Controlling the False Discovery Rate in Subspace Selection

Mateo Díaz[*]   Venkat Chandrasekaran[†]

Abstract

Controlling the false discovery rate (FDR) is a popular approach to multiple testing, variable selection, and related problems of simultaneous inference. In many contemporary applications, models are not specified by discrete variables, which necessitates a broadening of the scope of the FDR control paradigm. Motivated by the ubiquity of low-rank models for high-dimensional matrices, we present methods for subspace selection in principal components analysis that provide control on a geometric analog of FDR that is adapted to subspace selection. Our methods crucially rely on recently-developed tools from random matrix theory, in particular on a characterization of the limiting behavior of eigenvectors and the gaps between successive eigenvalues of large random matrices. Our procedure is parameter-free, and we show that it provides FDR control in subspace selection for common noise models considered in the literature. We demonstrate the utility of our algorithm with numerical experiments on synthetic data and on problems arising in single-cell RNA sequencing and hyperspectral imaging.

Keywords: multiple testing, principal components analysis, random matrix theory, rigidity of eigenvalues, spectral methods

1   Introduction

Controlling the false discovery rate (FDR) is a prominent approach to simultaneous testing that was pioneered by Benjamini and Hochberg [13]. They defined the false discovery rate associated with a multiple testing procedure $\widehat{R}$ as

$$\text{FDR}(\widehat{R}) = \mathbb{E}\left[\frac{|\widehat{R} \cap R^c|}{\max\{|\widehat{R}|, 1\}}\right], \tag{1}$$

where R denotes the set of true positives and $\widehat{R}$ denotes the set of rejected nulls. Furthermore, they also introduced an innovative methodology for FDR control [13], which is less conservative and yields more discoveries than classic methodologies that control familywise error. As a result, the paradigm of FDR control has played a prominent role in the practice of simultaneous testing and related problems such as variable selection, and these have had a profound impact in numerous scientific fields in recent decades. Due to the evolving nature of modern applications, models for high-dimensional data in many contemporary problem domains are not always specified by discrete variables. Selecting models that exhibit many discoveries and small false positive error in such contexts requires a broadening of the scope of the FDR control paradigm.

Low-dimensional subspaces offer perhaps the simplest and most popular model for low-complexity geometric structure in high-dimensional data. As a result, subspace estimation is a fundamental

---

[*]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA.

[†]Departments of Computing and Mathematical Sciences and of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA.

task in a wide array of applications [9, 18, 19, 20, 21, 23, 52, 54]. In this paper, we consider the problem of subspace estimation from a perspective motivated by multiple testing. Specifically, given noisy data about a subspace of interest $\mathcal{U} \subset \mathbb{R}^n$, we wish to obtain an estimate $\widehat{\mathcal{U}} \subset \mathbb{R}^n$ such that the following quantity is controlled at a desired level:

$$\text{FDR}(\widehat{\mathcal{U}}) = \mathbb{E}\left[\frac{\text{tr}\left(P_{\widehat{\mathcal{U}}} P_{\mathcal{U}^\perp}\right)}{\max\{\dim(\widehat{\mathcal{U}}), 1\}} \right]. \tag{2}$$

Here $\mathcal{U}^\perp$ is the orthogonal complement of $\mathcal{U}$, $P_{\widehat{\mathcal{U}}}$ and $P_{\mathcal{U}^\perp}$ denote projections onto the subspaces $\widehat{\mathcal{U}}$ and $\mathcal{U}^\perp$, and $\dim(\widehat{\mathcal{U}})$ is the dimension of the subspace $\widehat{\mathcal{U}}$. The numerator inside the expectation is the sum of the squares of the cosines of the principal angles between $\widehat{\mathcal{U}}$ and $\mathcal{U}^\perp$, and it evaluates the extent to which the estimate $\widehat{\mathcal{U}}$ is not aligned with $\mathcal{U}$. The denominator normalizes by the total amount of discovery contained in the estimate $\widehat{\mathcal{U}}$. Thus, the quantity $\text{FDR}(\widehat{\mathcal{U}})$ is a geometric analog of the original FDR notion (1); in particular, when the subspaces $\mathcal{U}$ and $\widehat{\mathcal{U}}$ are axis-aligned (2) reduces to (1).

The quantity (2) as well as its unnormalized analog $\text{FD}(\widehat{\mathcal{U}}) = \mathbb{E}\left[\text{tr}\left(P_{\widehat{\mathcal{U}}} P_{\mathcal{U}^\perp}\right)\right]$ were proposed as false positive error control measures for subspace estimation in a recent paper [43]. Thus, while subspace estimation has received considerable attention over several decades owing to the ubiquity of low-rank matrix models in applications, a viewpoint grounded in false positive error control is not common in this vast literature. In particular, we are not aware of any prior work that controls FDR in subspace estimation; the methodological focus of the recent effort [43] was only on controlling the unnormalized variant FD (see Section 1.2 for more details).

Concretely, we investigate the model problem of estimating the column space of a rank-r matrix $\boldsymbol{A}$ corrupted by additive noise $\boldsymbol{E}$:

$$\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{E}. \tag{3}$$

We consider two cases of this model, one in which $\boldsymbol{A} \in \mathbb{S}_+^n$ is positive semidefinite and $\boldsymbol{E} \in \mathbb{S}^n$ is symmetric, and a second in which both $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{n \times m}$ are general nonsquare matrices. For simplicity, our exposition and development will focus on the symmetric case, with extensions to the nonsquare case studied in Section 4. The model (3) was first introduced by Johnstone [31] and has been widely studied in the literature [2, 4, 6, 18, 19], although these results do not directly provide false positive error control (see Section 1.2 for more details). In contrast our objective in this paper is the following – given $\boldsymbol{X}$ and $\alpha \in (0, 1)$, we wish to obtain an estimate $\widehat{\mathcal{U}} \subset \mathbb{R}^n$ for the column space of $\boldsymbol{A}$ such that $\widehat{\mathcal{U}}$ has as large a dimension as possible while ensuring that $\text{FDR}(\widehat{\mathcal{U}}) \leq \alpha$. We consider the estimator $\widehat{\mathcal{U}} = \widehat{\mathcal{U}}_k$ corresponding to the span of the k leading eigenvectors of $\boldsymbol{X}$, which is the most natural and widely used in practice. Thus, our goal is to identify as large a threshold k as possible such that $\text{FDR}(\widehat{\mathcal{U}}_k) \leq \alpha$.

As false positive error control for subspace estimation is an objective that has only been articulated recently, it is natural to ask whether there are any benefits to controlling FDR compared to the unnormalized FD in the model problem (3). The next numerical illustration demonstrates that the virtues of controlling FDR over FD are indeed analogous to what one observes with variable selection. Specifically, Figure 1 provides plots of $\text{FD}(\widehat{\mathcal{U}}_k)$ and $\text{FDR}(\widehat{\mathcal{U}}_k)$ against k for two problem instances of size $n = 1000$ and true rank $r = 20, 40$, with the noise $\boldsymbol{E}$ being drawn from the Gaussian Orthogonal Ensemble (GOE). Comparing the plots of FDR and the unnormalized FD, we observe that controlling FDR in subspace estimation offers the prospect of more powerful selection procedures that are also robust to the (unknown) dimension of the true subspace, much like with variable selection.
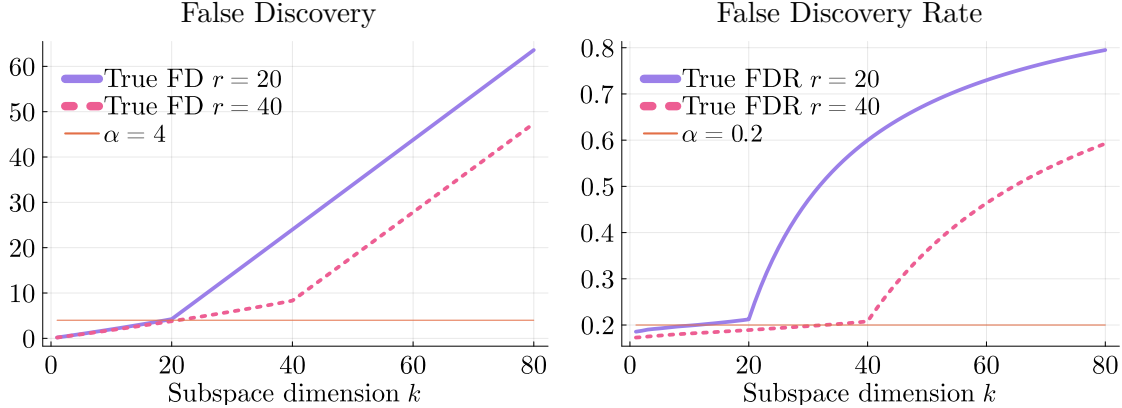
2

