Choi, Taylor and Tibshirani (2017, AoS)

Prelim. for Matsumura & Tachibana

Yasuyuki Matsumura (Kyoto University)

Last Updated: October 3, 2025

yasu0704xx.github.io

The Annals of Statistics
2017, Vol. 45, No. 6, 2590–2617
https://doi.org/10.1214/16-AOS1536

SELECTING THE NUMBER OF PRINCIPAL COMPONENTS: ESTIMATION OF THE TRUE RANK OF A NOISY MATRIX

BY YUNJIN CHOI*, JONATHAN TAYLOR†,1 AND ROBERT TIBSHIRANI†,2

*National University of Singapore** *and Stanford University*†

Principal component analysis (PCA) is a well-known tool in multivariate statistics. One significant challenge in using PCA is the choice of the number of principal components. In order to address this challenge, we propose distribution-based methods with *exact* type 1 error controls for hypothesis testing and construction of confidence intervals for signals in a noisy matrix with finite samples. Assuming Gaussian noise, we derive exact type 1 error controls based on the conditional distribution of the singular values of a Gaussian matrix by utilizing a post-selection inference framework, and extending the approach of [Taylor, Loftus and Tibshirani (2013)] in a PCA setting. In simulation studies, we find that our proposed methods compare well to existing approaches.

Choi, Taylor, and Tibshirani (2017, AoS)

- In using principal component analysis (PCA), the choice of the number of principal components is a significant challenge.
- Choi, Taylor and Tibshirani (2017) propose distribution-based methods with exact type 1 error controls for hypothesis testing and construction of confidence intervals for signals in a noisy matrix with finite samples.
- Assuming Gaussian noise, Choi, Taylor and Tibshirani (2017) derive exact type 1 error controls based on the conditional distribution of the singular values of a Gaussian matrix by utilizing a post-selection inference framework, and extending the approach of Taylor, Loftus and Tibshirani (2016) in a PCA setting.

Contents

Introduction

Principal Component Analysis (PCA)

- Principal Component Analysis (PCA) is a commonly used method in multivariate statistics.
  - a descriptive tool for examining the structure of a data matrix
  - a pre-processing step for reducing the dimension of the column space of the matrix (Josse and Husson, 2012)
  - matrix completion (Cai, Candés and Shen, 2010)
- One important challenge: How to determine the number of principal components to retain in PCA?
- A summary by Jolliffe (2002)
  1. subjective methods (e.g., scree plot)
  2. distribution-based test tools (e.g., Bartlett's test)
  3. computational procedures (e.g., cross-validation)

Example (Mardia, Kent and Bibby, 1979)

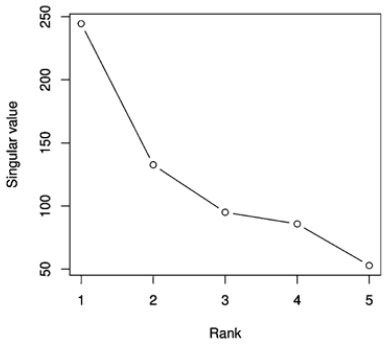


FIG. 1. *Singular values of the score data in decreasing order. The data consist of exam scores of* 88 *students on five different topics* (*Mechanics*, *Vectors*, *Algebra*, *Analysis and Statistics*).

This Paper

- Choi, Taylor and Tibshirani (2017) propose a class of statistical methods utilizing hypothesis testing framework for determining the rank of the signal matrix in a noisy matrix model.
- The estimated rank here corresponds to the number of principal to retain in PCA.
- Under Gaussian assumption, the proposed hypothesis testing method provides exact type 1 error controls along with exact confidence intervals of signal parameters in finite samples.

Related Works (Polyhedral Approach)

- Tibshirani, Taylor, Lockhart and Tibshirani (2017)
  - the truncated Gaussian test conditioning on the event of selecting active variables
  - utilizing the fact that the selection event can be characterized by an observed data vector y falling into a polyhedral set
- In PCA, the event of selecting principal components cannot be characterized by an observed data vector y falling into a polyhedral set.
  - Selecting a variable is a discrete event of forward stepwose regression.
  - On the other hand, a principal components in PCA are chosen from a continuum for Y being a matrix.
  - As the domain of selection event is a continuum, the resulting null distribution conditional on the selection event is defined on a measure zero domain rather than being a truncated Gaussian distribution.

Kac-Rice Test

- The *Kac-Rice test*: an exact method for testing and constructing confidence intervals for signals under a global null hypothesis in adaptive regression.
- Under the global null scenario, one of the proposed methods corresponds to the application of the *Kac-Rice test* to a penalized regression minimizing the Frobenius norm with a nuclear norm penalty.
- Choi, Taylor and Tibshirani (2017) extend the *Kac-Rice test* and the construction of confidence intervals to not only to the global null scenario but the general case.
- Also in the global null scenario, one of the extended methods provides stronger power than the *Kac-Rice test*.
- The exact property of the *Kac-Rice test* is preserved in extension to a general step by incorporating a post-selection inference framework.
- The resulting statistics use a conditional survival function of the eigenvalues of a Wishart matrix.

Literature & Contribution

- Inference based on the distribution of eigenvalues
  - Muirhead (1982, Theorem 9.6.2) : a likelihood ratio test with the asymptotic Chi-square distribution
  - Kritchman and Nadler (2008) : the asymptotic distribution of the largest eigenvalue of a Wishart matrix (the Tracy-Widom law), incorporating the result of Johnstone (2001)
- These test methods show conservative results and thus lose signal detection power in the general stage.
- Choi, Taylor and Tibshirani (2017) provide exact type 1 error controls and decent detection power at the same time, and additionally provide a method for constructing confidence intervals in addition to hypothesis testing.

Proposed Distribution-Based Methods

Setup

- Assume that the observed data matrix $Y \in \mathbb{R}^{N \times p}$ is the sum of a low-rank signal matrix $B \in \mathbb{R}^{N \times p}$ and a Gaussian noise matrix $E \in \mathbb{R}^{N \times p}$ as follows:

$$
\begin{aligned}
&Y = B + E, \\
&\mathsf{rank}(B) = \kappa < \mathsf{min}(N, p), \\
&E_{ij} \sim \mathsf{Normal}(0, \sigma^2) \text{ for } i \in \{1, \cdots, N\}, j \in \{1, \cdots, p\}
\end{aligned}
$$

  that is,

$$
Y \sim \mathsf{Normal}(B, \sigma^2 I_N \otimes I_p). \tag{1}
$$

- Assume $N > p$ w.l.o.g.

- Following the notation from Muirhead (1982, p.73), $Y \sim \text{Normal}(B, \sigma^2 I_N \otimes I_p)$ in (1) denotes

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim \text{Normal} \left(\begin{pmatrix} B_1 \\ \vdots \\ B_N \end{pmatrix}, \begin{pmatrix} \sigma^2 I_p & 0 & \cdots & 0 \\ 0 & \sigma^2 I_p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 I_p \end{pmatrix} \right),$$

  where Y_j and B_j represent the jth row vectors of Y and B, respectively, and the Kronecker product is denoted by $\otimes$.

