A Scalable Approach to Estimating the Rank of High-Dimensional Data

Wenlan Zang,¹, a) Jen-hwa Chu,¹, b) and Michael J. Kane¹

Department of Biostatistics, Yale University, New Haven, CT 06510, USA

(Dated: 2 August 2021)

A key challenge to performing effective analyses of high-dimensional data is finding a signal-rich, low-dimensional representation. For linear subspaces, this is generally performed by decomposing a design matrix (via eigenvalue or singular value decomposition) into orthogonal components, and then retaining those components with sufficient variations. This is equivalent to estimating the rank of the matrix and deciding which components to retain is generally carried out using heuristic or ad-hoc approaches such as plotting the decreasing sequence of the eigenvalues and looking for the “elbow” in the plot. While these approaches have been shown to be effective, a poorly calibrated or misjudged elbow location can result in an overabundance of noise or an under-abundance of signal in the low-dimensional representation, making subsequent modeling difficult. In this article, we propose a latent-space-construction procedure to estimate the rank of the detectable signal space of a matrix by retaining components whose variations are significantly greater than random matrices, of which eigenvalues follow a universal Marchenko-Pastur (MP) distribution.

a) wenlan.zang@yale.edu

b) Also at: Section of Pulmonary, Critical Care and Sleep Medicine (PCCSM), Yale University, New Haven, CT 06510, USA
I. INTRODUCTION AND BACKGROUND

High-dimensional data analysis, in which the number of variables or features in a data set is larger than the number of samples, is common in signal processing, computer vision, and genomic studies (White and Smyth, 2005; Yang, 1996). For these types of data sets, it is generally assumed that there is a low-dimensional and information-rich subspace within the data set (Hoff, 2007). Subsequent methods prescribe a variety of dimension-reduction techniques to find a linear vector subspace that provides a low-dimensional representation of high-dimensional data and retaining those components of the data carrying predictive (or otherwise significant) information. A standard construction for this class of problems considers decomposing an $n \times p$ matrix $X$ into a rank $K$ ($K \ll p$) signal matrix $S$ plus a noise matrix $N$. The determination of $K$ is paramount for dimension-reduction problems to avoid over- or under-fitting subsequent models (Carreira-Perpinán, 1997). However, due to the challenge of dimension estimation and relatively few effective approaches proposed, many models take rank $K$ as an analyst-specified parameter based on ill-defined or ad-hoc methods. For example, a standard method of estimating rank $K$ is to retain components that explain 70% of total variance or to plot the decreasing sequence of the matrix’s eigenvalues (i.e., scree plot) and having an individual pick the point at which the rate of decrease itself drops significantly (i.e., “elbow”) (Yong et al., 2013). The eigenvalue index at which the elbow occurs is denoted as $K$. An alternative approach (jackstraw) implements a statistical test to identify features with a statistically significant association with any linear combination of principal components (PCs) of interest based on a constructed null distribution by randomly
A Scalable Approach to Estimating the Rank of High-Dimensional Data

permuting the data set (Chung and Storey, 2015). Although jackstraw is widely applied in dimension estimation, it requires the number of PCs that capture “systematic variation from latent variables” as an input for the algorithm, and the test results from jackstraw only provide a subset of suggested cutoffs from input PCs instead of a clear dimension-estimation cutoff (Chung and Storey, 2015). Furthermore, jackstraw is computationally intensive and time-consuming when applied to large data sets.

Several attempts have been made previously to solve the rank estimation problem. Hoff (2007) proposed using the Bayesian rank-estimation procedure based on the signal-plus-Gaussian-noise model, providing prior and posterior distribution of rank \( K \) and posterior estimation of right- and left-singular vectors of matrix \( X \). Posterior inferences can be drawn from expectation \( E_X[p(K|X)] \) and rank can be determined by the posterior mode of \( p(K|X) \). The major weakness that prohibits application of Hoff’s method to large data sets (e.g., 1,000 nodes) is the computational complexity of the markov chain monte carlo (MCMC) algorithm (Hoff, 2007). In addition, in the markov chain, Hoff (2007) sampled eigenvalues and the corresponding eigenvectors from the full condition distribution instead of the marginal distribution. As a result, the probabilities of sampling nonzero eigenvalues do not mix well across different ranks of the mean matrix, as eigenvalues and the corresponding eigenvectors are dependent (Hoff, 2007). Recently, Luo and Li (2016) introduced a ladle estimator that combined the decreasing sequence of eigenvalues with the increasing pattern of variability of eigenvector directions. They observed that bootstrap variability of the eigenvectors decrease as the eigenvalues get further apart and increase as the eigenvalues get closer. Luo and Li (2016) proposed the ladle estimator based on this observation. They
demonstrated the consistency of the ladle estimator and its superior performance in solving various dimension-reduction problems in a low-dimensional setting. While Hoff’s method and ladle have been shown to be effective in several settings, if the focus is on large or high-dimensional data, then these two methods may not be the best choice for rank estimation.

We propose a procedure to estimate the rank of the detectable signal space of a high-dimensional matrix using the signal-plus-noise model under the assumption that the eigenvalues of the noise matrix follow the universal behavior described in Aparicio et al. (2018). Under this construction, the expected spectrum of the random matrix $N$ follows a universal MP distribution, which can be compared to the eigenvalues of $X$ (Marčenko and Pastur, 1967). In this context, components carrying detectable signal can be characterized as those whose eigenvalues are significantly larger than expected under the MP distribution. This construction provides a method for estimating the dimension of a data set and extracting those components of the data that may not be attributable to randomness. The major contribution of this chapter is to present this novel and practical method, which results in improved accuracy, robustness, and computational efficiency, especially for high-dimensional data.

The paper is organized as follows: section II describes the framework setup and provides a distributional signal-to-ratio bound. Section III illustrates a procedure of estimating detectable signal dimension and demonstrates its performance in relation to other standard approaches, including the ladle and Hoff’s estimators. Section IV presents an application of dimension-estimation method in scRNA-seq data analysis. Section V concludes with a discussion.
II. A BOUND ON THE DETECTABLE SIGNAL DIMENSION

A. Signal-Plus-Noise Model

Let $X_{n \times p}$ be a matrix with mean-centered columns consisting of a low-rank matrix $S$ plus a random matrix $N$. Without loss of generality, we assume $p > n$ for high-dimensional data. Following the construction of Livan et al. (2018), assume matrix $N$ is formed by $n \times p$ samples from $p$ random variables $[x_1, ..., x_p]$ drawn from an unknown joint distribution $p(x)$. Let $N_{ij}$ be independent and identically distributed (i.i.d.) entries with $E(N_{ij}) = 0$ and $E(N_{ij}^2) = \sigma^2$ (Götze et al., 2004). The scaled inner product of $X$ can be written as

$$
\frac{1}{p} XX^T = \frac{1}{p} (S + N)(S + N)^T
$$

$$
= \frac{1}{p} SS^T + \frac{1}{p} NN^T + \frac{1}{p} SN^T + \frac{1}{p} NS^T.
$$

