

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

BY YUNJIN CHOI, JONATHAN TAYLOR¹ AND ROBERT TIBSHIRANI²

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 components. In order to address this challenge, we propose an *exact* distribution-based method for hypothesis testing and construction of confidence intervals for signals in a noisy matrix. Assuming Gaussian noise, we use the conditional distribution of the singular values of a Wishart matrix and derive exact hypothesis tests and confidence intervals for the true signals. Our paper is based on the approach of [Taylor, Loftus and Tibshirani \(2013\)](#) for testing the global null: we generalize it to test for any number of principal components, and derive an integrated version with greater power. In simulation studies we find that our proposed methods compare well to existing approaches.

1. Introduction.

1.1. *Overview.* Principal component analysis (PCA) is a commonly used method in multivariate statistics. It can be used for a variety of purposes including as a descriptive tool for examining the structure of a data matrix, as a pre-processing step for reducing the dimension of the column space of the matrix ([Josse and Husson, 2012](#)), or for matrix completion ([Cai, Candès and Shen, 2010](#)).

One important challenge in PCA is how to determine the number of components to retain. [Jolliffe \(2002\)](#) provides an excellent summary of existing approaches to determining the number of components, grouping them into three branches: subjective methods (e.g., the scree plot), distribution-based test tools (e.g., Bartlett's test), and computational procedures (e.g., cross-validation). Each branch has advantages as well as disadvantages, and no single method has emerged as the community standard.

¹Supported by NSF Grant DMS-12-08857 and AFOSR Grant 113039.

²Supported by NSF Grant DMS-99-71405 and NIH Contract N01-HV-28183.
62J05, 62J07, 62F03.

Keywords and phrases: principal components, exact test, p-value.

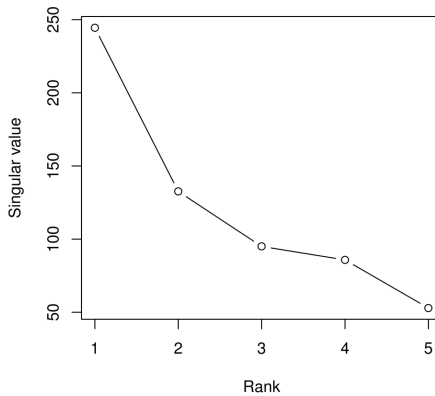


FIG 1. *Singular values of the score data in decreasing order. The data consist of exam scores of 88 students on five different topics (Mechanics, Vectors, Algebra, Analysis and Statistics).*

Figure 1 offers a scree plot as an example. The data are five test scores from 88 students (taken from [Mardia, Kent and Bibby \(1979\)](#)); the figure shows the five singular values in decreasing order. The “elbow” in this plot seems to occur at rank two or three, but it is not clearly visible. We revisit this example with our proposed approach in Section 5.3.

In this paper, we propose a statistical method for determining the rank of the signal matrix in a noisy matrix model. The estimated rank here corresponds to the number of components to retain in PCA. This method is derived from the conditional Gaussian-based distribution of the singular values, and yields exact p-values and confidence intervals.

1.2. Related work. Our method is motivated by the *Kac-Rice test* ([Taylor, Loftus and Tibshirani, 2013](#)), an exact method for testing and constructing confidence intervals for signals under the global null hypothesis in adaptive regression. Our work corresponds to the Kac-Rice test under the global null scenario, in a penalized regression minimizing the Frobenius norm with a nuclear norm penalty. In this paper, we extend the test and confidence intervals to the general case with improved power. The resulting statistic uses the survival function of the conditional distribution of the eigenvalues of a Wishart matrix.

In the context of inference based on the distribution of eigenvalues, [Muirhead \(1982\)](#) and, more recently, [Kritchman and Nadler \(2008\)](#) have pro-

posed methods for testing essentially the same hypothesis as in this paper. Both [Muirhead \(1982\)](#) and [Kritchman and Nadler \(2008\)](#) benefit from an asymptotic distribution of the test statistic: [Muirhead \(1982\)](#) forms a likelihood ratio test with the asymptotic Chi-square distribution. [Kritchman and Nadler \(2008\)](#) use the Tracy-Widom law, which is the asymptotic distribution of the largest eigenvalue of a Wishart matrix, incorporating the result of [Johnstone \(2001\)](#). We provide a method for constructing confidence intervals in addition to hypothesis testing, with the procedures being exact.

1.3. Organization of the paper. The rest of the paper is organized as follows. In Section 2, we propose methods based on the distribution of eigenvalues of a Wishart matrix. Section 2.1 introduces the main points of the *Kac-Rice test* ([Taylor, Loftus and Tibshirani, 2013](#)) from which we derive our proposals. Then, we suggest a procedure for hypothesis testing of the true rank of a signal matrix in Section 2.2. Our method for constructing exact confidence intervals of signals is described in Section 2.3.

In Section 3, we propose a method for estimating the rank of a signal matrix from the hypothesis tests. We illustrate sequential hypothesis testing procedure for determining the matrix rank, based on a proposal of [GSell et al. \(2013\)](#).

Section 4 introduces a data-driven method for estimating the noise level. Our suggested method is analogous to cross-validation. In order to apply cross-validation to a data matrix, we use a method suggested by [Mazumder, Hastie and Tibshirani \(2010\)](#).

Section 5 provides additional examples of the suggested methods. In Section 5.1, simulation results with estimated noise level are presented. Section 5.2 shows simulation results with non-Gaussian noise to check for robustness. We revisit the real data example introduced in Figure 1 in Section 5.3. The paper concludes with a brief discussion in Section 6.

2. A Distribution-Based Method. Throughout this paper, we 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:

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

that is,

$$(2.1) \quad Y \sim N(B, \sigma^2 I_N \otimes I_p).$$

In this paper, without loss of generality, we assume that $N > p$. We focus on the estimation of κ , the rank of the signal matrix B , and the construction of confidence intervals for the signals in B .

We first review the global null test and confidence interval construction of the first signal of [Taylor, Loftus and Tibshirani \(2013\)](#) and its application in matrix denoising problem in Section 2.1. Then we extend the global null test to a general test procedure for testing $H_{k,0} : \text{rank}(B) \leq k - 1$ versus $H_{k,1} : \text{rank}(B) \geq k$ for $k = 1, \dots, p-1$ in Section 2.2 and describe how to construct confidence intervals for the k^{th} largest signal parameters in Section 2.3.

2.1. Review of the Kac-Rice test. We briefly discuss the framework of [Taylor, Loftus and Tibshirani \(2013\)](#) and its application to a matrix denoising problem, which we extend further later in this paper. Section 2.1.1 covers the testing procedure for a class of null hypothesis. This null hypothesis corresponds to the global null in our matrix denoising problem. In Section 2.1.2, we construct a confidence interval for the largest signal.

2.1.1. Global null hypothesis testing. [Taylor, Loftus and Tibshirani \(2013\)](#) derived the *Kac-Rice test*, an exact test for a class of regularized regression problems of the following form:

$$(2.2) \quad \hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \cdot \mathcal{P}(\beta)$$

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 $y \in \mathbb{R}^p$ is generated from

$$y \sim N(X\beta, \Sigma),$$

the *Kac-Rice test* ([Taylor, Loftus and Tibshirani, 2013](#)) provides an exact method for testing

$$(2.3) \quad H_0 : \mathcal{P}(\beta) = 0$$

under the assumption that the penalty function \mathcal{P} is a support function of a convex set $\mathcal{C} \subseteq \mathbb{R}^p$. i.e.,

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

When applied to a matrix denoising problem of a popular form, (2.3) becomes a global null hypothesis:

$$(2.4) \quad H_0 : \Lambda_1 = 0 \equiv \text{rank}(B) = 0 \equiv B = 0_{N \times p}$$

where $\Lambda_1 \geq \Lambda_2 \geq \dots \geq \Lambda_p \geq 0$ denote the singular values of B . Here are the details. For an observed data matrix $Y \in \mathbb{R}^{N \times p}$, a widely used method to recover the signal matrix B in (2.1) is to solve the following criterion:

$$(2.5) \quad \hat{B} \in \operatorname{argmin}_{B \in \mathbb{R}^{N \times p}} \frac{1}{2} \|Y - B\|_F^2 + \lambda \|B\|_* \text{ where } \lambda > 0$$

where $\|\cdot\|_F$ and $\|\cdot\|_*$ denote a Frobenius norm and a nuclear norm respectively. The nuclear norm plays an analogous role as an ℓ_1 penalty term in lasso regression (Tibshirani, 1996). The objective function (2.5) falls into the class of regression problems described in (2.2), with the predictor matrix X being $I_N \otimes I_p$ and the penalty function $\mathcal{P}(\cdot)$ being

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

with $\mathcal{C} = \{A : \|A\|_{op} \leq 1\}$ where $\|\cdot\|_{op}$ denotes a spectral norm. We can therefore directly apply the *Kac-Rice test* with the resulting test statistic as follows, under the assumed model discussed in the beginning of Section 2:

$$(2.6) \quad \mathbb{S}_{1,0} = \frac{\int_{d_1}^{\infty} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^p (z^2 - d_j^2) dz}{\int_{d_2}^{\infty} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^p (z^2 - d_j^2) dz}$$

where $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$ denote the observed singular values of Y . The test statistic $\mathbb{S}_{1,0}$ in (2.6) is uniformly distributed under the null hypothesis (2.4) and provides a p-value for testing the global null hypothesis: the value $\mathbb{S}_{1,0}$ represents the probability of observing more extreme values than d_1 under the null hypothesis.

Viewed differently, the test statistic $\mathbb{S}_{1,0}$ corresponds to a conditional survival function of the largest observed singular value d_1 conditioned on all the other observed singular values d_2, \dots, d_p . The integrand of $\mathbb{S}_{1,0}$ coincides with the conditional distribution of the largest singular value of a central Wishart matrix up to a constant (James, 1964). Its denominator acts as a normalizing constant because the domain of the largest singular value d_1 is (d_2, ∞) . A small magnitude of $\mathbb{S}_{1,0}$ implies large d_1 compared to d_2 and thus supports $H_1 : \Lambda_1 > 0$.