- We focus on finding κ, the rank of signal matrix B, and the construction of confidence intervals for the signals in B.

For the Case of a Centered Data Matrix

- With a centered data matrix, the proposed approaches with this model assumption are valid for the popular spiked covariance model as well.
- Both traditional spiked covariance approaches and the proposed tests with a centered data matrix are examining basically the same statistics.
  - Note that the centering operation of a data matrix does not change the rank of B.
- The statistics in interest for the spiked covariance model approaches are the eigenvalues of a covariance matrix, and the proposed methods investigate the singular values of a data matrix.
  - The eigenvalues from the covariance matrix are the same as the squared singular values of a centered data matrix.

Taylor, Loftus and Tibshirani (2016)

- We first review the global null test and confidence interval construction of the first signal of Taylor, Loftus and Tibshirani (2016) and its application in matrix denoising problem, which corresponds to the PCA setting.
- Then we extend the global null test to a general test procedure for testing

$$\mathbb{H}_{k,0} : \text{rank}(B) \leq k-1 \text{ vs } \mathbb{H}_{k,1} : \text{rank}(B) \geq k$$
$$\text{for } k = 1, \cdots, p-1,$$

and describe how to construct confidence intervals for the kth largest signal parameters.

Review of the Kac-Rice Test: Global Null Hypothesis Testing

- The *Kac-Rice test* provides exact type 1 error controls for a class of regularized regression problems of the following form:

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2}||y - X\beta||_2^2 + \lambda \cdot \mathcal{P}(\beta) \tag{2}$$

  with an outcome $y \in \mathbb{R}^p$, a predictor matrix $X \in \mathbb{R}^{N \times p}$ and a penalty term $\mathcal{P}(\cdot)$ with a regularization parameter $\lambda \geq 0$.
- Assuming that the outcome $y \in \mathbb{R}^N$ is generated from $y \sim \text{Normal}(X\beta, \Sigma)$, the *Kac-Rice test* provides a method for testing

$$\mathbb{H}_0 : \mathcal{P}(\beta) = 0 \tag{3}$$

  that yields exact type 1 error controls under the assumption that the penalty function $\mathcal{P}$ is a support function of a convex set $\mathcal{C} \subseteq \mathbb{R}^p$, that is,

$$\mathcal{P}(\beta) = \max_{u \in \mathcal{C}} u^\top \beta.$$

- When applied to a matrix denoising problem of a popular form, (3) becomes a global null hypothesis:

$$\mathbb{H}_0 : \Lambda_1 = 0 \equiv \mathsf{rank}(B) = 0 \equiv B = 0_{N \times p}, \tag{4}$$

where $\Lambda_1 \geq \cdots \geq \Lambda_p \geq 0$ denote the singular values (SV) of B.

Details

- For an observed data matrix $Y \in \mathbb{R}^{N \times p}$, a widely used method to recover the signal matrix B in (1) is to solve the following criterion:

$$\widehat{B} \in \underset{B \in \mathbb{R}^{N \times p}}{\arg\min} \frac{1}{2}||Y - B||_F^2 + \lambda||B||_* \qquad \text{where } \lambda > 0, \tag{5}$$

where $||\cdot||_F$ denotes a Frobenius norm, and $||\cdot||_*$ denotes a nuclear norm. [1]

[1] The nuclear norm term plays an analogous role as an ℓ_1 penalty term in lasso regression (Tibshirani, 1996).

- The objective function (5) falls into the class of regression problems described in (2), with the predictor matrix X and the penalty function $\mathcal{P}(\cdot)$ respectively being

$$X = I_N \otimes I_p, \quad \mathcal{P}(B) = ||B||_* = \max_{u \in \mathcal{C}} \langle u, B \rangle$$

  with $\mathcal{C} = \{A : ||A||_{op} \leq 1\}$ where $|| \cdot ||_{op}$ denotes a spectral norm.
- Therefore we can directly apply the *Kac-Rice test* with the resulting test statistic as follows, under the assumed model discussed in (1):

$$\mathbb{S}_{1,0} = \frac{\int_{d_1}^{\infty} e^{-\frac{z^2}{2\sigma^2}} \, z^{N-p} \, \prod_{j=2}^{p} \left(z^2 - d_j^2 \right) \, dz}{\int_{d_2}^{\infty} e^{-\frac{z^2}{2\sigma^2}} \, z^{N-p} \, \prod_{j=2}^{p} \left(z^2 - d_j^2 \right) \, dz}, \tag{6}$$

  where $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ denote the observed singular values of Y.
- The test statistic $\mathbb{S}_{1,0}$ in (6) is uniformly distributed under the global null hypothesis (4), and provides exact type 1 error controls for testing the global null hypothesis.

More on the Kac-Rice Test Statistic in (6)

- The value $\mathbb{S}_{1,0}$ represents the probability of observing more extreme values than d_1 under the null hypothesis, which coincides with the traditional notion of p-value.
- Another view: The test statistic $\mathbb{S}_{1,0}$ corresponds to a conditional survival probability of the largest observed singular value d_1 conditioned on all the other observed singular values $d_2, \cdots, d_p$.
  - The integrand of $\mathbb{S}_{1,0}$ coincides with the conditional distribution of the largest eigenvalue of a central Wishart matrix (James, 1964) via a change of variables.
  - The denominator of $\mathbb{S}_{1,0}$ acts as a normalizing constant since the domain of the largest singular value d_1 conditioning on all the other singular values becomes (d_2, ∞).
- A small magnitude of $\mathbb{S}_{1,0}$ implies large d_1 compared to d_2, and thus supports $\mathbb{H}_1 : \lambda > 0$.

Review of the Kac-Rice Test: Confidence Intervals for the Largest Signal

- Along with the *Kac-Rice test* mentioned above, a procedure constructing an exact confidence interval for the leading signal in adaptive regression is proposed in Taylor, Loftus and Tibshirani (2016).
- By applying the result of Taylor, Loftus and Tibshirani (2016) to the matrix denoising setting, we can generate an exact confidence interval for

$$\tilde{\Lambda}_1 = \langle U_1 V_1^\top, B \rangle,$$

where

  - $Y = UDV^\top$ is a singular value decomposition (SVD) of Y with $D = \text{diag}(d_1, \cdots, d_p)$ for $d_1 \geq \cdots \geq d_p$
  - U_1 and V_1 are the first column vectors of U and V, respectively.
- It is desirable to directly find the confidence interval for Λ_1 (a leading SV of B). However, as B is unobservable, $U_1 V_1^\top$ is the "best guess" of the unit vector associated with Λ_1 in its direction.