We can specify finding the rank of $X$ in terms of eigenvalue decomposition as follows:

$$
XX^T = U_X \Lambda_X U_X^T,
$$

where $u_{X_i}$ denotes the $i$th column of the eigenvector $U_X$ and $\Lambda_X$ is a diagonal matrix with diagonal entries $\{\lambda_{X_1}, ..., \lambda_{X_n}\}$ as the non-negative eigenvalues sorted in a non-increasing order. When $n \geq p$, the $\Lambda_X = \text{diag}\{\lambda_{X_1}, ..., \lambda_{X_p}, 0, ..., 0\}$, where $n - p$ eigenvalues equals zero. Thus, we can rewrite the scaled eigenvalues of matrix $XX^T$ in terms of the following
A Scalable Approach to Estimating the Rank of High-Dimensional Data

approximation:

\[
\frac{1}{p} \Lambda_X(X) = \frac{1}{p} U_X^T (XX^T) U_X \\
= \frac{1}{p} U_X^T (SS^T + NN^T + SN^T + NS^T) U_X \\
\approx \frac{1}{p} U_X^T (U_S \Lambda_S(S) U_S^T + U_N \Lambda_N(N) U_N^T) U_X \\
\approx \frac{1}{p} \Lambda_S(S) + \frac{1}{p} \Lambda_N(N).
\] (3)

There are two challenges in the simplification provided in Equation 3. First, adding \( N \) to \( S \) can result in the underestimation of the elements in \( \Lambda_S(S) \) using \( \Lambda_X(S) \). Second, the eigen vectors of \( S \) and \( N \) are not generally well-aligned (\( U_N U_S \not\approx I \)). As a result, we provide rigorous justification below and in Appendix A. Given \( S \) and \( N \) are independent, we have \( \frac{1}{p} SN^T \) and \( \frac{1}{p} NS^T \) converge to 0 at a rate of \( \frac{\text{Tr}(S)}{p} \) when \( p \to \infty \) (see Appendix A). \( \Lambda_S(S) \) is a rank \( K \) diagonal matrix with non-zero entries corresponding to the rank of the signal space. When the signal’s magnitude is large compared to the noise, the corresponding vectors in \( U_X \) are aligned with \( U_S \) in the first \( K \) dimensions. \( U_X \) can therefore be regarded as as a perturbation on \( SS^T \), and \( U_X^T U_S \) will be approximately identical in the \( K \) leading eigenvectors. The remaining \((n-K)\) dimensions have eigenvalues close to zero. At the same time, \( W_p = \frac{1}{p} NN^T \) is a random matrix converging to \( I \) (see Appendix B) and \( U_X^T N^T NU_X \) can be seen as a rotation on the noise matrix, implying that \( U_X^T U_N \approx I \).

When \( n \geq p \), we calculate the scaled inner product of \( W_n = \frac{1}{n} X^T X \). Similarly, the scaled eigenvalues of \( X^T X \) can be rewritten as

\[
\frac{1}{n} \Lambda_X(X) \approx \frac{1}{n} \Lambda_S(S) + \frac{1}{n} \Lambda_N(N).
\] (4)
A Scalable Approach to Estimating the Rank of High-Dimensional Data

The top \( p \) eigenvalues of \( X^T X \) and \( XX^T \) are the same up to a constant \( n \) or \( p \) and the rest \( n - p \) eigenvalues are equal to zero for \( n \geq p \). We only focus on the consistent eigenvalues up to the minimum of \( n \) or \( p \). The eigenvalues in the null space are not considered in this paper. Similarly, we can demonstrate that \( \frac{1}{n} S^T N = \frac{1}{n} N^T S \approx 0 \) and that \( \frac{1}{n} N^T N \) converges to \( I \).

B. Empirical Spectral Distribution

By providing an estimate of \( \Lambda_N(N) \), we estimate the amount of detectable signal that can be distinguished from noise in the region where \( \Lambda_X(X) - \Lambda_N(N) \geq 0 \). We denote \( W = NN^T \) and its \( n \) non-negative eigenvalues as \( \lambda_1, \lambda_2, ..., \lambda_n \). The joint density of eigenvalues for the Wishart matrices is as follows:

\[
p(\lambda_1, \lambda_2, ..., \lambda_n) = Z_{n,\beta}^{(L)} e^{-\beta \sigma \sum_{i=1}^{n} \lambda_i} \prod_{i=1}^{n} \lambda_i^{\frac{\beta}{2} \left[p - n + 1\right]} \prod_{j<k} |\lambda_j - \lambda_k|^\beta
\]

\[= Z_{n,\beta=1}^{(L)} e^{-\frac{1}{2} \sum_{i=1}^{n} \lambda_i} \prod_{i=1}^{n} \lambda_i^{\frac{1}{2} \left[p - n - 1\right]} \prod_{j<k} |\lambda_j - \lambda_k|,\]

where \( \beta = 1 \) for real entries and \( \sigma = \frac{1}{2} \) for standard normal distribution (Götzte et al., 2004; Livan et al., 2018). When \( n \geq p \), we define it as Anti-Wishart ensemble \( \tilde{W} \), where we have \( n - p \) eigenvalues equal zero. The joint density function of the Anti-Wishart ensemble is similar to Equation 5, with part of the matrix elements being random and the remaining elements non-random. The determination of these non-random elements are related to the first \( p \) rows of \( \tilde{W} \) (Livan et al., 2018). The normalization factor can be calculated by Selberg’s
integral as

\[
Z_{n,\beta=1}^{(L)} = \left(\frac{\beta}{2}\right)^n \prod_{j=1}^n \frac{\Gamma\left(\frac{\beta}{2} + 1\right)}{\Gamma\left(\frac{\beta}{2}j + 1\right)\Gamma\left(\frac{\beta}{2}(j-1) + \alpha + 1\right)}
\]

\[
= \left(\frac{1}{2}\right)^n \prod_{j=1}^n \frac{\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{1}{2} + 1\right)\Gamma\left(\frac{1}{2}(p-n+j)\right)},
\]

where \(\gamma = n[\alpha + \frac{\beta}{2}(n-1) + 1]\) and \(\alpha = \frac{\beta}{2}(p-n+1) - 1\) (Dumitriu and Edelman, 2005; Forrester and Warnaar, 2008; Kumar, 2019; Livan et al., 2018). The empirical spectral distribution of \(W_p\) is defined as

\[
F_n(y) = \frac{1}{n} \sum_{k=1}^n I\{\lambda_k \leq y\}.
\]