2.1.2. Confidence intervals for the largest signal. Along with the *Kac-Rice test* mentioned in Section 2.1.1, a procedure for constructing an exact confidence interval for the leading signal in adaptive regression is suggested in Taylor, Loftus and Tibshirani (2013). As in Section 2.1.1, by applying the result of Taylor, Loftus and Tibshirani (2013) to our matrix denoising

setting, we can generate an exact confidence interval for $\tilde{\Lambda}_1$ which is defined as follows:

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

where $Y = UDV^T$ is a singular value decomposition of Y with $D = \text{diag}(d_1, \dots, d_p)$ for $d_1 \geq \dots \geq d_p$, and 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 instead of $\tilde{\Lambda}_1$, however, as B is unobservable, $U_1 V_1^T$ is the “best guess” of the unit vector associated with Λ_1 in its direction.

To discuss the procedure in detail, in the matrix denoising problem of (2.5), the result from Taylor, Loftus and Tibshirani (2013) yields an exact conditional survival function of d_1 around $\tilde{\Lambda}_1$ as follows:

$$(2.7) \quad \mathbb{S}_{1, \tilde{\Lambda}_1} = \frac{\int_{d_1}^{\infty} e^{-\frac{(z - \tilde{\Lambda}_1)^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^p (z^2 - d_j^2) dz}{\int_{d_2}^{\infty} e^{-\frac{(z - \tilde{\Lambda}_1)^2}{2\sigma^2}} z^{N-p} \prod_{j=2}^p (z^2 - d_j^2) dz} \sim \text{Unif}(0, 1).$$

The test statistic $\mathbb{S}_{1,0}$ for testing $H_0 : \Lambda_1 = 0$ in Section 2.1.1 conforms to (2.7) when it is true that $H_0 : \Lambda_1 = 0$. The suggested procedure for constructing the level α confidence interval is as follows:

$$(2.8) \quad CI = \{\delta : \min(\mathbb{S}_{1,\delta}, 1 - \mathbb{S}_{1,\delta}) > \alpha/2\}.$$

Since $\mathbb{S}_{1, \tilde{\Lambda}_1}$ in (2.7) is uniformly distributed, we observe that

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

and thus (2.8) generates an exact level α confidence interval.

2.2. General hypothesis testing. In this section, we extend the test for the global null in Section 2.1.1 to a general test which investigates whether there exists the k^{th} largest signal in B .

Suppose that we want to test the hypothesis

$$(2.9) \quad \begin{aligned} & H_{0,k} : \Lambda_k = 0 \text{ versus } H_{1,k} : \Lambda_k > 0 \\ \Leftrightarrow & H_{0,k} : \text{rank}(B) < k \text{ versus } H_{1,k} : \text{rank}(B) \geq k. \end{aligned}$$

for $k = 1, \dots, p-1$. For $k = 1$, the null hypothesis in (2.9) corresponds to a global null as in Section 2.1.1. With $k = p$ the signal matrix B is full rank under the alternative hypothesis, and the test becomes unidentifiable from a low rank signal matrix with higher noise level problem. Thus, we do not consider the case of $k = p$.

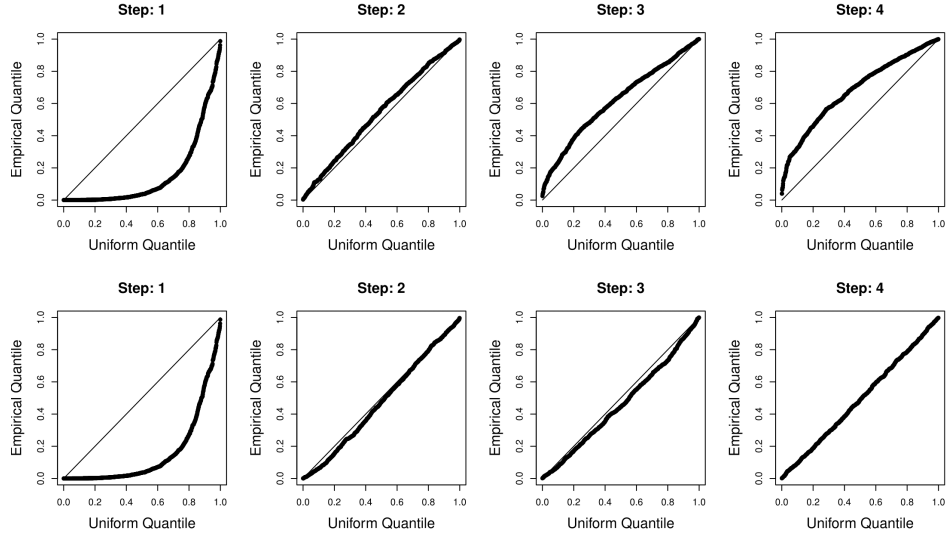


FIG 2. Quantile-quantile plots of the observed quantile of p -values versus the uniform quantiles. With $N = 20$ and $p = 10$, the true rank of B is $\text{rank}(B) = 1$. The top panels are from the sequential Kac-Rice test and the bottom panels are from the CSV. The k^{th} column represents quantile-quantile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

One of the most straightforward approaches for extending the global test (2.6) to testing (2.9) for $k = 1, \dots, p - 1$ would be to apply it sequentially. That is, we can remove the first $k - 1$ observed singular values of Y and then apply the test to the remaining $p - k + 1$ singular values, in analogy to other methods dealing with essentially the same hypothesis testing (Muirhead, 1982; Kritchman and Nadler, 2008).

How well does this work? Figure 2 shows an example. Here $N = 20$, $p = 10$ and there is a rank one signal of moderate size. The top panels show quantile-quantile plots of the p -values for the sequential Kac-Rice test versus the uniform distribution. We see that the p -values are small when the alternative hypothesis $H_{1,1}$ is true (step 1), and then fairly uniform for testing rank ≤ 1 versus > 1 (step 2) as desired, 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. One of the reasons for this conservativeness is that, at step $k = 3$, for example, the test does not consider that the two largest

singular values have been removed. The test instead plugs the 3rd largest singular value into the place of the 1st largest singular value. As the 1st and the 3rd largest singular values do not have the same distribution, the sequential *Kac-Rice test* at $k = 3$ is no longer uniformly distributed, and so results in conservative p-values.

The plots in the bottom panel come from our proposed conditional singular value (CSV) test, to be described in Section 2.2.1. It follows the uniform distribution quite well for all null steps. For testing $H_{0,k} : \Lambda_k = 0$ at the k^{th} step, the CSV method takes it into account that our interest is the k^{th} signal by conditioning on the first $k - 1$ singular values in its test statistic.

Section 2.2.1 presents the CSV test procedure in detail. In Section 2.2.2, we propose an integrated version of the CSV which has better power. Simulation results of the proposed procedures are illustrated in Section 2.2.3.

2.2.1. The Conditional Singular Value test. In this section, we introduce a test in which the test statistic has an “almost exact” null distribution under $H_{0,k}$ in (2.9) for $k \in \{1, \dots, p-1\}$. Writing the singular value decomposition of a signal matrix B as $B = U_B D_B V_B^T$, and submatrices of a $N \times p$ matrix X by $X = [X^r | X^{-r}]$ where $X^r \in \mathbb{R}^{N \times r}$ and $X^{-r} \in \mathbb{R}^{N \times (p-r)}$, the test of (2.9) can be rewritten as

$$(2.10) \quad \begin{aligned} &H_{0,k} : U_B^{-(k-1)} U_B^{-(k-1)T} B V_B^{-(k-1)} V_B^{-(k-1)T} = 0_{N \times p} \\ &\text{versus } H_{1,k} : U_B^{-(k-1)} U_B^{-(k-1)T} B V_B^{-(k-1)} V_B^{-(k-1)T} \neq 0_{N \times p} \end{aligned}$$

since $U_B^{-(k-1)} U_B^{-(k-1)T} B V_B^{-(k-1)} V_B^{-(k-1)T} = \text{diag}(0, \dots, 0, \Lambda_k, \dots, \Lambda_p)$.

As an alternative to (2.9) or (2.10), we derive an exact test statistic for the following hypothesis:

$$(2.11) \quad \begin{aligned} &H_{0,k} : U^{-(k-1)} U^{-(k-1)T} B V^{-(k-1)} V^{-(k-1)T} = 0_{N \times p} \\ &\text{versus } H_{1,k} : U^{-(k-1)} U^{-(k-1)T} B V^{-(k-1)} V^{-(k-1)T} \neq 0_{N \times p} \end{aligned}$$

where we write the singular value decomposition of the data matrix Y as $Y = U D V^T$. The test (2.11) investigates whether there remains signals in the residual space of Y of which the first $k - 1$ singular values are removed. Under $H_{0,k}$ of (2.10), when we have strong signals in B , $U^{-(k-1)} U^{-(k-1)T} B V^{-(k-1)} V^{-(k-1)T}$ approaches $0_{N \times p}$ as in (2.11).

The proposed test procedure is as follows:

TEST 2.1 (Conditional Singular Value test). *With a given level α , and*

the following test statistic,

$$(2.12) \quad \mathbb{S}_{k,0} = \frac{\int_{d_k}^{d_{k-1}} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j \neq k}^p |z^2 - d_j^2| dz}{\int_{d_{k+1}}^{d_{k-1}} e^{-\frac{z^2}{2\sigma^2}} z^{N-p} \prod_{j \neq k}^p |z^2 - d_j^2| dz},$$

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

Analogous to (2.6), $\mathbb{S}_{k,0}$ plays the role of a p-value. It 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 hypothesis $H_{1,k} : \Lambda_k > 0$. We refer to this procedure as the **conditional singular value test (CSV)**. Theorem 2.1 shows that this test is exact under (2.11). The test statistic $\mathbb{S}_{k,0}$ is a conditional survival function of the k^{th} singular value under $H_{0,k}$: the probability of observing larger values of the k^{th} singular value than the actually observed d_k , given $U^{(k-1)}, V^{(k-1)}$ and all the other singular values. The proofs of this and other results are given in the Appendix.

THEOREM 2.1.

$$\text{If } U^{-(k-1)} U^{-(k-1)^T} B V^{-(k-1)} V^{-(k-1)^T} = 0_{N \times p},$$

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

The bottom panels of Figure 2 confirm the claimed Type I error property of the procedure. After the true rank of one, the p-values are all close to uniform.