- In the matrix denoising problem of (5), the result from Taylor, Loftus and Tibshirani (2016) yields an exact conditional survival probability of d_1 as follows:

$$\mathbb{S}_{1,\tilde{\Lambda}_1} = \frac{\int_{d_1}^{\infty} e^{-\frac{(z-\tilde{\Lambda}_1)^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^{p} \left(z^2 - d_j^2\right) dz}{\int_{d_2}^{\infty} e^{-\frac{(z-\tilde{\Lambda}_1)^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^{p} \left(z^2 - d_j^2\right) dz} \sim \text{Unif}(0,1). \tag{7}$$

- A special case: When data is generated under $\Lambda_1 = 0$, $\tilde{\Lambda}_1 = 0$ holds. Then, $\mathbb{S}_{1,\tilde{\Lambda}_1}$ in (7) is the same as $\mathbb{S}_{1,0}$ for testing $\mathbb{H}_0 : \lambda_1 = 0$, and yields exact type 1 error controls due to its uniformity.

- To construct the level α confidence interval, let

$$\mathsf{CI} = \left\{ \delta : \min\left(\mathbb{S}_{1,\delta}, 1 - \mathbb{S}_{1,\delta}\right) > \frac{\alpha}{2} \right\}. \tag{8}$$

Since $\mathbb{S}_{1,\tilde{\Lambda}_1}$ in (7) is uniformly distributed, we observe that

$$\mathbb{P}\left(\tilde{\Lambda}_1 \in \mathsf{CI}\right) = 1 - \alpha,$$

and thus CI in (8) generates an exact level α confidence interval.

General Hypothesis Testing

- Choi, Taylor and Tibshirani (2017) extend the test for the global null in (4) to a genaral test which investigates whether there exists the kth largest signal in B.
- Consider the following hypothesis:

$$\begin{aligned} & \mathbb{H}_{0,k} : \Lambda_k = 0 \text{ vs } \mathbb{H}_{1,k} : \Lambda_k > 0 \\ \iff \quad & \mathbb{H}_{0,k} : \text{rank}(B) < k \text{ vs } \mathbb{H}_{1,k} : \text{rank}(B) \geq k, \end{aligned} \tag{9}$$

  for $k = 1, \cdots, p-1$.
  - For $k = 1$, the null hypothesis in (9) corresponds to a global null as in (4).
- We assume $\text{rank}(B) < p$ and do not consider the case of $k = p$.
  - At $k = p$, the signal matrix B is full rank under the alternative hypothesis with $\Lambda_p \neq 0$.
  - In this scenario, the problem of $\text{rank}(B) = p$ with the noise level of σ^2 becomes unidentifiable with the problem of $\text{rank}(B) = p - 1$ with the noise level of $\sigma^2 + \Lambda_p^2$.

- One of the most straightforward approaches for extending the global test (6) to testing (9) for $k = 1, \cdots, p-1$ would be to apply it sequentially, with removing the first observed $k-1$ singular values at the kth step, and start at the kth observed singular value.
- That is, at the kth step, we can ignore the first $k-1$ observed singular values of Y, and apply the test with plugging in the kth value to the place of the 1st value, the $(k+1)$th to the place of the 2nd value, and so on.
- This approach of ignoring the existence of the first $k-1$ singular values is analogous to other methods dealing with essentially the same hypothesis testing (Muirhead, 1982; Kritchman and Nadler, 2008).

How well does the "sequential Kac-Rice test" work?

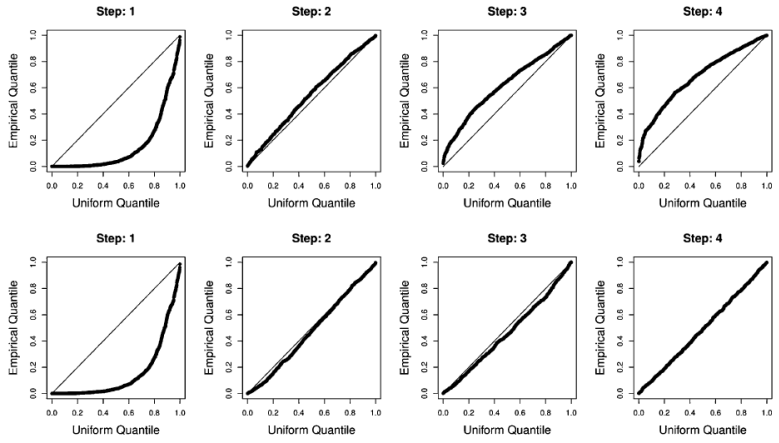


FIG. 2. *Quantile-quantile plots of the observed quantiles of p-values versus the uniform quantiles. With $N = 20$ and $p = 10$, the true rank of B is* $\mathrm{rank}(B) = 1$. *The top panels are from the sequential* Kac–Rice test *and the bottom panels are from the* CSV. *The kth column represents quantile-quantile plots of the p-values for testing* $H_{0,k} : \mathrm{rank}(B) \leq k - 1$ *for steps* $k = 1, 2, 3, 4$.

- An example in Figure 2 shows one with $N = 20$, $p = 10$, and a rank one signal of moderate size.
- The top panels: Quantile-quantile plots of the p-values for the sequential *Kac-Rice test* versus the uniform distribution
  - The p-values are small when the alternative hypothesis $\mathbb{H}_{1,1}$ is true (step 1), and then fairly uniform for testing rank ≤ 1 versus > 1 (step 2) as described, which is the first case in which the null hypothesis is true.
  - However, the test becomes more and more conservative for higher steps, although the p-values are generated under the null distributions.
  - The conservativeness of these p-values can lead to potential loss of power. [2]

[2] One of the reasons is that, at at $k = 3$, for example, the test does not consider taht the two largest singular values have been removed. The test instead ignores the existence of the 1st and 2nd singular values, and treats the 3rd one as the 1st, plugging it into the place of the 1st singular value. As the 1st and 3rd largest singular values do not have the same distribution with the density of the 1st singular value having more weights on large values, the sequential *Kac-Rice test* at $k = 3$ is no longer uniformly distributed; the test results in conservative p-values.

- The bottom panels: Quantile-quantile plots of the p-values for the conditional singular value (CSV) test to be described below
  - It follows the uniform distribution quite well for all "null" steps.
  - For testing $\mathbb{H}_{0,k} : \Lambda_k = 0$ at the kth step, the CSV method takes it into account that our interest is the kth signal by conditioning on the first $k-1$ singular values when deriving its test statistic.
- The following sections discuss the CSV test procedure in detail, and the ICSV test, an integrated version of the CSV which has better power.