As proven in Götze et al. (2004); Livan et al. (2018); Marčenko and Pastur (1967), the expected spectral distribution \(EF_n(y)\) and \(F_n(y)\) converge to the MP distribution in probability with density form:

\[
f_c(y) = \frac{1}{2\pi \sigma^2 cy} \sqrt{(y - c_-)(c_+ - y)} I_{[c_- , c_+]}(y) + I_{[1, \infty]}(c)(1 - c^{-1}) \delta(y),
\]

where rectangularity ratio \(l_{n,p=\infty}(n/p) \rightarrow c \in (0, \infty)\). Note that ratio of \(n/p\) approximates \(c\) and is not fixed. Denote \(\delta(y)\) as the Dirac delta function and \(c_\pm = \sigma^2(1 \pm \sqrt{c})^2\) (Götze et al., 2004; Livan et al., 2018; Vivo, 2008). When \(p > n\), we have the rectangularity ratio \(c < 1\) and the second term in the density form equals zero according to the index function. When \(n \geq p\), we have the rectangularity ratio \(c \geq 1\) and the second term of the density function does not equal zero only when the eigenvalues equal zero. Note that eigenvalues in the null space are not discussed, as we assume that the signal space is smaller than the minimum of \(n\) and \(p\). As a result, the Dirac delta function equals zero for both high- and
low-dimensional settings and share the same density form, which is as

\[
f_c(y) = \frac{1}{2\pi\sigma^2 cy} \sqrt{(y - c_-(c_+ - y))I_{[c_-,c_+]}(y)}.
\]  

\(9\)

C. Signal-to-Noise Ratio Bound

The empirical spectral distribution provides an estimate of the eigenvalues \(\Lambda_N(N)\) for the \(W_p\) matrix in high- and low-dimensional cases. We focus on the high-dimensional case where \(p > n\) in the in-text derivation and provide a distributional upper bound on the maximum eigenvalue of \(\Lambda_N(N)\) to determine the magnitude of the signal that can be distinguished from noise. The derivation of the signal-to-noise ratio (SNR) bound for low-dimensional cases can be derived similarly.

The proof of the SNR bound can be summarized as follows: (i) provide the distribution of the diagonal elements of the \(W_p\) matrix. (ii) Provide the distribution of the off-diagonal elements of the \(W_p\) matrix and show that the off-diagonal elements converge to zero. (iii) Set the bound for the eigenvalues of the \(W_p\) matrix by the Gershgorin circle theorem (Gershgorin, 1931). Intuitively, the eigenvalue of the \(j\)th column of the \(W_p\) square matrix lies within a circle of the \(j\)th diagonal elements at a radius of the sum of the absolute values of off-diagonal elements of the \(j\)th column (Gershgorin, 1931). (iv) Further set the bound for the maximum eigenvalue of the \(W_p\) matrix by the order statistics (Fréchet, 1927).

Assuming that the entries of the noise matrix \(N\) are i.i.d. random variables drawn from an unknown joint distribution (without normality assumption) and \(p\) is sufficiently large, we first note that by the central limit theorem, the diagonal and off-diagonal entries of \(W_p\)
matrix converge in distribution to normal distributions respectively, as \( n \) and \( m \) goes to infinity, where \( m = n(n - 1)/2 \) is the number of elements in the upper triangular matrix.

**Theorem 1.** Let \( N \in \mathcal{R}^{n \times p} \) with i.i.d. elements having mean 0, variance \( \sigma^2 \), and the fourth moment \( \gamma^4 < \infty \). The diagonal entries of the \( W_p \) matrix \( \xi_j \) \((j = 1, \ldots, n)\) converge in distribution to \( \mathcal{N}(\sigma^2, \frac{\gamma^4 - \sigma^4}{p}) \) when \( p \) is sufficiently large to have the central limit effect.

*Proof.* see Appendix C.

**Theorem 2.** Let \( N \in \mathcal{R}^{n \times p} \) with i.i.d. elements having mean 0, variance \( \sigma^2 \), and the fourth moment \( \gamma^4 < \infty \). The off-diagonal entries of the \( W_p \) matrix \( \varepsilon_{ik} \) \((i, k = 1, \ldots, n \text{ and } i < k)\) converge in distribution to \( \mathcal{N}(0, \frac{\gamma^4 + \sigma^4}{4p}) \) when \( p \) is sufficiently large to have the central limit effect.

*Proof.* see Appendix D.

Consider the special case where \( N \) follows a standard normal distribution, the variance of the main diagonal elements is \( \frac{\gamma^4 - \sigma^4}{p} = \frac{3\sigma^4 - \sigma^4}{p} = \frac{2}{p} \), or similarly by the main diagonal of variance for \( W_p \) as \( \frac{1}{p}(n + 1) = \frac{2}{p} \). Thus, the distribution of the main diagonal elements is \( \mathcal{N}(1, \frac{2}{p}) \), the justification for which is rigorously provided in Appendices B and C. Similarly, the distribution of the off-diagonal elements is \( \mathcal{N}(0, \frac{1}{p}) \), which is rigorously justified in Appendix D. As proved in Appendix B, the first moment of the \( W_p \) matrix for the normalized noise matrix is the identity matrix, with the off-diagonal entries converging to zero at a rate of \( \frac{1}{\sqrt{p}} \) when \( p \) is sufficiently large.

Theorems 1 and 2 provide the distribution of the diagonal and off-diagonal entries of the \( W_p \) matrix. We use the results of order statistics to estimate an upper bound of the
A Scalable Approach to Estimating the Rank of High-Dimensional Data

maximum diagonal element and the maximum radius of the Gershgorin circle of $W_p$ matrix. A similar upper bound can be estimated for the diagonal and the maximum radius of the Gershgorin circle for $W_n$ matrix when $n \geq p$.

**Theorem 3.** Let $\xi_j$ be the main diagonal elements of $W_p$, where $j = 1, ..., n$. The Fisher–Tippett–Gnedenko theorem shows that for the maximum diagonal element, $\xi_p = \max(\xi_j)$. We have

$$E\{\xi(n) > \alpha_1\} = e^{-e^{-\sqrt{\frac{\alpha_1 - \sigma_1}{\sqrt{\gamma^4 + \sigma^4}}}}}. \quad (10)$$

We can set the bound of $\xi(n)$ with the inequality $e^x \leq e^{x^2} + x$ for $x \in \mathbb{R}$ such that

$$E\{\xi(n) > \alpha_1\} \leq e^{e^{\frac{2\sqrt{(\alpha_1 - \sigma_1)}{\sqrt{\gamma^4 + \sigma^4}}}} - e^{\frac{(\alpha_1 - \sigma_1)}{\sqrt{\gamma^4 + \sigma^4}}}}. \quad (11)$$

**Proof.** see Appendix E.

**Theorem 4.** Let $\varepsilon_{ik}$ be the off-diagonal elements of $W_p$, which are normally distributed with mean 0 and variance $\frac{\gamma^4 + \sigma^4}{4p}$, where $i, k = 1, ..., n$ and $i < k$. Let $R_k = \sum_{i \neq k} |\varepsilon_{ik}|$ be the sum of the absolute values of the off-diagonal elements of the $k$th column of $W_p$ matrix with expectation $\mu_{GEV} = \sqrt{(n-1)(\gamma^4 + \sigma^4)}$ and variance $\sigma_{GEV}^2 = \frac{(n-1)(\pi^2 - 2)(\gamma^4 + \sigma^4)}{4\pi \gamma^4 + \sigma^4}$. The Fisher–Tippett–Gnedenko theorem shows that for the maximum entry, $R(n) = \max(R_k)$. We have

$$E\{R(n) > \alpha_2\} = e^{-e^{-b}}, \quad (12)$$

where $b = \frac{\alpha_2 - \mu_{GEV}}{\sigma_{GEV}}$. We can set the bound of $R(n)$ with the inequality $e^x \leq e^{x^2} + x$ for $x \in \mathbb{R}$ such that

$$e^{-e^{-b}} \leq e^{e^{-2b}} - e^{-b}. \quad (13)$$
The eigenvalue of the kth column of the $W_p$ matrix can be bound by the Gershgorin circle theorem. With the distributional upper bound of the $\xi_k$ and distributional upper bound of $R_k$, we have the upper bound of the largest eigenvalue of $W_p$ matrix, which lies within the Gershgorin discs $D(\xi(n), R(n))$ that are centered at $\xi(n)$ with a radius $R(n)$ of rate of convergence $\sqrt{\frac{n-1}{p}}$ for $p > n$.