2.2.2. The Integrated Conditional Singular Value test. As a potential improvement of the CSV, we introduce an integrated version of $\mathbb{S}_{k,0}$. Our aim is to achieve higher power in detecting signals in B compared to the ordinary CSV. Here we integrate out $\mathbb{S}_{k,0}$ with respect to $p - k$ small singular values (d_{k+1}, \dots, d_p) . The resulting statistic becomes a function of d_1, \dots, d_k , only the first k singular values of Y , while the ordinary CSV test statistic $\mathbb{S}_{k,0}$ is a function of all the singular values of Y . The idea is that conditioning on less can lead to greater power. The suggested test statistic is as follows:

$$(2.13) \quad \mathbb{V}_{k,0} = \frac{\int_{d_k}^{d_{k-1}} g(y_k; d_1, \dots, d_{k-1}) dy_k}{\int_0^{d_{k-1}} g(y_k; d_1, \dots, d_{k-1}) dy_k},$$

where

$$(2.14) \quad g(y_k; d_1, \dots, d_{k-1}) = \int \cdots \int \prod_{i=k}^p \left(e^{-\frac{y_i^2}{2\sigma^2}} y_i^{N-p} \right) \left(\prod_{i=k}^p \prod_{j>i} (y_i^2 - y_j^2) \right) \\ \cdot \left(\prod_{i=1}^{k-1} \prod_{j=k}^p (d_i^2 - y_j^2) \right) 1_{\{0 \leq y_p \leq y_{p-1} \leq \cdots \leq y_k \leq d_{k-1}\}} dy_{k+1} \cdots dy_p.$$

Our proposed **Integrated Conditional Singular Value (ICSV)** test is as follows:

TEST 2.2 (ICSV test). *With a given level α , we reject $H_{0,k}$ if $\mathbb{V}_{k,0} \leq \alpha$ and accept $H_{0,k}$ otherwise, where $\mathbb{V}_{k,0}$ is as defined in (2.13).*

As in the **CSV** test, $\mathbb{V}_{k,0}$ works as a p-value for the test, and also it is an exact test for (2.11), as is shown in Theorem 2.2. It is a survival function of the k^{th} singular value given $U^{(k-1)}, V^{(k-1)}$ and the first $k-1$ singular values under $H_{0,k}$.

THEOREM 2.2.

$$\text{If } U^{-(k-1)} U^{-(k-1)^T} B V^{-(k-1)} V^{-(k-1)^T} = 0_{N \times p},$$

$$\text{then } \mathbb{V}_{k,0} \sim \text{Unif}(0, 1)$$

Figure 3 demonstrates that the **ICSV** procedure achieves higher power than the ordinary **CSV**.

In this paper, we use importance sampling to evaluate the integral in (2.13) with samples drawn from the eigenvalues of a $(N-k+1) \times (p-k+1)$ Wishart matrix. As the computational cost increases sharply with large p , we are currently unable to compute this test for p 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. We leave this as future work.

2.2.3. Simulation Examples. In this section, we present results of the **CSV** and the **ICSV** on simulated examples. We compare the performance of these proposed methods with those in Kritchman and Nadler (2008) and Muirhead (1982)[Theorem 9.6.2] mentioned in Section 1.2, which we refer as the *pseudorank* and the *Muirhead's method* respectively:

TEST 2.3 (Pseudorank). *With a given level α , and following $\mu_{N,p}$ and*

$\sigma_{N,p}$,

$$\begin{aligned}\mu_{N,p} &= \left(\sqrt{N - \frac{1}{2}} + \sqrt{p - \frac{1}{2}} \right)^2 \\ \sigma_{N,p} &= \left(\sqrt{N - \frac{1}{2}} + \sqrt{p - \frac{1}{2}} \right) \left(\frac{1}{\sqrt{N - 1/2}} + \frac{1}{\sqrt{p - 1/2}} \right)^{1/3},\end{aligned}$$

we reject $H_{0,k}$ 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's method). With a given level α , and V_k defined as

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

we reject $H_{0,k}$ if

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

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

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

$$(2.15) \quad B = U_B D_B V_B^T, \quad \Lambda_i = m \cdot i \cdot \sigma \sqrt[4]{Np} \cdot I_{\{i \leq \text{rank}(B)\}}$$

where $D_B = \text{diag}(\Lambda_1, \dots, \Lambda_p)$ with $\Lambda_1 \geq \dots \geq \Lambda_p$, and U_B, V_B are rotation operators generated from a singular value decomposition of $N \times p$ 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).

We illustrate two cases of $p = 10$ and $p = 30$ with N fixed to $N = 50$. For both cases, we set $m = 1.5$ and $\text{rank}(B) = 0, 1, 2, 3$. For $p = 10$ and $p = 30$,

we evaluate the procedure through 3000 and 1000 repetitions respectively. The true value of the noise level $\sigma^2 = 1$ is used for all testing procedures.

Figures 3 and 4 present quantile-quantile plots of the expected (uniform) quantiles versus the observed quantiles of p-values. Under $H_{1,k}$, the ICSV test shows improved power compared to the CSV, and close to that of *pseudorank*. Both the CSV and the ICSV show stronger power than *Muirhead's test*. Under $H_{0,k}$, both the CSV and the ICSV quantiles nearly agree with the expected quantiles and provide almost exact p-value, as the theory predicts. *Pseudorank* estimation becomes strongly conservative for further steps, and the results of *Muirhead's test* depend on the size of N and p .

2.3. Confidence interval construction. Here we generalize the exact confidence interval construction procedure of the largest singular value in (2.8) to the k^{th} signal parameter for any $k = 1, \dots, p-1$.

We define the k^{th} signal parameter $\tilde{\Lambda}_k$ as follows:

$$(2.16) \quad \tilde{\Lambda}_k = \langle U_k V_k^T, B \rangle$$

where U_k and V_k are the k^{th} column vector of U and V respectively. We propose an approach to construct an exact level α confidence interval of $\tilde{\Lambda}_k$. Our proposed procedure is as follows:

$$(2.17) \quad CI_k(\mathbb{S}) = \{\delta : \min(\mathbb{S}_{k,\delta}, 1 - \mathbb{S}_{k,\delta}) > \alpha/2\},$$

where

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

We can find the boundary points of $CI_k(\mathbb{S})$ using bisection. Theorem 2.3 below shows that $CI_k(\mathbb{S})$ is an exact level α confidence interval. This procedure addresses the general case of (2.12): $\mathbb{S}_{k,0}$ tests whether $\tilde{\Lambda}_k = 0$.

THEOREM 2.3. $\mathbb{S}_{k,\delta}$ is uniformly distributed when $\delta = \tilde{\Lambda}_k$.

Figure 5 shows that the coverage rate of $CI_k(\mathbb{S})$. As expected, the coverage rate of the true parameter is close to the target $1 - \alpha$.

2.3.1. Simulation studies of the confidence interval construction. We illustrate the coverage rates of $CI_k(\mathbb{S})$ in (2.17) on simulated data. The simulation settings are the same as in Section 2.2.3 with $N = 50$ and $p = 10$.

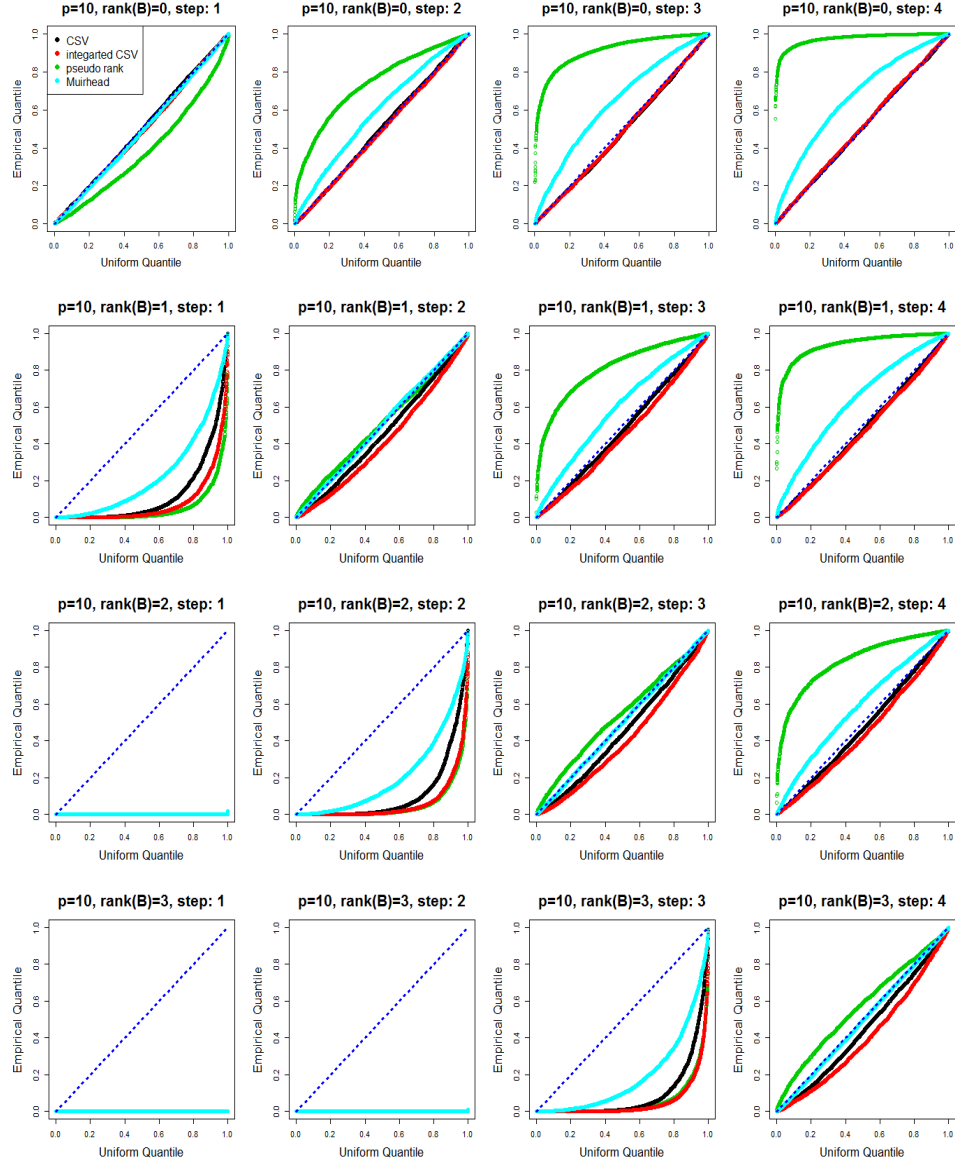


FIG 3. Quantile-quantile plots of the empirical quantile of p -values versus the uniform quantiles when $p = 10$ at $m = 1.5$. From the top to the bottom, each row represents the case of the true $\text{rank}(B)$ from 0 to 3. The columns represent the results for testing $H_{0,1}$ to $H_{0,4}$ from the left to the right.

