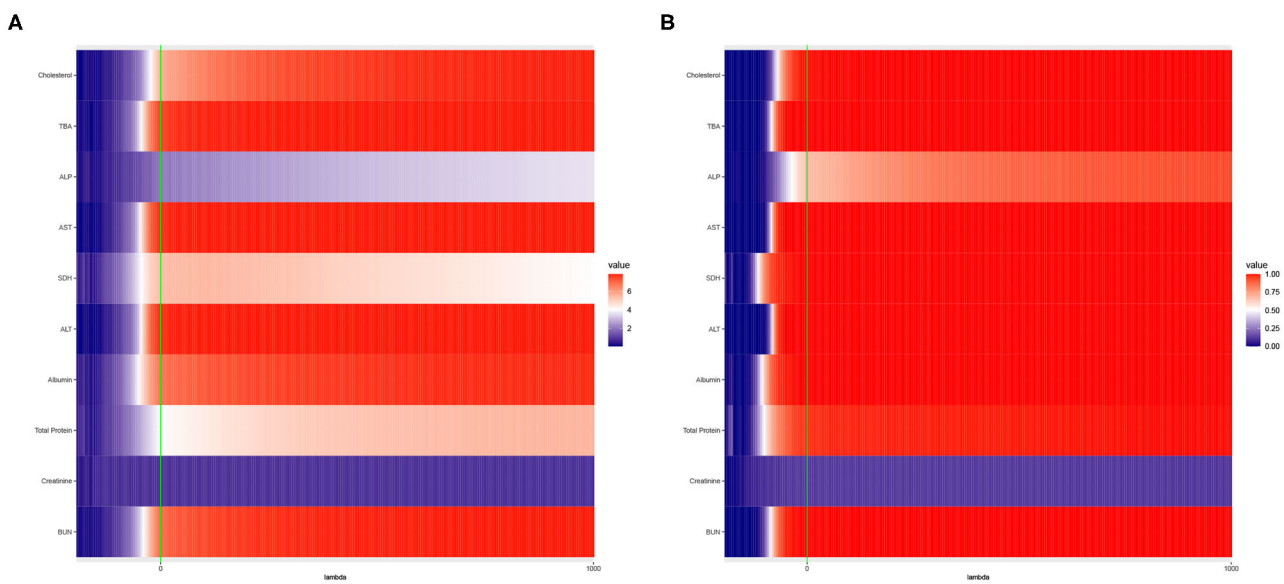


**FIGURE 2 |** Heat maps of **(A)** Average of $-\log_{10}$ $p$-values and **(B)** Empirical Power of the Adaptive Mantel test using the liver toxicity data with $n = 64$ observations, $p = 3,116$ genetic features. The green vertical line corresponds to $\lambda = 0$.

minimum eigenvalue is bounded below by a constant $\ell_{\min} > 0$ where $\ell_{\min}$ is independent of $p$. Patil et al. (2021) proposed that the smallest possible value of the regularization parameter wherein the GCV converges almost surely to LOOCV is given by $\lambda_{\min} = -(\sqrt{p/n} - 1)^2 \ell_{\min}$. However, Patil et al. (2021) did not clearly specify the value of $\ell_{\min}$ apart from the constraint that it is positive.

In our proposed method, we extend the work of Patil et al. (2021) by relaxing the assumption on the minimum eigenvalue and allowing a fraction of eigenvalues to accumulate near zero. Specifically, we define a threshold on the small but non-zero singular values of $X$. Let $d_1 \geq d_2 \ldots \geq d_r$ denote the non-negative singular values. To identify $\lambda_{\min}$, we first compute the adjusted singular values as follows

$$\widetilde{d}_j = \begin{cases} d_j & d_j \geq \tau \\ 0 & d_j < \tau \end{cases}$$

where $\tau$ is some threshold based on the quantiles. A generic choice for $\tau$ is the median of the singular values (Bühlmann and Ćevid, 2020). Next, we define $\ell_{\min} = \min\{\widetilde{d}_j : \widetilde{d}_j > 0\}$. The value of $\ell_{\min}$ coincides with $\tau$ if the value of the quantile is equal to one of the singular values. Otherwise, $\ell_{\min}$ is the singular value greater than and closest to the quantile. We then compute $\lambda_{\min} = -(\sqrt{p/n}-1)^2\ell_{\min}$ where the range of $\lambda$ allows for negative values including zero, when $p \neq n$. We will investigate several values of $\tau$ and how it affects both Type I error rate and empirical power in the numerical studies section.

## 3.2. Proposed Methods

We extend the Adaptive Mantel Test (AdaMant) by Pluta et al. (2021) to include the optimal selection of the ridge penalty parameter *via* cross-validation in this section. For simplicity, we will refer to this proposed procedure as AdaMantCV. By using a permutation procedure on the set of test statistics, this procedure can simultaneously test across a set of ridge penalty parameters without increasing the Type I error rate (Pluta et al., 2021). Prior to the analysis, centering and scaling the explanatory and response variables is necessary because they lead to potential computational efficiency and stability and conceptual simplicity. More importantly, performing this ensures that penalty term will have an similar effect on all coefficient estimates.

### 3.2.1. Adaptive Mantel Test With Cross-Validation

Under the null hypothesis of no association between the variation in a set of candidate SNPs and the variation in brain imaging data, say EEG coherence, we will implement a permutation procedure where $B$ permutations are generated from rows of $\mathbf{y}$ for a fixed matrix $\mathbf{X}$. Since the matrix $\mathbf{X}$ does not vary across the permutations generated, we only need to perform SVD once to obtain the vector of singular values for the calculation of the weights and specification of the interval of ridge penalty parameters.

We start by specifying the number of possible values of the ridge penalty parameter $\lambda \in \mathcal{I}$, denoted by $M$ and the threshold $\tau$. Using these inputs coupled with the singular values, we implement the proposed method in Section 3.1 to identify the interval $\mathcal{I} = (\lambda_{\min}, \infty)$. Unlike Pluta et al. (2021), the proposed algorithm only require the kernel $\mathcal{K}_\lambda^{\mathbf{X}}$ as input because we only consider the univariate response $\mathbf{y}$ in this study. For $\mathcal{K}_\lambda^{\mathbf{X}}$, we include a single family of kernels with varying ridge penalty parameters. For a given cross-validation measure, say GCV, we compute the optimal value of the ridge penalty parameter using $\mathbf{y}^{(b)}$ and $\mathbf{X}$ as

$$\widehat{\lambda}^{(b)} = \underset{\lambda\in\mathcal{I}}{\operatorname{argmin}}\,\mathrm{gcv}(\lambda), b = 0, 1, 2, \ldots, B \tag{14}$$

where $b = 0$ is associated to the original data set. For each $b = 0, 1, 2, \ldots, B$, we also compute the ridge regression score test statistic $T^{(b)}(\widehat{\lambda}^{(b)})$ given by $\mathrm{tr}(\mathbf{H}_{\widehat{\lambda}^{(b)}}\mathbf{K}^{(b)})$ where $\mathbf{H}_{\widehat{\lambda}^{(b)}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \widehat{\lambda}^{(b)}\mathbf{I}_p)^{-1}\mathbf{X}^\top$.

To account for varying magnitudes of the quantity $T^{(b)}(\widehat{\lambda}^{(b)})$ for different values of $\widehat{\lambda}^{(b)}$, our proposed AdaMantCV test statistic $P^{(0)}$ is transformed to take on values between 0 and 1, inclusive. This test statistic is computed as follows

$$P^{(0)} = \frac{1}{B+1}\sum_{b=0}^{B} \mathtt{I}\left(T^{(0)}(\widehat{\lambda}^{(0)}) \leq T^{(b)}(\widehat{\lambda}^{(0)})\right). \tag{15}$$

This indicates that for a fixed value of $\widehat{\lambda}^{(0)}$, we want to compare the magnitude of the observed test statistic from the original data vs. the computed test statistics using the permuted data. Finally, the analogous AdaMantCV $p$-value is the proportion of $P^{(b)}$ no greater than $P^{(0)}$, that is,

$$\begin{aligned} P_{\mathrm{CV}} &= \frac{1}{B+1}\sum_{b=0}^{B} \mathtt{I}\left(P^{(b)} \leq P^{(0)}\right), \text{where} \\ P^{(b)} &= \frac{1}{B+1}\sum_{c=0}^{B} \mathtt{I}\left(T^{(b)}(\widehat{\lambda}^{(b)}) \leq T^{(c)}(\widehat{\lambda}^{(b)})\right). \end{aligned} \tag{16}$$

The general pseudocode of the AdaMantCV algorithm is presented below. The limitation of the straightforward application of this procedure is that it is computationally expensive when $n \ll p$ (Pluta et al., 2021). To deal with this, following Pluta et al. (2021), we utilized the identity presented in Henderson and Searle (1981) and Kobak et al. (2020).