Proof. see Appendix F.

**Corollary 1.** Let $N \in \mathbb{R}^{n \times p}$ with i.i.d. elements following a standard normal distribution with mean 0 and variance 1. The diagonal entries of the $W_p$ matrix $\xi_j$ ($j = 1, ..., n$) converge in distribution to $\mathcal{N}(1, \frac{2}{p})$. The Fisher–Tippett–Gnedenko theorem shows that for the maximum diagonal element, $\xi(p) = \max(\xi_j)$. We have

$$E\{\xi(n) > \alpha_1\} = e^{-e^{-\frac{\sqrt{2(p-1)}}{\alpha_1}}}. \quad (14)$$

we can set the bound of $\xi(n)$ by

$$E\{\xi(n) > \alpha_1\} \leq e^{-e^{-\frac{\sqrt{2(p-1)}}{\alpha_1}}} - e^{-\frac{\sqrt{2(p-1)}}{4}}. \quad (15)$$

The off-diagonal entries of the $W_p$ matrix $\varepsilon_{ik}$ ($i, k = 1, ..., n$ and $i < k$) converge in distribution to $\mathcal{N}(0, \frac{1}{p})$. Let $R_k = \sum_{i \neq k} |\varepsilon_{ik}|$ be the sum of the absolute values of the off-diagonal elements of the kth column of the $W_p$ matrix with expectation $\mu_{GEV} = \sqrt{\frac{2(n-1)^2}{pn}}$ and variance $\sigma^2_{GEV} = \frac{(n-1)(\pi-2)}{pn}$. The Fisher–Tippett–Gnedenko theorem shows that for the maximum entry, $R(n) = \max(R_k)$. We have

$$E\{R(n) > \alpha_2\} = e^{-e^{-b}}, \quad (16)$$
A Scalable Approach to Estimating the Rank of High-Dimensional Data

where \( b = \frac{\alpha^2 - \mu_{GEV}}{\sigma_{GEV}} \). We can set the bound of \( R(n) \) with the inequality \( e^x \leq e^{x^2} + x \) for \( x \in \mathbb{R} \) such that

\[
e^{-e^{-b}} \leq e^{e^{-2b}} - e^{-b}.
\]

(17)

Thus, we can set the bound of the largest eigenvalue of \( W_p \) by \( D(\xi(n), R(n)) \).

III. ESTIMATING THE DETECTABLE SIGNAL DIMENSION IN PRACTICE

A. Algorithm

Equation 3 shows that the rank of the signal can be estimated by first estimating \( \Lambda_N(N) \) and then finding the zero crossing of the difference \( \Lambda_X(X) - \Lambda_N(N) \) corresponding to where any signal in the signal space cannot be distinguished from noise (see Figure 1).

As a more concrete example, consider the case of plotting scaled eigenvalues of \( X \) matrix with 100 samples, 500 features, and a true rank equals to 10 as well as random samples from the MP distribution in Figure 1. The red line plots the scaled eigenvalues of the matrix while the blue line plots the scaled eigenvalues of a similarly distributed noise-only matrix. The goal is to extract all the signal components up to \( K \) whose eigenvalues are significantly larger than the expected eigenvalues from the MP distribution. The dimension of the detectable signal is the first eigenvalue index where the corresponding eigenvalue is close to the largest eigenvalue sampled from the MP distribution, meaning it cannot be distinguished from noise. The algorithm is implemented as follows:
FIG. 1. Scaled eigenvalues of $X_{100\times500}$ with true rank 10 (red dotted line) compare to random samples from the MP distribution (blue dotted line).

**Algorithm 1** Algorithm for rank estimation

1: procedure **Algorithm**($X_{n\times p}$)

2: Calculate eigenvalue decomposition for scaled matrix $X$ to dimension $n'$ ($n' \ll n$). Denote eigenvalues as $\Sigma_x$.

3: Draw random samples from the MP distribution with a corrected variance. Denote eigenvalues as $\Sigma_n$.

4: If $\exists \Sigma_n > \Sigma_x$, $K = 0$.

5: Calculate posterior probability of change points for deviation ($\Delta = \Sigma_x - \Sigma_n$) with a prior equals to a decreasing linear sequence from 0.9 to 0.

6: Set $K$ to equal the largest dimension with the maximum posterior probability.
B. Implementation

An implementation of the dimension, ladle, and Hoff’s method as an open-source package for the R Programming environment can be found at https://github.com/WenlanzZ/dimension. We estimate the corrected variance of the MP distribution by the SNR bound to align the tails of the scaled eigenvalues of $X$ and noise. The test for deviation of scaled eigenvalues can be performed in several different ways. In practice, the univariate Bayesian change point (BCP) analysis, proposed by Barry and Hartigan (1993) has been found to be effective and is implemented in the R package bcp (Erdman et al., 2007). Frequently, there are several change points with their posterior probabilities approximating the maximum posterior probability. In order to guarantee the inclusion of all important signals, we propose a two-step procedure to estimate the dimension $K$ in practice. In the first step, we select candidates for $K$ by calculating the BCP on the posterior probability for a second time (double posterior). In the second step, among those candidates with the highest double posterior, we pick $K$ equal to the largest dimension where its posterior probability is larger than $\delta$ times the maximum posterior probability ($\delta \in \mathbb{R} : \delta \geq 0$ and $\delta \leq 1$ determine the range of approximation to the maximum posterior probability; by default set $\delta = 0.90$).

One limitation observed in simulation studies when implementing the dimension algorithm is related to R package bcp for calculating the posterior probabilities of the change points. The performance of R package bcp is not always stable at the edges of the sequence. Thus, in some extreme cases, a spike occurs at the edge after a long and flat pattern in the sequence of posterior probabilities. To address this problem, we adopted an alarm system
that detects flat and spike patterns and trims off sequence before spikes or after a long and flat pattern.