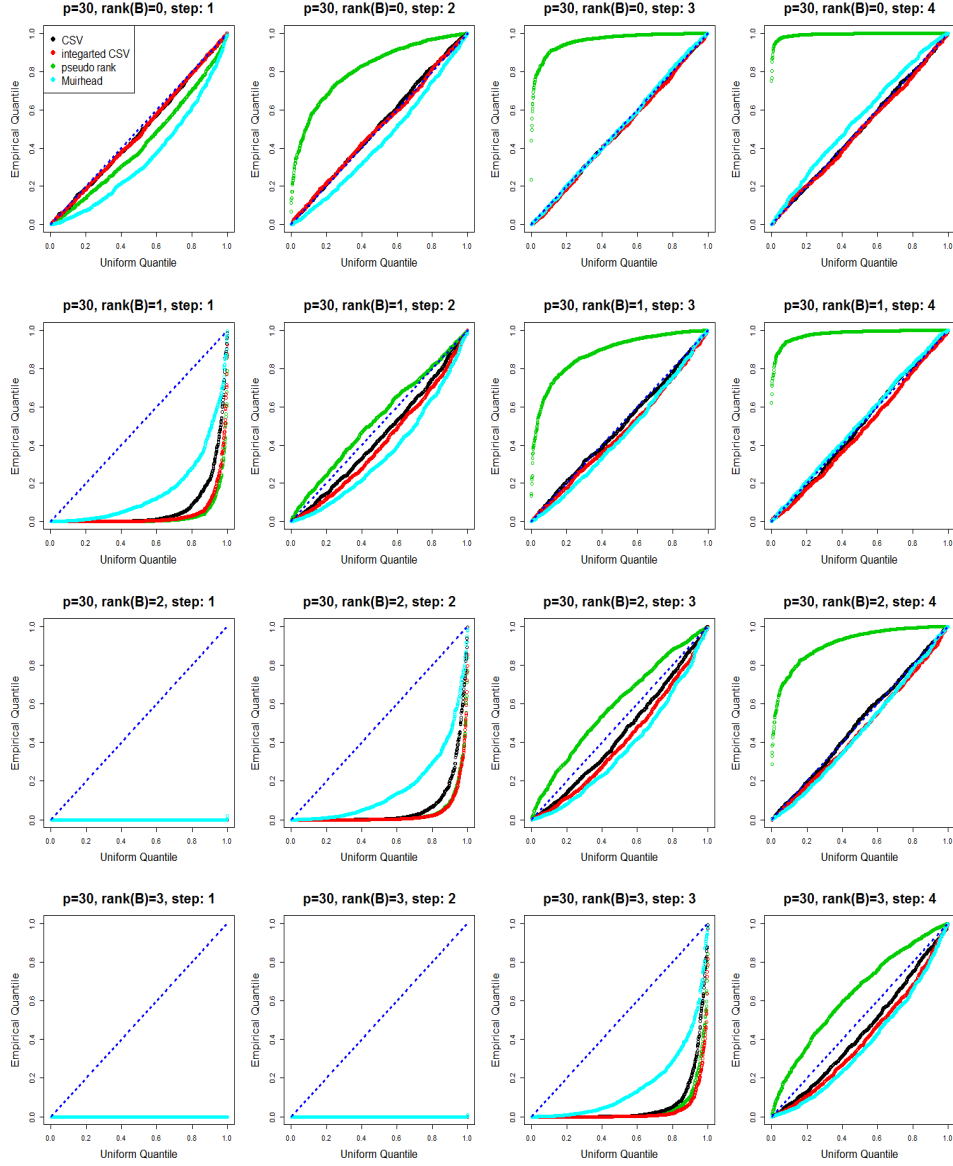


FIG 4. Quantile-quantile plots of the empirical quantile of p -values versus the uniform quantiles when $p = 30$ at $m = 1.5$. From the top to the bottom, each row represents the case of the true rank(B) from 0 to 3. The columns represent the results for testing $H_{0,1}$ to $H_{0,4}$ from the left to the right.

Figure 5 shows the coverage rate of the 95% confidence intervals for the first two signal parameters $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ in (2.16). Here, we vary m in (2.15) from 0 to 2. Large m leads to large magnitude of true $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$.

Figure 5 shows that regardless of $k = 1, 2$, true $\text{rank}(B)$, or the size of $\tilde{\Lambda}_k$, the constructed confidence intervals cover the parameters at the desired level.

3. Rank Estimation. This section discusses the selection of the number of principal components, or equivalently, the estimation of the true rank of B under our model assumptions. For determining the rank of B , here we investigate the *StrongStop* procedure (GSell et al., 2013), applying it to the tests developed in Section 2.2 (CSV and ICSV).

We determine the rank of B based on our testing results on $H_{0,k}$, since $H_{0,k}$ explicitly tests the range of the true rank. Given the sequence of hypothesis $H_{0,k} : \text{rank}(B) \leq k - 1$ with $k = 1, \dots, p - 1$, the rejection of these must be carried out in a sequential fashion such that once $H_{k,0}$ is rejected, all $H_{\gamma,0}$ for $\gamma \leq k$ should be rejected as well. Under such sequential testing framework of this kind, it is natural to choose $\text{rank}(B)$ to be the largest k that rejects $H_{0,k}$. The question here is how to choose the ‘stopping point’ for rejection.

One of the simplest methods is to choose the value k at which $H_{0,k}$ is rejected for the last time with a given level α :

$$\hat{\kappa}_{\text{simple}} = \max\{k \in \{1, \dots, p - 1\} : p_k \leq \alpha\},$$

which we refer as *SimpleStop*.

In this paper, instead of *SimpleStop*, we use *StrongStop*. This procedure takes sequential p-values as its input and controls family-wise error rate. When the p-values of the sequential tests are uniformly and independently distributed under the null, the *StrongStop* procedure controls the family-wise error rate under a given level of α (GSell et al., 2013)[Theorem 3]. For rank determination, by the nature of our hypothesis, $H_{0,\gamma} : \text{rank}(B) \leq \gamma - 1$ being true implies $H_{0,k} : \text{rank}(B) \leq k - 1$ also being true for all $k > \gamma$. The family-wise error rate control property in rank determination, therefore, becomes control of rank over-estimation with level α as follows:

$$\mathbb{P}(\hat{\kappa} > \kappa) \leq \alpha$$

where $\hat{\kappa}$ denotes the selected $\text{rank}(B) = \kappa$. The resulting procedure is as follows:

$$\hat{\kappa} = \max \left\{ k \in \{1, \dots, p - 1\} : \exp \left(\sum_{j=k}^{p-1} \frac{\log p_j}{j} \right) \leq \frac{\alpha k}{p - 1} \right\}.$$

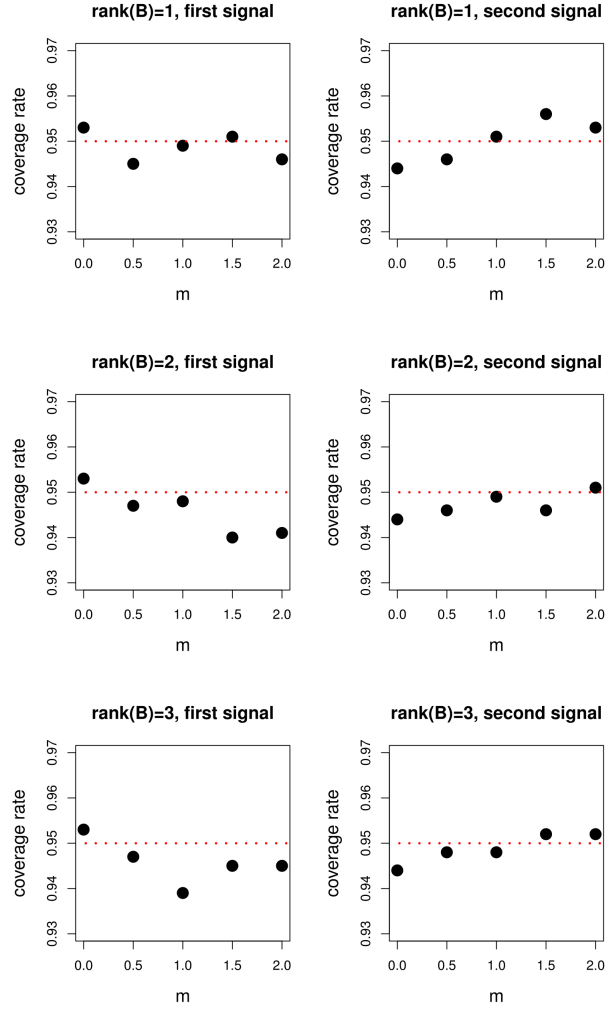


FIG 5. Coverage rate of $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ versus the magnitude m in (2.15) with $N = 50$ and $p = 10$. The first and the second signal correspond to $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ respectively. Coverage rate denotes the proportion of times that the constructed confidence intervals from $CI_k(\mathbb{S})$ covered the true parameter. From the top to the bottom, each row represents the case of the true rank(B) being 0 to 3. The columns illustrate the results regarding $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ from the left to the right.

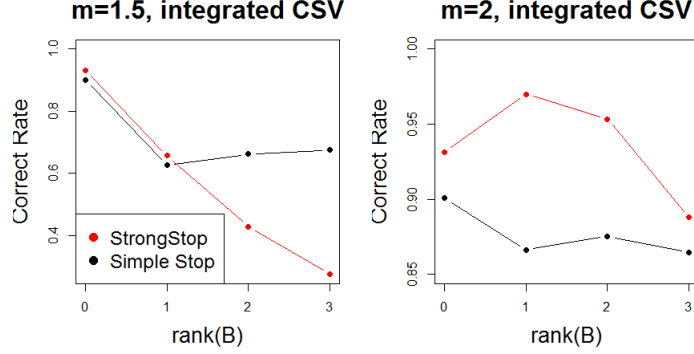


FIG 6. Rate of selecting the correct $\text{rank}(B)$ versus $\text{rank}(B)$ for the ICSV procedure with $N = 50$ and $p = 10$. From the left to the right, m increases from 1.5 to 2 where m determines the magnitude of the signals in (2.15). The black dots and the red dots represent the results from SimpleStop and StrongStop respectively. For both stopping rules, $\alpha = 0.1$ is used.