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1} \tag{17}$$

wherein the dimension of the matrix to be inverted is $n \times n$ instead of $p \times p$. Additionally, Pluta et al. (2021) has shown that the Mantel test statistic has a complexity of $O(n^2)$ and coupled with $B$ permutations, the total computational complexity is $O(n^2p + n^2B)$, which is less than the required computational complexity using SVD.

---

**procedure** ADAPTIVE MANTEL TEST WITH CV:$(\mathbf{X}, \mathbf{y}, \mathcal{K}_\lambda^{\mathbf{X}}, \mathrm{CV}(\lambda), \tau, M, B)$
  Specify the interval $\mathcal{I}$ as a function of $(\mathbf{X}, \tau, M)$
  Calculate $\widehat{\lambda}^{(0)} := \underset{\lambda\in\mathcal{I}}{\operatorname{argmin}}\,\mathrm{CV}(\lambda)$
  Calculate $T^{(0)}(\widehat{\lambda}^{(0)}) \leftarrow \mathrm{tr}[\mathbf{H}_{\widehat{\lambda}^{(0)}}\mathbf{K}]$ where $\mathbf{H}_{\widehat{\lambda}^{(0)}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \widehat{\lambda}^{(0)}\mathbf{I}_p)^{-1}\mathbf{X}^\top$ and denote $\mathbf{K} = \mathbf{y}\mathbf{y}^\top$
  Generate $B$ permutations of $\mathbf{y}$, labeled $\mathbf{y}^{(b)}$, and denote $\mathbf{K}^{(b)} = \mathbf{y}^{(b)}(\mathbf{y}^{(b)})^\top \forall b = 1, \ldots, B$
  Calculate $\widehat{\lambda}^{(b)} := \underset{\lambda\in\mathcal{I}}{\operatorname{argmin}}\,\mathrm{CV}(\lambda)$
  $T^{(b)}(\widehat{\lambda}^{(b)}) \leftarrow \mathrm{tr}[\mathbf{H}_{\widehat{\lambda}^{(b)}}\mathbf{K}^{(b)}]\ \forall b = 1, \ldots, B$ where $\mathbf{H}_{\widehat{\lambda}^{(b)}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \widehat{\lambda}^{(b)}\mathbf{I}_p)^{-1}\mathbf{X}^\top$
  $T^{(c)}(\widehat{\lambda}^{(b)}) \leftarrow \mathrm{tr}[\mathbf{H}_{\widehat{\lambda}^{(b)}}\mathbf{K}^{(c)}]\ \forall c = 0, 1, \ldots, B$
  $P^{(b)} \leftarrow \frac{1}{B+1}\sum_{c=0}^{B} \mathtt{I}\left(T^{(b)}(\widehat{\lambda}^{(b)}) \leq T^{(c)}(\widehat{\lambda}^{(b)})\right) \forall\ b, c = 0, 1, \ldots, B$
  $P_{\mathrm{CV}} \leftarrow \frac{1}{B+1}\sum_{b=0}^{B} \mathtt{I}\left(P^{(b)} \leq P^{(0)}\right)$
**end procedure**

---

### 3.2.2. Features of the Score Test Statistic in AdaMantCV

The ridge regression score test statistic $T^{(b)}(\widehat{\lambda}^{(b)}) = \text{tr}(\mathbf{H}_{\widehat{\lambda}^{(b)}}\mathbf{K}^{(b)})$ can be expressed equivalently as

$$T^{(b)}(\widehat{\lambda}^{(b)}) = \sum_{j=1}^{r} \frac{d_j^2}{d_j^2 + \widehat{\lambda}^{(b)}}[Z_j^{(b)}]^2 = \sum_{j=1}^{r} w_j(\widehat{\lambda}^{(b)})[Z_j^{(b)}]^2 \quad (18)$$

where $Z_j^{(b)} = \mathbf{U}_j^\top \mathbf{y}^{(b)}$, as described in (8). From the specification of the test statistic in (18), the weights are non-negative if the set of possible values of $\lambda$ is restricted to the positive range only. More importantly, $0 < w_j(\widehat{\lambda}^{(b)}) < 1$ for non-zero singular values $d_j$ which indicates that the test statistic components shrink toward zero if we choose a large positive value of the ridge penalty parameter. This means that for a fixed $[Z_j^{(b)}]^2 > 0$, the weight $w_j(\widehat{\lambda}^{(b)})$ corresponding to the $j$th component of the score test statistic $T^{(b)}(\widehat{\lambda}^{(b)})$ has the following features:

(i) $w_j(\widehat{\lambda}^{(b)}) \to 0$ if $\widehat{\lambda}^{(b)} \to \infty$. However, as argued in Pluta et al. (2021), the standardized version essentially indicates that $w_j(\infty) \propto d_j^2$;

(ii) $w_j(\widehat{\lambda}^{(b)}) = 1$, i.e., equal weights, if $\widehat{\lambda}^{(b)} = 0$; and

(iii) $w_j(\widehat{\lambda}^{(b)}) > 1$ and, smaller but a positive $d_j^2$ weight leads to a larger weight in a relative sense if $\widehat{\lambda}^{(b)} < 0$ provided $d_j^2 + \widehat{\lambda}^{(b)} > 0$.

This suggests that in the context of Adaptive Mantel Test in high-dimensional setting, a negative choice of penalty parameter has potential in achieving superior empirical power when $Z_j^2$ tends to associate with directions of low-variance where variance is measured by the eigenvalues of variance-covariance matrix of the covariates.

Another crucial consideration for the score test statistic in (18) is that the weights should be non-negative for all $j = 1, 2, \ldots r = \min(n, p)$ because the statistic is asymptotically distributed as a mixture of chi-squared random variables. If $\widehat{\lambda}^{(b)} \geq 0$, then there is no constraint about the form of the test statistic. However, if $\widehat{\lambda}^{(b)} < 0$, then we should ensure that the all the weights remain positive to satisfy the asymptotic distributional assumptions. A simple and straightforward strategy to handle negative weights is to use the adjusted weights defined as $\widetilde{w}_j = \max(w_j, 0)$. In addition, if $\widehat{\lambda}^{(b)} = 0$ and $d_j = 0$, we will utilize the formula for the fixed effects score test statistic mentioned in (5) to avoid the case where both the numerator and denominator of (18) is zero.

We can also show mathematically that for any $j = 1, 2, \ldots, r$, the optimal value of the ridge penalty parameter must satisfy the following condition

$$\widehat{\lambda}^{(b)} > \max_{1 \leq j \leq r} -d_j^2$$

to ensure that the weights are non-negative. This shows that a potential choice for $\lambda^\star_{\min} = -d_{(1)}^2 + \epsilon$, where $d_{(1)}^2$ is the smallest eigenvalue and $\epsilon > 0$. The value of $\epsilon$ is chosen to be the machine tolerance error in a statistical software. Formally, $\epsilon$ is the smallest positive floating-point number such that $1 + \epsilon \neq 1$. However, in the high-dimensional setting, the smallest eigenvalue is very close to zero and could even be lower than the machine tolerance $\epsilon$. This may lead to the case that $\lambda^\star_{\min} > 0$, i.e., the interval of ridge penalty parameters include positive values only. To address the objective of investigating the role of real-valued ridge penalty parameters in high-dimensional hypothesis testing, our focus in this study is on intervals $\mathcal{I}$ where the lower bound is negative.

### 3.2.3. Adaptive Mantel Test With Gamma Approximation and Cross-Validation

Alternatively, we can use the $B$ permutations to estimate the parameters of the asymptotic null distribution. As mentioned previously, test statistic described in (8) is asymptotically distributed as a mixture of chi-squared random variables. To characterize this null distribution, we will use the Gamma distribution instead because it captures a general family of distributions where the chi-squared distribution is a special case. For a fixed $\lambda \in \mathcal{I}$, suppose $T^{(1)}(\lambda), T^{(2)}(\lambda), \ldots, T^{(B)}(\lambda)$ is a random sample from Gamma distribution under the null hypothesis. We compute the parameter estimates $\widehat{\alpha}(\lambda)$ and $\widehat{\beta}(\lambda)$ using Method of Moments for each $\lambda \in \mathcal{I}$.

Following the classical definition of $p$-value $P^{(b)}(\widehat{\lambda}^{(b)})$, we compute for the probability of $T^{(b)}(\widehat{\lambda}^{(b)})$ is at least as large as the observed value $t^{(b)}(\widehat{\lambda}^{(b)})$ when the null hypothesis is true, that is,

$$P^{(b)}(\widehat{\lambda}^{(b)}) = \mathbb{P}_{H_0}[T^{(b)}(\widehat{\lambda}^{(b)}) \geq t^{(b)}(\widehat{\lambda}^{(b)})]. \quad (19)$$

The null distribution used in (19) is Gamma, with plug-in estimators for the shape and rate parameters denoted by $\widehat{\alpha}(\widehat{\lambda}^{(b)})$ and $\widehat{\beta}(\widehat{\lambda}^{(b)})$, respectively. The test statistic for this procedure is given by $P^{(0)}$ which is the $p$-value in (19) associated with the optimal ridge penalty parameter $\widehat{\lambda}^{(0)}$ for the original data. Similar to AdaMantCV, the $p$-value of this test can be computed using the proportion of $P^{(b)} = P^{(b)}(\widehat{\lambda}^{(b)})$ less than or equal to $P^{(0)}$ for all $b = 0, 1, \ldots, B$. The general pseudocode of the Adaptive Mantel Test with Gamma Approximation and Cross-Validation (AdaMantGACV) algorithm is presented below.