C. Simulation

We compared the performance of the proposed rank estimation procedure to the ladle and Hoff’s method. Consider two signal-plus-noise matrices whose true signal rank $K = 3$. For a fair comparison, the first matrix setting is the same as in Luo and Li (2016) and the second matrix setting is similar to Hoff (2007):

\begin{align}
X_1 &\sim N(0, \Sigma_{X_1}), \Sigma_{X_1} = M + 0.54^2 I_p, M = \text{diag}(2, 1, 1) + 0_{p,p}, \\
X_2 &= S + \epsilon, S[\cdot, 1 : K] \sim N(0, \Sigma), S[\cdot, (K + 1) : p] = 0, \Sigma = 3I_3.
\end{align}

We simulated 1,000 independent samples from the $X$ matrix and performed rank estimation with different methods. We allowed each method to run for up to 24 hours for 1,000 simulations. The percentage of correct estimation and elapsed time in seconds are reported in Table 1. For example, in the first case of the $X_1$ matrix (sigma separated by commas) with signal rank equal to three and 100 samples and 10 features, dimension had 94.8% of correct estimation with a mean absolute error of 0.056 and an average elapsed time of 0.040 seconds; ladle had 93% of correct estimation with a mean absolute error of 0.396 and an average elapsed time of 0.034 seconds; Hoff’s method had 78.8% of correct estimation with a mean absolute error of 0.367 and an average elapsed time of 0.538 seconds. In the sixth case of the $X_2$ matrix (sigma as a single integer) with signal rank equal to three and 10 samples and 100 features, dimension had 28.3% of correct estimation with a mean absolute error
A Scalable Approach to Estimating the Rank of High-Dimensional Data

of 0.980 and an average elapsed time of 0.041 seconds; ladle had 0% of correct estimation
with a mean absolute error of 5.930 and an average elapsed time of 0.044 seconds; Hoff’s
method had 0% of correct estimation with a mean absolute error of 7 and an average elapsed
time of 1.094 seconds. In Table 1, when we fixed \( p \) and let \( n \) grow for the \( n > p \) setting,
the dimension method achieved the highest accuracy and performed stably for both the \( X_1 \)
and \( X_2 \) matrices, while the ladle method seemed to work only with its own matrix setting.
When we fixed \( n \) and let \( p \) grow for the \( p > n \) setting, dimension excelled in both accuracy
and computational efficiency. In terms of computational efficiency, the dimension method
is more scalable to large data sets compared to the other methods. In extreme cases where
noise overwhelms signal and the signal space is too small to be detectable, dimension under-
estimated the true rank by a mean absolute error of around 6. When signal is moderate
in an extremely high-dimensional setting, dimension achieved an accuracy of 91.1% with a
mean absolute error of 0.207 and an average elapsed time of 3.767 seconds while the other
two methods are computationally inapplicable to providing an estimation.
TABLE I. Simulation results comparing dimension estimation methods under different settings.

| rank | unknown matrix | dimension | ladle | Hoff |
|------|----------------|-----------|-------|------|
| k    | n              | p         | sigma | acc  | mae  | time | acc  | mae  | time | acc  | mae  | time |
| 3    | 100            | 10        | 2, 1, 1 | 0.056| 0.040| 93   | 0.396| 0.034| 78.8 | 0.367| 0.538|      |
| 3    | 1000           | 10        | 2, 1, 1 | 0    | 0.585| 100  | 0.696| 96.3 | 0.044| 49.747|      |
| 3    | 10000          | 10        | 2, 1, 1 | 0    | 0.070| 100  | 0    | 15.442| -   | -   | >24h |
| 3    | 10             | 100       | 2, 1, 1 | 1.175| 0.041| 100  | 0    | 5.828| 0.043| 11.4 | 2.727| 0.591|
| 3    | 1000           | 10        | 2, 1, 1 | 1.420| 0.059| 100  | 0    | 3.057| 0.524| -   | -   | >24h |
| 3    | 10             | 100       | 6      | 0.980| 0.041| 100  | 0    | 5.930| 0.044| 0    | 7    | 1.094|
| 3    | 10             | 1000      | 6      | 1.696| 0.062| 100  | 0    | 3.123| 0.493| 0    | 7    | 116.187|
| 3    | 10             | 10000     | 6      | 1.914| 0.071| 100  | 0    | 15.211| -   | -   | >24h |
| 10   | 500            | 100       | 2, 2, 2 | 0.006| 0.268| 100  | 0    | 2.006| 99.9 | 0.001| 50.730|      |
| 10   | 500            | 100       | 2      | 99.9 | 0.000| 0.284| 0    | 70   | 15.933| 92.2 | 0.127| 23.333|      |
| 10   | 5000           | 100       | 2      | 100  | 0    | 0.614| -   | -   | >24h | -   | -   | >24h |
| 10   | 50000          | 100       | 2      | 100  | 0    | 3.902| -   | -   | >24h | -   | -   | >24h |
| 10   | 100            | 500       | 6      | 0.064| 4.434| 0.290| 0    | 10.008| 3.871| 23.4 | 1.693| 24.214|
| 10   | 100            | 5000      | 6      | 6.684| 0.646| -   | -   | >24h | -   | -   | >24h |
| 10   | 100            | 50000     | 6      | 8.553| 3.728| -   | -   | >24h | -   | -   | >24h |
| 10   | 100            | 500       | 10     | 85.9 | 0.250| 0.288| 0    | 10.971| 3.914| 69.2 | 0.689| 15.457|
| 10   | 100            | 5000      | 10     | 6.314| 0.699| -   | -   | >24h | -   | -   | >24h |
| 10   | 100            | 50000     | 100    | 91.1 | 0.207| 3.767| -   | -   | >24h | -   | -   | >24h |
IV. CASE STUDY: DIMENSION ESTIMATION IN SCRNA-SEQ ANALYSIS

The lung scRNA-seq data can be found on Gene Expression Omnibus (GEO) under accession code GSE136831. The data was examined by Adams et al. (2019) to build a single-cell atlas of Idiopathic Pulmonary Fibrosis (IPF). The data contains the gene expression of 312,928 cells profiled from 32 IPF lung samples, 18 chronic obstructive pulmonary disease (COPD) lung samples, and 29 control donor lung samples (Adams et al., 2019). We applied the dimension method to a subgroup of control lung sample (Patient ID 001C) with 2,000 highly variable genes and 127 cells. Figure 2 shows that among all the candidates with maximum double posterior probabilities (see Figure 2b), the eighth dimension is the largest with the maximum posterior probability of change point (see Figure 2a).