Here, p_k denotes the value of either $\mathbb{S}_{k,0}$ of the CSV or $\mathbb{V}_{k,0}$ of the ICSV, and conventionally $\max(\emptyset) = 0$. The independence of p-values from our proposed testing procedures for $H_{0,k}$ with $k > \kappa$ has yet been established; however *StrongStop* shows good performance for strong signals on simulated data.

Figure 6 illustrates the simulation results of *StrongStop* on p-values from the ICSV procedure with $\alpha = 0.1$. The simulation setting is the same as in Section 2.2.3 with $N = 50$ and $p = 10$. We compare *StrongStop* with *SimpleStop*. From Figure 6, we observe that for smaller signals, *SimpleStop* tends to choose the correct rank of B better than *StrongStop*. This can be explained by the strict overestimation control property of *StrongStop* which might lead to underestimation in low signal case. For strong signals, *StrongStop* is better at avoiding overfitting.

4. Estimating the Noise Level. For the testing procedures CSV and ICSV, and confidence interval $CI_k(\mathbb{S})$, we have assumed that the noise level σ^2 is known. In case the prior information of σ^2 is unavailable, the value of σ^2 needs to be estimated. In this section, we introduce a data-driven method for estimating σ^2 .

For the estimation of σ^2 , it is popular to assume that the rank of B is known. One of the simplest methods estimates σ^2 using mean of sum of

squared residuals by

$$\hat{\sigma}_{simple}^2 = \frac{1}{N(p - \kappa)} \sum_{j=\kappa+1}^p d_j^2$$

with known $\text{rank}(B) = \kappa$.

Instead of using the rank of B , [Gavish and Donoho \(2014\)](#) use the median of the singular values of Y as a robust estimator of σ^2 as follows:

$$(4.1) \quad \hat{\sigma}_{med}^2 = \frac{d_{med}^2}{N \cdot \mu_\beta}$$

where d_{med} is a median of the singular values of Y and μ_β is a median of a Marčenko-Pastur distribution with $\beta = N/p$. This estimator works under the assumption that $\text{rank}(B) \ll \min(N, p)$.

In this paper, we suggest three estimators, all of which make few assumptions. Our approach uses cross-validation and the sum of squared residuals as an extension of classical noise level estimator. For a fixed value of λ , we define our estimators $\hat{\sigma}_\lambda^2$, $\hat{\sigma}_{\lambda,df}^2$ and $\hat{\sigma}_{\lambda,df,c}^2$ as follows:

$$(4.2) \quad \begin{aligned} \hat{\sigma}_\lambda^2 &= \frac{1}{N \cdot p} \|Y - \hat{B}_\lambda\|_F^2 \\ \hat{\sigma}_{\lambda,df}^2 &= \frac{1}{N \cdot (p - df_{\hat{B}_\lambda})} \|Y - \hat{B}_\lambda\|_F^2 \\ \hat{\sigma}_{\lambda,df,c}^2 &= \frac{1}{N \cdot (p - c \cdot df_{\hat{B}_\lambda})} \|Y - \hat{B}_\lambda\|_F^2 \text{ for } c \in [0, 1] \end{aligned}$$

with

$$\begin{aligned} \hat{B}_\lambda &= \underset{B \in \mathbb{R}^{N \times p}}{\operatorname{argmin}} \frac{1}{2} \|Y - B\|_F^2 + \lambda \|B\|_* \\ df_{\hat{B}_\lambda} &= \sum_{k=1}^p 1\{l_{\lambda,k} > 0\} \end{aligned}$$

where $l_{\lambda,k}$ denotes the k^{th} singular value of \hat{B}_λ . The estimators $\hat{\sigma}_\lambda^2$ and $\hat{\sigma}_{\lambda,df}^2$ correspond to the ordinary mean squared residual, with the latter accounting for the degrees of freedom ([Reid, Tibshirani and Friedman, 2013](#)). With $c \in (0, 1)$, the estimator $\hat{\sigma}_{\lambda,df,c}^2$ lies between $\hat{\sigma}_\lambda^2$ and $\hat{\sigma}_{\lambda,df}^2$. We use cross-validation for choosing the appropriate value of the regularization parameter λ .

For these estimators, choosing an appropriate value for the regularization parameter λ is important, since \hat{B}_λ depends on λ . In penalized regression,

it is common to use cross-validation for this purpose, examining a grid of λ values. Unlike the regression setting, however, here there is no outcome variable and thus it is not clear how to make predictions on left-out data points.

In this paper, we use *softImpute* algorithm (Mazumder, Hastie and Tibshirani, 2010). In the presence of missing values in a given data matrix, *softImpute* carries out matrix completion with the following criterion:

$$(4.3) \quad \min_B \frac{1}{2} \|P_\Omega(Y) - P_\Omega(B)\|_F^2 + \lambda \|B\|_*$$

where Ω is an index set of observed data point with a function $P_\Omega(\cdot)$ such that $P_\Omega(Y)_{(i,j)} = Y_{i,j}$ if $(i,j) \in \Omega$ and 0 otherwise. We define the prediction error of the unobserved values Ω as follows:

$$err_\lambda(\Omega) = \|P_{\tilde{\Omega}}(Y) - P_{\tilde{\Omega}}(\hat{B}_\lambda^S(\Omega))\|_F^2$$

where $\hat{B}_\lambda^S(\Omega)$ denotes the estimator of B acquired from (4.3) and $\tilde{\Omega}$ denotes the index set of unobserved values ($\tilde{\Omega} = \Omega^c$). Using this prediction error, we carry out k-fold cross-validation, randomly generating k non-overlapping leave-out sets of size $\frac{N \cdot p}{k}$ from Y . For a grid of λ values, we compute the average of err_λ for each λ over the left-out data. We choose our λ to be the minimizer of the average err_λ as in usual cross-validation (see e.g. Hastie, Tibshirani and Friedman (2009)).

4.1. A study of noise level estimation. We illustrate simulation examples of noise level estimation of the proposed methods $\hat{\sigma}_\lambda^2$, $\hat{\sigma}_{\lambda,df}^2$ and $\hat{\sigma}_{\lambda,df,c}^2$ in (4.2) compared to $\hat{\sigma}_{med}^2$ in (4.1). For the estimator $\hat{\sigma}_{\lambda,df,c}^2$, we use $c = \frac{2}{3}$ in ad-hoc. These approaches do not require predetermined knowledge of $\text{rank}(B)$.

Simulation settings are the same as in Section 2.2.3 with $N = 50$ and $p = 10$. The true value of the noise level is $\sigma^2 = 1$ and for choosing λ , 20-fold cross-validation is used. Table 1 illustrates the simulation results for the three proposed estimators.

TABLE 1

Simulation results for estimating the noise level $\sigma^2 = 1$ with $N = 50$ and $p = 10$. We vary $\text{rank}(B)$ from 0 to 3, and m from 0.5 to 2.0. Each column represents the mean estimated value of σ^2 (“Est”), and standard error (“se”) of the corresponding estimator.

m	$\hat{\sigma}_{\lambda_{CV}}^2$		$\hat{\sigma}_{\lambda_{CV},df}^2$		$\hat{\sigma}_{\lambda_{CV},df,c}^2$		$\hat{\sigma}_{med}^2$	
	Est	se	Est	se	Est	se	Est	se
rank(B) = 0								
0.0	0.863	0.230	0.990	0.230	0.926	0.210	0.996	0.084
rank(B) = 1								
0.5	0.869	0.236	1.000	0.238	0.934	0.216	1.006	0.085
1.0	0.869	0.254	1.039	0.263	0.954	0.234	1.027	0.088
1.5	0.823	0.271	1.127	0.318	0.969	0.254	1.044	0.090
2.0	0.789	0.246	1.245	0.324	0.999	0.224	1.052	0.091
rank(B) = 2								
0.5	0.871	0.260	1.061	0.280	0.962	0.237	1.039	0.089
1.0	0.784	0.262	1.321	0.355	1.025	0.244	1.093	0.096
1.5	0.700	0.288	1.611	0.431	1.052	0.201	1.121	0.100
2.0	0.646	0.211	1.762	0.471	1.047	0.196	1.134	0.102
rank(B) = 3								
0.5	0.827	0.303	1.206	0.365	1.002	0.288	1.098	0.095
1.0	0.674	0.235	1.771	0.459	1.076	0.231	1.186	0.107
1.5	0.585	0.202	2.135	0.623	1.062	0.228	1.226	0.113
2.0	0.549	0.180	2.324	0.711	1.057	0.224	1.242	0.116

In this setting, $\hat{\sigma}_{\lambda_{CV}}^2$ decreases with larger $\text{rank}(B)$ and signals while $\hat{\sigma}_{\lambda_{CV},df}^2$ and $\hat{\sigma}_{med}^2$ increases. For large $\text{rank}(B)$ and signals, $\hat{\sigma}_{\lambda_{CV},df,c}^2$ shows good results, as compared to other methods. The poor performance of $\hat{\sigma}_{\lambda_{CV},df}^2$ may be caused by the use of an improper definition of df , the degrees of freedom. Following the definition of degrees of freedom by [Efron et al. \(2004\)](#), our simulation result shows that the number of non-zero singular values does not coincide with degrees of freedom under our setting. Further investigation into the degrees of freedom is needed in future work.

The competing method $\hat{\sigma}_{med}^2$ consistently shows a small standard deviation. However, with large $\text{rank}(B)$, especially when $\text{rank}(B) \geq p/2$, the procedures over-estimates σ^2 due to the effect of the signals.

5. Additional Examples. We discuss additional examples in this section. Section [5.1](#) presents results of the proposed methods when the esti-

mated noise level is used. In Section 5.2, hypothesis testing results with non-Gaussian noise are illustrated. Section 5.3 shows results on some real data.

5.1. *Simulation examples with unknown noise level.* In this section, we illustrate the results when estimated σ^2 value is used on simulated data. For the estimation of the noise level, we use $\hat{\sigma}_{\lambda_{CV},df,c}^2$ and $\hat{\sigma}_{med}^2$ which showed good performance in Section 4.1. As in Section 4.1, for the estimator $\hat{\sigma}_{\lambda_{CV},df,c}^2$, 20-fold cross-validation and $c = 2/3$ is used. The simulation settings are the same as in Section 2.2.3 with $N = 50$ and $p = 10$. We investigate the case of $m = 1.5$.