**procedure** ADAPTIVE MANTEL TEST WITH GAMMA APPROX AND CV:$(\mathbf{X}, \mathbf{y}, \mathcal{K}_\lambda^{\mathbf{X}}, \text{CV}(\lambda), \tau, M, B)$
  Specify the interval $\mathcal{I}$ as a function of $(\mathbf{X}, \tau, M)$
  Generate $B$ permutations of $\mathbf{y}$, labeled $\mathbf{y}^{(b)}$, $\forall b = 0, 1, \ldots, B$. Denote $\mathbf{K}^{(b)} = y^{(b)}(y^b)^\top$.
  Calculate $\widehat{\lambda}^{(b)} := \underset{\lambda \in \mathcal{I}}{\text{argmin}}\text{CV}(\lambda) \forall b = 0, 1, \ldots, B$
  $T^{(b)}(\widehat{\lambda}^{(b)}) \leftarrow \text{tr}[\mathbf{H}_{\widehat{\lambda}^{(b)}}\mathbf{K}^{(b)}]$ $\forall b = 0, 1, \ldots, B$
  $P^{(b)}(\widehat{\lambda}^{(b)}) \leftarrow \mathbb{P}_{H_0}[T^{(b)}(\widehat{\lambda}^{(b)}) \geq t^{(b)}(\widehat{\lambda}^{(b)})]$ $\forall b = 0, 1, \ldots, B$
  $P^{(b)} \leftarrow P^{(b)}(\widehat{\lambda}^{(b)}) \forall b = 0, 1, \ldots, B$
  $P_{\text{GACV}} \leftarrow \dfrac{1}{B+1} \displaystyle\sum_{b=0}^{B} \mathtt{I}\left(P^{(b)} \leq P^{(0)}\right)$
**end procedure**

**TABLE 1 |** Numerical comparison of the Type I error of Adaptive Mantel test (i) with cross-validation vs. (ii) with Gamma approximation and cross-validation using the simulated data with $n$ = 350, error standard deviation $\sigma = 0.50$ and $\boldsymbol{\beta} = \mathbf{0}$.

| | | | Adaptive Mantel Test | | | |
|---|---|---|---|---|---|---|
| | | | With CV | | With GA and CV | |
| **Covariance structure** | $p$ | $\tau$ | **GCV** | **LOOCV** | **GCV** | **LOOCV** |
| Compound symmetric | 500 | None | 0.042 | 0.042 | 0.036 | 0.036 |
| | | $Q_1$ | 0.030 | 0.028 | 0.032 | 0.028 |
| | | $Q_2$ | 0.048 | 0.050 | 0.052 | 0.052 |
| | 1,000 | None | 0.048 | 0.048 | 0.038 | 0.038 |
| | | $Q_1$ | 0.044 | 0.044 | 0.042 | 0.042 |
| | | $Q_2$ | 0.036 | 0.036 | 0.036 | 0.036 |
| Heteroskedastic | 500 | None | 0.038 | 0.038 | 0.038 | 0.038 |
| | | $Q_1$ | 0.026 | 0.026 | 0.028 | 0.028 |
| | | $Q_2$ | 0.046 | 0.046 | 0.050 | 0.050 |
| | 1,000 | None | 0.050 | 0.050 | 0.040 | 0.040 |
| | | $Q_1$ | 0.046 | 0.046 | 0.046 | 0.046 |
| | | $Q_2$ | 0.040 | 0.040 | 0.038 | 0.038 |

# 4. RESULTS

## 4.1. Simulation Studies

To gain insights regarding the performance of the proposed procedures in terms of the correct and incorrect rejections, we perform some simulation studies. The comparison is based on four simulation settings:

(i) Number of explanatory variables $p$
(ii) Covariance structure of the simulated design matrix $\mathbf{X}$
(iii) True linear model specification where $\mathbf{y}$ is generated from, and
(iv) Quantile of the singular values to be used as threshold $\tau$.

To mimic the characteristics of the real data set, we consider $n = 350$ subjects and number of explanatory variables $p$ as either 500 or 1000. The design matrix $\mathbf{X}$ is generated from multivariate normal distribution $N(\mathbf{0}_p, \boldsymbol{\Sigma}_X)$ where the covariance structure is either heteroskedastic or compound symmetric. For the heteroskedastic covariance structure is $\boldsymbol{\Sigma}_X = \mathbf{G}_p$ where the $j$th diagonal entry is $g_j = \log(j+1), j = 1, 2, \ldots, p$. Likewise, the compound symmetric structure $\boldsymbol{\Sigma}_X$ is characterized by $\rho_X = 0.025$.

As discussed in Section 2.2, we are also interested in comparing the empirical power of the proposed methods when the vector of responses are generated using either the fixed effects or random effects model assumption. For the fixed effects model under the alternative hypothesis, the coefficients are $\boldsymbol{\beta} = \xi \mathbf{1}_p$ while $\mathbf{b} \sim N(0, \sigma_b^2)$ for the random effects model where $\xi = 3$ and $\sigma_b = 0.50$. Finally, we also want to explore whether the threshold $\tau$ have an impact on both the empirical power and the Type I error. In particular, we compare the scenarios wherein $\tau = 0$, that is, no thresholding is implemented vs. the setting wherein we use the value of the first quartile ($Q_1$) as well as the median or second quartile ($Q_2$) of the singular values as the threshold.

There are four methods to be compared. For the Adaptive Mantel Test with Cross-Validation, we will implement it using GCV or LOOCV to select the optimal ridge penalty parameter. Similarly, we will implement the Adaptive Mantel Test with Gamma Approximation and Cross-Validation using these two cross-validation techniques, denoted as GAGCV and GALOOCV, respectively. For each simulation setting, 500 replications were run to estimate both the Type I error and the empirical power. A total of $B$ = 1000 permutations and $M$ = 250 values of $\lambda$ were considered for each replication. Relative to the simulation studies in Pluta et al. (2021), we consider a more exhaustive range and selection of the ridge penalty parameter.

When the null hypothesis is true, the Type I error rate is computed empirically as

$$\text{Type I error} = \frac{1}{S}\sum_{s=1}^{S} \mathrm{I}(P_{\mathrm{CV},s} \leq \alpha) \tag{20}$$

where $P_{\mathrm{CV},s}$ represents the $p$-value of the Adaptive Mantel test with either GCV, LOOCV, GAGCV, or GALOOCV for the $s$th replication. On the other hand, when the null hypothesis is not true, the empirical power is computed as the proportion of correct rejections. Throughout the simulations, we consider the level of significance $\alpha$ = 0.05.

Results from **Table 1** show the numerical comparison of the Type I error of the Adaptive Mantel test vs. the AdaMant with Gamma Approximation test implementing the optimal ridge penalty selection *via* generalized or leave-one-out cross-validation. Given that both algorithms are permutation-based,

**TABLE 2 |** Numerical comparison of the Empirical Power of Adaptive Mantel test (i) with cross-validation vs. (ii) with Gamma approximation and cross-validation using the simulated data with $n$ = 350, error standard deviation $\sigma = 1$ and $\boldsymbol{\beta} = \xi \mathbf{1}_p, \xi = 3$.

| | | | Adaptive Mantel Test | | | |
|---|---|---|---|---|---|---|
| | | | With CV | | With GA and CV | |
| **Covariance structure** | $p$ | $\tau$ | **GCV** | **LOOCV** | **GCV** | **LOOCV** |
| Compound symmetric | 500 | None | 0.062 | 0.062 | 0.104 | 0.104 |
| | | $Q_1$ | 0.342 | 0.290 | 0.378 | 0.326 |
| | | $Q_2$ | 0.348 | 0.288 | 0.382 | 0.324 |
| | 1,000 | None | 0.998 | 0.998 | 0.998 | 0.998 |
| | | $Q_1$ | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $Q_2$ | 1.000 | 1.000 | 1.000 | 1.000 |
| Heteroskedastic | 500 | None | 0.064 | 0.064 | 0.178 | 0.178 |
| | | $Q_1$ | 0.074 | 0.072 | 0.186 | 0.184 |
| | | $Q_2$ | 0.118 | 0.068 | 0.218 | 0.180 |
| | 1,000 | None | 0.996 | 0.996 | 0.998 | 0.998 |
| | | $Q_1$ | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $Q_2$ | 1.000 | 1.000 | 1.000 | 1.000 |

**TABLE 3 |** Numerical comparison of the Empirical Power of Adaptive Mantel test (i) with cross-validation vs. (ii) with Gamma approximation and cross-validation using the simulated data with $n$ = 350, error standard deviation $\sigma = 1$ and $\mathbf{b} \sim N(\mathbf{0}, \sigma_b^2 \mathbf{I}_p), \sigma_b = 0.50$.