FIG. 2. (a) The posterior means of the deviation and posterior probability plots. (b) The posterior probability and double posterior probability plots.
We then applied the Louvain clustering method implemented in the R package Seurat for PCs from Dimensions 1-50, 9-50, and 1-8 (see Figures 3a, 3b, 3c). The Uniform Manifold Approximation and Projection (UMAP) for each plot confirms that major features of cell-type clustering are associated with Dimensions 1-8 and there is no new cluster formed from Dimensions 9-50. Clustering from the identified signal subspace identified two additional clusters that are not identified by Dimensions 1-50. To validate the clustering results, we identified marker genes in each cluster via differential expression analysis and further matched the clustering results to known cell types with identified marker genes (see Figure 3c for cell-type identification). Comparing to the original identification published by Adams et al. (2019) (see Figure 3d), we found that most of the cells types were correctly annotated with the small subset using Dimensions 1-8, with the exception of the T cells that were clustered with the NK cells. However, NK cells and T cells are phenotypically similar and they both derive from lymphoid lineage cells. By contrast, when we included noises from Dimensions 1-50, T cells, fibroblast, and secretory cells were misclassified as one cell type. Therefore, an accurate estimation of the signal dimension is paramount for cell-type identification to avoid clustering with an overabundance of noise or signal omission, both of which lead to misleading conclusions.
FIG. 3. (a) UMAP visualization and cells labeled by Louvain clustering results with dimensions 1-50, (b) 9-50, and (c) 1-8. (d) Cell-type identification marked by Adams et al. (2019).
V. CONCLUSION

In this chapter, we posed the problem of dimension estimation in a signal-plus-noise regime as the eigenvalue index where the difference between a data matrix and its inherent noise, defined by a random matrix, were close to zero. Based on this regime, we proposed a procedure to estimate the dimension of detectable signal, along with a derivation of distributional bounds for the magnitude of detectable signal required to be distinguished from noise. We implemented the proposed procedure and related methods as per the open-source R package at https://github.com/WenlanZ/dimension. In simulation experiments, we demonstrated that the accuracy of the dimension method and its speed and computationally efficiency advantages over ladle and Hoff’s method, both of which are essential for large and high-dimensional data sets. In high-dimensional cases with extremely low signal, the dimension method provided a robust low-rank estimation while the other two methods provided either zero rank or full rank. Our results showed that the dimension method is more accurate in estimating the dimension while remaining computationally scalable compared to current methods in a wide range of scenarios under both high- and low-dimensional settings. Besides making performance comparisons with the simulated data, we also demonstrated that community detection in the identified signal subspace was beneficial in cell-type identification when analyzing scRNA-seq data sets. In addition to the applications to genomic data sets, the dimension-estimation method has more general applications, such as digital imaging, speech modelling, and neural networks.
ACKNOWLEDGMENTS
APPENDIX A:

Here we provide detailed proofs of the Equation 3. Based on the central limit theorem, we show that $\frac{1}{p}SN^T$ and $\frac{1}{p}NS^T \approx 0$. Given that $SN^T = NS^T$, we will focus on the proof for $\frac{1}{p}SN^T \approx 0$. The main technical challenge here comes from the case when the eigenvectors of $S$ and $N$ are not well-aligned ($U_N^T U_S \not\approx I$). Assume an extreme worst case when $S$ is a diagonal matrix with finite singular values ($\Sigma_s)_i$ for $i = 1, \ldots, p$ ($p - n$ rows are padded with zeros), we show that in general $\frac{1}{p}SN^T \not\approx 0$ for finite $p$. Let $N_{ij}$ be i.i.d. entries of matrix $N$ with $\text{E}(N_{ij}) = 0$ and $\text{E}(N_{ij}^2) = \sigma^2$. Without loss of generality, we normalize $N_{ij}$ so that it has mean zero and variance one.

\[
\frac{1}{p}SN^T = \frac{1}{p}
\begin{bmatrix}
S_{11} & 0 & \cdots & 0 \\
0 & S_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & S_{pp}
\end{bmatrix}
\begin{bmatrix}
N_{11} & N_{12} & \cdots & N_{1p} \\
N_{21} & S_{22} & \cdots & N_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
N_{p1} & 0 & \cdots & N_{pp}
\end{bmatrix}
\]

\[
= \frac{1}{p} \sum_{i=1}^{p} S_{ii} N_{ii}
\]

\[
= \frac{1}{p} \sum_{i=1}^{p} (\Sigma_s)_i N_{ii}
\]

\[
\leq \frac{\text{Tr} S}{p}.
\]

Thus we have $\frac{1}{p}SN^T \approx 0$ when $p \to \infty$ with a convergence rate $\frac{\text{Tr} S}{p}$. 

24
APPENDIX B:

Without loss of generality, we standardize the columns of random matrix $N_{n \times p}$ ($p > n$) by subtracting off the column-wise means and standardizing such that $\sigma^2 = 1$. We can first derive the first moment of $W$ as follows:

\[
NN^T = \sum_{j=1}^{p} N_j N_j^T = \begin{bmatrix}
\sum_{j=1}^{p} N_{1j}^2 & \sum_{j=1}^{p} N_{1j} N_{2j} & \cdots & \sum_{j=1}^{p} N_{1j} N_{nj} \\
\sum_{j=1}^{p} N_{2j} N_{1j} & \sum_{j=1}^{p} N_{2j}^2 & \cdots & \sum_{j=1}^{p} N_{2j} N_{nj} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{p} N_{nj} N_{1j} & \sum_{j=1}^{p} N_{nj} N_{2j} & \cdots & \sum_{j=1}^{p} N_{nj}^2 
\end{bmatrix} \tag{B1}
\]

where the main diagonals

\[
E \sum_{j=1}^{p} N_{jj}^2 = p, \tag{B2}
\]

for $j = 1, 2, \ldots, p$. and the off-diagonal we have

\[
\sum_{j=1}^{p} E N_{ij} N_{kj} = \sum_{j=1}^{p} E N_{ij} E N_{kj} = 0, \tag{B3}
\]

for $j = 1, 2, \ldots, p$ and $i \neq k$.

Thus we have the first moment of the $W_p$ matrix as

\[
EW_p = E \frac{1}{p} NN^T = I. \tag{B4}
\]
A Scalable Approach to Estimating the Rank of High-Dimensional Data

We then derive the second moment of the $W_p$ matrix as follows:

$$NN^TNN^T =$$

$$
\begin{bmatrix}
\sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{1k}N_{ik}N_{ij}N_{1j} & \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{1k}N_{ik}N_{ij}N_{2j} & \cdots & \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{1k}N_{ik}N_{ij}N_{nj} \\
\sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{2k}N_{ik}N_{ij}N_{1j} & \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{2k}N_{ik}N_{ij}N_{2j} & \cdots & \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{2k}N_{ik}N_{ij}N_{nj} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{nk}N_{ik}N_{ij}N_{1j} & \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{nk}N_{ik}N_{ij}N_{2j} & \cdots & \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{nk}N_{ik}N_{ij}N_{nj}
\end{bmatrix}
$$

\text{(B5)}

where the expectation of $d_i^{th}$ main diagonal elements is

$$E\sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{d_i k}N_{ik}N_{ij}N_{d_i j}
$$

$$= p E N_{d_i j}^4 + \sum_{i=d_i, j \neq k} N_{d_i k}N_{d_i k}N_{d_i j}N_{d_i j} + \sum_{i \neq d_i, j=k} N_{d_i k}N_{ik}N_{ik}N_{d_i k}
$$

\text{(B6)}

$$= p E N_{d_i j}^4 + \sum_{i=d_i, j \neq k} 1 + \sum_{i \neq d_i, j=k} 1
$$

$$= p(p + n + 1).
$$

Note that $E N_{d_i}^4$ is the second moment of the chi-square distribution and the off-diagonal elements are with expectation

$$E\sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{k=1}^{p} N_{d_k k}N_{ik}N_{ij}N_{d_j j} = 0,
$$

\text{(B7)}

where $d_i \neq d_j$.