Table 2 illustrates the estimated σ^2 values we used for the testing procedure. Figure 7 shows quantile-quantile plots of observed p-values obtained from using the estimated σ^2 versus the expected (uniform) quantiles. In quantile-quantile plots, both estimators of σ^2 show reasonable results in general, and for large $\text{rank}(B)$, $\hat{\sigma}_{\lambda_{CV},df,c}^2$ shows better result than $\hat{\sigma}_{med}^2$. In terms of coverage rate of confidence interval, we can see from Figure 8 that $\hat{\sigma}_{med}^2$ dominates for all cases, which might be due to small standard deviation of $\hat{\sigma}_{med}^2$ estimator. The estimation of $\text{rank}(B)$ is presented in Table 3. For the estimation, *StrongStop* is applied to the CSV p-values with level $\alpha = 0.05$. The estimation performance seems to vary with the quality of the estimate of σ^2 .

5.2. *Simulation example with non-Gaussian noise.* Our testing procedure is based on an assumption of Gaussian noise. Here we investigate the performance of the CSV test on simulated examples with the normality assumption of noise violated. Simulation settings are the same as in Section 2.2.3 with $N = 50$ and $p = 10$ except for the noise distribution. We study the case of $\text{rank}(B) = 1$ with $m = 1.5$ along with two sorts of noise distribution: heavy tailed and right skewed. The heavy tailed noise is drawn from $\sqrt{\frac{3}{5}}t_5$

TABLE 2

Simulation results for estimating the noise level $\sigma^2 = 1$ at $m = 1.5$ with $N = 50$ and $p = 10$. We vary the rank of B from 0 to 3. Shown are the mean estimated noise level (“Est”) and standard error (“se”) of the corresponding estimators.

rank(B)	$\hat{\sigma}_{\lambda_{CV},df,c}^2$		$\hat{\sigma}_{med}^2$	
	Est	se	Est	se
0	0.926	0.210	0.996	0.084
1	0.969	0.254	1.044	0.090
2	1.052	0.201	1.121	0.100
3	1.062	0.228	1.226	0.113

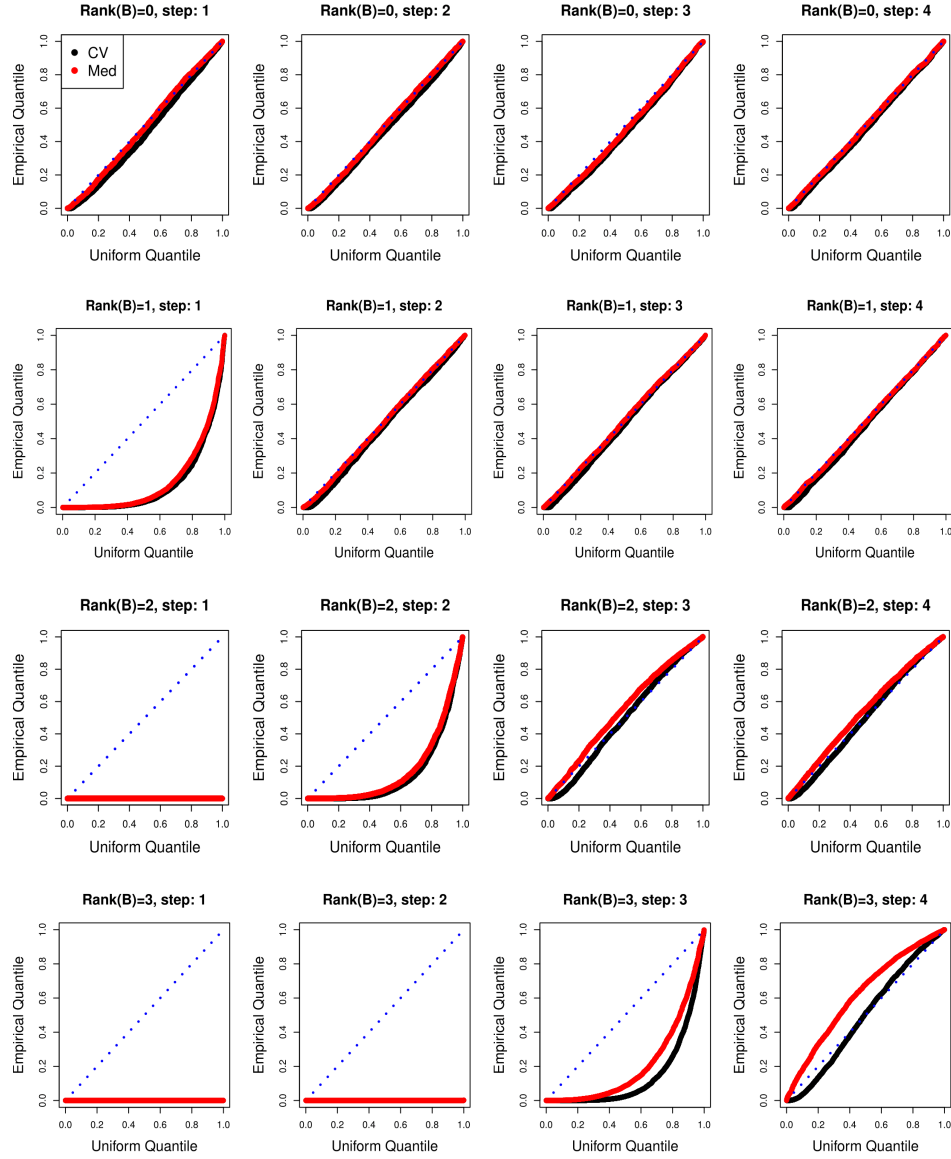


FIG 7. Quantile-quantile plots of the empirical quantiles of p -values with estimated σ^2 versus the uniform quantiles at $m = 1.5$ with $N = 50$ and $p = 10$. Empirical p -values are from the CSV test. Each row represents $\text{rank}(B) = 0$ to $\text{rank}(B) = 3$ from the top to the bottom. Each column represents the 1st test to the 4th test from the left to the right. The black dots (CV) and the red dots (med) represent p -values from using $\hat{\sigma}_{\lambda_{CV}, df, c}^2$ and $\hat{\sigma}_{med}^2$ respectively. For the estimator $\hat{\sigma}_{\lambda_{CV}}^2$, $c = 2/3$ is used.

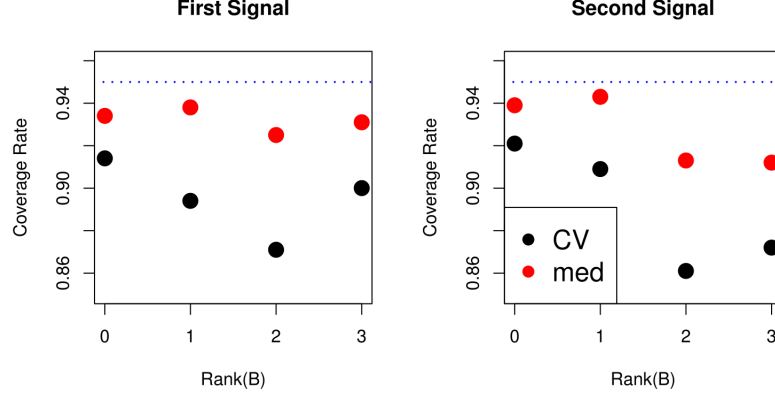


FIG 8. Coverage rate versus $\text{rank}(B)$ of the first two signal parameters at $m = 1.5$ using estimated σ^2 with $N = 50$ and $p = 10$. Coverage rate denotes proportion of times that the constructed confidence interval from $CI_k(\mathbb{S})$ covered the true parameter. The black dots (CV) and the red dots (med) represent the estimation using $\hat{\sigma}_{\lambda_{CV}, df, c}^2$ and $\hat{\sigma}_{med}^2$ respectively.

TABLE 3

Simulation results for selecting an exact $\text{rank}(B)$ from the CSV test using estimated σ^2 at $m = 1.5$ with $N = 50$ and $p = 10$. StrongStop is applied to sequential p -values with $\alpha = 0.05$. We vary the rank of B from 0 to 3. Shown are the rate of selecting the correct rank of B (“Rate”) and mean squared error (“MSE”) using the corresponding estimator of σ^2 .

$\text{rank}(B)$	$\hat{\sigma}_{\lambda_{CV}, df, c}^2$		$\hat{\sigma}_{med}^2$	
	Rate	MSE	Rate	MSE
0	0.894	3.704	0.948	0.063
1	0.455	4.243	0.486	0.514
2	0.231	1.080	0.157	0.853
3	0.181	0.845	0.026	0.975

where t_5 denotes t-distribution with degrees of freedom=5, and the right skewed noises are drawn from $\sqrt{\frac{3}{10}}t_5 + \sqrt{\frac{1}{2}}(\exp(1) - 1)$ where $\exp(1)$ denotes the exponential distribution with mean=1. In each case, noise entries are drawn i.i.d., and known value of $\sigma^2 = 1$ is used.