| | | | Adaptive Mantel Test | | | |
|---|---|---|---|---|---|---|
| | | | With CV | | With GA and CV | |
| **Covariance structure** | $p$ | $\tau$ | **GCV** | **LOOCV** | **GCV** | **LOOCV** |
| Compound symmetric | 500 | None | 0.070 | 0.070 | 0.112 | 0.112 |
| | | $Q_1$ | 0.196 | 0.152 | 0.230 | 0.192 |
| | | $Q_2$ | 0.150 | 0.136 | 0.188 | 0.176 |
| | 1,000 | None | 0.998 | 0.998 | 0.998 | 0.998 |
| | | $Q_1$ | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $Q_2$ | 1.000 | 1.000 | 1.000 | 1.000 |
| Heteroskedastic | 500 | None | 0.092 | 0.092 | 0.268 | 0.268 |
| | | $Q_1$ | 0.084 | 0.092 | 0.266 | 0.268 |
| | | $Q_2$ | 0.088 | 0.092 | 0.268 | 0.268 |
| | 1,000 | None | 0.996 | 0.996 | 0.996 | 0.996 |
| | | $Q_1$ | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $Q_2$ | 1.000 | 1.000 | 1.000 | 1.000 |

they naturally control the Type I error rate at any specified nominal level of significance $\alpha$. Even though the $p$-values $P_{\text{GCV}}$ and $P_{\text{LOOCV}}$ obtained using AdaMantCV vary, the resulting Type I error rates using either GCV or LOOCV are more or less similar, and the proportion of incorrect rejections is controlled. A similar pattern is observed for the $p$-values computed from the AdaMant with Gamma Approximation algorithm for both cross-validation methods. Overall, the results presented in **Table 1** verify that the proportion of incorrect rejections was controlled appropriately using the proposed methods.

**Tables 2**, **3** provide a comparison of empirical power when the response data are generated from the fixed effects model and random effects model, respectively. Results revealed that there is an improvement in the empirical power when Gamma approximation is added to the Adaptive Mantel test with cross-validation. Given that both class of methods, AdaMantCV and AdaMantGACV, control the proportion of false rejections, the higher empirical power exhibited by AdaMantGACV indicate that it is the better method. It is also apparent from both tables that either AdaMantCV or AdaMantGACV leads to a superior empirical power, i.e., approaching 1, when $p$ = 1,000. In contrast, both AdaMantCV and AdaMantGACV result to a lower empirical power when $p$ = 500, regardless of the covariance structure used for generating the design matrix $\mathbf{X}$. This result is supported by Hastie et al. (2019) and Patil et al. (2021) where statistical

inference is the most challenging when $p/n \approx 1$, compared to $p \ll n$ or $n \ll p$.

Using the AdaMantCV algorithm and $p = 1{,}000$, **Table 1** shows a decreasing Type I error rate when a threshold is imposed on the specification of the interval $\mathcal{I}$ as compared to using the unadjusted singular values. Also, using the median ($Q_2$) as the threshold compared to the first quartile ($Q_1$) leads to a lower proportion of incorrect rejections, regardless on whether GCV or LOOCV is utilized. It is important to note that in these cases, the empirical power are all approaching to 1 but the decreasing proportion of incorrect rejections for the AdaMantCV supports the proposed method in identifying the range of the ridge penalty parameters. In general, we observe an improved empirical power when thresholding is imposed.

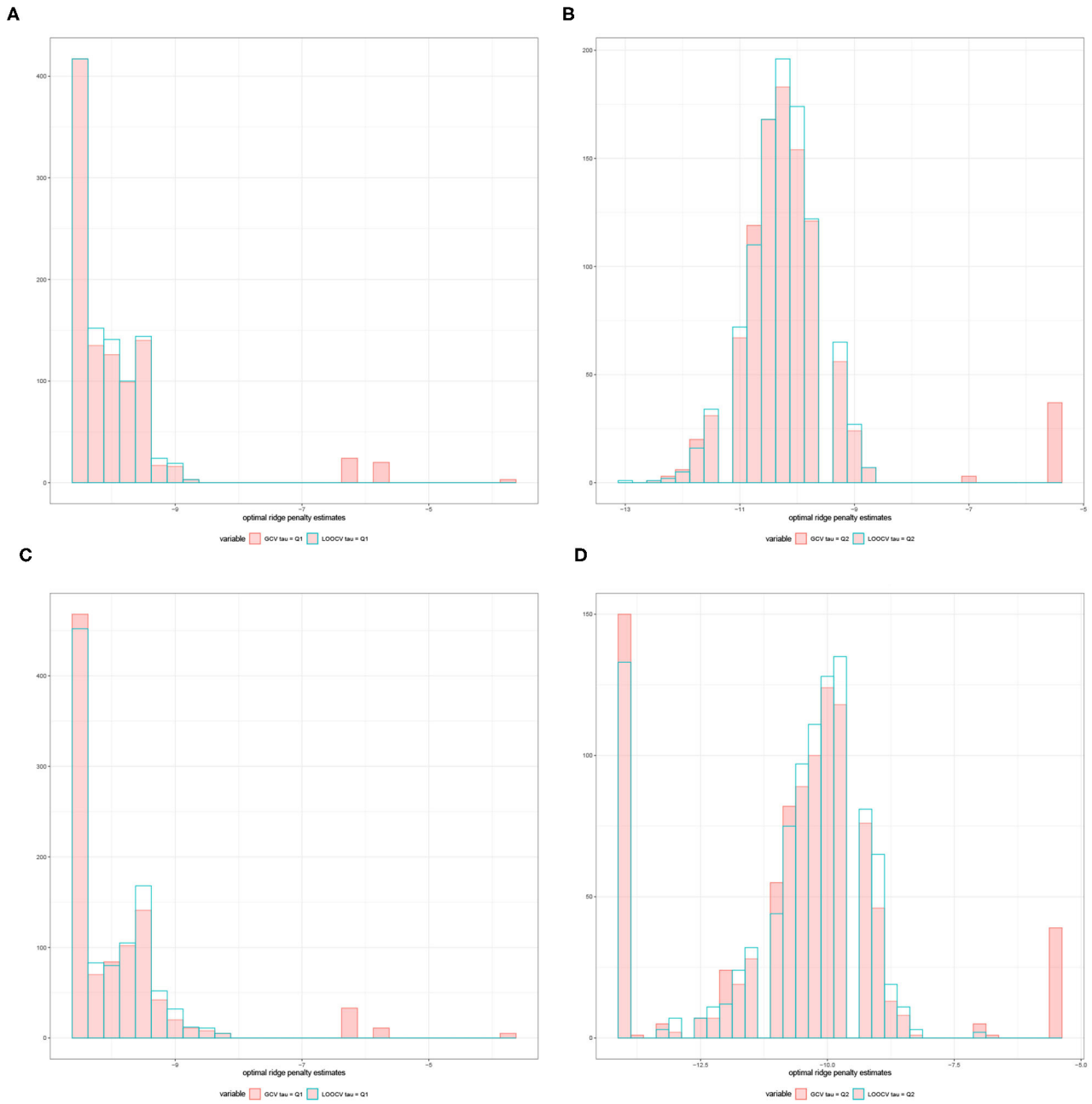


**FIGURE 3 |** Histograms comparing the optimal ridge penalty parameter in a replication with 1,000 permutations using GCV vs. LOOCV when the threshold $\tau$ is the **(A)** first quartile and **(B)** median. The simulated design matrix **X** for **(A,B)** has a compound symmetric covariance structure with $n = 350$ and $p = 1{,}000$, $\boldsymbol{\beta} = \mathbf{0}$ and error standard deviation $\sigma = 1$. The simulated design matrix **X** for **(C,D)** has a compound symmetric covariance structure with $n = 350$ and $p = 1{,}000$. The response is generated using fixed effects $\boldsymbol{\beta} = \xi \mathbf{1}_p, \xi = 3$ and error standard deviation $\sigma = 1$.

For a given replication, the optimal ridge penalty is obtained for the original data and the $B$ permutations of the data. A closer inspection of the distribution of $\widehat{\lambda}^{(b)}, b = 0, 1, \ldots, B$-values are provided in **Figure 3** for compound symmetric covariance, which was intensively studied in Kobak et al. (2020). In **Figure 3A** where the first quartile is used as a threshold, the average value of the ridge penalty parameter is $-8.927$ using GCV and $-9.114$ using LOOCV. Conversely, in **Figure 3B** where the median is the threshold $\tau$, the average value of the ridge penalty parameter is $-9.113$ and $-9.274$ using GCV and LOOCV, respectively. When no thresholding is imposed, the average value of the ridge penalty parameter is the maximum allowable value of $\lambda$ =1,000, using both GCV and LOOCV. These results were obtained from the setting where the simulated design matrix $\mathbf{X}$ has a compound symmetric covariance structure with $n = 350$ and $p = 1,000$, $\boldsymbol{\beta} = \mathbf{0}$ and error standard deviation $\sigma = 1$. In these numerical studies, we confirm that the optimal ridge penalty parameter is negative in some settings whenever a threshold is imposed. Furthermore, the results in **Table 1** ensure us that using a negative value of the ridge penalty parameter still leads to a controlled Type I error rate for a given level $\alpha$.