Thus we have the second moment of $W_p$ as

$$EW_p^2 = E \frac{1}{p^2} NN^TNN^T = (1 + \frac{1}{p}(n + 1))I.
$$

\text{(B8)}
APPENDIX C:

For a $N_{n \times p}$ matrix, where $p > n$. Let $N_{ij}$ be i.i.d. entries of $N$ with $E(N_{ij}) = 0$, $E(N^2_{ij}) = \sigma^2$ and $E(N^4_{ij}) = \gamma^4 < \infty$.

Consider the $j$th column of $N$, for $j = 1, 2, ..., p$. Let $E N^2_{ij} = \sigma^2$ and $Z_i = N^2_{ij}$, then we can write the main diagonal of the elements of the $W_p$ matrix as:

$$
\xi_j = \frac{1}{p} \langle N_j, N_j \rangle = \frac{1}{p} \sum_{i=1}^{p} Z_i, \quad (C1)
$$

where $E Z_i = \sigma^2$ and $\text{Var} (Z_i) = \gamma^4 - \sigma^4$. When $p$ sufficiently large to have the Lyapunov central limit effect, we have

$$
\xi_j \sim N \left( \sigma^2, \frac{\gamma^4 - \sigma^4}{p} \right). \quad (C2)
$$
APPENDIX D:

For a $N_{n \times p}$ matrix, where $p > n$. Let $N_{ij}$ be i.i.d. entries of $N$ with $E(N_{ij}) = 0$, $E(N^2_{ij}) = \sigma^2$ and $E(N^4_{ij}) = \gamma^4 < \infty$.

Consider the $i$th row and $k$th column of $N$, for $1 \leq i, k \leq p$ and $i < k$. We can write the off-diagonal of the elements of the $W_p$ matrix at the $i$th row and $k$th column as

$$\varepsilon_{ik} = \frac{1}{p} \sum_{j=1}^{p} N_{ij} N_{kj}.$$  \hfill (D1)

The product of two independent random variables can be written as

$$N_{ij} N_{kj} = \frac{1}{4} (N_{ij} + N_{kj})^2 - \frac{1}{4} (N_{ij} - N_{kj})^2.$$  \hfill (D2)

We rewrite $N_{ij} N_{kj}$ as

$$N_{ij} N_{kj} = \frac{1}{4} (a^2 - b^2),$$  \hfill (D3)

where $a = N_{ij} + N_{kj}$ and $b = N_{ij} - N_{kj}$. The moments of $a$ and $b$ is derived as follows:

$$E a = E b = \sum_{j=1}^{n} \mu_j = 0,$$  \hfill (D4)

$$E a^2 = E b^2 = \sum_{j=1}^{n} \mu_{2j} = 2\sigma^2,$$  \hfill (D5)

$$E a^4 = E b^4 = \sum_{j=1}^{n} \mu_{4j} + 6 \sum_{j=1}^{n-1} \sum_{k>j}^{n} \sigma^2_j \sigma^2_k = 2\gamma^4 + 6\sigma^4,$$  \hfill (D6)

$$\text{Var} a^4 = \text{Var} b^4 = 2\gamma^4 + 2\sigma^4.$$  \hfill (D7)
In the next step, we show that \( a \) and \( b \) are independent. Consider a vector
\[
U = \begin{bmatrix} N_{ij} \\ N_{kj} \end{bmatrix},
\]
\[
A = \begin{bmatrix} 1 & 1 \\ \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & -1 \\ \end{bmatrix}.
\]
According to a theorem in multivariate probability that two linear functions \( AU \) and \( BU \) of a Gaussian vector \( U \) are independent if and only if
\[
A \operatorname{Var}(U) B^T = 0.
\]
Thus, \( a \) and \( b \) are independent. So that we can calculate the expectation of \( N_{ij}N_{kj} \) as
\[
\mathbb{E} N_{ij}N_{kj} = \frac{1}{4} \mathbb{E} (a^2 - b^2) = 0,
\]
and the variance as
\[
\operatorname{Var} N_{ij}N_{kj} = \frac{1}{16} \operatorname{Var} (a^2 - b^2) = \frac{\gamma^4 + \sigma^4}{4}.
\]
Thus, for the off-diagonal entry \( \varepsilon_{ik} \), the expectation is
\[
\mathbb{E} \varepsilon_{ik} = \frac{1}{p} \mathbb{E} \sum_{j=1}^{p} N_{ij}N_{kj} = 0,
\]
and the variance as
\[
\operatorname{Var} \varepsilon_{ik} = \frac{1}{p^2} \operatorname{Var} \sum_{j=1}^{p} N_{ij}N_{kj} = \frac{\gamma^4 + \sigma^4}{4p}.
\]
When \( p \) sufficiently large to have the Lyapunov central limit effect, we have
\[
\varepsilon_{ik} \overset{\text{d}}{\sim} \mathcal{N} \left( 0, \frac{\gamma^4 + \sigma^4}{4p} \right).
\]
Theorem 1 indicates that the main diagonal entries $\xi_j$ ($j = 1, ..., n$) of $W_p$ are iid. normally distributed random variables with mean $\sigma^2$ and variance $\frac{\gamma^4 - \sigma^4}{p}$. The Fisher–Tippett–Gnedenko theorem shows that for the maximum diagonal element

$$\xi(n) = \max(\xi_j) \sim \text{GEV}(\sigma^2, \frac{\sqrt{p(\gamma^4 - \sigma^4)}}{\sigma^4}, 0).$$  \hspace{1cm} (E1)

The cumulative distribution function of the Generalized Extreme Value (GEV) distribution is

$$\Phi(\alpha_1) = E\{\xi(n) > \alpha_1\} = e^{-e^{-a}},$$  \hspace{1cm} (E2)

where $a = \sqrt{p(\alpha_1 - \sigma^2)} / \sqrt{\gamma^4 - \sigma^4}$.