Figure 1: $\mathrm{FD}(\widehat{\mathcal{U}}_k)$ as a function of subspace dimension k on the left and $\mathrm{FDR}(\widehat{\mathcal{U}}_k)$ as a function of subspace dimension k on the right. We consider two problem instances of size $n = 1000$ and true rank $r = \{20, 40\}$. The noise $\boldsymbol{E}$ is drawn from a GOE. The horizontal red lines correspond to $\alpha = 4$ for FD and $\alpha = 0.2$ for FDR.

1.1 Our Contributions

The core difficulty with our approach is that $\mathrm{FDR}(\widehat{\mathcal{U}}_k)$ cannot be computed directly from $\boldsymbol{X}$, and therefore it is unclear how to select a suitable threshold k. The point of departure for our development is to address this challenge by considering an asymptotic characterization of $\mathrm{FDR}(\widehat{\mathcal{U}}_k)$. Specifically, consider a setting in which we have a sequence of growing-sized problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \mid \boldsymbol{A}_n \in \mathbb{S}_+^n, \ \boldsymbol{E}_n \in \mathbb{S}^n\}$, and let $\widehat{\mathcal{U}}_k^{(n)}$ denote the preceding k-dimensional principal subspace estimator on a problem instance of size n. If the rank of $\boldsymbol{A}_n$ is bounded above by a fixed $r \in \mathbb{N}$ for all n, the noise $\boldsymbol{E}_n$ has a compactly supported limiting spectral measure $\mu_{\boldsymbol{E}}$ as $n \to \infty$,[1] and the spectrum of $\boldsymbol{A}_n$ is 'well-separated' from that of $\boldsymbol{E}_n$, then we have the following asymptotic expression for $\mathrm{FDR}(\widehat{\mathcal{U}}_k^{(n)})$:

$$\lim_{n\to\infty} \mathrm{FDR}(\widehat{\mathcal{U}}_k^{(n)}) = \lim_{n\to\infty} 1 + \frac{1}{k} \sum_{i=1}^{\min\{k,r\}} \frac{G_{\mu_{\boldsymbol{E}}}(\lambda_i(\boldsymbol{X}_n))^2}{G'_{\mu_{\boldsymbol{E}}}(\lambda_i(\boldsymbol{X}_n))} \tag{4}$$

where the quantity G_μ is the *Cauchy transform* of a compactly supported measure μ on the real line:

$$G_\mu(z) = \int \frac{1}{z-t} d\mu(t) \qquad \text{for all } z \notin \mathrm{supp}(\mu). \tag{5}$$

Intuitively, the eigenvalues of $\boldsymbol{X}_n$ associated with the noise spectrum $\lambda(\boldsymbol{E}_n)$ form a bulk, and by 'well-separated' we mean that the eigenvalues associated with the signal $\boldsymbol{A}_n$ do not 'mix' into the bulk. We formally quantify this separation in Section 2, and we adapt this formula to more general settings in which the spectra of $\boldsymbol{A}_n$ and of $\boldsymbol{E}_n$ become 'mixed.' The expression (4) is derived based on the results from [11, 12] on the spectra of fixed-rank perturbations of large random matrices; see Section 1.2 for a discussion of prior approaches to low-rank estimation that employ a similar approach.

Returning to the case of a fixed n (and suppressing the explicit subscript), the upshot of the expression (4) is that the formula $1 + \frac{1}{k} \sum_{i=1}^{\min(k,r)} \frac{G_{\mu_{\boldsymbol{E}}}(\lambda_i(\boldsymbol{X}))^2}{G'_{\mu_{\boldsymbol{E}}}(\lambda_i(\boldsymbol{X}))}$ on the right-hand-side is more amenable

---

[1]The existence of a limiting spectral measure might sound too constraining, but several natural distributions satisfy this assumption; e.g., if $\boldsymbol{G}_n$ is a standard Gaussian matrix, then the spectral distribution of the Gaussian Orthogonal Ensemble $\boldsymbol{E}_n = \frac{1}{\sqrt{2n}}(\boldsymbol{G}_n + \boldsymbol{G}_n^\top)$ converges to the semicircle law.

to computation and estimation from $\boldsymbol{X}$, with the potential to obtain accurate estimates for large problem sizes. However, there still remain two difficulties corresponding to a lack of information about the limiting spectral measure $\mu_{\boldsymbol{E}}$ and a lack of knowledge about the true rank r. We describe how to address these challenges next, which yields an algorithm for controlling FDR in subspace estimation.

Assuming first that an estimate $\widehat{r}$ of the rank r is available, the Cauchy transform $G_{\mu_{\boldsymbol{E}}}$ and its derivative $G'_{\mu_{\boldsymbol{E}}}$ may be estimated empirically via $\widehat{G}(y) = \frac{1}{n-\widehat{r}}\sum_{j=\widehat{r}+1}^{n}\frac{1}{y-\lambda_j(\boldsymbol{X})}$ and $\widehat{G'}(y) = -\frac{1}{n-\widehat{r}}\sum_{j=\widehat{r}+1}^{n}\frac{1}{(y-\lambda_j(\boldsymbol{X}))^2}$, where the eigenvalues $\lambda(\boldsymbol{X})$ are ordered as $\lambda_1(\boldsymbol{X}) \geq \cdots \geq \lambda_n(\boldsymbol{X})$. With this estimate, we summarize our method for controlling FDR in subspace estimation in Algorithm 1; here, Steps 2 and 3 employ the above empirical estimates.

Algorithm 1 FDR control for eigensubspaces

Input: Observations $\boldsymbol{X}$, level α, and (observable) rank estimator $\widehat{r}$, e.g., via Algorithm 2.
Output: Estimate $\widehat{k}$ of number of top components to select.