General Hypothesis Testing: The Conditional Singular Value Test

- In this section, we introduce a test in which the test statistic has an "almost exact" null distribution under $\mathbb{H}_{0,k}$ in (9) for $k \in \{1, \cdots, p-1\}$.
- Notations:
  - We write the singular value decomposition of a signal matrix B as $B = U_B D_B V_B^\top$, the singular value decomposition of an observed data matrix Y as $Y = UDV^\top$, and an $N \times r$ and an $N \times (p-r)$ column-wise submatrices of an $N \times p$ matrix M by $M_{[r]}$ and by $M_{[-r]}$ where $M = \left[M_{[r]}, M_{[-r]}\right]$.
  - Also, P_Q denotes a projection matrix onto a column-space of an $n_1 \times n_2$ matrix Q, and $P_Q^\perp$ denotes a projection matrix onto a kernel of P_Q, that is, assuming $n_1 \geq n_2$ and Q is of full-rank, we have

$$P_Q = Q(Q^{\top Q})^{-1}Q^\top,$$
$$P_Q^\perp = I_{n_1} - P_Q.$$

- With these notations, the hypothesis in (9) can be rewritten as

$$\mathbb{H}_{0,k} : P^{\perp}_{U_{B[k-1]}} B P^{\perp}_{V_{B[k-1]}} = 0_{N\times p} \text{ vs } \mathbb{H}_{1,k} : P^{\perp}_{U_{B[k-1]}} B P^{\perp}_{V_{B[k-1]}} \neq 0_{N\times p}. \quad (10)$$

  Note that $P^{\perp}_{U_{B[k-1]}} B P^{\perp}_{V_{B[k-1]}} = 0_{N\times p}$ is equivalent to $\Lambda_k = \cdots \Lambda_p = 0$.
- The hypothesis in (10) examines whether the column spaces of $U_{B[k-1]}$ and $V_{B[k-1]}$ capture all nontrivial signals in B, or equivalently, whether the deflated residual space $(U_{B[-(k-1)]}, V_{B[-(k-1)]})$ contains any signals.
- The proposed test procedure which we refer to as the conditional singular value (CSV) test is as follows:

Test 2.1 (Conditional Singular Value test)

With a given level α, and the following test statistic:

$$\mathbb{S}_{k,0} = \frac{\int_{d_k}^{d_{k-1}} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j\neq k}^{p} \left| z^2 - d_j^2 \right| dz}{\int_{d_{k+1}}^{d_{k-1}} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j\neq k}^{p} \left| z^2 - d_j^2 \right| dz}, \tag{11}$$

where $d_0 = \infty$, we reject $\mathbb{H}_{0,k}$ if $\mathbb{S}_{k,0} \leq \alpha$ and accept $\mathbb{H}_{0,k}$ otherwise.

- Analogous to $\mathbb{S}_{1,0}$ in (6), $\mathbb{S}_{k,0}$ compares the relative size of d_k ranging between (d_{k+1}, d_{k-1}), and a small value of $\mathbb{S}_{k,0}$ implies a large value of d_k, supporting the alternative hypothesis $\mathbb{H}_{1,k} : \Lambda_k > 0$.
- Likewise, the test statistic $\mathbb{S}_{K,0}$ plays the role of a p-value: it is the probability of observing larger values of the kth singular value than the actually observed d_k, under the distribution that is close to the null scenario.

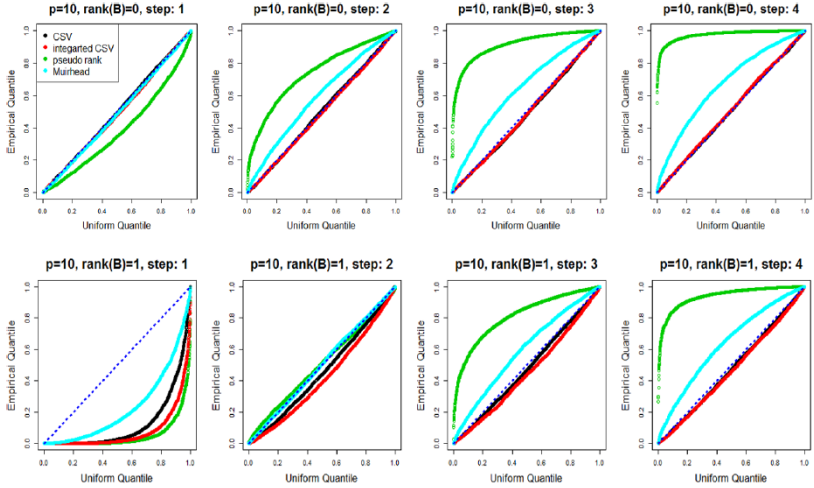
- Here, $\mathbb{S}_{k,0}$ represents a survival probability of the kth singular value conditional on the observed $U_{[k-1]}, V_{[k-1]}$ and all the other singular values.
- As aforementioned, the distribution of the kth singular value when the following situation holds:

$$P^{\perp}_{U_{[k-1]}} B P^{\perp}_{V_{[k-1]}} = 0_{N \times p}. \tag{12}$$

- Here, (12) adresses the situation that $(U_{[k-1]}, V_{[k-1]})$ captures all the signals in B.
- Though (12) is close to $\mathbb{H}_{0,k}$ in (10), there exists slight descrepancy between these two scenarios.
  - As Y is drawn from a noisy matrix of B, $(U_{[k-1]}, V_{[k-1]})$ is also perturbed and does not coincide with $(U_{B[k-1]}, V_{B[k-1]})$.
  - As a result, even though $\mathbb{H}_{0,k}$: $\mathrm{rank}(B) < k$ is true, the $(U_{[k-1]}, V_{[k-1]})$ cannot capture the entire signals and the bits of left-over remains in the deflated residuals.
  - This happens especially when $k - 1 = \mathrm{rank}(B)$.
  - However, as the signals grow stronger and the step k advances further, the singular values $(U_{[k-1]}, V_{[k-1]})$ capture most of the signals, and thus the descrepancy between $\mathbb{H}_{0,k}$ in (10) and the situation as in (12) becomes slimmer.

Null Distribution of the Test Statistic $\mathbb{S}_{k,0}$

- Here, the data singular vectors $(U_{[k-1]}, V_{[k-1]})$ involved in (12) can be viewed as a component of a data-driven model of B with its rank being $k-1$.
- As k goes beyond rank(B), there is little information remaining to estimate the later singular vectors of B as their corresponding population singular value is 0.
- Nevertheless, the validity of the test statistic $\mathbb{S}_{k,0}$ depends only on finding candidate subspaces that contain the true singular vector subspaces of B.
- Continuing well beyond rank(B) we will have added several left and right singular vectors that have littele to do with B (cf. Theorem 4 of Paul, 2007).
- The fact that (12) remains valid even well beyond rank(B) at least partially explains why our tests continue to have roughly exact size while pseudo-rank becomes more and more conservative, as seen in Figure 3.


p=10, rank(B)=0, step: 1
p=10, rank(B)=0, step: 2
p=10, rank(B)=0, step: 3
p=10, rank(B)=0, step: 4
p=10, rank(B)=1, step: 1
p=10, rank(B)=1, step: 2
p=10, rank(B)=1, step: 3
p=10, rank(B)=1, step: 4
CSV
integarted CSV
pseudo rank
Muirhead
Empirical Quantile
Uniform Quantile