We can bound the maximum diagonal element $\xi(n)$ with the inequality $e^x \leq e^{x^2} + x$ for $x \in \mathbb{R}$ such that

$$e^{-e^{-a}} \leq e^{e^{-2a}} - e^{-a}.$$  \hspace{1cm} (E3)

Thus, we have

$$E\{\xi(n) > \alpha_1\} \leq e^{e^{-\frac{2\sqrt{p(\alpha_1 - \sigma^2)}}{\sqrt{\gamma^4 - \sigma^4}}}} - \frac{\sqrt{p(\alpha_1 - \sigma^2)}}{\sqrt{\gamma^4 - \sigma^4}}.$$  \hspace{1cm} (E4)
APPENDIX F:

Theorem 2 indicates that the off-diagonal entries \( \varepsilon_{ik} \) \( (i, k = 1, \ldots, n) \) and \( i < k \) are iid. normally distributed random variables with mean 0 and variance \( \frac{\gamma^4 + \sigma^4}{4p} \). Given the distribution of the \( i \)th row and \( k \)th column of the off-diagonal element of the \( W_p \) matrix. We can calculate the sum of the absolute values of the off-diagonal elements of the \( k \)th column of the \( W_p \) matrix as \( R_k = \sum_{i \neq k} |\varepsilon_{ik}| \). The absolute value of \( \varepsilon_{ik} \) follows a half-normal distribution and the off-diagonal elements are independent, thus we can derive the expectation of the sum of the absolute off-diagonal values as follows:

\[
\mu_{\text{GEV}} = E R_k = \sum_{i \neq k} E |\varepsilon_{ik}| = (n - 1) \sqrt{\frac{2(\gamma^4 + \sigma^4)}{4p\pi}} = \sqrt{\frac{(n - 1)^2(\gamma^4 + \sigma^4)}{2p\pi}}, \tag{F1}
\]

and the variance as

\[
\sigma^2_{\text{GEV}} = \text{Var} R_k = \sum_{i \neq k} \text{Var} |\varepsilon_{ik}| = \frac{(n - 1)(\pi - 2)(\gamma^4 + \sigma^4)}{4p\pi}. \tag{F2}
\]

The Fisher–Tippett–Gnedenko theorem shows that for the maximum entry of \( R_k \)

\[
R_{(n)} = \max(R_k) \sim \text{GEV}(\mu_{\text{GEV}}, \sigma_{\text{GEV}}, 0), \tag{F3}
\]

The cumulative distribution function of the Generalized Extreme Value (GEV) distribution is

\[
\Phi(\alpha_2) = E\{R_{(n)} > \alpha_2\} = e^{-e^{-b}}, \tag{F4}
\]

where \( b = \frac{\alpha_2 - \mu_{\text{GEV}}}{\sigma_{\text{GEV}}} \). We can bound the \( R_{(n)} \) with the inequality \( e^x \leq e^{x^2} + x \) for \( x \in \mathbb{R} \) such that

\[
e^{-e^{-b}} \leq e^{e^{-2b}} - e^{-b}. \tag{F5}
\]
Furthermore, we can bound the eigenvalue of the $k$th column of the $W_p$ matrix by the Gershgorin circle theorem. Known that the eigenvalue of the $W_p$ matrix lies within at least one of the Gershgorin discs $D(\xi_k, R_k)$, where $D(\xi_k, R_k) \subseteq \mathbb{C}$ is a closed disc centered at the $k$th diagonal element of $W_p$ (denote as $\xi_k$) with radius $R_k$. With the distributional upper bound of the $\xi_k$ and distributional upper bound of $R_k$, we have the upper bound of the largest eigenvalue of $W_p$, which lies within the Gershgorin discs $D(\xi(n), R(n))$ that are centered at $\xi(n)$ with a radius $R(n)$ of rate of convergence $\sqrt{\frac{n-1}{p}}$ for $p > n$. 
Adams, T. S., Schupp, J. C., Poli, S., Ayaub, E. A., Neumark, N., Ahangari, F., Chu, S. G., Raby, B., Deluliis, G., Januszyk, M., Duan, Q., Arnett, H. A., Siddiqui, A., Washko, G. R., Homer, R., Yan, X., Rosas, I. O., and Kaminski, N. (2019). “Single cell rna-seq reveals ectopic and aberrant lung resident cell populations in idiopathic pulmonary fibrosis,” bioRxiv https://www.biorxiv.org/content/early/2019/09/06/759902, doi: 10.1101/759902.

Aparicio, L., Bordyuh, M., Blumberg, A. J., and Rabadan, R. (2018). “Quasi-universality in single-cell sequencing data,” arXiv preprint arXiv:1810.03602.

Barry, D., and Hartigan, J. A. (1993). “A bayesian analysis for change point problems,” Journal of the American Statistical Association 88(421), 309–319.

Carreira-Perpinán, M. A. (1997). “A review of dimension reduction techniques,” Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09 9, 1–69.

Chung, N. C., and Storey, J. D. (2015). “Statistical significance of variables driving systematic variation in high-dimensional data,” Bioinformatics 31(4), 545–554.

Dumitriu, I., and Edelman, A. (2005). “Eigenvalues of hermite and laguerre ensembles: large beta asymptotics,” in Annales de l’IHP Probabilités et statistiques, Vol. 41, pp. 1083–1099.

Erdman, C., Emerson, J. W. et al. (2007). “bcp: an r package for performing a bayesian analysis of change point problems,” Journal of Statistical Software 23(3), 1–13.

Forrester, P., and Warnaar, S. (2008). “The importance of the selberg integral,” Bulletin of the American Mathematical Society 45(4), 489–534.
A Scalable Approach to Estimating the Rank of High-Dimensional Data

Fréchet, M. (1927). “Sur la loi de probabilité de l’écart maximum,” Ann. Soc. Math. Polon. 6, 93–116.

Gerschgorin, S. (1931). “On bounding the eigenvalues of a matrix,” Izv. Akad. Nauk. SSSR Otd Mat. Estest 1, 749–754.

Götze, F., Tikhomirov, A. et al. (2004). “Rate of convergence in probability to the marchenko-pastur law,” Bernoulli 10(3), 503–548.

Hoff, P. D. (2007). “Model averaging and dimension selection for the singular value decomposition,” Journal of the American Statistical Association 102(478), 674–685.

Kumar, S. (2019). “Recursion for the smallest eigenvalue density of $\beta$-wishart–laguerre ensemble,” Journal of Statistical Physics 175(1), 126–149.

Livan, G., Novaes, M., and Vivo, P. (2018). Introduction to random matrices: theory and practice (Springer).

Luo, W., and Li, B. (2016). “Combining eigenvalues and variation of eigenvectors for order determination,” Biometrika 103(4), 875–887, https://doi.org/10.1093/biomet/asw051, doi: 10.1093/biomet/asw051.

Marčenko, V. A., and Pastur, L. A. (1967). “DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES,” Mathematics of the USSR-Sbornik 1(4), 457–483, https://doi.org/10.1070/Fsm1967v001n04abeh001994.

Vivo, P. (2008). “From wishart to jacobi ensembles: statistical properties and applications,” Ph.D. thesis, Brunel University, School of Information Systems, Computing and Mathematics.
White, S., and Smyth, P. (2005). “A spectral clustering approach to finding communities in graphs,” in Proceedings of the 2005 SIAM international conference on data mining, SIAM, pp. 274–285.

Yang, B. (1996). “Asymptotic convergence analysis of the projection approximation subspace tracking algorithms,” Signal processing 50(1-2), 123–136.

Yong, A. G., Pearce, S. et al. (2013). “A beginner’s guide to factor analysis: Focusing on exploratory factor analysis,” Tutorials in quantitative methods for psychology 9(2), 79–94.