Step 1 Obtain an eigenvalue factorization $\boldsymbol{X} = \widehat{\boldsymbol{U}}\boldsymbol{\Lambda}\widehat{\boldsymbol{U}}^{\top}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$.
Step 2 Define estimates of the Cauchy transform and its derivative

$$\widehat{G}(y) = \frac{1}{n-\widehat{r}}\sum_{j=\widehat{r}+1}^{n}\frac{1}{y-\lambda_j} \quad \text{and} \quad \widehat{G'}(y) = -\frac{1}{n-\widehat{r}}\sum_{j=\widehat{r}+1}^{n}\frac{1}{(y-\lambda_j)^2}.$$

Step 3 Let $\widehat{\mathcal{U}}_k$ be the span of the first k columns of $\boldsymbol{U}$ and estimate the FDR for different k:

$$\widehat{\mathrm{FDR}}(\widehat{\mathcal{U}}_k) = 1 + \frac{1}{k}\sum_{i=1}^{\min\{k,\widehat{r}\}}\frac{\widehat{G}(\lambda_i)^2}{\widehat{G'}(\lambda_i)} \tag{6}$$

Step 4 Output

$$\widehat{k} = \max\left\{k \in [n] \,\middle|\, \widehat{\mathrm{FDR}}(\widehat{\mathcal{U}}_k) \leq \alpha\right\}.$$

To ensure that Algorithm 1 works as desired, the approximation of the Cauchy transforms needs to be accurate, which in turn requires a good estimate of the rank r. Figure 2 illustrates the results of FDR estimates using (6) with different rank estimates $\widehat{r}$. We consider two problem instances of sizes $n = 500, 2000$, and in each case, the rank of $\boldsymbol{A}$ is $r = 20$ and the noise $\boldsymbol{E}$ is drawn from the GOE. In the right-hand-side of (4), the rank r is replaced by the estimates $\widehat{r} = r/2, r, 2r$. As the figures demonstrate, an overestimate of the rank r yields an underestimate of the true FDR, which would ultimately result in subspace selection procedures that do not provide control at the desired level; however, these FDR underestimates appear to become less significant for larger n. An underestimate of the rank r yields an overestimate of the true FDR, which would ultimately yield a procedure that provides control at the desired level but with nontrivial loss in power. Therefore, high-quality estimates of the correct rank are crucial for obtaining a powerful procedure for subspace estimation that provides the desired FDR control.

To estimate the rank of $\boldsymbol{A}$ from $\boldsymbol{X}$, we consider the same setup as in Figure 2 and we plot the eigenvalues of $\boldsymbol{X}$ in Figure 3; the true rank $r = 20$ is hardly a distinctive feature and is therefore difficult to estimate directly from these scree plots. To address this challenge, we again appeal to a basic observation from random matrix theory – the gaps between consecutive eigenvalues of random matrices are controlled as n grows large. Unlike i.i.d. random variables, the eigenvalues of random matrices are known to repel each other and form a rather rigid grid with gaps of similar size. This
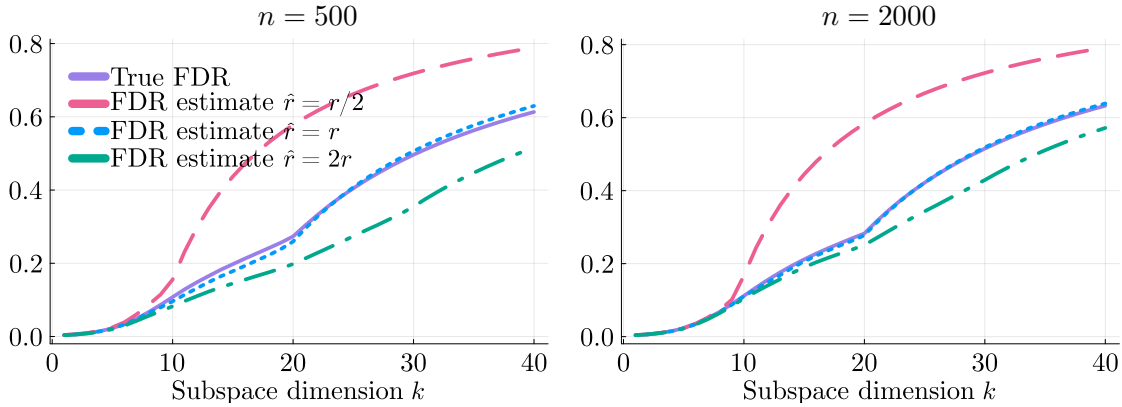
Figure 2: FDR estimates using (6) against subspace dimension k for different rank estimates $\widehat{r}$. The noise is drawn from a GOE, and the true rank of $\boldsymbol{A}$ is $r = 20$. The two figures display results for problem sizes $n = 500$ (left) and $n = 2000$ (right).
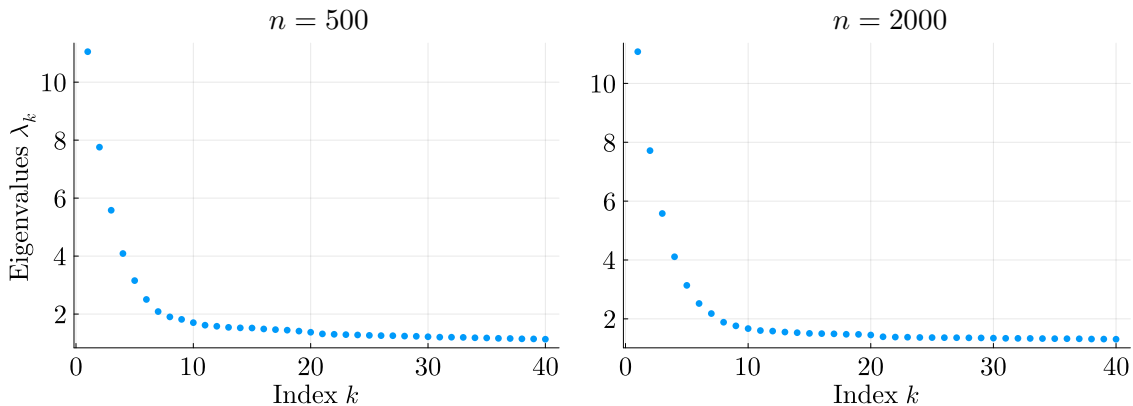

Figure 3: Eigenvalues against subspace dimension k. Same setting as in Figure 2.

phenomenon was observed at the inception of random matrix theory. Indeed, Wigner's seminal work on resonance absorption identified eigenvalue spacings as key objects of study [55]:

> *Perhaps I am now too courageous when I try to guess the distribution of the distances between successive levels (of energies of heavy nuclei). ... The question is simply, what are the distances of the characteristic values of a symmetric matrix with random coefficients?*

In that work, Wigner proposed a distribution for the spacings — now known as Wigner's surmise — which aimed to capture the repelling behavior. Subsequently, several works found connections with other problems in mathematics and the sciences [33, 49, 58]; famously, Montgomery established the relation of eigenvalue spacings to the zeros of the Riemann zeta function [34]. One of the main conclusions from the extensive line of work [10, 16, 17, 25, 37, 39, 46, 47, 48] studying eigenvalue spacings is that they decrease at a controlled rate when n goes to infinity. In particular, for eigenvalues in the interior of the bulk, the spacings are more or less uniform, i.e., $\lambda_i - \lambda_{i+1} = O(n^{-1})$. In comparison, spacings close to an edge of the bulk converge asymptotically to the Tracy-Widom distribution and tend to be more spread out, i.e., $\lambda_i - \lambda_{i+1} = O(n^{-2/3})$. In both cases — whether within the bulk or at the edge — spacings decrease at a much faster rate than $n^{-1/2}$, and this is the property we exploit to identify the rank.
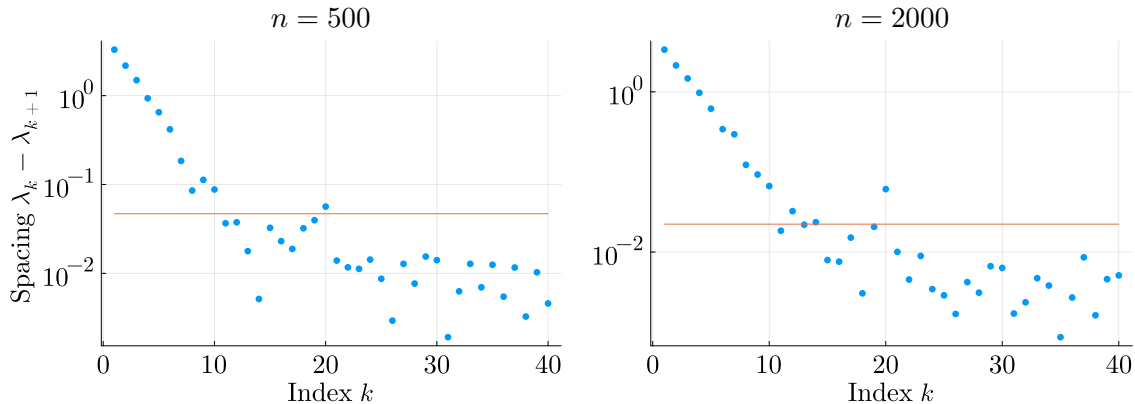
Figure 4: Eigenvalue spacings (log-scale) against subspace dimension k. Same setting as in Figure 2. The red line shows the threshold that we use in our procedure (Algorithm 2).