Consequently, **Figure 3C** displays the comparison of the optimal ridge penalty parameters across all permutations when $\tau$ is equal to the first quartile. The average value of the ridge penalty parameter is $-9.867$ using GCV and $-10.023$ using LOOCV in this setting. Meanwhile, when the median is utilized as threshold in **Figure 3D**, the average value of the ridge penalty parameter is $-10.623$ and $-10.689$ using GCV and LOOCV, respectively. When no thresholding is imposed, the average value of the ridge penalty parameter is 989.1 and 991.9, using GCV and LOOCV, respectively. Results were obtained using a simulated design matrix $\mathbf{X}$ with compound symmetric covariance structure, $n = 350$ and $p = 1,000$. The response is generated using fixed effects $\boldsymbol{\beta} = \xi \mathbf{1}_p, \xi = 3$ and $\sigma = 1$. We observe that some negative ridge penalty parameters have empirical power approaching 1. In these simulations, we were able to verify that this phenomenon can be observed in high-dimensional hypothesis testing where the empirical power approaches 1 as shown in **Table 2**.

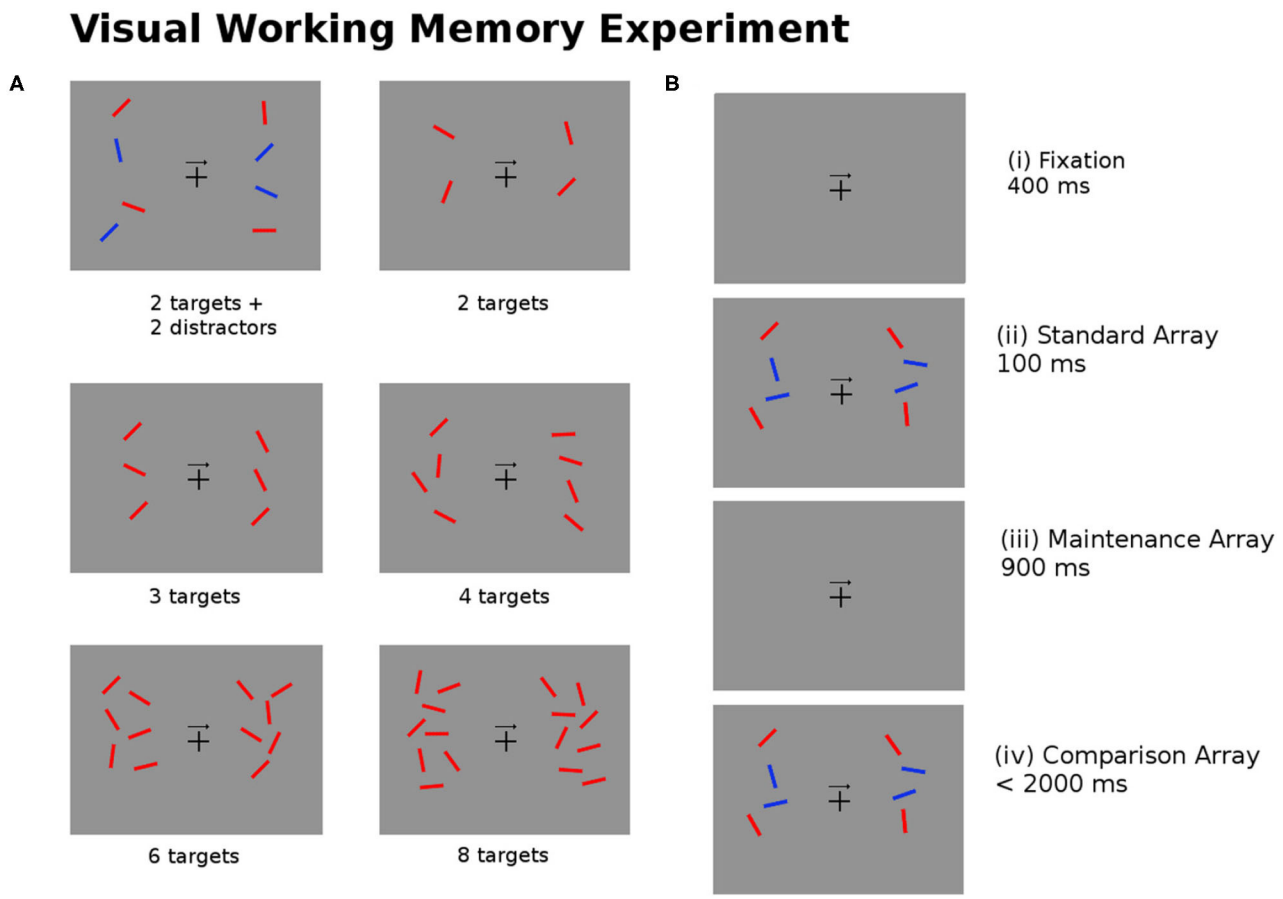


**FIGURE 4 | (A)** The experimental task required subjects to recall the position of the red bars (targets) on the left or right side of the image as indicated by the arrow in the center of the image. The six images show examples of the test image for different numbers of targets and distractors (blue bars). **(B)** For every trial of the experiment the subjects are shown (i) a fixation image (400 ± 200 ms); (ii) the standard array to be memorized (100 ms); (iii) the maintenance image (900 ms); (iv) the comparison array, at which point subjects are asked to respond within 2,000 ms if the targets in the comparison arrays match or mismatch the standard array (Pluta et al., 2021).

In summary, our simulation studies provide evidence in favor of the proposed thresholding procedure discussed in Section 3.1. More importantly, **Figure 3** illustrate the almost sure convergence results in Patil et al. (2021) wherein the distribution of the optimal ridge penalty estimates obtained using GCV and LOOCV coincide and exhibit the same pattern. This almost convergence result for GCV and LOOCV also applies to the Type I error and empirical power of the proposed association tests. Lastly, the Gamma approximation and cross-validation incorporated in the Adaptive Mantel test yields superior power while maintaining the proportion of false positives. This empirically justifies the use of Gamma distribution to approximate the null distribution of the ridge-penalized test statistic.

## 4.2. Application to Imaging Genetics Study

In this study, we consider data from 350 healthy college students from Beijing Normal University (BNU) who participated in an

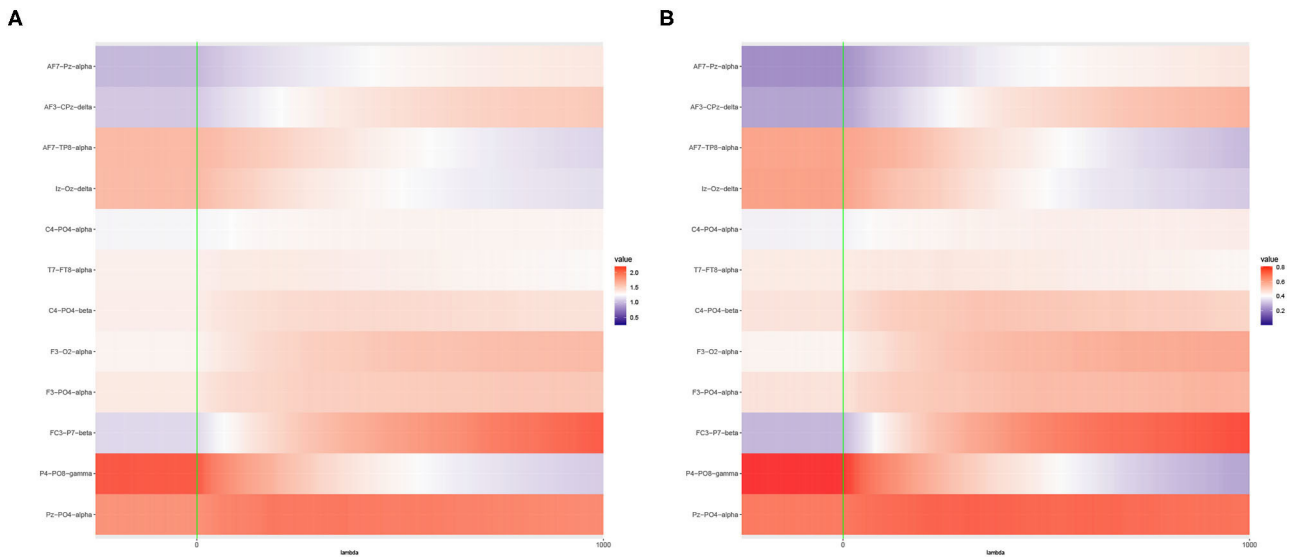


**FIGURE 5 |** Heat maps of **(A)** Average of -$\log_{10}$ $p$-values and **(B)** Empirical Power of the Adaptive Mantel test using the imaging genetics data with $n = 350$ observations, $p = 497$ genetic features and coherence information from top channel-channel-band triples. The green vertical line corresponds to $\lambda = 0$.

