Multiple-Splitting Projection Test for High-Dimensional Mean Vectors

Wanjun Liu   WALIU@LINKEDIN.COM
LinkedIn Corporation
Sunnyvale, CA 94085, USA

Xiufan Yu   XIUFAN.YU@ND.EDU
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
Notre Dame, IN 46556, USA

Runze Li   RZLI@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Abstract

We propose a multiple-splitting projection test (MPT) for one-sample mean vectors in high-dimensional settings. The idea of projection test is to project high-dimensional samples to a 1-dimensional space using an optimal projection direction such that traditional tests can be carried out with projected samples. However, estimation of the optimal projection direction has not been systematically studied in the literature. In this work, we bridge the gap by proposing a consistent estimation via regularized quadratic optimization. To retain type I error rate, we adopt a data-splitting strategy when constructing test statistics. To mitigate the power loss due to data-splitting, we further propose a test via multiple splits to enhance the testing power. We show that the p-values resulted from multiple splits are exchangeable. Unlike existing methods which tend to conservatively combine dependent p-values, we develop an exact level α test that explicitly utilizes the exchangeability structure to achieve better power. Numerical studies show that the proposed test well retains the type I error rate and is more powerful than state-of-the-art tests.

Keywords: Exchangeable p-values; High-dimensional mean tests; Multiple data-splitting; Optimal projection direction; Regularized quadratic optimization.

1. Introduction

Hypothesis testing on mean vectors is a fundamental problem in statistical inference theory and attracts considerable interest in numerous scientific applications. For example, neuroscientists make inferences on the average signals of fMRI data to monitor brain activities and diagnose abnormal tissues (Ginestet et al., 2017). Geneticists analyze gene expression levels to understand the mechanism of how genes are related to diseases (Wang et al., 2015). In these applications, the data dimension p is typically comparable with or much larger than sample size n, making traditional tests ineffective or practically infeasible. In this work, we study the problem of testing whether a population mean $\boldsymbol{\mu}$ equals to some known vector

$\boldsymbol{\mu}_0$ under high-dimensional regime where $p > n$. Without loss of generality, we set $\boldsymbol{\mu}_0 = \mathbf{0}$ throughout the paper. To formally formulate the problem, let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ be a random sample from a p-dimensional population $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Of interest is to test

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}. \tag{1}$$

The Hotelling's T^2 test has been well studied when $p < n$ and p is fixed. As p exceeds n, the sample covariance matrix becomes singular and hence T^2 is not well-defined. Even in the case $p < n$, the testing power of T^2 is largely defective if $p/n \to c \in (0, 1)$ (Bai and Saranadasa, 1996).

Three types of tests have been developed in efforts to handle the high-dimensional challenge. The first type is quadratic-form test, which replaces the singular sample covariance matrix with an invertible matrix (e.g., identity matrix) (Bai and Saranadasa, 1996; Chen et al., 2011; Chen and Qin, 2010). These tests tend to neglect the dependence among covariates and may suffer from low power when covariates are strongly correlated. The second type is known as extreme-type test, which utilizes the extreme value of a sequence of marginal test statistics, see Cai et al. (2014); Zhong et al. (2013). Such extreme-type statistics typically converge to some extreme value distribution and are generally disadvantaged by slow convergence, making it hard to control the type I error when n is small. The third type is projection test (Lopes et al., 2011; Huang, 2015; Liu and Li, 2020; Li and Li, 2021), which maps the high-dimensional sample to a low-dimensional space, and subsequently applies traditional methods (e.g., Hotelling's T^2) to the projected sample. Intuitively, the projection procedure seeks to transform the data in such a way that the dimension is reduced, while the statistical distance between H_0 and H_1 is mostly preserved through the transformed distributions.

Recently, Huang (2015) proved that the optimal choice of projection direction is $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. To facilitate a data-driven decision regarding the projection direction, Huang (2015) also proposed a projection test based on a data-splitting procedure, i.e., half of the sample is employed to estimate the optimal projection direction, while the other half is used to perform the test. However, there are two main drawbacks with this data-splitting projection test. First, a ridge-type estimator is used to estimate the projection direction. Their power analysis relies on the assumption that the ridge-type estimator is consistent, which is no longer true in high-dimensional settings. Secondly, the single data-splitting procedure is often criticized as only half of the sample is used to perform the test, which inevitably results in power loss. These two drawbacks actually reveal two existing unsolved issues with the projection test based on a data-splitting procedure:

1. How to estimate the optimal projection direction with statistical guarantee?
2. How to mitigate the power loss caused by the data-splitting procedure?

In this paper, we propose a multiple-splitting projection test for high-dimensional mean vectors. Our proposed test addresses the aforementioned issues in the following two ways: (1) the optimal projection is estimated via a regularized quadratic optimization such that a consistent estimator is obtained; and (2) a multiple data-splitting procedure is proposed to improve the testing power. The main contributions can be summarized in three folds.

First, we propose a consistent estimation of the optimal projection direction via non-convex regularized quadratic programming. Non-asymptotic error bounds are established,

which hold for *all stationary* points with high probability. In other words, we do not need to solve the global solution to the nonconvex optimization problem as any stationary point has desirable statistical guarantee.

Second, we prove that p-values constructed from a multiple data-splitting procedure are *exchangeable*. Furthermore, we generalize the exchangeability of p-values proposition to a more general permutation framework. As an extension, the methodology proposed in this work can be further applied to many statistical inference problems.

Third, an *exact* level α test is proposed to combine multiple p-values which explicitly utilizes the exchangeability of these p-values. Such exchangeability is often neglected in traditional combination approaches. By doing so, our test is more powerful than the single-splitting test as well as existing combination approaches. To the best of our knowledge, this is the first work that exploits the exchangeability of p-values and utilizes such exchangeability in developing high-dimensional hypothesis testing.

The rest of this paper is organized as follows. In Section 2, we introduce a new estimation of the optimal projection direction via regularized quadratic programming. In Section 3, we investigate the dependency structure of p-values resulted from a multiple-splitting procedure and propose an exact level α multiple-splitting projection test. In Section 4, we conduct numerical studies to compare the proposed MPT with existing tests as well as other p-value combination methods. We conclude this paper with discussion on potential applications of this multiple-splitting framework to other statistical inference problems in Section 5.

2. Estimation of Optimal Projection Direction

In this section, we introduce a consistent estimation of the optimal projection direction for projection tests. Section 2.1 provides a brief introduction to projection tests. Section 2.2 presents the estimator as a stationary point of a regularized quadratic optimization problem and establishes its non-asymptotic error bounds.

2.1 Background on Projection Tests

The idea of projection test is to project the high-dimensional vector $\mathbf{x} \in \mathbb{R}^p$ onto a space of low dimension such that traditional tests can be applied. Let $\mathbf{P}$ be a $p \times q$ full column-rank projection matrix (or vector if $q = 1$) with $q < n$ and define $\mathbf{y}_i = \mathbf{P}^\top \mathbf{x}_i \in \mathbb{R}^q, i = 1, \ldots, n$. Under H_0, $\mathrm{E}(\mathbf{y}_i) = \mathbf{0}$ and Hotelling's T^2 test can be applied to the q-dimensional projected sample $\mathbf{y}_i$'s,

$$T_{\mathbf{P}}^2 = n\bar{\mathbf{x}}^\top \mathbf{P}(\mathbf{P}^\top \widehat{\boldsymbol{\Sigma}} \mathbf{P})^{-1} \mathbf{P}^\top \bar{\mathbf{x}},$$

where $\bar{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the sample mean and sample covariance matrix. Under H_0, $T_{\mathbf{P}}^2$ converges to χ_q^2 distribution as $n \to \infty$.

The projection test pivots the attention to the question on how to effectively construct the projection matrix $\mathbf{P}$. Various approaches have been developed with respect to different choices of $\mathbf{P}$. A data-dependent method was proposed in Lauter (1996) by setting $\mathbf{P} = \mathbf{d}$, where $\mathbf{d}$ is a $p \times 1$ vector depending on data only through $\mathbf{X}^\top \mathbf{X}$. Lopes et al. (2011) proposed a random projection test in which the entries in $\mathbf{P}$ are randomly drawn from standard normal distribution. More recently, Huang (2015) proved that under normality assumption, the optimal choice q is 1 and the optimal projection direction is of the form

$\mathbf{P} = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ in the sense that the power of $T_{\mathbf{P}}^2$ is maximized. For non-Gaussian samples, the direction $\mathbf{P} = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ is still asymptotically optimal as long as the sample mean of projected sample is asymptotically normal.

The estimation of the optimal projection direction has not been systematically studied yet, leaving a gap between theory and practice for projection tests. In what follows, we propose a new estimating procedure such that a consistent estimator is obtained.

Notations: Before proceeding, we first set up some notations. For a vector $\mathbf{v} = (v_j)_{j=1}^p \in \mathbb{R}^p$, let $\|\mathbf{v}\|_k$ be its ℓ_k norm, $k = 1, 2$. Its ℓ_0 norm $\|\mathbf{v}\|_0$ is the number of nonzero entries in $\mathbf{v}$ and ℓ_∞ norm is $\|\mathbf{v}\|_\infty = \max|v_j|$. For a matrix $\mathbf{M} = (m_{ij})_{i,j=1}^p \in \mathbb{R}^{p\times p}$, its elementwise ℓ_∞ norm is $\|\mathbf{M}\|_{\max} = \max|m_{ij}|$. For a set $\mathcal{D}$, $|\mathcal{D}|$ denotes its cardinality. We use $a \vee b$ to denote the larger one of a and b.

2.2 Estimation via Regularized Quadratic Optimization

In this subsection, we aim to bridge the gap between theoretical analysis and practical implementation regarding the optimal projection direction. The empirical performance of a data-driven projection test relies heavily on the estimation accuracy of the optimal projection direction. However, in high-dimensional settings, there is no statistical guarantee for the ridge-type estimator introduced in Huang (2015).

We propose a new consistent estimator to improve the test performance with the assumption that $\mathbf{w}^\star$ is sparse. Observing that $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ is the minimizer of $\frac{1}{2}\mathbf{w}^\top\mathbf{\Sigma}\mathbf{w} - \boldsymbol{\mu}^\top\mathbf{w}$, we propose to estimate $\mathbf{w}^\star = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ using the following regularized quadratic optimization

$$\underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}\mathbf{w}^\top\widehat{\mathbf{\Sigma}}\mathbf{w} - \bar{\mathbf{x}}^\top\mathbf{w} + P_\lambda(\mathbf{w}), \tag{2}$$

where $P_\lambda(\mathbf{w}) = \sum_{j=1}^p P_\lambda(w_j)$ is a penalty function satisfying the following conditions

   (i) $P_\lambda(0) = 0$ and $P_\lambda(t)$ is symmetric around 0,

   (ii) $P_\lambda(t)$ is differentiable for $t \neq 0$ and $\lim_{t\to 0^+} P_\lambda'(t) = \lambda$,

   (iii) $P_\lambda(t)$ is a non-decreasing function on $t \in [0, \infty)$,

   (iv) $P_\lambda(t)/t$ is a non-increasing function on $t \in [0, \infty)$,

   (v) There exists $\gamma > 0$ such that $P_\lambda(t) + \frac{\gamma}{2}t^2$ is convex.