To estimate the rank r of $\boldsymbol{A}$, Algorithm 2 computes the spacings between eigenvalues of $\boldsymbol{X}$ and finds the largest index for which the spacing exceeds $p \cdot n^{-1/2}$ where $p > 0$ is a user-provided constant. Intuitively, the gaps associated with the noise $\boldsymbol{E}$ should be much smaller than $p \cdot n^{-1/2}$ as discussed above, while the gaps associated with the signal $\boldsymbol{A}$ have constant size; thus, the largest index for which the associated gap is above the threshold of $p \cdot n^{-1/2}$ should yield the components of $\boldsymbol{X}$ that correspond to the signal $\boldsymbol{A}$. Figure 4 provides an illustration of this approach in the problem instances from Figure 3.

Algorithm 2 RANKESTIMATE

Input: Nonincreasing spectrum $\lambda(\boldsymbol{X})$, *optional* $p > 0$ (default to $p = 1$).
Output: A rank estimate $\widehat{r}$.

Step 1 For each $j \leq n - 1$ compute the eigenvalue spacing

$$\Delta_j \leftarrow \lambda_{j-1}(\boldsymbol{X}) - \lambda_j(\boldsymbol{X}).$$

Step 2 Compute

$$\widehat{r} \leftarrow \max \left\{ j \in [n/2] \,\middle|\, \Delta_{j+1} > p \cdot n^{-1/2} \right\}.$$

In summary, Algorithms 1 and 2 constitute our methodology for FDR control in subspace estimation. Our procedure is parameter-free (the choice of p in Algorithm 2 is optional and can default to $p = 1$), and as we demonstrate in this paper, our method provably provides FDR control in subspace estimation under reasonable statistical assumptions.

Outline. The remainder of this paper is organized as follows. In Section 2, we discuss the technical assumptions that form the basis of our theoretical guarantees. Our main results on FDR control in subspace estimation are provided in Section 3; we also provide examples of natural matrix ensembles that satisfy the conditions required for our theorems. Section 4 provides extensions of our methods, assumptions, and results to the asymmetric case. We describe the results of numerical experiments on synthetic and real data in Section 5. Section 6 presents proofs of our theorems from Sections 3 and 4. Finally, Section 7 closes the paper with future directions.

1.2 Related Work

False discovery rate control beyond classic settings. We are not aware of methods for controlling FDR in problems beyond multiple testing, variable selection, and related variants [3, 8, 13, 14, 27, 28, 29, 30, 53, 60]. A line of work initiated by [15] studies FDR control in the setting of a continuum of binary hypotheses. The work of Taeb et al. [43] first proposed the FDR notion for subspace estimation, yet the methodological focus was on controlling the unnormalized variant FD of Figure 1. Moreover, even for this unnormalized variant, the results in [43] do not provide control at a user-specified level. In contrast, the present work provides control on FDR at a desired level for the particular case of a low-rank matrix corrupted by additive noise. Finally, we remark that our methods share conceptual similarity with the work of Storey [42] on multiple testing in that FDR control is provided by first estimating FDR and then controlling the estimate. Indeed, the key steps of Algorithm 1 center on estimating FDR, and much of our analysis is focused on establishing the accuracy of these estimates.

Principal component analysis. The stylized model we study (3) was first introduced by Johnstone [31] and has since been extensively investigated [2, 4, 6, 18, 19]. There has been recent progress on the information-theoretic limits of PCA in detecting a low-rank signal [38]. A related line of work [2, 4, 18, 19] has focused on error bounds for the recovered eigenvectors and singular vectors. There is also a line of work initiated by [22] in which a testing approach is employed to threshold the singular values. In broad overview, our work departs qualitatively from this prior literature by providing methods that control false discovery rate in subspace estimation, and we are not aware of prior efforts in this direction.

Matrix recovery via asymptotics. Building on the random matrix theory results of Beynach-Georges and Nadakuditi [11, 12], several authors, including Donoho, Gavish, and Romanov [24] and Nadakuditi [35] developed methodology for denoising low-rank matrices corrupted by additive noise. Broadly, the approach in these papers is based on identifying exact formulas that hold in the limit of large n and employing data-driven estimates of these asymptotic quantities for the case of finite n. The approach in the present paper is similar to this prior body of work but with the important distinction that we seek subspace estimates with FDR control guarantees rather than a low-rank matrix estimate with the smallest mean-squared error (MSE). The precise asymptotic formulas corresponding to these two criteria are different. Moreover, controlling the MSE for low-rank approximation is a distinct goal than controlling the FDR for subspace estimation; in some applications identifying such a subspace with control on false positive error is significant for subsequent steps in the data analysis pipeline such as dimension reduction.

Eigenvalue spacings. Several papers have studied eigenvalue gaps in random matrices [16, 17, 25, 37, 46, 47, 48], and the underlying phenomemon that the eigenvalue gaps decay at a predictable rate is arguably a universal one, i.e., it holds in many settings. To our knowledge, this phenomenon has not been explicitly employed in low-rank or subspace estimation previously. As described above, the problem of subspace estimation with false positive error control brings to the fore the challenge of estimating the rank of the underlying low-rank signal well. In this context, we crucially leverage the eigenvalue rigidity of large random matrices [10, 26, 39].

1.3 Notation

Given a real-valued function $f \colon \mathbb{R} \to \mathbb{R}$ and a constant $c \in \mathbb{R}$, we denote the right-sided limit by $f(c^+) = \lim_{t \downarrow c} f(t)$ and the left-sided limit by $f(c^-) = \lim_{t \uparrow c} f(t)$. We use δ_{ij} to denote the indicator function of $i = j$. We employ the shorthand $[n] = \{1, \ldots n\}$. The symbols $\mathbb{S}^n$ and $\mathbb{S}^n_+$ denote the sets of symmetric and (symmetric) positive semidefinite matrices, respectively. We index eigenvalues and singular values in nonincreasing order, e.g., $\lambda_1(\boldsymbol{M}) \geq \cdots \geq \lambda_n(\boldsymbol{M})$ for any symmetric matrix $\boldsymbol{M} \in \mathbb{S}^n$. Given a sequence $\{\boldsymbol{M}_n\}$ of symmetric random matrices of increasing dimension, we use $\mu_{\boldsymbol{M}_n}$ to denote the empirical eigenvalue distribution of $\boldsymbol{M}_n$, i.e.,

$$\mu_{\boldsymbol{M}_n} = \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_i(\boldsymbol{M}_n)}. \tag{7}$$

Similarly, for a sequence of asymmetric matrix $\{\boldsymbol{M}_n\} \subseteq \mathbb{R}^{n \times m}$ we use $\mu_{\boldsymbol{M}_n}$ to denote the empirical singular value distribution of $\boldsymbol{M}_n$. Given sequences $(a_n)_n$ and $(b_n)_n$, we write $a_n = o(b_n)$ if $\lim_{n \to \infty} a_n / b_n = 0$. The L^p norm of a random variable $\mathbf{Z}$ is denoted $\|\mathbf{Z}\|_p := \mathbb{E}[|\mathbf{Z}|^p]^{1/p}$.