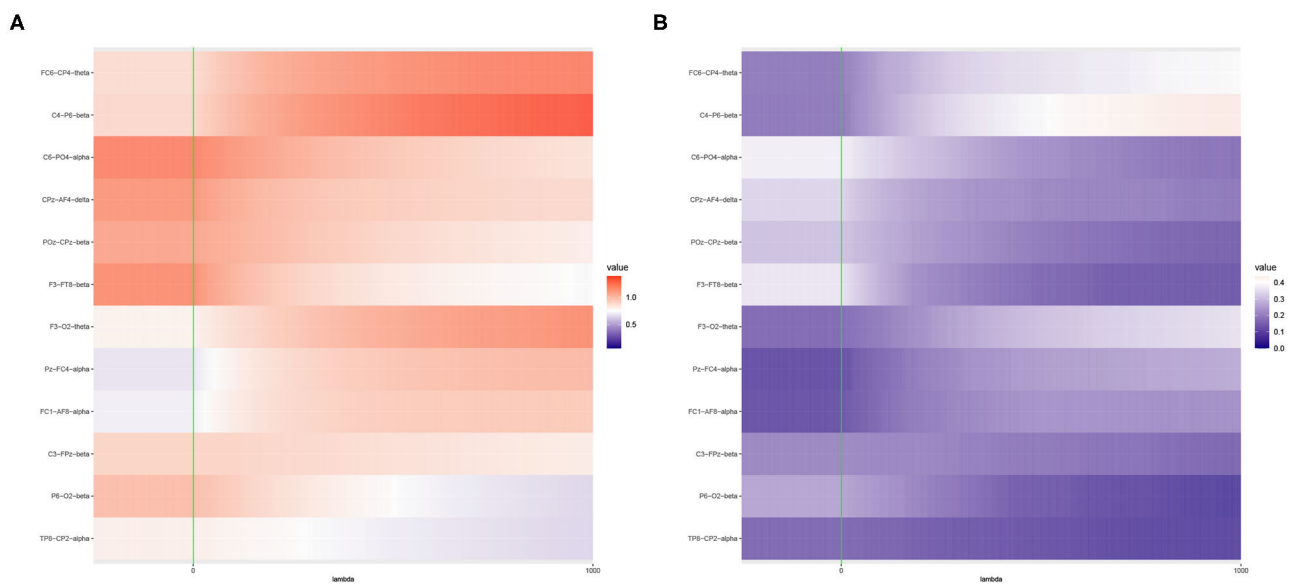


**FIGURE 6 |** Heat maps of **(A)** Average of -$\log_{10}$ $p$-values and **(B)** Empirical Power of the Adaptive Mantel test using the imaging genetics data with $n = 350$ observations, $p = 497$ genetic features and coherence information from some channel-channel-band triples. The green vertical line corresponds to $\lambda = 0$.

experiment involving visual working memory. For every trial, a 64-channel EEG was recorded at 1 kHz. These data constitute a subset of imaging, genetics, and behavioral data collected with the purpose of identifying neurological characteristics of brain connectivity that are significantly associated with both genetic variations and cognitive performance (Pluta et al., 2021).

The task for the participants involves remembering the position of the targets (red bars) on either the left or right side of the image as indicated by the arrow in the center of the image. There are six experimental conditions used for 2, 3, 4, 6, and 8 targets, and for 2 targets added with 2 blue distractors. **Figure 4** describes the experimental task undertaken by the subjects.

The hypothesis being tested is whether the variation in a set of candidate SNPs is associated to the variation in EEG coherence. The main objective of association testing is to determine whether the heritability of EEG coherence in the delta (2–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–50 Hz) bands is significantly greater than zero. In the initial study by Pluta et al. (2021), the matrix of explanatory variables corresponds to the genotype data including 484,496 autosomal SNPs which satisfies the minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) $p$-value quality control thresholds. In addition, the matrix of response variables includes 20,480 features which correspond to the pairwise coherence for 64 EEG channels and 5 frequency bands.

However, with only $n = 350$ subjects, we expect that association between the $q = 20{,}480$ channel-channel-frequency imaging features and $p \approx 500{,}000$ SNPs to be challenging to detect. As mentioned in the previous sections, genetic variants are notorious for contributing fairly weak effects to disease risk due to heterogeneity, among others. Hence, we start with a candidate set of SNPs identified in a genome-wide association study on educational attainment by Okbay et al. (2016). Furthermore, we narrowed down the pairwise channel and frequency band features using the top results presented in Pluta et al. (2021). The resulting number of SNPs considered is $p = 497$ while the total number of brain connectivity features is 250. As mentioned previously, we will consider one response variable at a time, that is, the univariate scenario where $q = 1$.

### 4.2.1. Optimal Ridge Penalty

The liver toxicity data set used in Section 2.4 is well-known for displaying strong signals and is typically used to demonstrate the predictive performance of newly developed vs. existing procedures. However, existing literature on imaging genetics studies suggest that the signals are weaker and sparse. In this section, we are interested in investigating the characteristics of the empirical power for several values of the ridge penalty parameter. For a fixed value of $\lambda$, we implement the empirical

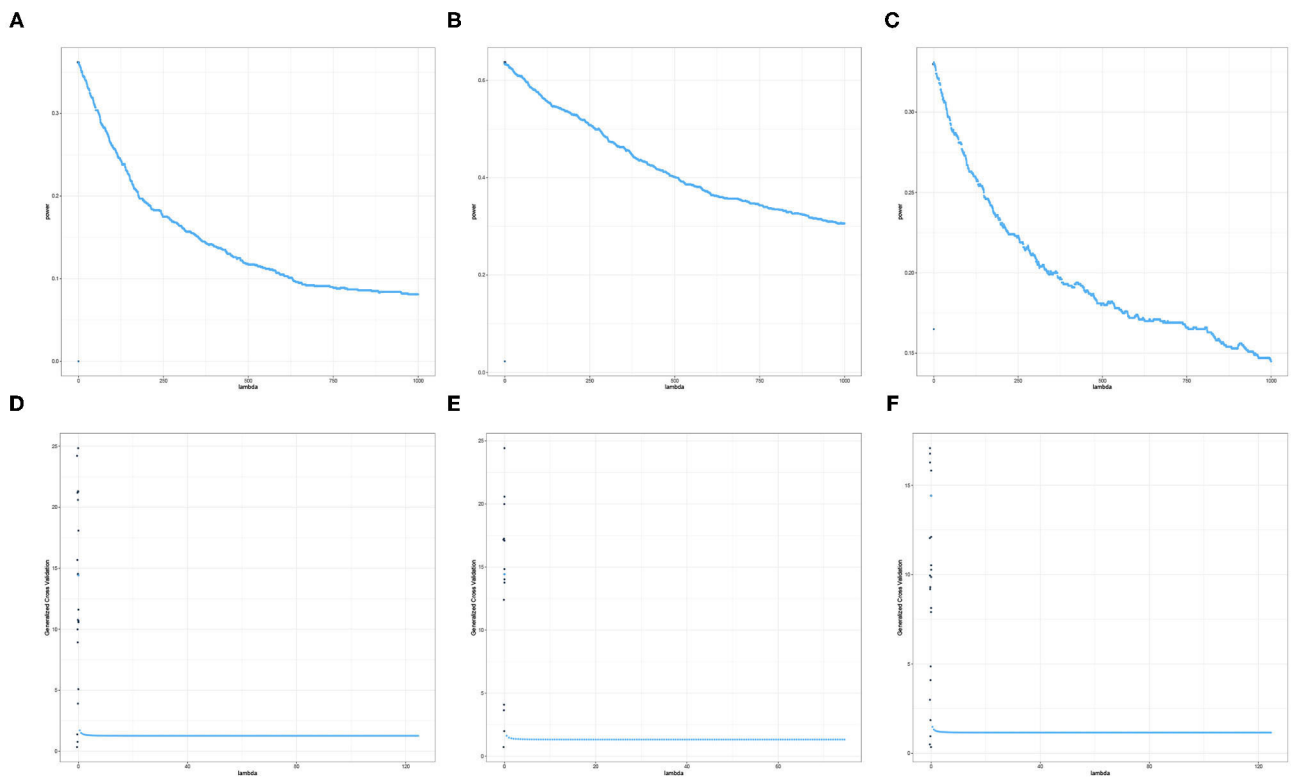


**FIGURE 7 |** Empirical Power and GCV plots of the Adaptive Mantel test for the imaging genetics data with $n = 350$ observations, $p = 497$ genetic features and coherence information from selected channel-channel-band triples with decreasing power trend. **(A)** Power: F3-FT8-beta. **(B)** Power: Iz-Oz-delta. **(C)** Power: CPz-AF4-delta. **(D)** GCV: F3-FT8-beta. **(E)** GCV: Iz-Oz-delta. **(F)** GCV: CPz-AF4-delta.