Such conditions on P_λ are mild (Loh and Wainwright, 2015) and are satisfied by a wide variety of penalties including the Lasso (Tibshirani, 1996) and nonconvex regularizers such as the SCAD (Fan and Li, 2001), and the MCP (Zhang, 2010). We further assume that the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ satisfies the following restricted strong convexity (RSC) condition,

$$\mathbf{\Delta}^\top\widehat{\mathbf{\Sigma}}\mathbf{\Delta} \geqslant \nu\|\mathbf{\Delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\mathbf{\Delta}\|_1 \ \text{for} \ \mathbf{\Delta} \in \mathbb{R}^p \ \text{and} \ \|\mathbf{\Delta}\|_1 \geqslant 1, \tag{3}$$

where $\nu > 0$ is a strictly positive constant and $\tau \geqslant 0$ is a non-negative constant. When $p < n$, $\widehat{\mathbf{\Sigma}}$ is positive definite, one can set $\tau = 0$ and ν be the smallest eigenvalue of $\widehat{\mathbf{\Sigma}}$. In the high-dimensional setting where $p > n$, $\widehat{\mathbf{\Sigma}}$ is semi-positive definite and $\mathbf{\Delta}^\top\widehat{\mathbf{\Sigma}}\mathbf{\Delta} \geqslant 0$ for all $\mathbf{\Delta} \in \mathbb{R}^p$.

Thus the RSC condition (3) holds trivially for $\{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_1/\|\boldsymbol{\Delta}\|_2^2 > c\}$ with $c = \frac{\nu}{\tau}\sqrt{\frac{n}{\log p}}$. As a result, we only require the RSC condition to hold in the set $\{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_1/\|\boldsymbol{\Delta}\|_2^2 \leqslant c, \|\boldsymbol{\Delta}\|_1 \geqslant 1\}$. The RSC condition (3) is imposed on $\widehat{\boldsymbol{\Sigma}}$ only for $\|\boldsymbol{\Delta}\|_1 \geqslant 1$, and it turns out the condition actually holds for all $\boldsymbol{\Delta} \in \mathbb{R}^p$, see Lemma 10. Such RSC-type condition is widely used in establishing the non-asymptotic error bounds in high-dimensional statistics and is satisfied with high probability under sub-Gaussianity assumption (Agarwal et al., 2012; Loh and Wainwright, 2015, 2017). Alternatively, the RSC-type condition can also be replaced by a similar condition known as restricted eigenvalue (RE) condition (Bickel et al., 2009; Van De Geer and Bühlmann, 2009).

It is quite challenging to obtain the global solution to the optimization problem (2) if a nonconvex penalty P_λ is used. Instead of searching for the global solution, we establish the non-asymptotic error bounds for any stationary point $\widehat{\mathbf{w}}$ that satisfies the following first-order condition,

$$\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} - \bar{\mathbf{x}} + \nabla P_\lambda(\widehat{\mathbf{w}}) = \mathbf{0}, \tag{4}$$

where ∇P_λ denotes the sub-gradient of P_λ. The condition (4) is a necessary condition for $\widehat{\mathbf{w}}$ to achieve a local minimum. Therefore the set of $\widehat{\mathbf{w}}$ satisfying (4) includes all local minimizers as well as the global one.

Lots of efficient algorithms have been developed to attain stationary points even when the objective function is nonconvex. These algorithms include local linear approximation (Fan et al., 2014; Wang et al., 2013; Zou and Li, 2008), composite gradient descent method (Loh and Wainwright, 2015; Nesterov, 2013), and proximal-gradient method (Wang et al., 2014). In practice, we may choose the tuning parameter λ in the penalty function by cross-validation or the high-dimensional BIC criterion proposed in Wang et al. (2013). We impose the following conditions,

(C1) $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are identically and independently distributed sub-Gaussian vectors.

(C2) The sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies the RSC condition in (3) with $3\gamma \leqslant 4\nu$.

(C3) There exists constant $C_1 > 0$ such that $\|\mathbf{w}^\star\|_1 \leqslant C_1$.

Remark 1 *Condition (C3) is posited to ensure a good estimation of $\mathbf{w}^\star$. By the definition of $\mathbf{w}^\star$, $\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star$ should be somewhat close to $\boldsymbol{\mu}$. Note that $\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \boldsymbol{\mu}\|_\infty = \|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \boldsymbol{\Sigma}\mathbf{w}^\star\|_\infty \leqslant \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \cdot \|\mathbf{w}^\star\|_1$. A diverging $\|\mathbf{w}^\star\|_1$ would amplify the estimation error of $\widehat{\boldsymbol{\Sigma}}$.*

The following theorem establishes the ℓ_1 and ℓ_2 error bounds for all stationary points $\widehat{\mathbf{w}}$ under the alternative hypothesis.

Theorem 2 *Suppose conditions (C1)-(C3) hold. Let $\widehat{\mathbf{w}}$ be any stationary point of the problem (2) with $\lambda = C\sqrt{\log p/n}$ for some large constant C. Then under H_1 (i.e., $\mathbf{w}^\star \neq \mathbf{0}$), with probability at least $1 - cp^{-1}$ for some absolute constant c, we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 = O\left(s\sqrt{\frac{\log p}{n}}\right) \quad and \quad \|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2 = O\left(\sqrt{\frac{s\log p}{n}}\right),$$

where $s = \|\mathbf{w}^\star\|_0$ is the number of nonzero entries in $\mathbf{w}^\star$.

Remark 3 *Though inspired by Loh and Wainwright (2015), we would like to clarify the difference between Theorem 2 and the results in Loh and Wainwright (2015). The optimization problem in Loh and Wainwright (2015) requires an additional constraint $\|\mathbf{w}\|_1 \leqslant R$ for some tuning parameter R to ensure $\|\widehat{\mathbf{w}}\|_1$ is bounded by R. R needs to be chosen carefully such that $\mathbf{w}^\star$ is feasible and both the penalty parameter λ and sample size n also depend on R. However, how to choose R is not clear in practice. In our work, we modify the RSC condition by substituting $\|\mathbf{\Delta}\|_2^2$ for $\|\mathbf{\Delta}\|_2$ in the RSC condition (3) so that the constraint $\|\widehat{\mathbf{w}}\|_1 \leqslant R$ is no longer needed.*

Note that the error bounds in Theorem 2 hold for all stationary points. In other words, any local solution is guaranteed with desirable statistical accuracy and a global one is unnecessary if it is too challenging to achieve. Theorem 2 implies $\widehat{\mathbf{w}}$ is a consistent estimator under H_1 if $\mathbf{w}^\star$ is sparse (more precisely, $\sqrt{s \log p/n} \to 0$). Such consistency under H_0 is not guaranteed as there is no signal in the true parameter $\mathbf{w}^\star = \mathbf{0}$. Fortunately, with the data-splitting technique, we will see in Section 3 that the size of the proposed projection test is always controlled regardless of the consistency of the estimator $\widehat{\mathbf{w}}$.

3. Data-Splitting Based Projection Test

In this section, we present full methodological details of our proposed multiple-splitting projection test (MPT) together with its theoretical properties. After carefully studying the dependency of p-values resulted from a multiple-splitting procedure, we introduce a new combination framework that makes use of the dependency structure. Section 3.1 demonstrates a single-splitting projection test using the estimator introduced in the previous section. Section 3.2 studies the exchangeability of p-values. Section 3.3 provides a brief overview of traditional approaches for combining multiple p-values. Section 3.4 formally presents our combination framework and the proposed MPT.

3.1 Single-Splitting Projection Test (SPT)

Data-splitting technique has a long history in statistical applications and remains attractive in modern statistics (Wasserman and Roeder, 2009; Barber and Candès, 2019). We begin with one single data-splitting. Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ denote the set of full sample and we partition the full sample into two disjoint sets $\mathcal{D}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n\}$ with $|\mathcal{D}_1| = n_1$ and $|\mathcal{D}_2| = n_2 = n - n_1$. The idea is to use $\mathcal{D}_1$ to estimate the optimal projection direction while use $\mathcal{D}_2$ to conduct the test with projected sample. To be more specific, we estimate the optimal projection direction $\mathbf{w}^\star$ using a stationary point $\widehat{\mathbf{w}}$ of the following regularized quadratic optimization problem

$$\underset{\mathbf{w}}{\text{minimize}} \; \frac{1}{2}\mathbf{w}^\top \widehat{\mathbf{\Sigma}}_1 \mathbf{w} - \bar{\mathbf{x}}_1^\top \mathbf{w} + P_\lambda(\mathbf{w}), \tag{5}$$

where $\bar{\mathbf{x}}_1$ and $\widehat{\mathbf{\Sigma}}_1$ are sample mean and sample covariance matrix computed from $\mathcal{D}_1$. Then we project the observations in $\mathcal{D}_2$ to a 1-dimensional space as follows: $y_i = \widehat{\mathbf{w}}^\top \mathbf{x}_i, i = n_1 + 1, \ldots, n$. The one-sample t-test is readily applied to the projected sample and the resulting test statistic is

$$T_{\widehat{\mathbf{w}}} = \sqrt{n_2}\bar{y}/s_y, \tag{6}$$

where $\bar{y}$ and s_y^2 are the sample mean and sample variance of $\{y_{n_1+1}, \ldots, y_n\}$. Due to the data-splitting, the estimator $\widehat{\mathbf{w}}$ is independent of $\mathcal{D}_2$. As a result, the test $T_{\widehat{\mathbf{w}}}$ is an exact one-sample t-test if $\mathbf{x}_i$'s are normally distributed, and the p-value of test $T_{\widehat{\mathbf{w}}}$ is given by

$$p(T_{\widehat{\mathbf{w}}}) = 2\left(1 - G_{n_2-1}(|T_{\widehat{\mathbf{w}}}|)\right), \tag{7}$$

where G_{n_2-1} is the cdf of t_{n_2-1} distribution. Without normality assumption, $T_{\widehat{\mathbf{w}}}$ has an asymptotic standard normal distribution, and the p-value is given by $p(T_{\widehat{\mathbf{w}}}) = 2\left(1 - \Phi(|T_{\widehat{\mathbf{w}}}|)\right)$. We reject the null hypothesis at significance level α whenever $p(T_{\widehat{\mathbf{w}}}) < \alpha$.

We refer to $T_{\widehat{\mathbf{w}}}$ as the single-splitting projection test (SPT). Ideally, one would like to use full sample to estimate $\mathbf{w}^\star$ and use full sample to perform the test, which makes the limiting distribution challenging to derive since the projection $\widehat{\mathbf{w}}$ and sample are dependent. Thanks to the data-splitting procedure, an exact t-test can be achieved as $\widehat{\mathbf{w}}$ is independent of $\mathcal{D}_2$. Furthermore, we would like to point out that the size of the SPT is well controlled regardless of how $\widehat{\mathbf{w}}$ is estimated, but a consistent $\widehat{\mathbf{w}}$ ensures high power under the alternative. The following theorem demonstrates the asymptotic power of the SPT.

Theorem 4 *Suppose that the conditions in Theorem 2 hold. Further assume* $(1 \vee \|\boldsymbol{\mu}\|_\infty)s\sqrt{\log p/n} \to 0$ *and* $n_2/n \to \kappa \in (0,1)$, *where* n_2 *is the sample size of* $\mathcal{D}_2$. *Let* $\zeta = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ *and* $z_{\alpha/2}$ *be the upper* $\alpha/2$ *quantile of* $N(0,1)$, *then the asymptotic power of the proposed SPT at a given significance level* α *is*

$$\beta(T_{\widehat{\mathbf{w}}}) = \Phi(-z_{\alpha/2} + \sqrt{n\kappa\zeta}).$$

Remark 5 *The term* ζ *can be interpreted as the signal strength of alternative hypothesis. As long as* $n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \to \infty$, *the proposed SPT has asymptotic power approaching 1. Under local alternative* $\boldsymbol{\mu} = \boldsymbol{\delta}/\sqrt{n}$ *for some fixed* $\boldsymbol{\delta} \neq \mathbf{0}$, *the asymptotic power of the SPT is* $\Phi(-z_{\alpha/2} + \sqrt{\kappa\boldsymbol{\delta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}})$. *To achieve high power empirically, we adopt the same strategy as in Huang (2015) and recommend to take* $n_2 = \lfloor \kappa n \rfloor$ *with* $\kappa \in [0.4, 0.6]$.