2 Background and Assumptions

In this section we describe the main assumptions that form the basis of our development in the symmetric case. Theoretical guarantees for the symmetric case are presented in Section 3, with the proofs given in Section 6. The nonsquare case is discussed in Section 4.

As described in Section 1, the core of Algorithm 1 is an estimate-then-control approach for selecting a subspace – we estimate the FDR and use our estimate as a proxy to select the number of top components k. The key insight underlying this approach comes from random matrix theory: for a sequence of problem instances of increasing size $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \mid \boldsymbol{A}_n \in \mathbb{S}^n_+, \boldsymbol{E}_n \in \mathbb{S}^n\}$ satisfying certain conditions, the FDR associated to the estimator that selects the principal eigenspaces of $\boldsymbol{X}_n$ converges asymptotically to a quantity that is amenable to estimation from the spectrum of $\boldsymbol{X}_n$. We describe next the assumptions on problem sequences that facilitate our development. The first assumption primarily concerns the signal sequence $\{\boldsymbol{A}_n\}$.

Assumption 1. The following conditions hold for the sequence $\{\boldsymbol{A}_n\}$ for all n.

(**Independence**) The distributions of $\boldsymbol{A}_n \in \mathbb{S}^n_+$ and $\boldsymbol{E}_n \in \mathbb{S}^n$ are independent.

(**Spectrum of the low-rank signal**) There exists a constant $r \in \mathbb{N}$ and deterministic real numbers $\theta_1 \geq \cdots \geq \theta_r > 0$ such that $\mathrm{rank}(\boldsymbol{A}_n) = r$ and $\lambda_i(\boldsymbol{A}_n) = \theta_i$ for all $i \leq r$.

(**Invariance**) The eigenvector distribution of $\boldsymbol{A}_n$ or of $\boldsymbol{E}_n$ is orthogonally invariant.

The second assumption centers around properties of the noise sequence $\{\boldsymbol{E}_n\}$.

Assumption 2. The following conditions hold for the sequence $\{\boldsymbol{E}_n\}$.

(**Spectrum**) The empirical eigenvalue distribution $\mu_{\boldsymbol{E}_n}$, defined as in (7), converges almost surely weakly to a deterministic compactly supported measure $\mu_{\boldsymbol{E}}$ as n goes to infinity.

(**Edge convergence**) Let the infimum and supremum of $\mathrm{supp}(\mu_{\boldsymbol{E}})$ be denoted by a and b. The smallest and largest eigenvalues of $\boldsymbol{E}_n$ converge almost surely to a and to b, respectively.

(**Decay at the edge**) The limiting distribution $\mu_{\boldsymbol{E}}$ has a density $f_{\boldsymbol{E}}$ such that as $t \to b$ with $t < b$, we have that $f_{\boldsymbol{E}}(t) \sim c \cdot (b - t)^\alpha$ for some constant $c > 0$ and exponent $\alpha \in (0, 1]$. An analogous decay holds for the lower edge a.

We comment briefly on these assumptions. The requirement that $\boldsymbol{A}_n$ is positive semidefinite is not strict; one could drop this condition by defining a method that considers components at both ends of the spectrum. However, we keep this assumption since it simplifies the presentation. The invariance condition in Assumption 1 is necessary for our analysis, but it could potentially be slightly relaxed; see [12, Section 2]. The three requirements in Assumption 2 might appear restrictive at first sight, but they hold for several natural random matrix ensembles such as the GOE from the introduction; we refer the reader to Section 3.2 for further examples.

Assumptions 1 and 2 underpin the success of Algorithm 1; see Theorem 3.1 in Section 3. Under these assumptions, the asymptotic behavior of the eigenvalues and eigenvector of $\boldsymbol{X}_n$ can be pinned down in terms of the distribution $\mu_{\boldsymbol{E}}$ and the spectrum of $\boldsymbol{A}_n$ based on recent advances in random matrix theory [5, 7, 11, 12]. Broadly speaking, the spectrum of $\boldsymbol{X}_n$ has two components: a bulk of eigenvalues coming from the noise and a small set of eigenvalues corresponding to those of $\boldsymbol{A}_n$. While the empirical distribution of the bulk converges towards $\mu_{\boldsymbol{E}}$, the eigenvalues associated to those of $\boldsymbol{A}_n$ exhibit a sharp phase transition. In particular, for any $i \in [r]$,

$$\lambda_i(\boldsymbol{X}_n) \to \begin{cases} G_{\mu_{\boldsymbol{E}}}^{-1}(1/\theta_i) & \text{if } \theta_i > 1/G_{\mu_{\boldsymbol{E}}}(b^+), \\ b & \text{otherwise,} \end{cases} \tag{8}$$

where $G_{\mu_{\boldsymbol{E}}}$ is the Cauchy transform (5) of $\mu_{\boldsymbol{E}}$. This type of phenomenon is known as the *BBP phase transition*, named after Baik, Ben Arous, and Peche, whose pioneering work [7] unveiled this behavior for the first time. Intuitively, if an eigenvalue of $\boldsymbol{A}_n$ is above the noise level, i.e., $\theta_i > 1/G_{\mu_{\boldsymbol{E}}}(b^+)$, then the associated component of $\boldsymbol{X}_n$ is not subsumed by the bulk. On the other hand, if an eigenvalue is below the noise level, then its associated eigenvalue is absorbed inside the bulk. Further, correlations between the eigenvectors of $\boldsymbol{A}_n$ and $\boldsymbol{X}_n$ can be characterized in terms of the Cauchy transform of $\mu_{\boldsymbol{E}}$ and the spectrum of $\boldsymbol{A}_n$; we defer the additional details to Appendix B. By understanding such asymptotic properties of $\boldsymbol{X}_n$, one can characterize the limiting FDR for any fixed k as follows:

$$\limsup_{n\to\infty} \text{FDR}(\widehat{\mathcal{U}}_k^{(n)}) \leq \lim_{n\to\infty} 1 + \frac{1}{k} \sum_{i=1}^{\min\{k,r^\star\}} \frac{G_{\mu_{\boldsymbol{E}}}(\lambda_i(\boldsymbol{X}_n))^2}{G'_{\mu_{\boldsymbol{E}}}(\lambda_i(\boldsymbol{X}_n))} \tag{9}$$

where $r^\star$ denotes the so-called *observable rank*, which corresponds to the number of components that are not subsumed by the noise:

$$r^\star := \#\{j \mid \theta_j > 1/G_{\mu_{\boldsymbol{E}}}(b^+)\}. \tag{10}$$

Expression (9) is a generalization of (4) from Section 1 as it allows for problem instances in which the signal spectrum is not well-separated from that of the noise; see Proposition 6.2 for a proof of (9). We shall see in Section 3 that when $r = r^\star$, the above inequality holds with equality and the lim-sup can be replaced by a limit.