Figure 9 shows quantile-quantile plots of the observed p -values versus expected (uniform) quantiles. The top panels correspond to the heavy tailed noise from $\sqrt{\frac{3}{5}}t_5$ and the bottom panels correspond to the right skewed noise from $\sqrt{\frac{3}{10}}t_5 + \sqrt{\frac{1}{2}}(\exp(1) - 1)$. Quantile-quantile plots from both types of noise show that the p -values deviate slightly from a uniform distribution

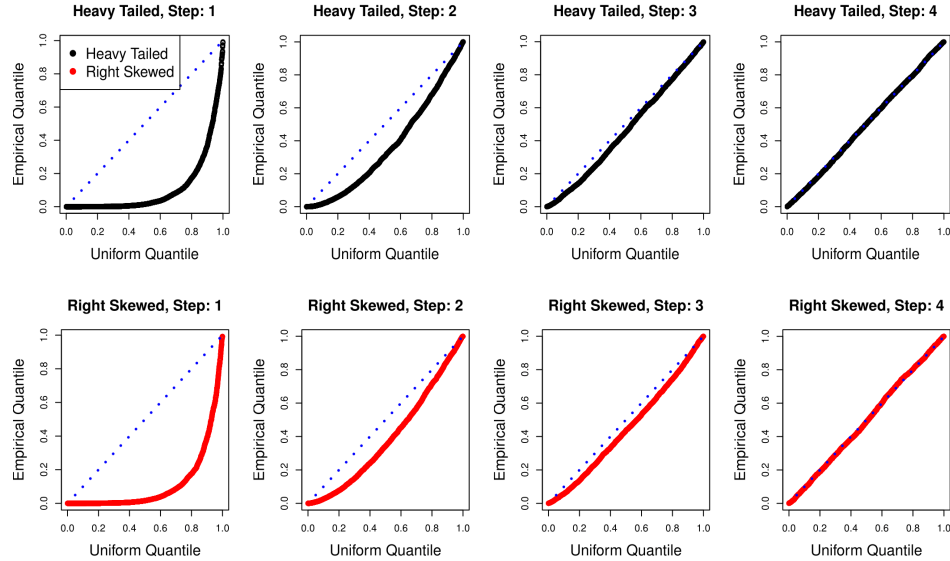


FIG 9. Quantile-quantile plots of empirical quantile of p -values versus uniform quantile when noises are drawn i.i.d. from the heavy tailed or the right skewed distribution with known value of $\sigma^2 = 1$ with $N = 50$ and $p = 10$. The true rank of B is $\text{rank}(B) = 1$ with $m = 1.5$. The top panels correspond to the heavy tailed noise from $\sqrt{\frac{3}{5}}t_5$ and the bottom panels correspond to the right skewed noise from $\sqrt{\frac{3}{10}}t_5 + \sqrt{\frac{1}{2}}(\exp(1) - 1)$. Each column represents the 1st test to the 4th test from the left to the right.

under the null $H_{0,2} : \text{rank}(B) \leq 1$. For further steps, p -values lie closer to the reference line.

The nonconformity shown in early steps under the null hypothesis is not surprising considering the construction of the CSV procedure based on Gaussian noise. In future work, we will investigate whether the procedures introduced here can be extended to a method robust to non-normality using data-oriented method such as bootstrap.

5.3. Real data example. In this section, we revisit the real data example mentioned in Figure 1. We apply the CSV test to the data of examination marks of 88 students on 5 different topics of Mechanics, Vectors, Algebra, Analysis and Statistics (Mardia, Kent and Bibby, 1979, p. 3-4), and determine the number of principal components to retain for PCA.

In this data, Mechanics and Vectors were closed book exams while the other topics were open book exam. We use $\hat{\sigma}_{\lambda_{CV}, df, c}^2 = 75.957$ and $\hat{\sigma}_{med}^2 =$

131.332 for the estimated noise level. For $\hat{\sigma}_{\lambda_{CV},df,c}^2$, 20-fold cross-validation is used with $c = 2/3$. The CSV test results are presented in Table 4. The estimated $\text{rank}(B)$ is 2 with $\hat{\sigma}_{\lambda_{CV},df,c}^2$ estimator and 1 with $\hat{\sigma}_{med}^2$ with level $\alpha = 0.05$ using *StrongStop*. Thus, in PCA we may use one or two principal components depending on our choice of the noise level. In this example, one or two principal components makes sense as these 5 topics cover closely related areas.

TABLE 4

P-values at each step (“Step”) and the selected number of principal components (“Selected”) from the CSV test using the estimated noise level on the examination marks of 88 students on five different topics (Mechanics, Vectors, Algebra, Analysis and Statistics). StrongStop is applied to select the number of principal components at level $\alpha = 0.05$. The estimated value of σ^2 is 75.957 for the estimator $\hat{\sigma}_{\lambda_{CV},df,c}^2$ and 131.332 for the estimator $\hat{\sigma}_{med}^2$.

Step	$\hat{\sigma}_{\lambda_{CV},df,c}^2$	$\hat{\sigma}_{med}^2$
1	0.000	0.000
2	0.000	0.015
3	0.001	0.573
4	0.093	0.940
<hr/>		
	$\hat{\sigma}_{\lambda_{CV},df,c}^2$	$\hat{\sigma}_{med}^2$
Selected	2	1

6. Conclusions. In this paper, we have proposed distribution-based methods for choosing the number of principal components of a data matrix. We have suggested novel methods both for hypothesis testing and the construction of confidence intervals of the signals. The methods have exact type I error control and show promising results in simulated examples. We have also introduced data-based methods for estimating the noise level.

There are many topics that deserve further investigation. In following studies, the analysis of power of the suggested tests and the width of the constructed confidence interval will be investigated. Also, application of the methods to high dimensional data using numerical approximations will be explored. For multiple hypothesis testing corrections to be properly applied, we will study the dependence structure of the p-values in different steps. In addition, for robustness to non-Gaussian noise, bootstrap versions of this procedure will be investigated. Future work may involve a notion of degrees of freedom of the spectral estimator of the signal matrix. These extensions may lead to improvement in noise level estimation.

Variations of these procedures can potentially be applied to canonical correlation analysis (CCA) and linear discriminant analysis (LDA), and these are topics for future work.

APPENDIX

A.1. Lemma 1 and Proof.

LEMMA 1. For $Y \in \mathbb{R}^{N \times p}$ where $Y \sim N(B, \sigma^2 I_N \otimes I_p)$, we write the singular value decomposition of Y by $Y = U_Y D_Y V_Y^T$ where $D_Y = \text{diag}(y_1, \dots, y_p)$ with $y_1 \geq \dots \geq y_p \geq 0$. Without loss of generality we assume $N \geq p$. As in Section 2.2.1, U_Y^r, U_Y^{-r}, V_Y^r and V_Y^{-r} denote submatrices of U_Y and V_Y . Writing the law by $\mathcal{L}(\cdot)$, the density of $\mathcal{L}(\cdot)$ by $d\mathcal{L}$, $\text{diag}(y_1, \dots, y_r)$ by D_Y^r , and $\text{diag}(y_{r+1}, \dots, y_p)$ by D_Y^{-r} , we have

$$\frac{d\mathcal{L}(D_Y^{-r}, U_Y^{-r}, V_Y^{-r} | U_Y^r, V_Y^r, D_Y^r)}{d\mathcal{L}(D_Y^{-r}, U_Y^{-r}, V_Y^{-r} | U_Y^r, V_Y^r, B = 0)} \propto e^{\text{tr}(U_Y^{-r} D_Y^{-r} V_Y^{-r})^T B}.$$

PROOF. Without loss of generality, we assume $\sigma^2 = 1$. Writing the density function of Y as $p_{Y,B}(\cdot)$, when $B = 0$, we have

$$\begin{aligned} d\mathcal{L}(D_Y, U_Y, V_Y | B = 0) &= p_{Y,0}(U_Y D_Y V_Y^T) |J(D_Y, U_Y, V_Y)| \\ &\propto e^{-\frac{1}{2} \text{tr}((U_Y D_Y V_Y^T)^T (U_Y D_Y V_Y^T))} \prod_{i=1}^p y_i^{N-p} \prod_{i < j} (y_i^2 - y_j^2) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \\ \text{(A.1)} \quad &\propto F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \end{aligned}$$

where

$$\text{(A.2)} \quad F_{N,p}(y_1, \dots, y_p) = C_{N,p} e^{-\frac{\sum_{i=1}^p y_i^2}{2}} \prod_{i=1}^p y_i^{N-p} \prod_{i < j} (y_i^2 - y_j^2)$$

denotes the density of singular values of a $N \times p$ Gaussian matrix from $N(0, I_N \otimes I_p)$ with a normalizing constant $C_{N,p}$, $J(\cdot)$ is the Jacobian of mapping $Y \rightarrow (D_Y, U_Y, V_Y)$, $\mu_{N,p}$ denotes Haar probability measure on $O(N)$ under the map $U \rightarrow U^p$, and $\mu_{p,p}$ denotes Haar probability measure on $O(p)$. Here, $O(m)$ denotes an orthogonal group of $m \times m$ matrices.

Note that, $d\mathcal{L}(D_Y, U_Y, V_Y | B = 0) \propto F_{N,p} d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y)$ can be alternatively derived from $D_Y \perp (U_Y, V_Y)$, and $U_Y \perp V_Y$. U_Y and V_Y have Haar probability measure on $O(n)$ under mapping $U \rightarrow U^p$ and on $O(p)$ respectively, since the distribution of U_Y and V_Y are rotation invariant. The determinant of the Jacobian $|J(D_Y, U_Y, V_Y)|$ can be calculated either explicitly or from this relation.

As in the case of $B = 0$, for general B we have,

$$\begin{aligned}
d\mathcal{L}(D_Y, U_Y, V_Y) &= p_{Y,B}(U_Y D_Y V_Y^T) |J(D_Y, U_Y, V_Y)| \\
&\propto e^{-\frac{1}{2} \text{tr}((U_Y D_Y V_Y^T - B)^T (U_Y D_Y V_Y^T - B))} |J(D_Y, U_Y, V_Y)| \\
&\propto e^{\text{tr}(U_Y D_Y V_Y^T)^T B} \cdot e^{-\frac{1}{2} \text{tr} B^T B} p_{Y,0}(U_Y D_Y V_Y^T) |J(D_Y, U_Y, V_Y)| \\
\text{(A.3)} \quad &\propto e^{\text{tr}(U_Y D_Y V_Y^T)^T B} \cdot F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \text{ from (A.1)}.
\end{aligned}$$

Therefore, with given U_Y^r , V_Y^r , and D_Y^r , we have

$$\begin{aligned}
d\mathcal{L}(D_Y, U_Y, V_Y | U_Y^r, V_Y^r, D_Y^r) &\propto e^{\text{tr}(U_Y D_Y V_Y^T)^T B} \cdot F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \\
&= e^{\text{tr}(U_Y^r D_Y^r V_Y^{rT})^T B} \cdot e^{\text{tr}(U_Y^{-r} D_Y^{-r} V_Y^{-rT})^T B} \cdot F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \\
&\propto e^{\text{tr}(U_Y^{-r} D_Y^{-r} V_Y^{-rT})^T B} F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y)
\end{aligned}$$

since $Y = U_Y D_Y V_Y^T = U_Y^r D_Y^r V_Y^{rT} + U_Y^{-r} D_Y^{-r} V_Y^{-rT}$, and thus

$$\begin{aligned}
d\mathcal{L}(D_Y, U_Y, V_Y | U_Y^r, V_Y^r, D_Y^r) &\propto \\
\text{(A.4)} \quad &e^{\text{tr}(U_Y^{-r} D_Y^{-r} V_Y^{-rT})^T B} F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y)
\end{aligned}$$

and equivalently,

$$\frac{d\mathcal{L}(D_Y^{-r}, U_Y^{-r}, V_Y^{-r} | U^r, V^r, D^r)}{d\mathcal{L}(D_Y^{-r}, U_Y^{-r}, V_Y^{-r} | U^r, V^r, D^r, B = 0)} \propto e^{\text{tr}(U_Y^{-r} D_Y^{-r} V_Y^{-rT})^T B}.$$

□

A.2. Proof of Theorem 2.1.