3.2 Exchangeability of p-values

To compensate the power loss due to the data-splitting procedure, we consider a multiple-splitting procedure (formally presented in Section 3.4), which repeats the data-splitting multiple times and aggregates all the information in p-values to make inferences on H_0. More specifically, we consider m times of data-splitting for some fixed integer m. Let π_k, $k = 1, \ldots, m$, be a random permutation of $\{1, \ldots, n\}$. Accordingly, let $\mathcal{D}^{\pi_k} = \{\mathbf{x}_{\pi_k(1)}, \ldots, \mathbf{x}_{\pi_k(n)}\}$ denote the permutated sample. For each $k = 1, \ldots, m$, we apply the SPT to $\mathcal{D}^{\pi_k}$ and obtain the p-value p_k according to (7). Before proceeding with the combination of p-values, we first investigate the dependence structure of these p-values. The following theorem establishes the exchangeability among the p-values.

Theorem 6 *The p-values* $(p_1, p_2, \ldots, p_m)$ *resulted from the multiple-splitting procedure are exchangeable, i.e.,* $(p_1, \ldots, p_m) \overset{d}{=} (p_{\pi(1)}, \ldots, p_{\pi(m)})$ *for any* π, *a permutation of* $\{1, \ldots, m\}$.

The exchangeability of p-values holds no matter it is under H_0 or H_1. We would like to point out that such exchangeability structure of p-values holds for a general permutation

framework. In many statistical problems, with the technique of data-splitting, the first half sample $\mathcal{D}_1^{\pi_k}$ can be used to learn the underlying model (e.g., parameter estimation, variable selection). We denote the acquired knowledge by $f(\mathcal{D}_1^{\pi_k})$. Then together with the second half sample $\mathcal{D}_2^{\pi_k}$, a p-value (or some other statistic) can be derived for some specific inference problem $p_k = g(f(\mathcal{D}_1^{\pi_k}), \mathcal{D}_2^{\pi_k})$. With fixed mappings f and g, it can be shown that the p_k's are also exchangeable. For instance, let us consider the high-dimensional linear regression problem and of interest is to test whether some coefficient, say β_j, is 0 or not. With data-splitting, one can use $\mathcal{D}_1^{\pi_k}$ to select a set of important covariates $\widehat{\mathcal{A}}$ such that $|\widehat{\mathcal{A}}| < n-1$. Then we can fit ordinary least squares with the covariate set $\widehat{\mathcal{A}} \cup \{j\}$ and obtain the p-value p_k of a test regarding whether $\beta_j = 0$. Since the p-values are exchangeable, the MPT introduced in Section 3.4 is readily to combine the p-values.

In fact, such exchangeability holds even without data-splitting. The key of exchangeability lies in that conditioning on the dataset $\mathcal{D}$, its m permutations $\mathcal{D}^{\pi_1}, \ldots, \mathcal{D}^{\pi_m}$ are independent from each other. The following theorem generalizes the results in Theorem 6.

Theorem 7 *Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample and $\pi_1, \cdots, \pi_m$ be m permutations of $\{1, \cdots, n\}$. $\mathcal{D}^{\pi_1}, \ldots, \mathcal{D}^{\pi_m}$ denote the m permutated samples of $\mathcal{D}$. Let g be a mapping from $\mathcal{D}^{\pi_k}$ to a statistic: $T_k = g(\mathcal{D}^{\pi_k})$, then $T_1, \ldots, T_m$ are exchangeable.*

3.3 Traditional Combination of p-values

One popular strategy to enhance testing power is via the combination of p-values (Romano and DiCiccio, 2019; Yu et al., 2019, 2020, 2022). In fact, combining multiple p-values from a set of hypothesis tests has been widely used in statistical literature. Let $p_1, \ldots, p_m$ denote m valid p-values. That is, under H_0,

$$\Pr(p_k \leqslant u) = u, \ 0 < u < 1 \quad \text{for } k = 1, \ldots, m.$$

A natural question is how we can decide whether H_0 should be rejected based on the m p-values such that the type I error rate is still retained.

Classical approaches require independence assumption among p-values. Examples include the Fisher's method, the Pearson's method, the Stouffer's method, the Tippett's method, and many others. In the meantime, significant efforts were made to combine dependent p-values. Rüschendorf (1982) proved that twice the average p-values remains a valid p-value and proposed an average-based combination test, which rejects H_0 at level α if the average of $p_1, \ldots, p_m$ is less than or equal to $\alpha/2$. Romano and DiCiccio (2019) introduced a quantile-based combination test. A special case is we reject H_0 at level α if the median of p-values is less than or equal to $\alpha/2$. Under H_0, we know $Z_k = \Phi^{-1}(p_k) \sim N(0,1)$, where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Assuming $(Z_1, \ldots, Z_m)^\top$ follows a multivariate normal distribution, Romano and DiCiccio (2019) proposed a Z-average test based on the sample mean of Z_k's, that rejects H_0 if $|\sum_{k=1}^{m} Z_k| \geqslant m z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of $N(0,1)$. More recently, Liu and Xie (2020) introduced a new combination test based on the Cauchy transformation which is insensitive to dependencies among p-values. The test rejects H_0 at level α if $\sum_{k=1}^{m} \tan\{(0.5 - p_i)\pi\} \geqslant m c_\alpha$, where c_α is the upper α-quantile of standard Cauchy distribution.

In general, these methods tend to be over-conservative in order to control Type I error rate without taking advantage of certain dependence structure (e.g., exchangeability). This

can be regarded as a trade-off between potential size inflation and possible power loss. The traditional combination methods generally ignore the dependency structure, therefore tend to make unnecessarily large comprise in testing power in order to retain a correct size for a less favorable scenario. In Section 4.1, we use simulation studies to compare the size and power of the proposed MPT with those of traditional combination methods. On the one hand, the numerical studies show that the size of MPT is slightly below the level $\alpha = 0.05$ while the size of traditional combination methods are almost 0. On the other hand, the power of MPT is much higher than those traditional combination methods. In summary, compared to the traditional ones, the proposed MPT is less conservative in terms of testing size, and exhibits much higher testing power.

3.4 Multiple-Splitting Projection Test (MPT)

Based on the m exchangeable p-values $\{p_1, \ldots, p_m\}$, the question is how we can make a decision on whether H_0 should be rejected or not. To improve upon the traditional methods, we propose a new framework to combine p-values obtained from multiple splits. The proposed framework takes advantage of the exchangeability structure among those p-values, as a result, achieving higher testing power than existing commonly used combination approaches.
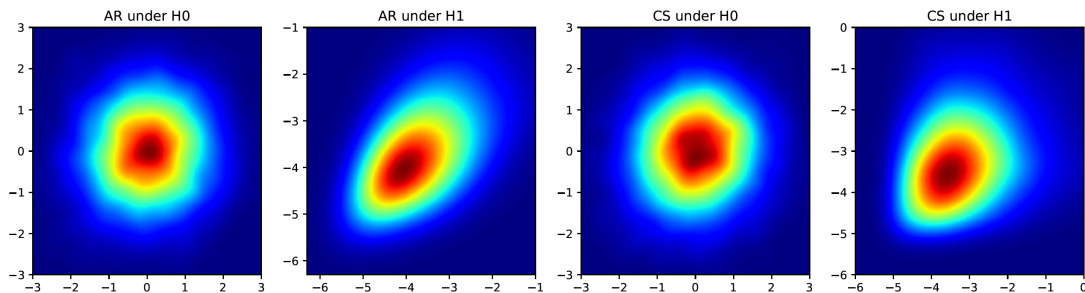

Figure 1: Density plots of (Z_1, Z_2) under H_0 and H_1 with autocorrelation (AR) covariance structure and compound symmetry (CS) covariance structure when $n = 40$ and $p = 1000$.

Let $Z_k = \Phi^{-1}(p_k)$, $k = 1, \ldots, m$. Under H_0, $Z_1, \ldots, Z_m$ are exchangeable standard normal random variables, with correlation $\text{corr}(Z_i, Z_j) = \rho \geqslant 0$ for any pair $(i, j), i \neq j$ due to exchangeability. Figure 1 depicts the density of (Z_1, Z_2) with $m = 2$ under different covariance structures (see Section 4 for detailed descriptions of simulation settings). It shows that (Z_1, Z_2) are clearly exchangeable (symmetric). Under H_0, (Z_1, Z_2) are approximately normally distributed centering at $(0, 0)$. Under H_1, the joint distribution of (Z_1, Z_2) is not normal and its center is far away from $(0, 0)$.

Let $\bar{Z}$ be the sample mean, then we have $\text{E}(\bar{Z}) = 0$ and $\text{Var}(\bar{Z}) = (1 + (m-1)\rho)/m$ under H_0. If $(Z_1, \ldots, Z_m)$ are jointly normally distributed, then the standardized statistic of $\bar{Z}$, $M_\rho = \bar{Z}/\sqrt{(1 + (m-1)\rho)/m} \sim N(0, 1)$. In general, by the central limit theorem for exchangeable random variables (e.g., see Klass and Teicher (1987)), we know

$$M_\rho = \bar{Z}/\sqrt{(1 + (m-1)\rho)/m} \xrightarrow{d} N(0, 1). \tag{8}$$

9

The correlation ρ is typically unknown and needs to be estimated from the sample. Two approaches to estimate ρ will be provided later in this subsection. Let $\widehat{\rho}$ denote an estimator of ρ. We first present the asymptotic distribution of $M_{\widehat{\rho}}$ under H_0.

Theorem 8 *Let $\widehat{\rho}$ be a consistent estimator of $\rho > 0$. Under H_0, we have as $m \to \infty$,*

$$M_{\widehat{\rho}} = \bar{Z}/\sqrt{(1 + (m-1)\widehat{\rho})/m} \xrightarrow{d} N(0,1). \tag{9}$$

Remark 9 *When $\rho = 0$, the p-value combination is reduced to the independent case. Hence M_ρ converges to the standard normal distribution at rate $\sqrt{m}$ as $m \to \infty$. However, when $\rho \neq 0$, $M_{\widehat{\rho}}$ converges at a slower rate m as the variance of $\bar{Z}$ is not asymptotically degenerate. Hypothesis testing based on such a slow convergence rate is more likely to fail in controlling the type I error in finite-sample performance.*

Remark 9 suggests that the asymptotic distribution in (9) does not serve as a good cornerstone to test H_0. The slow convergence rate and potential size inflation become major concerns for practitioners. In practice, one may conduct a large number of splits which bring in extra computational burden. This motivates us to seek an exact level α test to ensure the finite-sample performance. Let $c(\rho, m, \alpha/2)$ be the upper $\alpha/2$ quantile of the distribution of $M_{\widehat{\rho}}$ and we reject H_0 if $|M_{\widehat{\rho}}| > c(\rho, m, \alpha/2)$. Given $\widehat{\rho}$, the exact distribution of $M_{\widehat{\rho}}$ depends on ρ and is very difficult to derive, so is $c(\rho, m, \alpha/2)$. Instead, we use the critical value $c(m, \alpha)$ that is chosen against the least favorable ρ such that type I error is controlled regardless ρ, i.e., $c(m, \alpha/2) = \sup_{\rho \in (0,1)} c(\rho, m, \alpha/2)$. Then we reject H_0 at level α if

$$|M_{\widehat{\rho}}| > c(m, \alpha/2). \tag{10}$$

We refer to the test (10) as multiple-splitting projection test (MPT) and summarize full methodological details in Algorithm 1. Note that the critical value $c(m, \alpha/2)$ depends on the way you estimate ρ. With the choice $c(m, \alpha/2)$, the MPT is still an exact level α test but no longer a size α test.