Turning our attention next to assumptions under which Algorithm 2 provides high-quality rank estimates, we begin by elaborating on the rationale provided in Section 1. Recall that Algorithm 2 chooses the largest index for which the associated eigenvalue gap is greater than $p \cdot n^{-1/2}$. As articulated in Section 1, the reasoning behind this choice is based on the properties of the eigenvalue spacings of the noise $s_i = \lambda_i(\boldsymbol{E}_n) - \lambda_{i+1}(\boldsymbol{E}_n)$. These spacings decrease at a rate of $s_i = O(n^{-1})$ inside the bulk and more slowly $O(n^{-2/3})$ near the edges [10, 16, 25, 39, 47, 48]. In principle, we could have picked $n^{-\omega}$ with any $\omega < 2/3$ in Step 2 of Algorithm 2, but we have observed that the more aggressive $n^{-1/2}$ exhibits good practical performance. The next assumption codifies this discussion, and it provides the basis for the success of Algorithm 2.

Assumption 3. The following conditions hold for the sequence $\boldsymbol{E}_n$.

(**Connected support**) The support of the asymptotic spectral measure $\mu_{\boldsymbol{E}}$ is connected.

(**Spacings**) Let $s_j^{(n)} = \lambda_j(\boldsymbol{E}_n) - \lambda_{j+1}(\boldsymbol{E}_n)$, for all large enough n and all $i \leq n$

$$s_i^{(n)} = o\left(n^{-1/2}\right) \qquad \text{almost surely.}$$

Without the connected support assumption, there could be large eigenvalue spacings within the noise, and thus, our method could fail to identify the right rank. One way to bypass the connectedness assumption is to assume that we have access to an upper bound on the rank R of the signal sequence $\{\boldsymbol{A}_n\}$. With this information, one could only focus on the first $R - 1$ eigenvalue spacings, and for sufficiently large n, Algorithm 2 would exclude large spacings produced by the noise. In Section 3.2, we describe natural matrix ensembles that satisfy Assumptions 2 and 3.

3 Main Results

We present here the main theoretical results of our work for the symmetric case. Our first set of results in Section 3.1 pertain to guarantees that Algorithms 1 and Algorithm 2 are correct under the assumptions described in Section 2 for all large n, i.e., they provide FDR control at the desired level as well as accurate rank estimates. Next, we show in Section 3.2 that a number of natural random matrix ensembles satisfy Assumptions 2 and 3 of Section 2.

3.1 Guarantees for FDR Control

Recall that we consider a sequence of growing-sized problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \mid \boldsymbol{A}_n \in \mathbb{S}_+^n, \ \boldsymbol{E}_n \in \mathbb{S}^n\}$. We denote the column space of $\boldsymbol{A}_n$ by $\mathcal{U}^{(n)}$, and we denote the k-dimensional principal subspace of $\boldsymbol{X}_n$, i.e., the span of the k eigenvectors of $\boldsymbol{X}_n$ corresponding to the k largest eigenvalues, by $\widehat{\mathcal{U}}_k^{(n)}$.

One challenge with characterizing $\text{tr}\left(P_{\widehat{\mathcal{U}}_k^{(n)}} P_{\mathcal{U}^\perp}\right)$ in the definition of $\text{FDR}(\widehat{\mathcal{U}}_k^{(n)})$ from (2) – beyond the case of well-separated signal and noise spectra as in (4) – is that the eigenvectors of $\boldsymbol{X}_n$ associated with eigenvalues that are subsumed by the noise in the BBP transition, i.e., $\theta_j \leq 1/G_{\mu_{\boldsymbol{E}}}(b^+)$, are not clearly characterized in terms of their correlation with the corresponding eigenvectors of $\boldsymbol{A}_n$. Intuitively, one would expect that these pairs of eigenvectors are asymptotically uncorrelated; whether this is the case is a significant open question in random matrix theory that has only recently been answered in more restrictive settings [5, Theorem 1.3], and might require additional assumptions [11, Remark 2.6]. We circumvent this difficulty by considering only the span $\bar{\mathcal{U}}^{(n)}$ of the first $r^\star$ components of $\boldsymbol{A}_n$ corresponding to the observable rank (10), and defining the following quantities:

$$\overline{\text{FDR}}(\widehat{\mathcal{U}}_k^{(n)}) = \mathbb{E}\left[\frac{\text{tr}\left(P_{\widehat{\mathcal{U}}_k^{(n)}} P_{\bar{\mathcal{U}}^{(n)\perp}}\right)}{k}\right] \quad \text{and} \quad \text{FDR}^\infty(k) = \lim_{n \to \infty} \overline{\text{FDR}}(\widehat{\mathcal{U}}_k^{(n)}). \tag{11}$$

We prove in Appendix 6.1.1 that FDR^∞ is well-defined. As $\bar{\mathcal{U}}^{(n)} \subset \mathcal{U}^{(n)}$, we have that $\text{FDR}(\widehat{\mathcal{U}}_k^{(n)}) \leq \overline{\text{FDR}}(\widehat{\mathcal{U}}_k^{(n)})$. We leverage this observation in our next result by proving that Algorithm 1 provides control on $\overline{\text{FDR}}(\widehat{\mathcal{U}}_k^{(n)})$ at the desired level, which in turn implies control on $\text{FDR}(\widehat{\mathcal{U}}_k^{(n)})$. The proof of this result is deferred to Section 6.1.1.

Theorem 3.1 (**FDR control**). *Consider a sequence of problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \mid \boldsymbol{A}_n \in \mathbb{S}_+^n, \ \boldsymbol{E}_n \in \mathbb{S}^n\}$ for which Assumptions 1 and 2 hold. Further, suppose the rank estimate $\widehat{r}$ required for Algorithm 1 satisfies*

$$r^\star \leq \widehat{r} \leq r$$

almost surely for all large n. For a generic $\alpha \in (0, 1)$ and for all large n, the output of Algorithm 1 satisfies $\overline{\mathrm{FDR}}(\widehat{\mathcal{U}}_k^{(n)}) \leq \alpha$, which in turn implies that $\mathrm{FDR}(\widehat{\mathcal{U}}_k^{(n)}) \leq \alpha$. Moreover, if $\alpha \leq \mathrm{FDR}^\infty(r)$, the output $\widehat{k}$ of Algorithm 1 satisfies

$$\widehat{k} = \max\{k \in [r] \mid \overline{\mathrm{FDR}}(\widehat{\mathcal{U}}_k^{(n)}) \leq \alpha\}.$$

In words, this theorem states that the true FDR is controlled by Algorithm 1 for large enough n. Moreover, Algorithm 1 maximizes power for sufficienty small α in the sense that it selects the largest-dimensional principal subspaces of $\boldsymbol{X}_n$ that satisfy $\overline{\mathrm{FDR}}(\widehat{\mathcal{U}}_k^{(n)}) \leq \alpha$.

We comment on several aspects of the result. First, the requirement that α is generic is to ensure that α does not equal $\mathrm{FDR}^\infty(k)$ for any $k \in \mathbb{N}$, and this is arguably an artifact of our analysis. Nonetheless, our theorem holds for any value of α outside of this set of measure zero. Second, in terms of power maximization, the bound $\alpha \leq \mathrm{FDR}^\infty(r)$ is appropriate as any value of α greater than this quantity could lead to the selection of more components than the rank of $\boldsymbol{A}_n$; we expect these components to arise from the noise $\boldsymbol{E}_n$ and to have little correlation with the signal. Finally, it is natural to wonder whether the inequality $\mathrm{FDR}(\widehat{\mathcal{U}}_k^{(n)}) \leq \overline{\mathrm{FDR}}(\widehat{\mathcal{U}}_k^{(n)})$ is loose, which would lead to our method being too conservative. Numerical experiments suggest that this is not the case, and in fact, it appears that $\mathrm{FDR}(\widehat{\mathcal{U}}_k^{(n)})$ converges to $\mathrm{FDR}^\infty(k)$ as $n \to \infty$. It would be interesting to establish such convergence in the future; such an analysis would rely on resolving the open question we described in the discussion preceding (11) on obtaining a characterization of the correlation between eigenvectors of $\boldsymbol{X}_n$ and $\boldsymbol{A}_n$ below the BBP threshold.