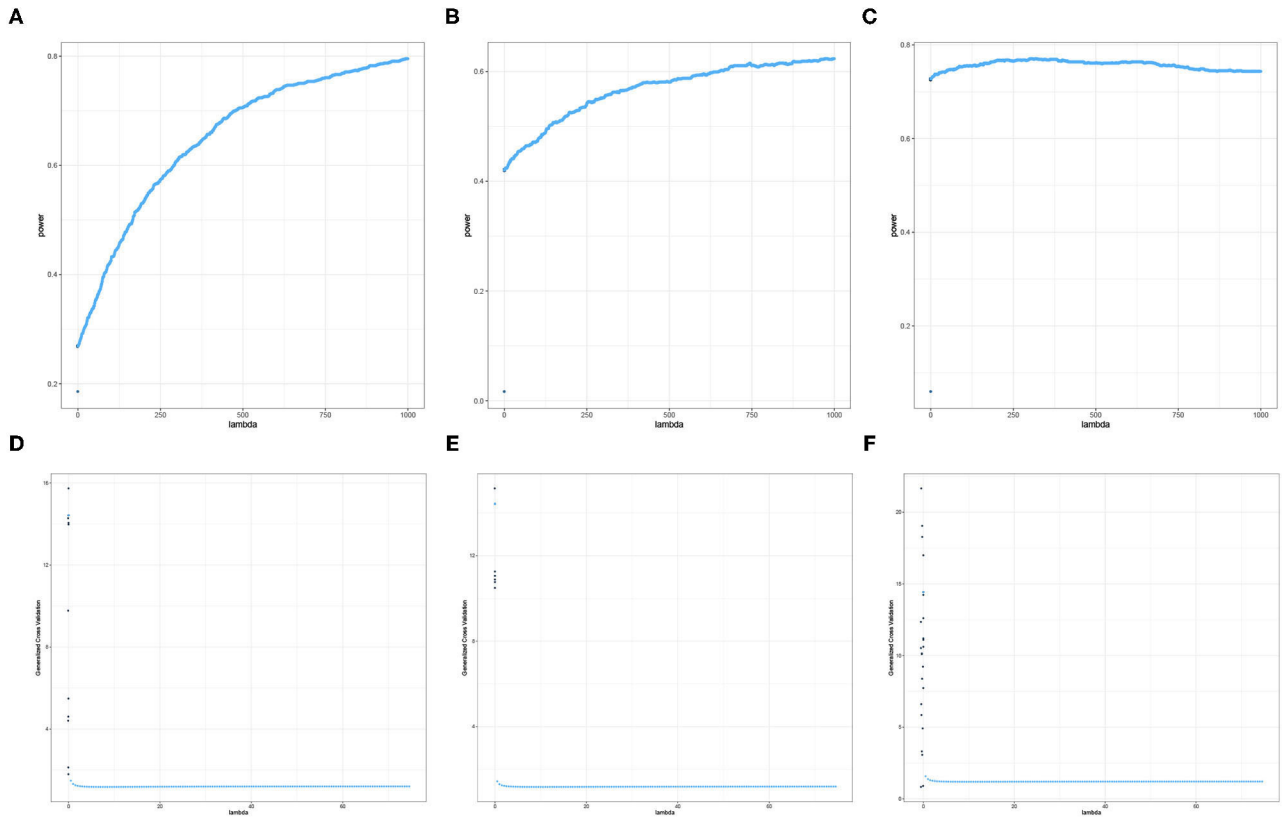


**FIGURE 8 |** Empirical Power and GCV plots of the Adaptive Mantel test for the imaging genetics data with $n = 350$ observations, $p = 497$ genetic features and coherence information from selected channel-channel-band triples with increasing power trend. **(A)** Power: FC3-P7-beta. **(B)** Power: F3-O2-alpha. **(C)** Power: Pz-PO4-alpha. **(D)** GCV: FC3-P7-beta. **(E)** GCV: F3-O2-alpha. **(F)** GCV: Pz-PO4-alpha.

power calculation for the Adaptive Mantel test discussed in Section 2.4. The overall results are summarized in **Figures 5**, **6**.

It is apparent that the brain imaging features considered in **Figure 5** resulted to a higher empirical power compared to the features displayed in **Figure 6**. However, instead of just reporting channel-channel-band triples with high empirical power, we are interested in studying the underlying characteristics which prompted the results. In fact, the pattern of results observed in **Figures 5**, **6** can be categorized into three main groups of power trends. The cool to warm hue of the heat map indicate that there is an increasing trend in the empirical power as the value of ridge penalty parameter chosen increases. However, the warm to cool hue in the heat map describes the opposite trend. The third case corresponds to the almost constant hue for any ridge penalty included in the interval.

The results displayed in **Figures 7**, **8** illustrate clearly the increasing or decreasing power trends for some pairwise channel-band triples as compared to the subtle differences observed in **Figures 5**, **6**. Among these overall cluster of results, we will explore further how the optimal ridge penalty parameter impacts the empirical power. Visually, we can deduce using **Figures 7A–C** that the optimal value of the ridge penalty parameter should be negative or close to zero to arrive at the highest value of empirical power. In contrast, **Figures 8A,B** suggest that the optimal ridge penalty should be chosen as high as possible, i.e., $\lambda \rightarrow \infty$ to achieve empirical power approaching 1. Lastly, **Figure 8C** displays an almost horizontal trend where $\lambda$ can be chosen anywhere in the interval and yield comparable empirical power with any other $\lambda$. The corresponding GCV plots for both **Figures 7**, **8** are provided to evaluate the pattern exhibited by the GCV for different values of $\lambda$. It is clear that when the ridge penalty parameter is positive, the value of the GCV is a smooth function. However, since this is the high-dimensional setting and multiple factors are at play simultaneously, it is not clear which dominating factor dictates the power trend and GCV plots displayed by the real data. We will probe into the theoretical justifications of this phenomena applied to high-dimensional or ultra-high-dimensional settings in future research.

### 4.2.2. Adaptive Mantel Test With Cross-Validation

Among the 250 brain connectivity features studied, we identified five features which indicate that the variation in a set of candidate SNPs is associated to the variation in EEG coherence. Using the AdaMantCV and AdaMantGACV methods, we determine that the heritability of EEG coherence in the delta (2–4 Hz), alpha (8–12 Hz), and gamma (30–50 Hz) bands is significantly greater than zero. The $p$-values are presented in **Table 4**.

**TABLE 4 |** Comparison of the *p*-values using AdaMantCV and AdaMantGACV methods where heritability of EEG coherence is significantly greater than zero.

| | | Adaptive Mantel Test | | | |
|---|---|---|---|---|---|
| | | With CV | | With GA and CV | |
| **Band** | **Channels** | **GCV** | **LOOCV** | **GCV** | **LOOCV** |
| Delta | FP1 - T8 | 0.016 | 0.015 | 0.012 | 0.012 |
| Delta | CPz - F8 | 0.009 | 0.008 | 0.007 | 0.007 |
| Delta | FC5 - O2 | 0.004 | 0.003 | 0.005 | 0.004 |
| Alpha | F3 - FC1 | 0.037 | 0.038 | 0.006 | 0.007 |
| Gamma | P4 - PO8 | 0.034 | 0.034 | 0.014 | 0.015 |

In addition, we were able to identify 21 other channel-channel-band triples which wherein the variation in a set of candidate SNPs is associated to the variation in EEG coherence using AdaMant with Gamma Approximation and cross-validation but not using AdaMantCV. The real data analysis results are aligned with the simulation studies wherein the Adaptive Mantel test with Gamma approximation and cross-validation have superior power while maintaining the proportion of false positives. Consequently, we have identified that there are more significant variations in the alpha and delta band frequencies using AdaMant with Gamma approximation and cross-validation. These results are consistent with the existing literature by Smit et al. (2005) where heritability is generally highest around the alpha peak frequency. According to Shaw (2003), the variation in the alpha rhythm has been posited to reflect individual differences in working memory, attentional demands and/or arousal, and also cognitive preparedness.

## 5. DISCUSSION

For over several decades, ridge regression has proved to be a valuable tool for use by researchers and it has recently been intensively explored in the high-dimensional context to better understand more complicated models. Even though there are several available methods for choosing an optimal value of the ridge penalty parameter, the ultimate choice of $\lambda$ for a specific application still remains to be unsolved. One contributing reason to this is the difficulty in characterizing the unknown data generating distribution, which usually influences the optimal level of regularization.

In this study, we examine the role that ridge penalization plays in hypothesis testing by conducting an empirical power study of an imaging genetics data set. Our results confirm that in high-dimensional settings, overfitting might provide higher power, in addition to good generalization in predictive problems. One noticeable difference is that while no penalty, i.e., $\lambda = 0$ often works well for predictions, it does not have any power in hypothesis testing, as typical test statistics reach their extreme values or they do not change over permutations when $p > n$. While an empirical study provides helpful guides for practical purposes, it is only the first step toward a more rigorous and holistic understanding of the broader scenarios. We are working on theoretical justifications and hope they will provide further insights to hypothesis testing problems into high-dimensional or ultra-high-dimensional settings.