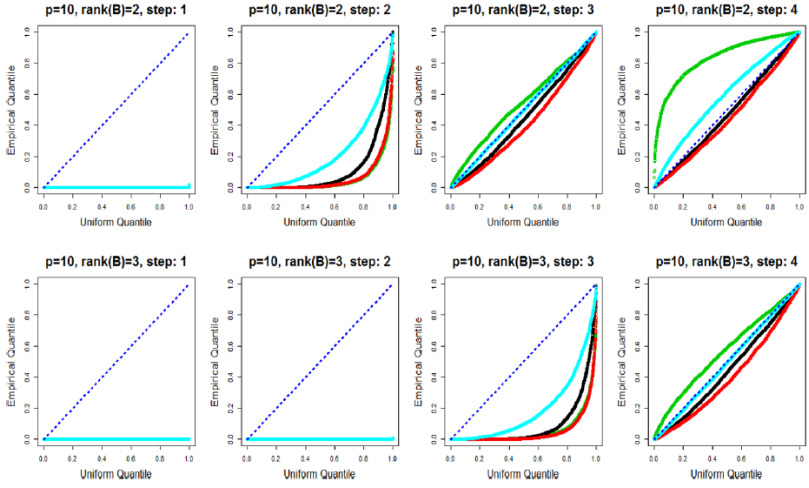


FIG. 3. *Quantile-quantile plots of the empirical quantiles of p-values versus the uniform quantiles when* $p = 10$ *and* $N = 50$ *at* $m = 1.5$. *From the top to the bottom, each row represents the case of the true* $\mathrm{rank}(B)$ *from* 0 *to* 3. *The columns represent the results for testing* $H_{0,1}$ *to* $H_{0,4}$ *from the left to the right. The six plots including* (2*nd row*, 1*st column*), (3*rd row*, 1*st column*), (3*rd row*, 2*nd column*), (4*th row*, 1*st column*), (4*th row*, 2*nd column*) *and* (4*th row*, 3*rd column*) *represent the cases under the alternative. The rest of the plots are under the null.*

- Also, the bottom panels of Figure 2 confirm the claimed exact type 1 error control property of the procedure as after the true rank of one, the p-values are all close to uniform.
- Empirical results show close to exact type 1 error controls regardless of the magnitude of the signals even at $k-1 = \text{rank}(B)$.
- Theorem 2.1 shows that the test statistic $\mathbb{S}_{k,0}$ is uniformly distributed when (12) holds, and thus yields close to exact type 1 error controls.

Theorem 2.1

If Y is drawn from Normal$(B, I_n \otimes I_p)$ and $P^{\perp}_{U_{[k-1]}} B P^{\perp}_{V_{[k-1]}} = 0_{N \times p}$ hold, then

$$\mathbb{S}_{k,0} \sim \text{Uniform}(0,1).$$

- In constructing the test statistic $\mathbb{S}_{k,0}$, conditioning on $(U_{[k-1]}, V_{[k-1]})$ and $(d_1, \cdots, d_{k-1})$ represents the selection event of choosing $k-1$ active principal components at step k, and consitioning additionally on $d_{k+1}, \cdots d_p$ leads to an inference of a saturated model.
- Though the test statistic $\mathbb{S}_{k,0}$ and its associated conditional density do not involve $(U_{[k-1]}, V_{[k-1]})$ since the singular vectors are independent from the singular values, it is incorporated on the test procedure as a component of the selection event of choosing $k-1$ active principal components.
- By accounting for the selection event of active principal components along with their associated singular vectors, the proposed test procedure achieves (approximately) exact type 1 error controls; by constraining the test space to a deflated distribution reflecting the current stage led by the selection procedure.
- This framework involves the post-selection inference as in the work of Tibshirani, Taylor, Lockhart and Tibshirani (2017).
- The resulting conditional density of the singular values used in thet test statistic reflects the fact that $(U_{[k-1]}, V_{[k-1]})$ is a sufficient statistic for B under (12).

General Hypothesis Testing: The Integrated Conditional Singular Value Test

- As a potential improvement of the CSV, we introduce as integrated version of $\mathbb{S}_{k,0}$, which we refer to as Integrated Conditional Singular Value (ICSV) test. Our aim is to achieve higher power in detecting signals in B compared to the ordinaty CSV.
- The idea is that conditioning on less can lead to greater power.
  - The ordinary CSV test statistic $\mathbb{S}_{k,0}$ assumes a saturated model, conditioning on all the observed singular values except for the kth one.
  - Here, we condition on only the first $k-1$ observed singular values, and integrate out the last $p-k$ singular values with respect to the null distribution conditional on the first $k-1$ observed singular values.
  - This can be considered as averaging the last $p-k$ singular values across all the possible values of those with proper weights where the proper weights correspond to the conditional null distribution of the last $p-k$ singular values.

- The resulting statistics becomes a function of $d_1, \cdots, d_k$, only the first k observed singular values of Y, which are associated with the selection event of active principal components that lead to step k.
- While the ordinary CSV test statistic $\mathbb{S}_{k,0}$ utilizes a saturated model by conditioning on all the observed singular values except for the kth one, the ICSV conditions only on the selection event and can be viewed as as inference on nonsaturated model.
- In its construction, the ICSV test utilizes a post-selection inference framework as in the work of Tibshirani, Taylor, Lockhart and Tibshirani (2017) in the same manner as the ordinary CSV test.

ICSV Test Statistic

- The proposed test statistic is as follows:

$$\mathbb{V}_{k,0} = \frac{\int_{d_k}^{d_{k-1}} g(y_k; d_1, \cdots, d_{k-1}) dy_k}{\int_{0}^{d_{k-1}} g(y_k; d_1, \cdots, d_{k-1}) dy_k} \tag{13}$$

where

$$\begin{aligned} & g(y_k; d_1, \cdots, d_{k-1}) \\ & = \int \cdots \int \prod_{i=k}^{p} \left(e^{-\frac{y_i^2}{2\sigma^2}} y_i^{N-p} \right) \left(\prod_{i=k}^{p} \prod_{j>i} (y_i^2 - y_j^2) \right) \\ & \cdot \left(\prod_{i=1}^{k-1} \prod_{j=k}^{p} (d_i^2 - y_j^2) \right) 1(0 \le y_p \le y_{p-1} \le \cdots \le y_k \le d_{k-1}) \, dy_{k+1} \cdots dy_p \end{aligned} \tag{14}$$

Test 2.2 (ICSV test)

With a given level α, we reject $\mathbb{H}_{0,k}$ if $\mathbb{V}_{k,0} \leq \alpha$ and accept $\mathbb{H}_{0,k}$ otherwise, where $\mathbb{V}_{k,0}$ is as defined in (13).

- As in the CSV test, $\mathbb{V}_{k,0}$ performs as a p-value for the test.
  - It examines the relative size of of a gap from from d_k to d_{k-1}.
  - A small value of $\mathbb{V}_{k,0}$ implies a large value of d_k, supporting the alternative hypothesis $\mathbb{H}_{1,k} : \Lambda_k > 0$.
- In the ICSV test, the range of d_k is emerged from $(d_k + 1, d_{k-1})$ to $(0, d_{k-1})$ compared to the CSV test.
- The test statistic $\mathbb{V}_{k,0}$ is a survival probability of the kth singular value conditioned on the observed $U_{[k-1]}, V_{[k-1]}$ and the forst $k-1$ singular values with the distribution under (12), where the conditions reflect the data-driven selection event from the previous steps.