Next, we recall that Algorithm 1 must be supplied with an accurate rank estimate. The following result shows that for noise matrices with controlled eigenvalue spacings, Algorithm 2 produces accurate estimates.

Theorem 3.2 (**Rank estimate**). *Consider a sequence of problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \mid \boldsymbol{A}_n \in \mathbb{S}_+^n, \ \boldsymbol{E}_n \in \mathbb{S}^n\}$ for which Assumptions 1, 2, and 3 hold. Then, the output of Algorithm 2, $\textsc{RankEstimate}(\boldsymbol{X}_n)$, almost surely satisfies*

$$r^\star \leq \textsc{RankEstimate}(\boldsymbol{X}_n) \leq r \qquad \text{for all large enough } n.$$

This result follows from an application of Weyl's interlacing inequalities, and its proof is presented in Section 6.1.2. Numerical experiments suggest that Algorithm 2 outputs $r^\star$ even when $r \neq r^\star$; nonetheless, existing results from random matrix theory do not appear to be strong enough to establish this equality. We leave this as an interesting open question for future work.

3.2 Matrix Ensembles satisfying Assumptions 2 and 3

In this section, we identify two ensembles — Wigner and Wishart matrices — for which Assumptions 2 and 3 hold. For Assumption 3 in particular, we leverage results from random matrix theory concerning the rigidity of the empirical spectrum [26, 39]. Intuitively, rigidity amounts to the i'th eigenvalue of an $N \times N$ random matrix being anchored around the i/N-quantile of their asymptotic spectral distribution; this phenomenon ensures that eigenvalue spacings shrink at a controlled rate.

We start by describing Wigner matrices, a broad family of symmetric random matrices.

Definition 3.3. A *Wigner ensemble* is a sequence of symmetric random matrices $\{\boldsymbol{H}_n \in \mathbb{S}^n\}$ for which the following conditions hold for all n (we suppress the subscript n in stating these conditions).

11

1. (**Independence**) The upper-triangular entries $\{H_{ij} \mid 1 \le i \le j \le n\}$ are independent.

2. (**Mean and variance**) For each i, j we have $\mathbb{E}\, H_{ij} = 0$ and $\mathbb{E}\, |H_{ij}|^2 = \frac{1}{n}(1 + O(\delta_{ij}))$; the factor $(1 + O(\delta_{ij}))$ requires the off-diagonal entries to have variances equal to $\frac{1}{n}$ but it allows for the diagonal entries to have nonuniform variances that scale as $\frac{1}{n}$ as n grows large.

3. (**Moment bounds**) For each $p \in \mathbb{N}$, there exists a constant $C_p > 0$ such that $\|\sqrt{n} H_{ij}\|_p \le C_p$ for all indices i, j.

The first two conditions are standard. The last condition entails some degree of concentration on the entries of $\boldsymbol{H}$. In particular, with $C_p \asymp \sqrt{p}$ the third condition reduces to subgaussian decay of the entries, while with $C_p \asymp p$ the condition reduces to subexponential decay [51]. A popular example of Wigner matrices is the GOE, that is, $(\boldsymbol{G} + \boldsymbol{G}^\top)/\sqrt{2n}$ with $\boldsymbol{G}$ being a random matrix with i.i.d. standard Gaussian entries.

Wigner ensembles have been widely studied over the last century, beginning with the seminal work of Eugene Wigner on energy levels of heavy nuclei [56, 57]. It is now well-known that the empirical distribution of the spectrum of Wigner matrices converges weakly almost surely to the semicircle law, i.e., to the distribution with density

$$\varrho(t) = \frac{1}{2\pi} \sqrt{(4 - t^2)_+},$$

which is supported on the interval $(-2, 2)$. Wigner matrices satisfy Assumption 2 whenever the distribution of the eigenvectors is rotationally invariant, e.g., when the entries are Gaussian. The next result shows that the spacings of Wigner matrices satisfy Assumption 3. The proof follows from an eigenvalue rigidity result first established in [26], and we present it in Section 6.1.3.

Proposition 3.4 (**Wigner matrix**). *Suppose the sequence of matrices $\{\boldsymbol{E}_n \in \mathbb{S}^n\}$ is a Wigner ensemble. We have that Assumptions 2 and 3 hold.*

Next, we turn to another salient class of random matrices known as Wishart ensembles. These arise commonly in applications by virtue of being sample covariance matrices of data with independent coordinates.

Definition 3.5. A *Wishart ensemble* is a sequence of symmetric matrices $\{\boldsymbol{W}_n = \boldsymbol{Y}_{n \times m} \boldsymbol{Y}_{n \times m}^\top \in \mathbb{S}^n\}$ with $\{\boldsymbol{Y}_{n \times m} \in \mathbb{R}^{n \times m}\}$ being a sequence of random matrices for which the following hold for all n, m (we suppress the subscripts n, m in stating these conditions).

1. (**Independence**) The entries $\{Y_{ij} \mid 1 \le i \le n, 1 \le j \le m\}$ are independent.

2. (**Mean and variance**) For all indices i, j we have $\mathbb{E}\, Y_{ij} = 0$ and $\mathbb{E}\, |Y_{ij}|^2 = \frac{1}{\sqrt{m}}$.

3. (**Tail decay**) There exists a constant $c > 0$ such that for all i, j we have that $\mathbb{P}(|Y_{ij}| \ge t) \le c \exp\left(-(t\sqrt{m})^c\right)$.

Finally, we have the following requirement on the sizes n, m.

4. (**Limiting size**) The asymptotic ratio of the matrix dimensions $\vartheta := \lim_{n,m \to \infty} n/m$ satisfies

$$0 < \vartheta < \infty.$$

All these conditions are rather standard in the literature. The third condition plays an analogous role to that of the moment bound in the third condition of Definition 3.3, and it reduces to the tails being subgaussian and subexponential when $\vartheta = 2$ and $\vartheta = 1$, respectively [51]. A common

example of Wishart ensembles are sample covariance matrices of standard Gaussian vectors, i.e., $\boldsymbol{W} = \frac{1}{m}\boldsymbol{G}\boldsymbol{G}^\top$ with the entries of $\boldsymbol{G}$ being i.i.d. standard Gaussians.

Vladimir Marchenko and Leonid Pastur's influential work [32] established the asymptotic convergence of the spectrum of certain Wishart ensembles to the Marchenko-Pastur distribution, which is characterized by the following density:

$$\nu(t) = \frac{1}{2\pi t}\sqrt{((c_+ - t)(t - c_-))_+} \quad \text{with} \quad c_\pm = \left(1 \pm \sqrt{\vartheta}\right)^2. \tag{12}$$

Subsequent efforts showed the universality of this distribution for general Wishart ensembles [39]. Unlike Wigner matrices, Wishart ensembles might fail to satisfy the decay condition Thus, to satisfy Assumption 2, the matrices $\boldsymbol{Y}$ have to be asymptotically rectangular so that $\vartheta \neq 1$. Interestingly, an asymptotically rectangular shape is also required to obtain control of the eigenvalue spacings at the lower end of the spectrum.

Proposition 3.6 (Wishart matrices). *Suppose the sequence of matrices $\{\boldsymbol{E}_n \in \mathbb{S}^n\}$ is a Wishart ensemble with $\vartheta \neq 1$. We have that the Assumptions 2 and 3 hold.*

The proof of this proposition is completely analogous to that of Proposition 3.4 and leverages an eigenvalue rigidity result established in [39]. We give a proof sketch in Section 6.1.4.