Algorithm 1 Multiple-splitting Projection Test (MPT)

1: **Input:** dataset $\mathcal{D}$, the number of splits m, n_1, and significance level α
2: **Step 1:** randomly generate m permutations of $\{1, \ldots, n\}$, denoted by π_k, $k = 1, \ldots, m$
3: **Step 2:** obtain multiple p-values
4: **for** $k = 1$ **to** m **do**
5:   (1) partition the permuted sample $\mathcal{D}^{\pi_k}$ into $\mathcal{D}_1^{\pi_k}$ and $\mathcal{D}_2^{\pi_k}$ and obtain $\bar{\mathbf{x}}_1^k$, $\widehat{\mathbf{\Sigma}}_1^k$ from $\mathcal{D}_1^{\pi_k}$
6:   (2) estimate $\widehat{\mathbf{w}}^k$ using a stationary point of $\underset{\mathbf{w}}{\text{minimize}} \frac{1}{2}\mathbf{w}^\top \widehat{\mathbf{\Sigma}}_1^k \mathbf{w} - \bar{\mathbf{x}}_1^{k\top} \mathbf{w} + P_\lambda(\mathbf{w})$
7:   (3) project $\mathcal{D}_2^{\pi_k}$ and obtain $y_i^k = \widehat{\mathbf{w}}^{k\top} \mathbf{x}_{\pi_k(i)}, i = n_1 + 1, \ldots, n$
8:   (4) $T_{\widehat{\mathbf{w}}^k} = \sqrt{n_2}\bar{y}^k/s_y^k$, where $\bar{y}^k$ and $(s_y^k)^2$ are the sample mean and variance of $\{y_{n_1+1}^k, \ldots, y_n^k\}$

9:   (5) compute the p-values by $p_k = 2\left(1 - \Phi(|T_{\widehat{\mathbf{w}}^k}|)\right)$
10: **end for**
11: **Step 3:** combine the p-values
12:   (1) compute the sample mean $\bar{Z}$ and variance s_Z^2 of $\{Z_k = \Phi^{-1}(p_k), k = 1, \ldots, m\}$
13:   (2) compute test statistic $M_{\widehat{\rho}} = \bar{Z}/\sqrt{(1 + (m-1)\widehat{\rho})/m}$
14: **Return:** Reject H_0 at level α if $|M_{\widehat{\rho}}| > c(m, \alpha/2)$

Two estimators of ρ are introduced in Follmann and Proschan (2012). Let s_Z^2 is the sample variance of Z_i's. The first estimator is given by $\widehat{\rho}_1 = \max(0, 1 - s_Z^2)$. The second estimator, which is quantile based, is given by $\widehat{\rho}_2 = \max(0, 1 - (m-1)s_Z^2/\chi_{m-1}^2(1-\beta))$, where $\chi_{m-1}^2(1-\beta)$ is the upper $(1-\beta)$ quantile of χ_{m-1}^2. An appealing approach is to choose β as large as possible so that the test with $c(m, \alpha/2) = z_{\alpha/2}$ remains level α for all ρ. We refer to Follmann and Proschan (2012) for more details. Follmann and Proschan (2012) also provides the table of critical values $c(m, \alpha/2)$ and β for different m, see Tables 2 and 3 in Appendix A. We see that $c(m, \alpha/2)$ increases dramatically as m increases, leading to low power when m is large. Hence the quantile approach is preferred when m is relatively large.

As for the choice of m, Figure 2 shows how the testing power changes with the number of splits m under settings with different correlation structure, samples sizes and dimensions. We would like to point out that the proposed MPT is a valid level α test regardless the dependence structure since the critical value is chosen such that Type I error is controlled for all dependence structure (i.e., for all ρ). In other words, the proposed MPT is able to control Type I error regardless the number of splits m and data characteristics (e.g., sample size, data dimensionality). The main purpose of conducting multiple splitting (i.e., choosing $m > 1$) is to mitigate power loss brought by single data-splitting procedure and improve the testing power over SPT. Under alternative hypothesis, the testing power from each split depends on the original data characteristics. Hence theoretically the exact relationship between m and the testing power of MPT also depend on original data characteristics. As shown in the plot, the testing power increases as m increases but the improvement of power becomes insignificant when m is relatively large. Considering the fact that a large number of splits will increase computational cost, we recommend to set $m \in [30, 60]$ as a reasonable choice in practice considering the trade-off between testing power and computational cost.

4. Numerical Studies

In this section, we conduct numerical studies to demonstrate the finite-sample performance of the proposed MPT through both Monte Carlo simulation and a real data example.

4.1 Monte Carlo Simulation

We compare the proposed MPT with other state-of-the-art tests and p-value combination approaches. In particular, we include the following tests in our experiments:

- **Projection test**: our proposed SPT and MPT (with $m = 40$), ridge projection test (Ridge) (Huang, 2015), and random projection test (RPT) (Lopes et al., 2011).

- **Combining p-values**: Median-based combination (Median) (Romano and DiCiccio, 2019), average-based combination (Average) (Rüschendorf, 1982), average-based combination using $\Phi^{-1}(p_k)$ (Z-average) (Romano and DiCiccio, 2019), and Cauchy combination (Cauchy) (Liu and Xie, 2020).

- **Quadratic-form test**: CQ test (Chen and Qin, 2010).
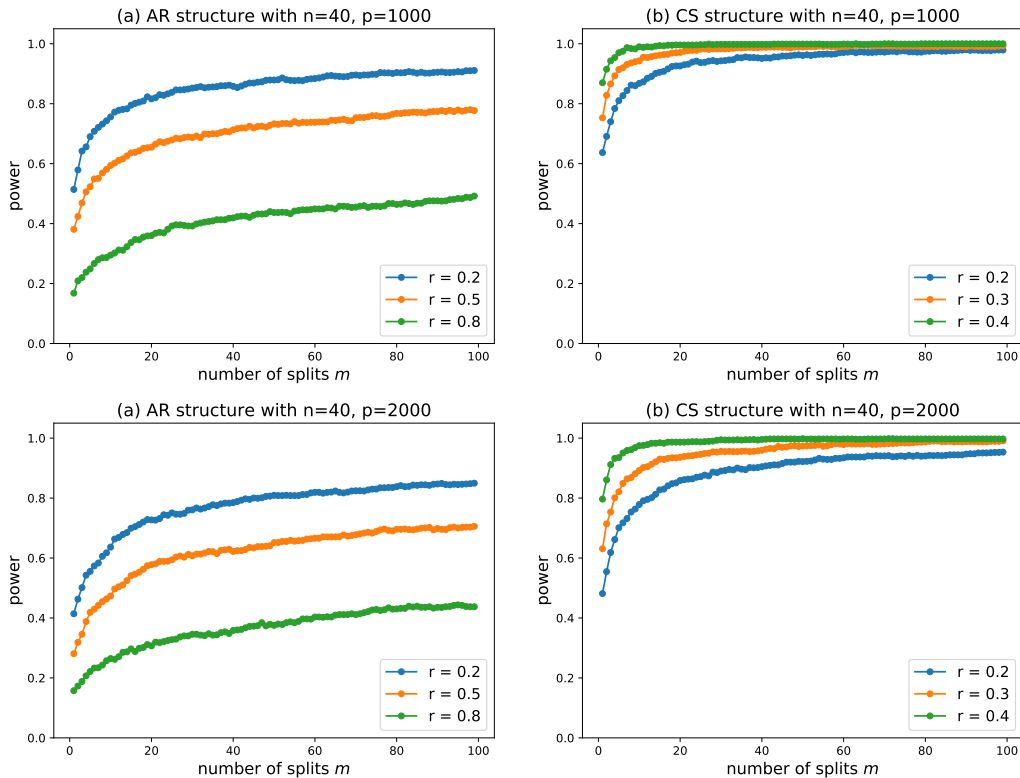
- **Extreme-type test**: CLX test (Cai et al., 2014).

Figure 2: Testing power versus number of splits m for autocorrelation (AR) structure and compound symmetry (CS) structure with different choices of n and p. Different colors correspond to different correlations r.

We generate a random sample of size n from $N_p(c\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mathbf{1}_{10}^\top, \mathbf{0}_{p-10}^\top)^\top$. We set $c = 0, 0.5$ to examine the size and the power of these tests, respectively. To examine the test robustness to non-normally distributed data, we also generate random samples from a multivariate t_6-distribution. Let σ_{ij} be the (i, j) entry in $\boldsymbol{\Sigma}$. For $r \in (0, 1)$, we consider the following two covariance matrices: (1) compound symmetry (CS) with $\sigma_{ij} = r$ if $i \neq j$ and $\sigma_{ij} = 1$ if $i = j$ and (2) autocorrelation (AR) with $\sigma_{ij} = r^{|i-j|}$. We vary r from 0.1 to 0.9 with step size 0.1 to examine the impact of correlation on size and power. We set sample size $n = 40, 100$ and dimension $p = 1000$.

In the above settings, the optimal projection direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is sparse or approximately sparse. When $\boldsymbol{\Sigma}$ has the compound symmetry structure, $\boldsymbol{\Sigma}^{-1}$ is an approximately sparse matrix in the sense that the off-diagonal entries are of order p^{-1} and dominated by its diagonal entries. Then optimal projection direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is also approximately sparse since the first 10 entries dominate the rest entries. When $\boldsymbol{\Sigma}$ has the autocorrelation structure, $\boldsymbol{\Sigma}^{-1}$ is a 3-sparse matrix, meaning that at most three entries in each row or column are nonzero, and the resulting optimal projection direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is sparse as well. We set $\kappa = 0.5$ when implementing the SPT and the MPT, i.e., half of the sample is used to estimate the projection direction and the other half is used to perform the test. The quantile approach $\widehat{\rho}_2$

is used to estimate pairwise correlation among Z_k's. We set the type I error rate $\alpha = 0.05$. All simulation results are based on 10,000 independent replications.
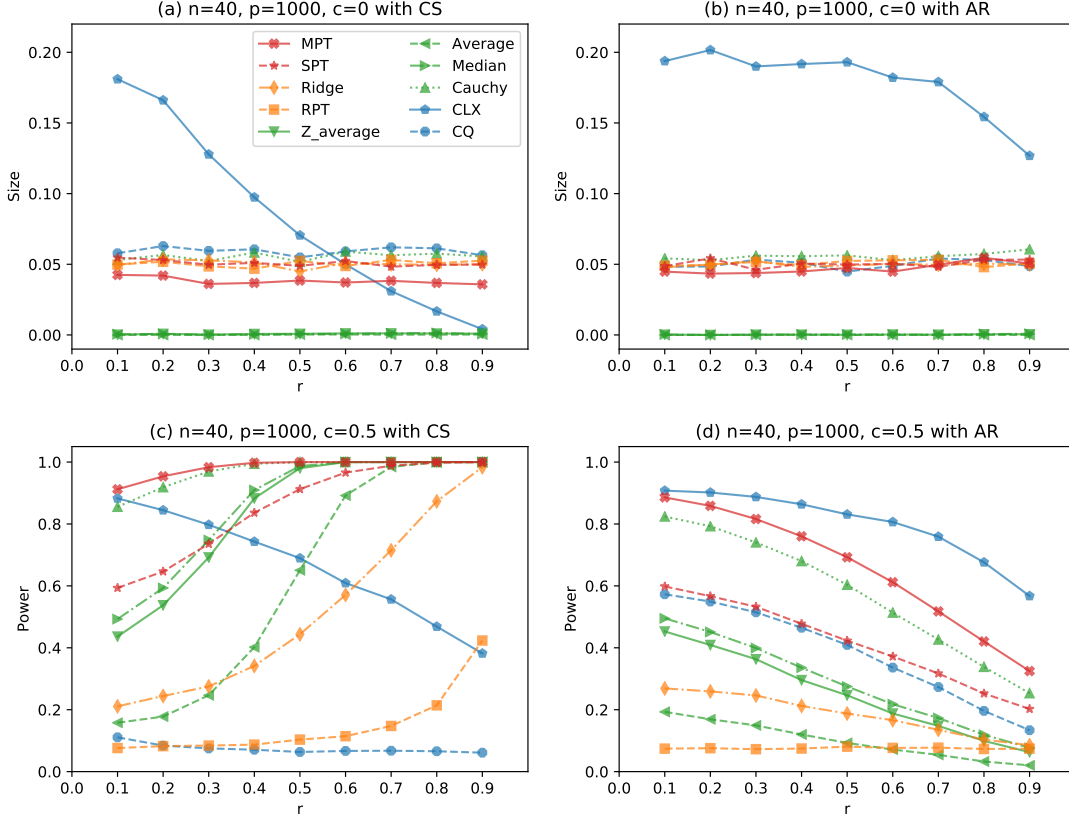

Figure 3: Size and power of different tests for normally distributed samples with $n = 40$. Panels (a) and (b) show size ($c = 0$) under the null hypothesis for the CS and AR structure, respectively. Panels (c) and (d) show power ($c = 0.5$) under the alternative hypothesis for the CS and AR structure, respectively.