- In accordance with the CSV test, the test statistic of the ICSV test is uniformly distributed under (12), and thus provides close to exact type 1 error controls, which is shown in Theorem 2.2.
- Here, fixing the observed $(U_{[k-1]}, V_{[k-1]}, d_1, \cdots, d_{k-1})$ corresponds to the selection of a data-driven model, and the proposed procedure investigates the deflated residual resulting from conditioning on the selection event.

Theorem 2.2

If Y is drawn from Normal$(B, I_N \otimes I_p)$ and $P^{\perp}_{U_{[k-1]}} B P^{\perp}_{V_{[k-1]}} = 0_{N \times p}$ hold, then

$$\mathbb{V}_{k,0} \sim \text{Unifrom}(0, 1).$$

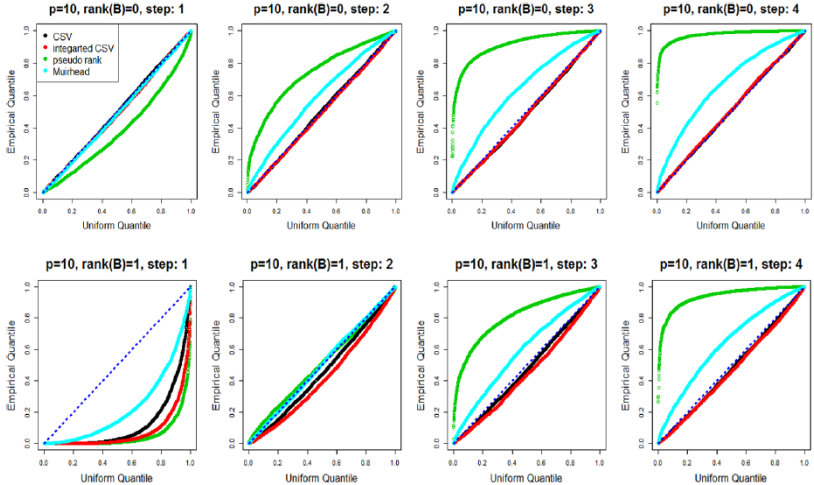
- Along with controlling type 1 errors to exact target levels, the ICSV test yields independent p-values when p-values are generated under (12). This independence relation between p-values is shown in Theorem 2.3, which corresponds to a special case of Theorem 4 in Fithian, Taylor, Tibshirani and Tibshirani (2015). The independence between p-values under the null hypothesis is a sufficient condition for a number of multiple hypothesis testing cirrection procedures (Benjamini and Hochberg, 1995; G'Sell, Wager, Chouldechova and Tibshirani, 2013; Simes, 1986).

Theorem 2.3

If Y is drawn from Normal$(B, I_N \otimes I_p)$ and $P^{\perp}_{U_{[k-1]}} B P^{\perp}_{V_{[k-1]}} = 0_{N \times p}$ hold, then

$$\{\mathbb{V}_{i,0} | i \in \{k, \cdots, p-1\}\}$$

are independent to each other.


p=10, rank(B)=0, step: 1
p=10, rank(B)=0, step: 2
p=10, rank(B)=0, step: 3
p=10, rank(B)=0, step: 4
p=10, rank(B)=1, step: 1
p=10, rank(B)=1, step: 2
p=10, rank(B)=1, step: 3
p=10, rank(B)=1, step: 4
CSV
integarted CSV
pseudo rank
Muirhead
Empirical Quantile
Uniform Quantile

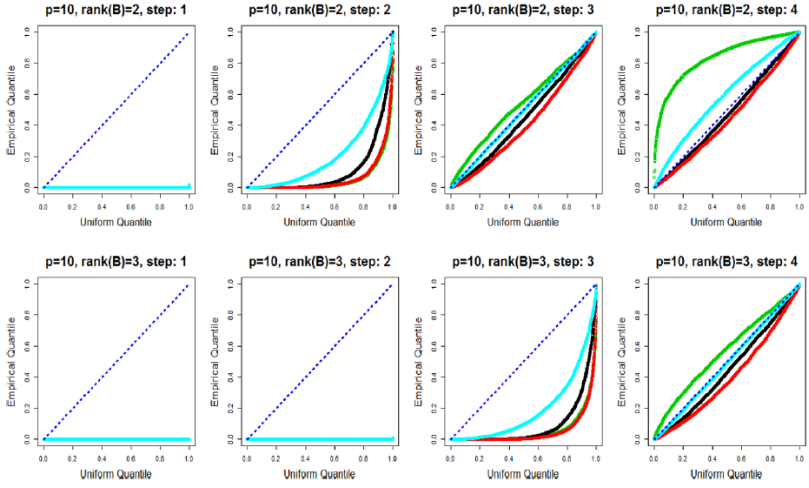


FIG. 3. *Quantile-quantile plots of the empirical quantiles of p-values versus the uniform quantiles when* $p = 10$ *and* $N = 50$ *at* $m = 1.5$. *From the top to the bottom, each row represents the case of the true* $\mathrm{rank}(B)$ *from* 0 *to* 3. *The columns represent the results for testing* $H_{0,1}$ *to* $H_{0,4}$ *from the left to the right. The six plots including* (2*nd row*, 1*st column*), (3*rd row*, 1*st column*), (3*rd row*, 2*nd column*), (4*th row*, 1*st column*), (4*th row*, 2*nd column*) *and* (4*th row*, 3*rd column*) *represent the cases under the alternative. The rest of the plots are under the null.*

- Figure 3 demonstrates that the ICSV procedure achives higher power than the ordinary CSV, and controls type 1 error for testing $\mathbb{H}_{k,0}$: rank$(B) \leq k-1$ near an exact target level as desired.
- In this paper, Choi, Taylor and Tibshirani (2017) use importance sampling to evaluate the integral in (13) with samples drawn from the eigenvalues of a $(N-k+1) \times (p-k+1)$ Wishart matrix.
- As the computational cost increases shraply with large p, we are currently unable to compute this test for beyond say 30 or 40.
- An interesting open problem is the numerical approximation of this integral, in order to scale the test to larger problems.