PROOF. We follow the notations in Lemma 1. Without loss of generality we assume $\sigma^2 = 1$. When $U_Y^{-(k-1)} U_Y^{-(k-1)T} B V_Y^{-(k-1)} V_Y^{-(k-1)T} = 0_{N \times p}$, then $e^{\text{tr}(U_Y^{-(k-1)} D_Y^{-(k-1)} V_Y^{-(k-1)T})^T B} = 1$. Thus, from Lemma 1 and (A.4), we have

$$\begin{aligned}
d\mathcal{L}(D_Y^{-(k-1)}, U_Y^{-(k-1)}, V_Y^{-(k-1)} | D_Y^{(k-1)}, U_Y^{(k-1)}, V_Y^{(k-1)}) & \\
&\propto d\mathcal{L}(D_Y^{-(k-1)}, U_Y^{-(k-1)}, V_Y^{-(k-1)} | D_Y^{(k-1)}, U_Y^{(k-1)}, V_Y^{(k-1)}, B = 0) \\
&\propto F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y)
\end{aligned}$$

where $F_{N,p}$ denotes the density of singular values of a $N \times p$ Gaussian matrix from $N(0, I_N \otimes I_p)$ as in (A.2). Therefore, we have

$$\begin{aligned}
d\mathcal{L}(y_k | y_i, i \neq k, U_Y^{(k-1)}, V_Y^{(k-1)}) & \\
\text{(A.5)} \quad &\propto \prod_{i \neq k} |y_i^2 - y_k^2| e^{-\frac{y_k^2}{2}} y_k^{N-p} I_{\{y_k \in (y_{k+1}, y_{k-1})\}} d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y).
\end{aligned}$$

Note that (A.5) is the integrand of the CSV test statistic $\mathbb{S}_{k,0}$ with some cancellation from the fraction, and thus $\mathbb{S}_{k,0}$ is the probability of having the k^{th} singular value that is bigger than the observed one given all the other singular values, $U_Y^{(k-1)}$ and $V_Y^{(k-1)}$.

Thus, if the observation Y is actually generated under $U_Y^{-(k-1)} U_Y^{-(k-1)^T} B V_Y^{-(k-1)} V_Y^{-(k-1)^T} = 0_{N \times p}$, then the density of the k^{th} singular value becomes (A.5) up to a constant. Consequently, when writing the observed singular values by d_1, \dots, d_p , and the observed singular value decomposition of Y by $U_Y D_Y V_Y^T$ without confusion,

$$\begin{aligned} \mathbb{S}_{k,0} &= \int_{d_k}^{d_{k-1}} d\mathcal{L}(y_k | d_i, i \neq k, U_Y^{(k-1)}, V_Y^{(k-1)}) dy_k \\ &= P(y_k \geq d_k | d_i, i \neq k, U_Y^{(k-1)}, V_Y^{(k-1)}) \\ &\sim \text{Unif}(0, 1). \end{aligned}$$

The denominator of $\mathbb{S}_{k,0}$ works as a normalizing constant. \square

A.3. Proof of Theorem 2.2.

PROOF. We follow the notation of Lemma 1 and without loss of generality, assume $\sigma^2 = 1$. Under $U^{-(k-1)} U^{-(k-1)^T} B V^{-(k-1)} V^{-(k-1)^T} = 0_{N \times p}$, we have $e^{\text{tr}(U_Y^{-(k-1)} D_Y^{-(k-1)} V_Y^{-(k-1)^T})^T B} = 1$. Then, as in Theorem 2.1, we have

$$\begin{aligned} &d\mathcal{L}(D_Y^{-(k-1)}, U_Y^{-(k-1)}, V_Y^{-(k-1)} | D_Y^{(k-1)}, U_Y^{(k-1)}, V_Y^{(k-1)}) \\ &\propto F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \end{aligned}$$

and therefore,

$$\begin{aligned} &d\mathcal{L}(y_k, U_Y^{-(k-1)}, V_Y^{-(k-1)} | D_Y^{(k-1)}, U_Y^{(k-1)}, V_Y^{(k-1)}) \\ &\propto \left(\int \dots \int F_{N,p}(y_1, \dots, y_p) dy_p \dots dy_{k+1} \right) \cdot d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \\ \text{(A.6)} \quad &\propto g(y_k) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \end{aligned}$$

where $g(\cdot)$ is defined in (2.14). Note that (A.6) is the integrand of the ICSV test statistic $\mathbb{V}_{k,0}$, and $\mathbb{V}_{k,0}$ is the conditional survival function of the k^{th} singular value given $D_Y^{(k-1)}$, $U_Y^{(k-1)}$ and $V_Y^{(k-1)}$.

Thus, if the observation is under $U_Y^{-(k-1)} U_Y^{-(k-1)^T} B V_Y^{-(k-1)} V_Y^{-(k-1)^T} = 0_{N \times p}$, then its k^{th} singular value has the conditional density of (A.6) upto a constant. Consequently, when writing the observed k^{th} singular value by d_k ,

and the observed singular value decomposition of Y by $U_Y D_Y V_Y^T$ without confusion,

$$\begin{aligned}\mathbb{V}_{k,0} &= \int_{d_k}^{d_{k-1}} d\mathcal{L}(y_k, U_Y^{-(k-1)}, V_Y^{-(k-1)} | D_Y^{(k-1)}, U_Y^{(k-1)}, V_Y^{(k-1)}) dy_k \\ &= P(y_k \geq d_k | D_Y^{(k-1)}, U_Y^{(k-1)}, V_Y^{(k-1)}) \\ &\sim \text{Unif}(0, 1).\end{aligned}$$

The denominator of $\mathbb{V}_{k,0}$ works as a normalizing constant. \square

A.4. Proof of Theorem 2.3.

PROOF. We follow the notations in Lemma 1 and without loss of generality, assume $\sigma^2 = 1$. From (A.3), we have

$$\begin{aligned}d\mathcal{L}(D_Y, U_Y, V_Y) &\propto e^{\text{tr}(U_Y D_Y V_Y^T)^T B} \cdot F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \\ &= e^{\sum_{k=1}^p y_k \tilde{\Lambda}_k} \cdot F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y)\end{aligned}$$

where $\tilde{\Lambda}_k = \langle U_{Y,k} V_{Y,k}^T, B \rangle$ as defined in (2.16). Therefore, we have

$$\begin{aligned}d\mathcal{L}(y_k, |Y_i, i \neq k, U_Y, V_Y) &\propto e^{y_k \tilde{\Lambda}_k} \cdot F_{N,p}(y_1, \dots, y_p) d\mu_{N,p}(U_Y) d\mu_{p,p}(V_Y) \\ (A.7) \quad &\propto e^{-\frac{1}{2}(y_k - \tilde{\Lambda}_k)^2} y_k^{N-p} \prod_{j \neq k}^p |y_k^2 - y_j^2| 1\{0 \leq y_p \leq \dots \leq y_1\}.\end{aligned}$$

As (A.7) is the integrand of $\mathbb{S}_{k, \tilde{\Lambda}_k}$, when writing the observed singular values by d_1, \dots, d_p , and the observed singular value decomposition of Y by $U_Y D_Y V_Y^T$ without confusion, we have

$$\begin{aligned}\mathbb{S}_{k, \tilde{\Lambda}_k} &= \int_{d_k}^{d_{k-1}} d\mathcal{L}(y_k, |d_i, i \neq k, U_Y, V_Y) dy_k \\ &\sim \text{Unif}(0, 1).\end{aligned}$$

Here, the denominator of $\mathbb{S}_{k, \tilde{\Lambda}_k}$ works as a normalizing constant. \square

ACKNOWLEDGEMENTS

We would like to thank Boaz Nadler and Iain Johnstone for helpful conversations. Robert Tibshirani was supported by NSF grant DMS-9971405 and NIH grant N01-HV-28183.

REFERENCES

- CAI, J.-F., CANDÈS, E. J. and SHEN, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20** 1956–1982. [MR2600248 \(2011c:90065\)](#)
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. [MR2060166 \(2005d:62116\)](#)
- GAVISH, M. and DONOHO, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory* **60** 5040–5053. [MR3245370](#)
- GSELL, M. G., WAGER, S., CHOULDECHOVA, A. and TIBSHIRANI, R. (2013). Sequential Selection Procedures And False Discovery Rate Control. Preprint. Available at [arXiv:1309.5352](#).
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The elements of statistical learning*, second ed. *Springer Series in Statistics*. Springer, New York Data mining, inference, and prediction. [MR2722294 \(2012d:62081\)](#)
- JAMES, A. T. (1964). Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Statist.* **35** 475–501. [MR0181057 \(31 #5286\)](#)
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961 \(2002i:62115\)](#)
- JOLLIFFE, I. T. (2002). *Principal component analysis*, second ed. *Springer Series in Statistics*. Springer-Verlag, New York. [MR2036084 \(2004k:62010\)](#)
- JOSSE, J. and HUSSON, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Statist. Data Anal.* **56** 1869–1879. [MR2892383](#)
- KRITCHMAN, S. and NADLER, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems* **94** 19–32.
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. [MR560319 \(81h:62003\)](#)
- MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. [MR2719857 \(2011m:62184\)](#)
- MUIRHEAD, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons, Inc., New York Wiley Series in Probability and Mathematical Statistics. [MR652932 \(84c:62073\)](#)
- NADLER, B. (2008). Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. [MR2485013 \(2010g:62190\)](#)
- REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2013). A study of error variance estimation in lasso regression. Preprint. Available at [arXiv:1311.5274](#).
- TAYLOR, J., LOFTUS, J. and TIBSHIRANI, RYAN (2013). Tests in adaptive regression via the Kac-Rice formula. Preprint. Available at [arXiv:1308.3020](#).
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242 \(96j:62134\)](#)

Y. CHOI
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: yunjin@stanford.edu

J. TAYLOR
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: jonathan.taylor@stanford.edu

R. TIBSHIRANI
DEPARTMENT OF HEALTH, RESEARCH & POLICY
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: tibs@stanford.edu