Figure 3 reports the size and power of different tests for normally distributed samples with $n = 40$. In terms of size, the proposed MPT successfully controls the type I error rate below α. It is slightly conservative since the critical value is chosen against the worst ρ. The size of Cauchy test and CQ test are slightly inflated. The Median test, Average test and Z-average test are too conservative and their size are very close to 0. The CLX test completely fails to control the type I error rate due to the slow convergence rate of the limiting distribution. As for power analysis, under the CS structure, the MPT outperforms all other tests. Cauchy test is slightly less powerful than the MPT but more powerful than other conservative combination approaches. The power of CLX test and CQ test decreases as the correlation increases since both tests ignore the dependence among variables. In addition, CLX test and CQ test require the largest eigenvalue of Σ is upper bounded by some constant, which is not satisfied in the CS structure. Under the AR structure, note that the CLX test cannot control the size at all under H_0 (can be as large as 0.20). The size

inflation makes the testing power artificially high and hence it is not trustworthy. Excluding the CLX test, the proposed MPT has the best performance, followed by the Cauchy test. Other conservative combination tests are even less powerful than the SPT, which indicates such conservative combination methods do not necessarily improve the testing power.
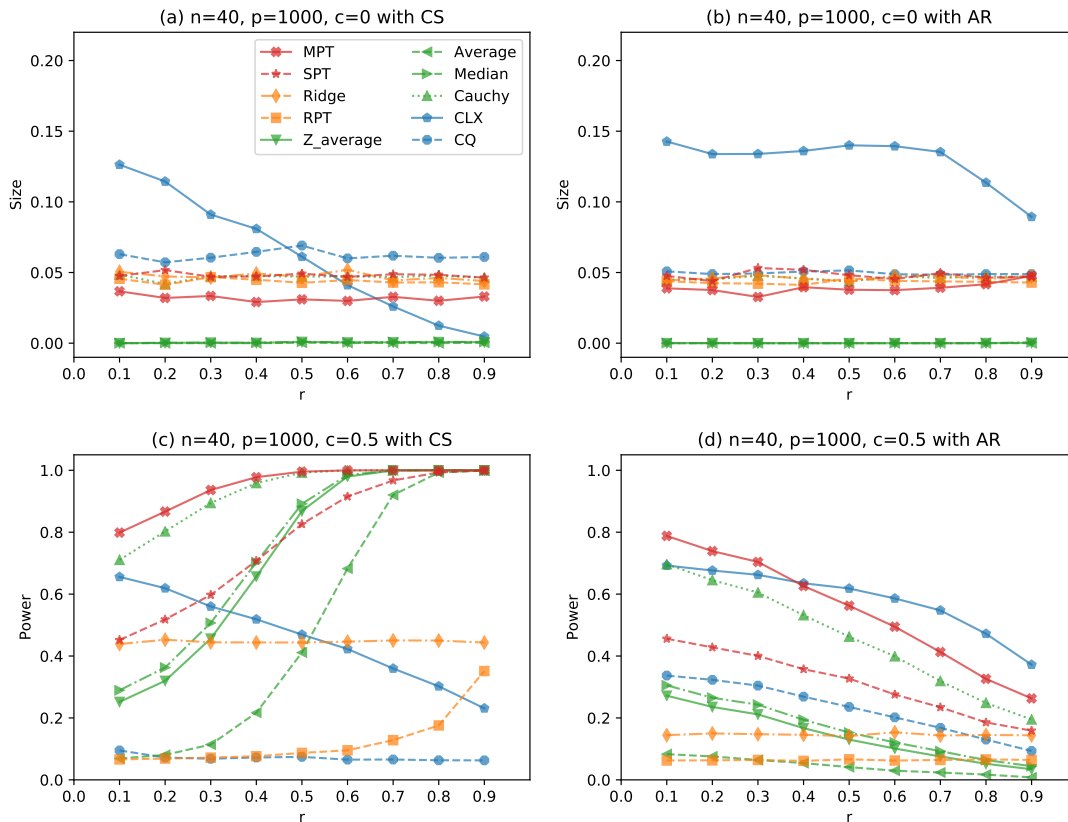

Figure 4: Size and power of different tests for samples following multivariate t_6 distribution with $n = 40$. Panels (a) and (b) show size ($c = 0$) under the null hypothesis for the CS and AR structures, respectively. Panels (c) and (d) show power ($c = 0.5$) under the alternative hypothesis for the CS and AR structures, respectively.

We also examine the finite-sample performance of the proposed MPT when the normality assumption is not satisfied. Figure 4 shows the size and power comparisons of different tests when $\mathbf{x}_i$'s are generated from multivariate t_6 distribution with AR and CS covariance structure. The results show a similar pattern as those in the normal settings, which provide numerical evidences on the robustness of the MPT to non-Gaussianity. When $n = 100$, the patterns of size and power are similar to that of $n = 40$. Due to the limit of space, we relegate the results for $n = 100$ to the appendix, see Figures 5 and 6 in Appendix C.

The numerical results in this subsection emphasize that the MPT greatly improves the testing power upon the SPT thanks to the multiple splits. In a brief summary, our proposed MPT successfully controls the type I error rate and achieves the highest testing power in

comparison with all other state-of-the-arts level α tests. In addition, the studies reveal that the proposed MPT is quite robust to non-Gaussianity.

4.2 Real Data Analysis

We apply the proposed MPT and SPT together with other tests introduced above to a real dataset of high resolution micro-computed tomography (Percival et al., 2014). This dataset contains skull bone densities of $n = 29$ mice with genotype "T0A1" in a genetic mutation study. For each mouse, bone density is measured for 16 different areas of its skull at density levels between 130 - 249. In this empirical analysis, we are interested in comparing the bone density patterns of two areas in the skull, namely "Mandible" and "Nasal". We use all density levels between 130 - 249 for our analysis, and hence dimension $p = 120$. Since the two areas come from the same mouse, we first take the difference of bone density in the two areas at the corresponding density level for each observation. Then we normalize the bone density in the sense that $\frac{1}{29} \sum_{i=1}^{29} X_{ij}^2 = 1$ for all $1 \leqslant j \leqslant 120$. The null hypothesis is the density patterns of two skull areas are the same.

Table 1: Decisions on whether null hypothesis should be rejected or not at significance level $\alpha = 0.05$ based on different tests for the bone density dataset with various signal strengths. The numbers in the parentheses in the p-values if applicable.

δ	1.0	0.8	0.6	0.4	0.3	0.2	0.18
MPT	✔	✔	✔	✔	✔	✔	✔
Cauchy	✔	✔	✔	✔	✔	✔	✔
Median	✔	✔	✔	✔	✔	✘	✘
Average	✔	✔	✔	✔	✘	✘	✘
Z-average	✔	✔	✔	✔	✔	✘	✘
SPT	✔ (10^{-10})	✔ (10^{-9})	✔ (10^{-7})	✔ (10^{-6})	✔ (10^{-4})	✔ (0.042)	✘ (0.246)
Ridge	✔ (10^{-8})	✔ (10^{-7})	✔ (10^{-5})	✔ (0.001)	✔ (0.014)	✘ (0.146)	✘ (0.387)
RPT	✔ (10^{-9})	✔ (10^{-8})	✔ (10^{-6})	✔ (10^{-4})	✔ (0.010)	✘ (0.203)	✘ (0.347)
CQ	✔ (0)	✔ (0)	✔ (0)	✔ (10^{-4})	✘ (0.081)	✘ (0.772)	✘ (0.945)
CLX	✔ (0)	✔ (10^{-14})	✔ (10^{-8})	✔ (0.004)	✘ (0.189)	✘ (0.965)	✘ (0.994)

We apply the proposed MPT and SPT together with other tests to the bone density dataset. The decisions as well as p-values if applicable (in the parentheses) are reported in the first column in Table 1. All tests are able to reject the null hypothesis, implying that the bone density patterns are significantly different. To compare the power of different tests, we further conduct tests as we decrease the signal strength in the bone density difference. To be specific, let $\bar{\mathbf{x}}$ be the sample mean and $\mathbf{r}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ be the residual for the ith subject. Then a new observation $\mathbf{z}_i = \delta \bar{\mathbf{x}} + \mathbf{r}_i$ is constructed for the ith subject for some $\delta \in [0, 1]$. By this construction, a smaller δ leads to a weaker signal strength and would make the test more challenging. Table 1 also reports the decisions and p-values (in the parentheses) for $\delta = 1.0, 0.8, 0.6, 0.4, 0.3, 0.2, 0.18$. When $\delta \geqslant 0.4$, all the tests perform well and reject the null hypothesis at level 0.05. When δ decreases to 0.3, the CQ test, the CLX test and the average based combination test start to fail to reject the null hypothesis. When $\delta = 0.2$, the proposed MPT and SPT and Cauchy combination test are able to reject the null. Further

decreasing δ to 0.18, only the MPT and Cauchy combination test can detect the density difference. This real data example demonstrates that the proposed MPT is more powerful than existing tests and performs well even when the signal is very weak.

5. Discussion

In this work, we study the hypothesis test for one-sample mean vectors in high dimensions. We first study the question of estimating optimal projection direction and provide statistical guarantee. Furthermore, we propose the multiple-splitting projection test, which makes use of the exchangeability of multiple p-values, to mitigate the power loss arising from the single data-splitting procedure. The proposed multiple data-splitting framework can be easily extended to a two-sample problem in which the optimal projection direction is $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ (Huang, 2015). Sharing the same spirit, half of the sample can be used to estimate the projection direction and the remaining half is used to perform the two-sample t-test. Then resultant multiple p-values can be combined similarly to the MPT. As pointed out in Theorem 7, the exchangeability phenomenon generally holds for a permutation framework. This work can be extended to many other statistical inference problems, such as testing the coefficients for a high-dimensional regression model. We hope such insight provides new ideas for researchers from related areas. Another interesting extension is to develop more refined combination methods which better handle the exchangeability. We leave these interesting questions as future work.

Acknowledgments

Appendices

The appendices provide additional materials for the main manuscript. Appendix A provides the tables of critical values for the proposed MPT. Appendix B presents technical lemmas and complete proofs of theoretical results. Appendix C reports additional numerical results of size and power comparisons for $n = 100$ to serve as a complementary to the numerical studies in Section 4.