General Hypothesis Testing: Simulation Examples

- This section presents results of the CSV and the ICSV for testing the general hypothesis $H_{k,0}$: rank$(B) \leq k-1$ in (9) on simulated examples.
- Comparisons:
  - the pseudorank method by Kritchman and Nadler (2008)
  - the Muirhead's method by Muirhead (1982), Theorem 9.6.2

Test 2.3 (Pseudorank)

With a given level α, and follwing $\mu_{N,p}$ and $\sigma_{N,p}$:

$$\mu_{N,p} = \left(\sqrt{N - \frac{1}{2}} + \sqrt{p - \frac{1}{2}}\right)^2$$

$$\sigma_{N,p} = \left(\sqrt{N - \frac{1}{2}} + \sqrt{p - \frac{1}{2}}\right)\left(\frac{1}{\sqrt{\frac{N-1}{2}}} + \frac{1}{\sqrt{\frac{p-1}{2}}}\right)^{\frac{1}{3}},$$

we reject $H_{k,0}$: $\mathrm{rank}(B) \leq k - 1$ if

$$\frac{d_k^2 - \mu_{N,p-k}}{\sigma_{N,p-k}} > s(\alpha),$$

where $s(\alpha)$ is the upper α-quantile of the Tracy-Widom distribution.

Test 2.4 (Muirhead)

With a given level α, and V_k defined as

$$V_k = \frac{(N-1)^{q-1} \prod_{i=k}^{p} d_i^2}{\left(\frac{1}{q} \sum_{i=k}^{p} d_i^2\right)^2},$$

we reject $H_{k,0}$: $\text{rank}(B) \leq k-1$ if

$$-\left(N - k - \frac{2q^2+q+2}{6q} + \sum_{i=1}^{k-1} \frac{\bar{l}_q^2}{\left(d_i^2 - \bar{l}_q\right)^2}\right) \log V_k > \chi_{\frac{(q+2)(q-1)}{2}}(\alpha),$$

where $q = p-k+1$, $\bar{l}_q = \sum_{i=k}^{p} \frac{d_i^2}{q}$ and $\chi_m^2(\alpha)$ denotes the upper α quantile of the χ^2 distribution with degree m.

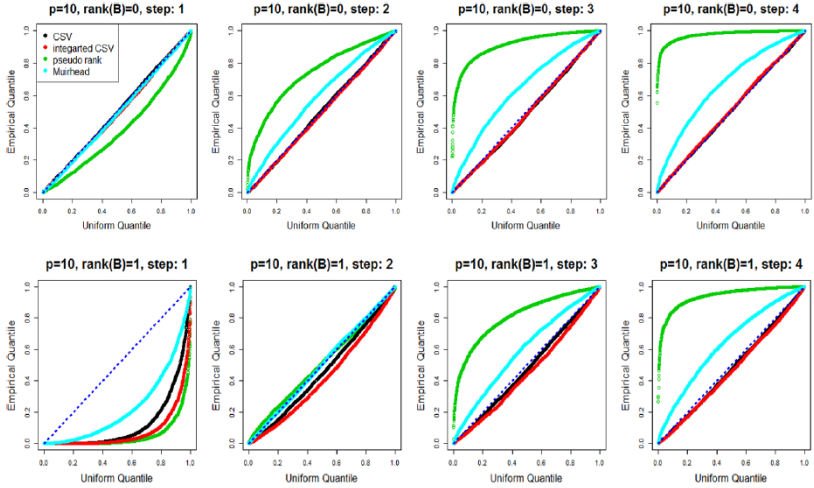
DGP

- We investigate cases with i.i.d. Gaussian noise entries with $\sigma^2 = 1$.
- An $N \times p$ data matrix Y has the signal matrix B formed as follows:

$$B = U_B D_B V_B^\top, \qquad \Lambda_i = m \cdot i \cdot \sigma \sqrt[4]{Np} \cdot 1\{i \leq \text{rank}(B)\}, \tag{15}$$

  where $D_B = \text{diag}(\Lambda_1, \cdots, \Lambda_p)$ with $\Lambda_1 \geq \cdots \geq \Lambda_p$, and U_B, V_B are rotation operators generated from a singular value decomposition of an $N \times p$ random Gaussian matrix with i.i.d. entries.
- The signals of B increase linearly.
- The constant m determines the magnitude of the signals.
- From $m = 1$, a phase transition phenomenon is observed when $\text{rank}(B) = 1$ in which the expectation of the largest singular value of Y starts to reflect the signal (Nadler, 2008).
- Case.1: $(N, p) = (50, 10)$ with $m = 1.5$ and $\text{rank}(B) = 0, 1, 2, 3$. 1000 rep.
- Case.2: $(N, p) = (120, 100)$ with $m = 1.3$ and $\text{rank}(B) = 3, 5, 10, 50$. 1000 rep.

Figures 3 and 4 present quantile-quantile plots of the expected (uniform) quantiles versus the observed quantiles of p-values for testing the general hypothesis in (2.9). Under $H_{1,k}$, the ICSV test shows improved power compared to the CSV, and close to that of the *pseudorank* in Figure 3 [the six plots including (2nd row, 1st column), (3rd row, 1st column), (3rd row, 2nd column), (4th row, 1st column), (4th row, 2nd column) and (4th row, 3rd column)]. The first column of Figure 4 is under $H_{1,k}$ and illustrates that the CSV shows stronger power relative to the other methods as the true rank of B increases. For the instance of $N = 120$ and $p = 100$ with small true rank of B, the *Muirhead's* test shows anti-conservative performance under the null at the early steps which increases the risk of false discovery. Under $H_{0,k}$, both the CSV and the ICSV quantiles nearly agree with the expected quantiles and provide almost exact type 1 error controls as the theory predicts for all null steps. The results from both the *Pseudorank* and the *Muirhead's* test become strongly conservative for further steps.


p=10, rank(B)=0, step: 1
p=10, rank(B)=0, step: 2
p=10, rank(B)=0, step: 3
p=10, rank(B)=0, step: 4
p=10, rank(B)=1, step: 1
p=10, rank(B)=1, step: 2
p=10, rank(B)=1, step: 3
p=10, rank(B)=1, step: 4
CSV
integarted CSV
pseudo rank
Muirhead
Empirical Quantile
Uniform Quantile