We also propose a thresholding procedure to allow the set of candidate values of $\lambda$ to include negative values and investigate how these negative penalty parameters can affect the empirical power of the Mantel Test. Furthermore, we extend the Adaptive Mantel Test (AdaMant) algorithm to incorporate Gamma approximation and the optimal selection of the ridge penalty parameter *via* generalized and leave-one-out cross-validation. We compare the resulting optimal choice between the two cross-validation procedures and note that they coincide. We also applied the proposed method to imaging genetics study of visual working memory measured by EEG coherence in healthy college students. Overall, we have encountered an interesting statistical phenomenon and gained some insights regarding ridge regression, especially as it applies to imaging genetics data.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

IG and ZY drafting and refining the manuscript. ZY and HO critical reading of the manuscript. All authors contributed to the manuscript preparation of the article and approved the submitted version.

## REFERENCES


Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics* 13, 469–475. doi: 10.1080/00401706.1971.10488811

Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv*. 4, 40–79. doi: 10.1214/09-SS054

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. U.S.A*. 117, 30063–30070. doi: 10.1073/pnas.1907378117

Batmanghelich, N. K., Dalca, A. V., Sabuncu, M. R., and Golland, P. (2013). "Joint modeling of imaging and genetics," in *International Conference on Information Processing in Medical Imaging* (Asilomar, CA: Springer), 766–777. doi: 10.1007/978-3-642-38868-2_64

Bühlmann, P., and Ćevid, D. (2020). Deconfounding and causal regularisation for stability and external validity. *Int. Stat. Rev*. 88, S114–S134. doi: 10.1111/insr.12426

Bushel, P. R., Wolfinger, R. D., and Gibson, G. (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Syst. Biol*. 1, 15. doi: 10.1186/1752-0509-1-15

Cule, E., and De Iorio, M. (2013). Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet. Epidemiol*. 37, 704–714. doi: 10.1002/gepi.21750

Cule, E., Vineis, P., and De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC Bioinform*. 12, 372. doi: 10.1186/1471-2105-12-372

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 Genet*. 9, e1003608. doi: 10.1371/journal.pgen.1003608

Delaney, N. J., and Chatterjee, S. (1986). Use of the bootstrap and cross-validation in ridge regression. *J. Bus. Econ. Stat*. 4, 255–262. doi: 10.1080/07350015.1986.10509520

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P., et al. (2013). Mantel test in population genetics. *Genet. Mol. Biol*. 36, 475–485. doi: 10.1590/S1415-47572013000400002

Dobriban, E., and Wager, S. (2018). High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Stat*. 46, 247–279. doi: 10.1214/17-AOS1549

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21, 215–223. doi: 10.1080/00401706.1979.10489751

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. doi: 10.48550/arXiv.1903.08560

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Edn*. New York, NY: Springer Series in Statistics. doi: 10.1007/978-0-387-84858-7

Hayes, B., Goddard, M., et al. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Heinloth, A. N., Irwin, R. D., Boorman, G. A., Nettesheim, P., Fannin, R. D., Sieber, S. O., et al. (2004). Gene expression profiling of rat livers reveals indicators of potential adverse effects. *Toxicol. Sci*. 80, 193–202. doi: 10.1093/toxsci/kfh145

Henderson, H. V., and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Rev*. 23, 53–60. doi: 10.1137/1023004

Hoerl, A. E. (1962). Applications of ridge analysis to regression problems. *Chem. Eng. Prog*. 58, 54–59.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Hsu, D., Kakade, S. M., and Zhang, T. (2012). "Random design analysis of ridge regression," in *Conference on Learning Theory, JMLR Workshop and Conference Proceedings* (Edinburgh).

Karoui, N. E. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv [Preprint] arXiv*:1311.2445.

Kobak, D., Lomond, J., and Sanchez, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res*. 21, 169–161.

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013). "Near-optimal bounds for cross-validation via loss stability," in *International Conference on Machine Learning* (Atlanta, GA), 27–35.

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 14, 667–681. doi: 10.1093/biostatistics/kxt006

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., et al. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* 72, 156–164. doi: 10.1111/biom.12368

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* 63, 1079–1088. doi: 10.1111/j.1541-0420.2007.00799.x

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

Marenco, S., and Radulescu, E. (2010). Imaging genetics of structural brain connectivity and neural integrity markers. *Neuroimage* 53, 848–856. doi: 10.1016/j.neuroimage.2009.11.030

Meijer, R. J., and Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometr. J*. 55, 141–155. doi: 10.1002/bimj.201200088

Meyer-Lindenberg, A., Nicodemus, K. K., Egan, M. F., Callicott, J. H., Mattay, V., and Weinberger, D. R. (2008). False positives in imaging genetics. *Neuroimage* 40, 655–661. doi: 10.1016/j.neuroimage.2007.11.058

Nathoo, F. S., Kong, L., Zhu, H., and Alzheimer's Disease Neuroimaging Initiative (2019). A review of statistical methods in imaging genetics. *Can. J. Stat*. 47, 108–131. doi: 10.1002/cjs.11487

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp*. 15, 1–25. doi: 10.1002/hbm.1058

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533, 539–542. doi: 10.1038/nature17671

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. (2021). "Uniform consistency of cross-validation estimators for high-dimensional ridge regression," in *International Conference on Artificial Intelligence and Statistics* (PMLR), 3178–3186.

Peper, J. S., Brouwer, R. M., Boomsma, D. I., Kahn, R. S., and Hulshoff Pol, H. E. (2007). Genetic influences on human brain structure: a review of brain imaging studies in twins. *Hum. Brain Mapp*. 28, 464–473. doi: 10.1002/hbm.20398

Pluta, D., Shen, T., Xue, G., Chen, C., Ombao, H., and Yu, Z. (2021). Ridge-penalized adaptive mantel test and its application in imaging genetics. *Stat. Med*. 40, 5313–5332. doi: 10.1002/sim.9127

Randolph, T. W., Harezlak, J., and Feng, Z. (2012). Structured penalties for functional linear models–partially empirical eigenvectors for regression. *Electron. J. Stat*. 6, 323. doi: 10.1214/12-EJS676

Rao, C. R. (1948). "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation," in *Mathematical Proceedings of the Cambridge Philosophical Society* (Cambridge University Press), 50–57. doi: 10.1017/S0305004100023987

Richards, D., Mourtada, J., and Rosasco, L. (2021). "Asymptotics of ridge (less) regression under general source condition," in *International Conference on Artificial Intelligence and Statistics* (PMLR), 3889–3897.

Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *J. R. Stat. Soc. Ser. C Appl. Stat*. 25, 257–265. doi: 10.2307/2347233

Shaw, J. C. (2003). *The Brain's Alpha Rhythms and the Mind*. Chichester: BV Elsevier Science.

Shaw, P. A., and Proschan, M. A. (2013). Null but not void: considerations for hypothesis testing. *Stat. Med*. 32, 196–205. doi: 10.1002/sim.5497

Smit, D., Posthuma, D., Boomsma, D., and De Geus, E. (2005). Heritability of background eeg across the power spectrum. *Psychophysiology* 42, 691–697. doi: 10.1111/j.1469-8986.2005.00352.x

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* 36, 111–133. doi: 10.1111/j.2517-6161.1974.tb00994.x

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35, 2769–2794. doi: 10.1214/009053607000000505

Thompson, P. M., Ge, T., Glahn, D. C., Jahanshad, N., and Nichols, T. E. (2013). Genetics of the connectome. *Neuroimage* 80, 475–488. doi: 10.1016/j.neuroimage.2013.05.013

Tikhonov, A. N. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* 39, 195–198.

Tost, H., Bilek, E., and Meyer-Lindenberg, A. (2012). Brain connectivity in psychiatric imaging genetics. *Neuroimage* 62, 2250–2260. doi: 10.1016/j.neuroimage.2011.11.007

Wu, D., and Xu, J. (2020). On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. doi: 10.48550/arXiv.2006.05800

Xu, Z., Xu, G., Pan, W., and Alzheimer's Disease Neuroimaging Initiative (2017). Adaptive testing for association between two random vectors in moderate to high dimensions. *Genet. Epidemiol.* 41, 599–609. doi: 10.1002/gepi.22059

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). "Understanding deep learning requires rethinking generalization," in *The 5th International Conference on Learning Representations* (Toulon).

Zhao, B., and Zhu, H. (2019). Cross-trait prediction accuracy of high-dimensional ridge-type estimators in genome-wide association studies. doi: 10.48550/arXiv.1911.10142

Zhou, C., Zwilling, C. E., Calhoun, V. D., and Wang, M. Y. (2014). Efficient blockwise permutation tests preserving exchangeability. *Int. J. Stat. Med. Res.* 3, 145. doi: 10.6000/1929-6029.2014.03.02.8

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2022 Gauran, Xue, Chen, Ombao and Yu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*