Appendix A. Tables of critical values

In this section, we present the tables of critical values for the proposed MPT. Follmann and Proschan (2012) derives the critical values of $c(m, \alpha/2)$ and β for tests $M_{\widehat{\rho}_1}$ and $M_{\widehat{\rho}_2}$ at level $\alpha = 0.05$, respectively. We summarize the critical values in Tables 2-3.

Table 2: Critical value $c(m, \alpha/2)$ with respect to m for the test $M_{\widehat{\rho}_1}$ at level $\alpha = 0.05$

Method	Value	number of splits m									
		2	3	4	5	10	20	40	100	1000	10000
$M_{\widehat{\rho}_1}$	$c(m, \alpha/2)$	1.988	2.058	2.133	2.204	2.489	2.865	3.126	4.115	7.17	12.66

Table 3: The smallest value β with respect to m to control the type I error of the test $M_{\widehat{\rho}_2}$ with $c(m, \alpha/2) = z_{\alpha/2}$ at level $\alpha = 0.05$

Method	Value	number of splits m									
		2	3	4	5	10	20	40	100	1000	10000
$M_{\widehat{\rho}_2}$	β	0.25	0.25	0.25	0.25	0.20	0.20	0.15	0.15	0.10	0.05

Appendix B. Lemmas and Proofs

B.1 Technical Lemmas

In this subsection, we introduce a few technical lemmas to help establish theoretical results. Before proceeding, we first introduce some notations for sub-Gaussian and sub-exponential random variables. The sub-Gaussian norm of a sub-Gaussian random variable X is

$$\|X\|_{\psi_2} = \sup_{p \geqslant 1} p^{-\frac{1}{2}} (\mathrm{E}|X|^p)^{1/p}.$$

The sub-exponential norm of a sub-exponential random variable X is

$$\|X\|_{\psi_1} = \sup_{p \geqslant 1} p^{-1} (\mathrm{E}|X|^p)^{1/p}.$$

Lemma 10 *If the RSC condition (3) holds, then*

$$\boldsymbol{\Delta}^\top \mathbf{W} \boldsymbol{\Delta} \geqslant \nu \|\boldsymbol{\Delta}\|_2^2 - \tau \sqrt{\frac{\log p}{n}} \|\boldsymbol{\Delta}\|_1 \ \text{for all } \boldsymbol{\Delta} \in \mathbb{R}^p.$$

Proof For any $\|\boldsymbol{\Delta}\|_1 < 1$, the L_1 norm of $\boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_1$ is 1 and hence satisfies the RSC condition in (3). We have

$$\frac{\boldsymbol{\Delta}^\top}{\|\boldsymbol{\Delta}\|_1} \mathbf{W} \frac{\boldsymbol{\Delta}}{\|\boldsymbol{\Delta}\|_1} \geqslant \nu \frac{\|\boldsymbol{\Delta}\|_2^2}{\|\boldsymbol{\Delta}\|_1^2} - \tau \sqrt{\frac{\log p}{n}} \frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_1}$$

$$\frac{\boldsymbol{\Delta}^\top}{\|\boldsymbol{\Delta}\|_1} \mathbf{W} \frac{\boldsymbol{\Delta}}{\|\boldsymbol{\Delta}\|_1} \geqslant \nu \frac{\|\boldsymbol{\Delta}\|_2^2}{\|\boldsymbol{\Delta}\|_1^2} - \tau \sqrt{\frac{\log p}{n}} \frac{\|\boldsymbol{\Delta}\|_1^2}{\|\boldsymbol{\Delta}\|_1^2}$$

$$\boldsymbol{\Delta}^\top \mathbf{W} \boldsymbol{\Delta} \geqslant \nu \|\boldsymbol{\Delta}\|_2^2 - \tau \sqrt{\frac{\log p}{n}} \|\boldsymbol{\Delta}\|_1^2.$$

17

Since $\|\boldsymbol{\Delta}\|_1 < 1$, then $\|\boldsymbol{\Delta}\|_1^2 \leqslant \|\boldsymbol{\Delta}\|_1$, implying

$$\boldsymbol{\Delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Delta} \geqslant \nu \|\boldsymbol{\Delta}\|_2^2 - \tau \sqrt{\frac{\log p}{n}} \|\boldsymbol{\Delta}\|_1.$$

The proof of Lemma 10 is complete. ∎

Lemma 11 *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are independent and identically distributed sub-Gaussian random vectors. Let $\bar{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})_{p \times p}$ be the sample mean and sample covariance matrix. If $\log p < n$, then with probability at least $1 - 2p^{-1}$, we have*

(i) $\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty \leqslant C\sqrt{\log p/n}$ for some large C.

(ii) $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \leqslant C\sqrt{\log p/n}$ for some large C.

Proof Let $\bar{\mathbf{x}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$ be the sample mean and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top$ be the sample covariance matrix. Without loss of generality, we assume $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$ and the sub-Gaussian parameter for $\mathbf{x}_i$ is σ^2. Write $\mathbf{x}_k = (\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kp})^\top$ and each X_{kj} is a sub-Gaussian random variable with parameter $\boldsymbol{\sigma}^2$ and let $K = \max_{1 \leqslant j \leqslant p} \|\mathbf{x}_{kj}\|_{\psi_2}$. Obviously, $\bar{\mathbf{x}}$ is also sub-Gaussian random vector with parameter $\sigma^2/\sqrt{n}$. For any $t > 0$, we have

$$\Pr(\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty > t) \leqslant 2p \exp\left\{-cnt^2/K^2\right\}.$$

Take $t = C\sqrt{\log p/n}$ for some large $C > 0$, we have

$$\Pr(\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty < C\sqrt{\log p/n}) \geqslant 1 - 2p^{-1}. \tag{11}$$

The sample covariance $\widehat{\boldsymbol{\Sigma}}$ can be decomposed as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top.$$

Hence we know,

$$\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \leqslant \max_{i,j} |\frac{1}{n} \sum_{k=1}^n \mathbf{x}_{ki}\mathbf{x}_{kj} - \sigma_{ij}| + \max_{i,j} |\bar{\mathbf{x}}_i \bar{\mathbf{x}}_j|.$$

In addition, we have

$$\|\mathbf{x}_{ki}\mathbf{x}_{kj}\|_{\psi_1} \leqslant 2\|\mathbf{x}_{ki}\|_{\psi_2}\|\mathbf{x}_{kj}\|_{\psi_2} \leqslant 2K^2.$$

Hence $\|\mathbf{x}_{ki}\mathbf{x}_{kj} - \sigma_{ij}\|_{\psi_1} \leqslant 4K^2$. According to the inequality of tail probability for sub-exponential variables, we have

$$\Pr\left(\max_{i,j} \left|\frac{1}{n} \sum_{k=1}^n \mathbf{x}_{ki}\mathbf{x}_{kj} - \sigma_{ij}\right| > t\right) \leqslant \max\left(2p^2 \exp\left\{-cn\frac{t^2}{16K^4}\right\}, 2p^2 \exp\left\{-cn\frac{t}{4K^2}\right\}\right).$$

It is easy to verify that $\|\bar{\mathbf{x}}_i\|_{\psi_2} \leqslant K/\sqrt{1/n}$ and $\|\bar{\mathbf{x}}_i\bar{\mathbf{x}}_j\|_{\psi_1} \leqslant 2\|\bar{\mathbf{x}}_i\|_{\psi_2}\|\bar{\mathbf{x}}_j\|_{\psi_2} \leqslant 2K^2/n$, we have

$$\Pr(\max_{i,j} |\bar{\mathbf{x}}_i\bar{\mathbf{x}}_j| > t) \leqslant 2p^2 \exp\left\{-\frac{cnt}{2K^2}\right\}.$$

By the choice of $t = \frac{C}{2}\sqrt{\frac{\log p}{n}}$ for some large $C > 0$, we have $\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}| \leqslant C\sqrt{\log p/n}$ with probability at least $1 - 2p^{-1}$, which completes the proof. ∎

Lemma 12 (Loh and Wainwright (2015)) *Assume penalty function $P_\lambda(t)$ satisfies conditions (i)-(v), then*

(a) $|P_\lambda(t_1) - P_\lambda(t_2)| \leqslant \lambda|t_1 - t_2|$ *for any $t_1, t_2 \in \mathbb{R}$.*

(b) *For any $\mathbf{w} \in \mathbb{R}^p$, we have $\lambda\|\mathbf{w}\|_1 \leqslant P_\lambda(\mathbf{w}) + \frac{\nu}{2}\|\mathbf{w}\|_2^2$.*

(c) *Suppose $\|\mathbf{w}^\star\|_0 = s > 0$, then for any $\mathbf{w} \in \mathbb{R}^p$ such that $cP_\lambda(\mathbf{w}^\star) - P_\lambda(\mathbf{w}) \geqslant 0$ with $c \geqslant 1$, we have $cP_\lambda(\mathbf{w}^\star) - P_\lambda(\mathbf{w}) \leqslant \lambda(c\|\boldsymbol{\delta}_\mathcal{A}\|_1 - \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1)$, where $\boldsymbol{\delta} = \mathbf{w} - \mathbf{w}^\star$ and $\mathcal{A}$ is the index set of the s largest elements of $\boldsymbol{\delta}$ in magnitude.*

(d) *Define $J_\lambda(t) = \lambda|t| - P_\lambda(t)$. Then the function $J_\lambda(t) - \frac{\mu}{2}t^2 = \lambda|t| - P_\lambda(t) - \frac{\mu}{2}t^2$ is concave and differentiable.*

B.2 Proof of Theorem 2

Lemma 10 shows that the RSC condition in (3) actually holds for all $\boldsymbol{\Delta} \in \mathbb{R}^p$. Now we are ready to prove Theorem 2. Define $\mathbf{w}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\hat{\boldsymbol{\Delta}} = \hat{\mathbf{w}} - \mathbf{w}^\star$. The first order necessary condition (4) implies that

$$\hat{\boldsymbol{\Delta}}^\top \hat{\boldsymbol{\Sigma}}\hat{\mathbf{w}} + \langle \nabla P_\lambda(\hat{\mathbf{w}}) - \bar{\mathbf{x}}, \hat{\boldsymbol{\Delta}}\rangle = 0. \tag{12}$$

By the RSC condition (3), we have

$$\hat{\boldsymbol{\Delta}}^\top \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Delta}} \geqslant \nu\|\hat{\boldsymbol{\Delta}}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\hat{\boldsymbol{\Delta}}\|_1. \tag{13}$$

Adding (12) to (13), we have

$$-\hat{\boldsymbol{\Delta}}^\top \hat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \langle \nabla P_\lambda(\hat{\mathbf{w}}) - \bar{\mathbf{x}}, \hat{\boldsymbol{\Delta}}\rangle \geqslant \nu\|\hat{\boldsymbol{\Delta}}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\hat{\boldsymbol{\Delta}}\|_1. \tag{14}$$

Lemma 12 shows that $P_{\lambda,\gamma}(\mathbf{w}) = P_\lambda(\mathbf{w}) + \frac{\gamma}{2}\|\mathbf{w}\|_2^2$ is a convex function, hence

$$P_{\lambda,\gamma}(\mathbf{w}^\star) - P_{\lambda,\gamma}(\hat{\mathbf{w}}) \geqslant \langle \nabla P_\lambda(\hat{\mathbf{w}}) + \gamma\hat{\mathbf{w}}, \mathbf{w}^\star - \hat{\mathbf{w}}\rangle,$$

which implies

$$-\langle \nabla P_\lambda(\hat{\mathbf{w}}), \hat{\boldsymbol{\Delta}}\rangle \leqslant P_\lambda(\mathbf{w}^\star) - P_\lambda(\hat{\mathbf{w}}) + \frac{\gamma}{2}\|\hat{\boldsymbol{\Delta}}\|_2^2. \tag{15}$$