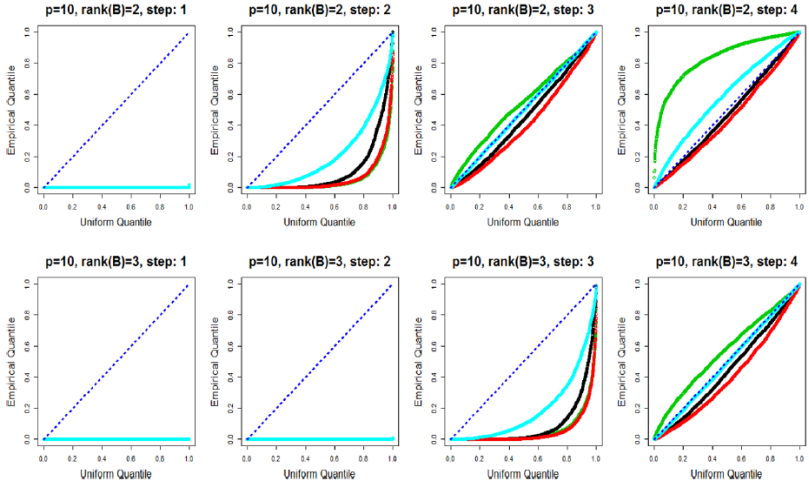
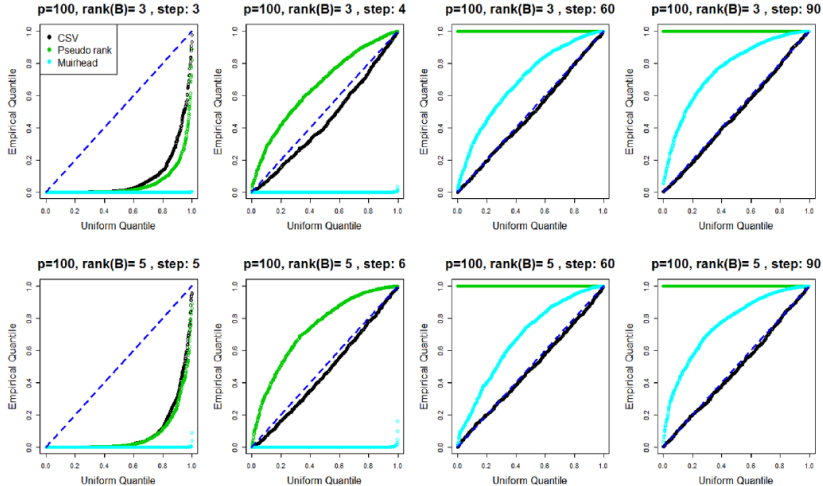


FIG. 3. *Quantile-quantile plots of the empirical quantiles of p-values versus the uniform quantiles when* $p = 10$ *and* $N = 50$ *at* $m = 1.5$. *From the top to the bottom, each row represents the case of the true* $\mathrm{rank}(B)$ *from* 0 *to* 3. *The columns represent the results for testing* $H_{0,1}$ *to* $H_{0,4}$ *from the left to the right. The six plots including* (2*nd row*, 1*st column*), (3*rd row*, 1*st column*), (3*rd row*, 2*nd column*), (4*th row*, 1*st column*), (4*th row*, 2*nd column*) *and* (4*th row*, 3*rd column*) *represent the cases under the alternative. The rest of the plots are under the null.*


p=100, rank(B)= 3 , step: 3
p=100, rank(B)= 3 , step: 4
p=100, rank(B)= 3 , step: 60
p=100, rank(B)= 3 , step: 90
p=100, rank(B)= 5 , step: 5
p=100, rank(B)= 5 , step: 6
p=100, rank(B)= 5 , step: 60
p=100, rank(B)= 5 , step: 90
CSV
Pseudo rank
Muirhead
Empirical Quantile
Uniform Quantile

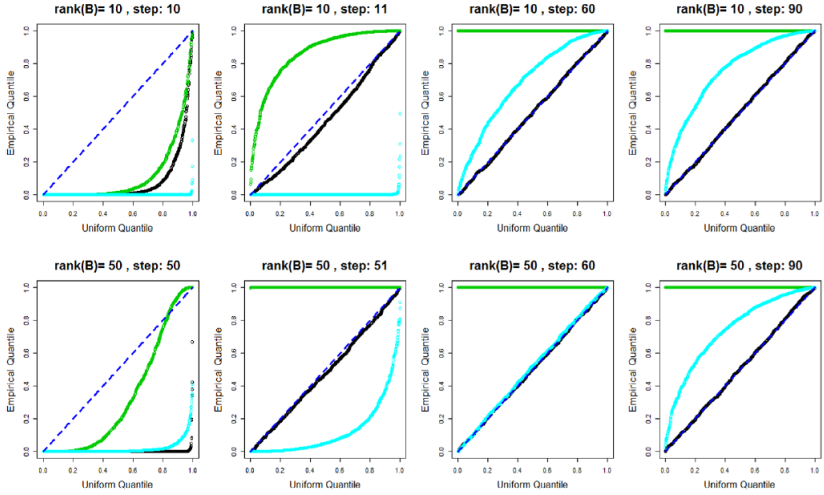


FIG. 4. *Quantile-quantile plots of the empirical quantiles of p-values versus the uniform quantiles when* $p = 100$ *and* $N = 120$ *at* $m = 1.3$. *From the top to the bottom, each row represents the case of* κ, *the true rank of* B, *equals* 3, 5, 10 *and* 50. *The columns represent the results for testing* $H_{0,\kappa}$, $H_{0,\kappa+1}$, $H_{0,60}$ *and* $H_{0,90}$ *from the left to the right. The first column represents the instances under the alternative. The rest of the columns are under the null.*

Confidence Interval Construction

- Here we generalize the exact confidence interval construction procedure of the largest singular value in (8) to the kth signal parameter for any $k = 1, \cdots, p-1$.
- Define the kth siganl parameter $\tilde{\Lambda}_k$ as follows:

$$\tilde{\Lambda}_k = \langle U_k V_k^\top, B \rangle, \tag{16}$$

where U_k and V_k are the kth column vector of U and V, respectively.

- We propose an approach to construct an exact level α confidence interval of $\tilde{\Lambda}_k$.
- The proposed procedure is as follows:

$$\mathsf{CI}_k(\mathbb{S}) = \left\{ \delta : \min\left(\mathbb{S}_{k,\delta}, 1 - \mathbb{S}_{k,\delta}\right) > \frac{\alpha}{2} \right\}, \tag{17}$$

where

$$\mathbb{S}_{k,\delta} = \frac{\int_{d_k}^{d_{k-1}} e^{-\frac{(z-\delta)^2}{2\sigma^2}} z^{N-p} \prod_{j \neq k}^{p} \left| z^2 - d_j^2 \right| dz}{\int_{d_{k+1}}^{d_{k-1}} e^{-\frac{(z-\delta)^2}{2\sigma^2}} z^{N-p} \prod_{j \neq k}^{p} \left| z^2 - d_j^2 \right| dz}.$$

Confidence Interval Construction: Simulation Studies

2.3.1

Rank Estimation

Estimating the Noise Level

Additional Examples

Conclusions

Conclusions

- Choi, Taylor and Tibshirani (2017) have proposed distribution-based methods for choosing the number of principal components of a data matrix.
  - Both for hypothesis testing anf the construction of confidence intervals of the signals
- The methods have exact type 1 error control and show promising results in simulated examples.
- Choi, Taylor and Tibshirani (2017) have also introduced data-based methods for estimating the noise level.

Further Investigation

- The analysis of the power of the proposed tests
- The width of the constructed confidence interval
- Application to high-dimensional data using numerical approximations
- Multiple hypothesis testing corrections: Required to Study the dependence structure of the p-values between different steps
- Robustness to non-Gaussian noise: Bootstrap version of the proposed procedure
- Degrees of freedom of the spectral estimator of the signal matrix (?)
- Canonical correlation analysis (CCA)
- Linear discriminant analysis (LDA)

Appendix A: Additional Details

Additional Details of the Proposed Tests

Appendix B: Technical Proofs

Lemma 1

Theorem 2.1

Theorem 2.2

Theorem 2.3

Theorem 2.4