Combining (14) and (15),

$$\nu\|\widehat{\boldsymbol{\Delta}}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\Delta}}\|_1 \leqslant -\widehat{\boldsymbol{\Delta}}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star + \bar{\mathbf{x}}^\top\widehat{\boldsymbol{\Delta}} + P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) + \frac{\gamma}{2}\|\widehat{\boldsymbol{\Delta}}\|_2^2$$

$$(\nu - \gamma/2)\|\widehat{\boldsymbol{\Delta}}\|_2^2 \leqslant P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) + \|\tilde{\boldsymbol{\Sigma}}\mathbf{w}^\star - \bar{\mathbf{x}}\|_\infty\|\widehat{\boldsymbol{\Delta}}\|_1 + \tau\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\Delta}}\|_1$$

$$(\nu - \gamma/2)\|\widehat{\boldsymbol{\Delta}}\|_2^2 \leqslant P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) + \left(\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \bar{\mathbf{x}}\|_\infty + \tau\sqrt{\frac{\log p}{n}}\right)\|\widehat{\boldsymbol{\Delta}}\|_1.$$

By triangle inequality, we know $\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \bar{\mathbf{x}}\|_\infty \leqslant \|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \boldsymbol{\mu}\|_\infty + \|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty$. According to Lemma 11, there exists $M_1, M_2 > 0$ such that

$$\Pr(\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty \leqslant M_1\sqrt{\log p/n}) \geqslant 1 - 2p^{-1}. \tag{16}$$

$$\Pr(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \leqslant M_2\sqrt{\log p/n}) \geqslant 1 - 2p^{-1}.$$

Then with probability at least $1 - 2p^{-1}$,

$$\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \boldsymbol{\mu}\|_\infty = \|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \boldsymbol{\Sigma}\mathbf{w}^\star\|_\infty \leqslant \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty\|\mathbf{w}^\star\|_1 \leqslant M_2 C_1\sqrt{\log p/n}. \tag{17}$$

Combining (16) and (17), we know that with probability at least $1 - 4p^{-1}$, we have

$$\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \bar{\mathbf{x}}\|_\infty \leqslant M'\sqrt{\log p/n},$$

with $M' = M_1 + M_2 C_1$. Take $\lambda = M\sqrt{\log p/n}$ with $M = 4\max\{M', \tau\}$, we have

$$\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}^\star - \bar{\mathbf{x}}\|_\infty + \tau\sqrt{\log p/n} \leqslant \lambda/2.$$

Hence

$$(\nu - \gamma/2)\|\widehat{\Delta}\|_2^2 \leqslant P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) + \frac{\lambda}{2}\|\widehat{\Delta}\|_1$$

$$\leqslant P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) + \frac{1}{2}P_\lambda(\widehat{\Delta}) + \frac{\gamma}{4}\|\widehat{\Delta}\|_2^2,$$

where the second inequality is because $\frac{\lambda}{2}\|\widehat{\Delta}\|_1 \leqslant \frac{1}{2}P_\lambda(\widehat{\Delta}) + \frac{\gamma}{4}\|\widehat{\Delta}\|_2^2$ by Lemma 12(b). By the subadditivity of P_λ, we have $P_\lambda(\widehat{\Delta}) = P_\lambda(\widehat{\mathbf{w}} - \mathbf{w}^\star) \leqslant P_\lambda(\widehat{\mathbf{w}}) + P_\lambda(\mathbf{w}^\star)$. Then

$$(\nu - \gamma/2)\|\widehat{\Delta}\|_2^2 \leqslant P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) + \frac{1}{2}P_\lambda(\widehat{\mathbf{w}}) + \frac{1}{2}P_\lambda(\mathbf{w}^\star) + \frac{\gamma}{4}\|\widehat{\Delta}\|_2^2$$

$$(\nu - 3\gamma/4)\|\widehat{\Delta}\|_2^2 \leqslant \frac{3}{2}P_\lambda(\mathbf{w}^\star) - \frac{1}{2}P_\lambda(\widehat{\mathbf{w}})$$

$$(2\nu - 3\gamma/2)\|\widehat{\Delta}\|_2^2 \leqslant 3P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}).$$

By Lemma 12(c), we have $3\lambda\|\widehat{\Delta}_{\mathcal{I}}\|_1 - \lambda\|\widehat{\Delta}_{\mathcal{I}^c}\|_1 \geqslant 3P_\lambda(\mathbf{w}^\star) - P_\lambda(\widehat{\mathbf{w}}) \geqslant 0$, where $\mathcal{I}$ denotes the index set of the s largest elements of $\widehat{\Delta}$ in magnitude. Since $\nu \geqslant 3\gamma/4$, we have

$$0 \leqslant (2\nu - 3\gamma/2)\|\widehat{\Delta}\|_2^2 \leqslant 3\lambda\|\widehat{\Delta}_{\mathcal{I}}\|_1 - \lambda\|\widehat{\Delta}_{\mathcal{I}^c}\|_1.$$

As a result, we have $\|\widehat{\Delta}_{\mathcal{I}^c}\|_1 \leqslant 3\|\widehat{\Delta}_{\mathcal{I}}\|_1$ and

$$\left(2\nu - \frac{3}{2}\gamma\right)\|\widehat{\Delta}\|_2^2 \leqslant 3\lambda\|\widehat{\Delta}_{\mathcal{I}}\|_1 - \lambda\|\widehat{\Delta}_{\mathcal{I}^c}\|_1 \leqslant 3\lambda\|\widehat{\Delta}_{\mathcal{I}}\|_1 \leqslant 3\lambda\sqrt{s}\|\widehat{\Delta}_{\mathcal{I}}\|_2,$$

from which we conclude that

$$\|\widehat{\Delta}\|_2 \leqslant \frac{6\lambda\sqrt{s}}{4\nu - 3\gamma} = O\left(\sqrt{\frac{s\log p}{n}}\right).$$

The ℓ_1 norm bound follows immediately from the ℓ_2 norm bound

$$\|\widehat{\Delta}\|_1 = \|\widehat{\Delta}_{\mathcal{I}}\|_1 + \|\widehat{\Delta}_{\mathcal{I}^c}\|_1 \leqslant 4\|\widehat{\Delta}_{\mathcal{I}}\|_1 \leqslant 4\sqrt{s}\|\widehat{\Delta}_{\mathcal{I}}\|_2 \leqslant \frac{24\lambda s}{4\nu - 3\gamma} = O\left(s\sqrt{\frac{\log p}{n}}\right).$$

B.3 Proof of Theorem 4

According to Theorem 2, we know $\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 = O(s\sqrt{\log p/n_1}) = o(1)$ with high probability. Let $\bar{\mathbf{x}}_2$ and $\widehat{\mathbf{\Sigma}}_2$ be the sample mean and sample covariance matrix based on $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n\}$. On one hand,

$$\begin{aligned}
|\widehat{\mathbf{w}}^\top \mathbf{\Sigma}\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^\top \mathbf{\Sigma}\mathbf{w}^\star| &= |\widehat{\mathbf{w}}^\top \mathbf{\Sigma}(\widehat{\mathbf{w}} - \mathbf{w}^\star)| \\
&\leqslant \|\mathbf{\Sigma}\|_{\max}\|\widehat{\mathbf{w}}\|_1\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 \\
&\leqslant \|\mathbf{\Sigma}\|_{\max}(\|\mathbf{w}^\star\|_1 + \|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1)\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 \\
&= O\left(s\sqrt{\log p/n_1}\right).
\end{aligned}$$

One the other hand,

$$\begin{aligned}
|\widehat{\mathbf{w}}^\top \mathbf{\Sigma}\mathbf{w}^\star - \mathbf{w}^{\star\top} \mathbf{\Sigma}\mathbf{w}^\star| &= |(\widehat{\mathbf{w}} - \mathbf{w}^\star)^\top \mathbf{\Sigma}\mathbf{w}^\star| \\
&\leqslant \|\mathbf{\Sigma}\|_{\max}\|\mathbf{w}^\star\|_1\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 \\
&= O\left(s\sqrt{\log p/n_1}\right).
\end{aligned}$$

Hence by triangle inequality,

$$\begin{aligned}
|\widehat{\mathbf{w}}^\top \mathbf{\Sigma}\widehat{\mathbf{w}} - \mathbf{w}^{\star\top} \mathbf{\Sigma}\mathbf{w}^\star| &\leqslant |\widehat{\mathbf{w}}^\top \mathbf{\Sigma}\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^\top \mathbf{\Sigma}\mathbf{w}^\star| + |\widehat{\mathbf{w}}^\top \mathbf{\Sigma}\mathbf{w}^\star - \mathbf{w}^{\star\top} \mathbf{\Sigma}\mathbf{w}^\star| \\
&= O\left(s\sqrt{\log p/n_1}\right).
\end{aligned} \tag{18}$$

Given $\widehat{\mathbf{w}}$, we know that $y_{n_1+1}, \ldots, y_n$ are i.i.d. random variables with mean $\boldsymbol{\mu}^\top\widehat{\mathbf{w}}$ and variance $\widehat{\mathbf{w}}^\top\mathbf{\Sigma}\widehat{\mathbf{w}}$. By central limit theorem and $\widehat{\mathbf{w}}^\top\mathbf{\Sigma}\widehat{\mathbf{w}} - \mathbf{w}^{\star\top}\mathbf{\Sigma}\mathbf{w}^\star = o(1)$, we know

$$\frac{\sqrt{n_2}(\bar{y} - \boldsymbol{\mu}^\top\widehat{\mathbf{w}})}{\sqrt{\mathbf{w}^{\star\top}\mathbf{\Sigma}\mathbf{w}^\star}} \xrightarrow{d} N(0,1).$$

The test statistic of the SPT is $T_{\widehat{\mathbf{w}}} = \sqrt{n_2}\bar{y}/s_y$ and we reject H_0 whenever $|T_{\widehat{\mathbf{w}}}| > z_{\alpha/2}$. The power function for the SPT is

$$
\Pr\left(\left|\frac{\sqrt{n_2}\bar{y}}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}}\right| > z_{\alpha/2}\right) = \Pr\left(\frac{\sqrt{n_2}\bar{y}}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}} > z_{\alpha/2}\right) + \Pr\left(\frac{\sqrt{n_2}\bar{y}}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}} < -z_{\alpha/2}\right)
$$

$$
= \Pr\left(\frac{\sqrt{n_2}(\bar{y} - \boldsymbol{\mu}^\top\widehat{\mathbf{w}})}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}} > z_{\alpha/2} + \frac{\sqrt{n_2}\boldsymbol{\mu}^\top\widehat{\mathbf{w}}}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}}\right) +
$$

$$
\Pr\left(\frac{\sqrt{n_2}(\bar{y} - \boldsymbol{\mu}^\top\widehat{\mathbf{w}})}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}} < -z_{\alpha/2} + \frac{\sqrt{n_2}\boldsymbol{\mu}^\top\widehat{\mathbf{w}}}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}}\right)
$$

$$
\simeq \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n_2}\boldsymbol{\mu}^\top\widehat{\mathbf{w}}}{\sqrt{\mathbf{w}^{\star\top}\boldsymbol{\Sigma}\mathbf{w}^\star}}\right) + o(1)
$$

$$
\simeq \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n_2}\boldsymbol{\mu}^\top\widehat{\mathbf{w}}}{\sqrt{\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}\right).
$$

Notice that

$$
\boldsymbol{\mu}^\top\widehat{\mathbf{w}} - \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}^\top\widehat{\mathbf{w}} - \boldsymbol{\mu}^\top\mathbf{w}^\star = \boldsymbol{\mu}^\top(\widehat{\mathbf{w}} - \mathbf{w}^\star)
$$

$$
\leqslant \|\boldsymbol{\mu}\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_1 = O\left(\|\boldsymbol{\mu}\|_\infty s\sqrt{\frac{\log p}{n_1}}\right) \to 0.
$$

As a result, we know the asymptotic power is

$$
\beta(T_{\widehat{\mathbf{w}}}) = \Phi\left(-z_{\alpha/2} + \sqrt{n_2\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right).
$$

B.4 Proof of Theorems 6 and 7

Theorem 6 is a direct corollary of Theorem 7 by setting $T_k = p_k$, we only prove Theorem 7 here. Conditioning on the observed data $\mathcal{D}$, we know its random permutations $\mathcal{D}^{\pi_1}, \mathcal{D}^{\pi_2}, \ldots$ are independent from each other. Therefore, the resultant statistics $T_k = g(\mathcal{D}^{\pi_k})$ are independent and identically distributed conditioning on $\mathcal{D}$. By the de Finete theorem (Aldous, 1985) which states that a mixture of independent and identically distributed sequences are exchangeable, we know is $(T_1, T_2, \dots)$ is an exchangeable sequence, and hence $(T_1, \ldots, T_m)$ is exchangeable for any finite m.

B.5 Proof of Theorem 8

According to the central limit theorem for exchangeable random variables, we have

$$
M_\rho = \bar{Z}/\sqrt{(1 + (m-1)\rho)/m} \xrightarrow{d} N(0,1).
$$

Let $\widehat{\rho}$ be a consistent estimator for $\rho \neq 0$, i.e., $\widehat{\rho} \xrightarrow{p} \rho$. Hence,

$$
\frac{\sqrt{(1 + (m-1)\rho)/m}}{\sqrt{(1 + (m-1)\widehat{\rho})/m}} \xrightarrow{p} 1.
$$

As a result, we know

$$
M_{\widehat{\rho}} = \bar{Z}/\sqrt{(1 + (m-1)\widehat{\rho})/m} \xrightarrow{d} N(0,1).
$$

Appendix C. Additional Numerical Results

Figures 5 and 6 reports the size and power of different tests with $n = 100$ for samples following multivariate normal distribution and t_6 distribution, respectively. The pattern of size and power for different tests is similar to that of $n = 40$. The proposed MPT can control the type I error rate below the pre-specified significance level $\alpha = 0.05$ while the CLX test completely fails to control the size. Among those tests which can retain the type I error rate, the proposed MPT is the most powerful one for both CS and AR covariance structure. The studies also reveal that the proposed MPT is quite robust to non-Gaussianity.
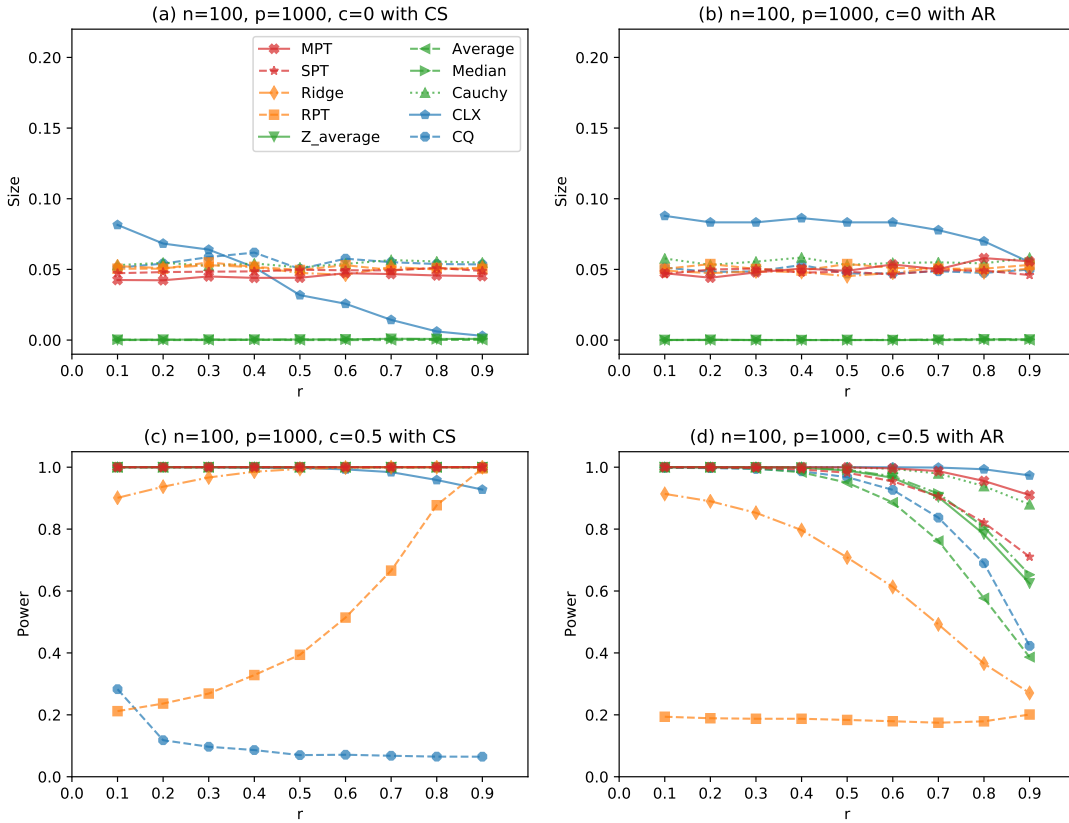

Figure 5: Size and power of different tests for normally distributed samples with $n = 100$. Panels (a) and (b) show size ($c = 0$) under the null hypothesis for the CS and AR structure, respectively. Panels (c) and (d) show power ($c = 0.5$) under the alternative hypothesis for the CS and AR structure, respectively.
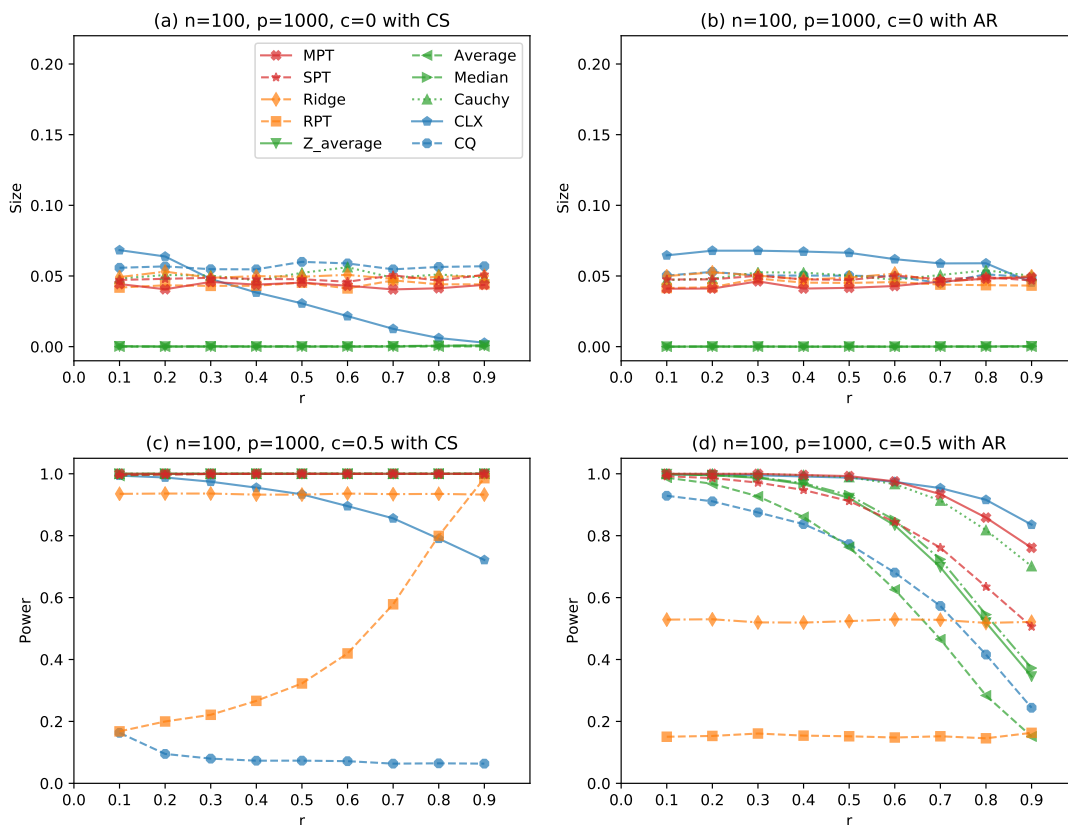
Figure 6: Size and power of different tests for samples following multivariate t_6 distribution with $n = 100$. Panels (a) and (b) show size ($c = 0$) under the null hypothesis for the CS and AR structure, respectively. Panels (c) and (d) show power ($c = 0.5$) under the alternative hypothesis for the CS and AR structure, respectively.

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40 (5):2452–2482, 2012.

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII-1983*, pages 1–198. Springer, 1985.

Zhidong Bai and Hewa Saranadasa. Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, 6:311–329, 1996.

Rina Foygel Barber and Emmanuel J Candès. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

T Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372, 2014.

Lin S Chen, Debashis Paul, Ross L Prentice, and Pei Wang. A regularized hotelling's T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, 106(496):1345–1360, 2011.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 2014.

Dean Follmann and Michael Proschan. A test of location for exchangeable multivariate normal data with unknown correlation. *Journal of Multivariate Analysis*, 104(1):115–125, 2012.

Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750, 2017.

Yuan Huang. Projection test for high-dimensional mean vectors with optimal direction. *Department of Statistics, The Pennsylvania State University at University Park*, 2015.

Michael Klass and Henry Teicher. The central limit theorem for exchangeable random variables without moments. *The Annals of Probability*, 15(1):138–153, 1987.

Jurgen Lauter. Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics*, 52(3):964–970, 1996.

Changcheng Li and Runze Li. Linear hypothesis testing in linear models with high-dimensional responses. *Journal of the American Statistical Association*, forthcoming: 1–13, 2021.

Wanjun Liu and Runze Li. Projection test with sparse optimal direction for high-dimensional one sample mean problem. In *Contemporary Experimental Design, Multivariate Analysis and Data Mining*, pages 295–309. Springer, 2020.

Yaowu Liu and Jun Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(1):559–616, 2015.

Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125, 2013.

Christopher J Percival, Yuan Huang, Ethylin Wang Jabs, Runze Li, and Joan T Richtsmeier. Embryonic craniofacial bone volume and bone mineral density in Fgfr2+/P253R and nonmutant mice. *Developmental Dynamics*, 243(4):541–551, 2014.

Joseph P Romano and Cyrus DiCiccio. Multiple data splitting for testing. Technical report, 2019.

Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41(5):2505, 2013.

Lan Wang, Bo Peng, and Runze Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669, 2015.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164–2201, 2014.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201, 2009.

Xiufan Yu, Jiawei Yao, and Lingzhou Xue. Innovated power enhancement for testing multifactor asset pricing models. *Available at SSRN 3809369*, 2019.

Xiufan Yu, Danning Li, and Lingzhou Xue. Fisher's combined probability test for high-dimensional covariance matrices. *arXiv preprint arXiv:2006.00426*, 2020.

Xiufan Yu, Danning Li, Lingzhou Xue, and Runze Li. Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing. *Journal of the American Statistical Association*, forthcoming:1–39, 2022.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Ping-Shou Zhong, Song Xi Chen, and Minya Xu. Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *The Annals of Statistics*, 41(6):2820–2851, 2013.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.