4 Extension to Asymmetric Matrices

In this section, we extend our methods to asymmetric matrices. The results and algorithms that we present are conceptually analogous to those in the preceding sections concerning the symmetric case. However, the lack of symmetry makes several of the statements slightly more involved. We consider the problem of estimating the column and row spaces of a rank-r matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ given the observation

$$\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{E}$$

where $\boldsymbol{E}$ represents additive noise. We denote the column and row spaces of $\boldsymbol{A}$ by $\mathcal{U}$ and $\mathcal{V}$, respectively, and these two subspaces represent the signal we wish to recover. As before, our approach is to our goal is to control the false discovery rate incurred by picking the top k components of the singular value decomposition (SVD) of $\boldsymbol{X} = \widehat{\boldsymbol{U}}\boldsymbol{\Sigma}\widehat{\boldsymbol{V}}^\top$. Namely, we wish to control

$$\text{FDR}_{\text{l}}\left(\widehat{\mathcal{U}}^{(k)}\right) = \mathbb{E}\left[\frac{\text{tr}\left(P_{\widehat{\mathcal{U}}_k}P_{\mathcal{U}^\perp}\right)}{\max\{k, 1\}}\right] \quad \text{and} \quad \text{FDR}_{\text{r}}\left(\widehat{\mathcal{V}}^{(k)}\right) = \mathbb{E}\left[\frac{\text{tr}\left(P_{\widehat{\mathcal{V}}_k}P_{\mathcal{V}^\perp}\right)}{\max\{k, 1\}}\right] \tag{13}$$

where $\widehat{\mathcal{U}}^{(k)} = \text{span}\{\widehat{\boldsymbol{u}}_1, \dots \widehat{\boldsymbol{u}}_k\}$ and $\widehat{\mathcal{V}}^{(k)} = \text{span}\{\widehat{\boldsymbol{v}}_1, \dots \widehat{\boldsymbol{v}}_k\}$ are the subspaces spanned by the top k left and right singular vectors of $\boldsymbol{X}$, respectively. We note here that in some contexts a data analyst may only seek FDR control for one of these subspaces; for example, one might only wish to perform dimension reduction of the feature space in some applications. Accordingly, the algorithm we describe in the section can control just one or both FDRs concurrently. Throughout this section, we assume without loss of generality that $n \leq m$.

4.1 FDR Control for Column and Row Spaces

We begin by presenting extensions of Algorithms 1 and 2 based on a similar template, but with appropriate adaptations to use singular values and to estimate the relevant quantities for the asymmetric FDR (13).

Algorithm 3 FDR control for singular subspaces
***

Input: Observation $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, level α, and rank estimator $\widehat{r}$ (e.g., via Algorithm 4).
Output: Estimate $\widehat{k}$ of the number of top components to be selected.

Step 1 Obtain the SVD of $\boldsymbol{X} = \widehat{\boldsymbol{U}}\boldsymbol{\Sigma}\widehat{\boldsymbol{V}}^\top$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n$.
Step 2 For any $q \in (0, 1]$ define the following transform estimates and their derivatives

$$\widehat{\varphi}(y; q) = \frac{q}{n - \widehat{r}} \cdot \sum_{j=\widehat{r}+1}^{n} \frac{y}{y^2 - \sigma_j^2} + \frac{1-q}{y}, \qquad \widehat{\varphi}'(y; q) = -\frac{q}{n - \widehat{r}} \cdot \sum_{j=\widehat{r}+1}^{n} \frac{y^2 + \sigma_j^2}{(y^2 - \sigma_j^2)^2} - \frac{1-q}{y^2},$$

$$\widehat{D}(y) = \widehat{\varphi}(y; 1) \cdot \widehat{\varphi}\left(y; \frac{n}{m}\right), \quad \text{and} \quad \widehat{D}'(y) = \widehat{\varphi}'(y; 1) \cdot \widehat{\varphi}\left(y; \frac{n}{m}\right) + \widehat{\varphi}(y; 1) \cdot \widehat{\varphi}'\left(y; \frac{n}{m}\right)$$

Step 3 Let $\widehat{\mathcal{U}}^{(k)}$ (resp. $\widehat{\mathcal{V}}^{(k)}$)be the span of the first k left (resp. right) singular vectors, set $q = 1$ (resp. $q = \frac{n}{m}$), and estimate the left FDR (resp. right) for different k's

$$\widehat{\mathrm{FDR}}_{\mathtt{l}}(\widehat{\mathcal{U}}^{(k)}) = 1 - \frac{2}{k} \sum_{i=1}^{\min\{k,\widehat{r}\}} \frac{\widehat{D}(\sigma_i) \cdot \varphi(\sigma_i; q)}{\widehat{D}'(\sigma_i)}.$$

Step 4 To control the left FDR (resp. right), output

$$\widehat{k} = \max\left\{ k \in [n] \,\middle|\, \widehat{\mathrm{FDR}}_{\mathtt{l}}(k) \leq \alpha \right\}.$$

To simultaneously control left and right FDR, use $\max\left\{ \widehat{\mathrm{FDR}}_{\mathtt{l}}(k), \widehat{\mathrm{FDR}}_{\mathtt{r}}(k) \right\} \leq \alpha$ instead.
***

To facilitate the analysis of this algorithm, we consider as before a sequence of growing-size problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \in \mathbb{R}^{n \times m}\}$. The second dimension $m = m_n$ grows with n; we omit the subscript to avoid cluttering notation. We make the following two assumptions.

Assumption 4. The sequence $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \in \mathbb{R}^{n \times m}\}$ satisfies the following for all n.

(**Independence**) The distributions of $\boldsymbol{A}_n$ and $\boldsymbol{E}_n$ are independent.

(**Spectrum of the low-rank signal**) There exists a constant $r \in \mathbb{N}$ and deterministic real numbers $\theta_1 \geq \cdots \geq \theta_r > 0$ such that $\mathrm{rank}(\boldsymbol{A}_n) = r$ and $\sigma_i(\boldsymbol{A}_n) = \theta_i$ for all $i \leq r$.

(**Invariance**) For either $\boldsymbol{A}_n$ or $\boldsymbol{E}_n$, the distributions of left and right singular vectors are independent of each other and orthogonally invariant.

(**Asymptotic size**) The ratio between the dimensions n, m converges $n/m \to \vartheta$ with $\vartheta \in (0, 1]$.

Assumption 5. The sequence $\{\boldsymbol{E}_n \in \mathbb{S}^n\}$ satisfies the following.

(**Spectrum**) The empirical singular value distribution $\mu_{\boldsymbol{E}_n}$, defined as

$$\mu_{\boldsymbol{E}_n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\sigma_i(\boldsymbol{E}_n)} \tag{14}$$

converges almost surely weakly to a deterministic compactly supported measure $\mu_{\boldsymbol{E}}$ as $n \to \infty$.

(**Edge convergence**) Let the infimum and supremum of $\mathrm{supp}(\mu_{\boldsymbol{E}})$ be denoted by a and b. The smallest and largest singular values of $\boldsymbol{E}_n$ converge almost surely to a and to b, respectively.

(**Decay at the edge**) The limiting distribution $\mu_{\boldsymbol{E}}$ has a density $f_{\boldsymbol{E}}$ such that as $t \to b$ with $t < b$, we have that $f_{\boldsymbol{E}}(t) \sim c \cdot (b - t)^{\alpha}$ for some constant $c > 0$ and exponent $\alpha \in (0, 1]$.

These assumptions closely parallel Assumptions 1 and 2. The only additional assumption that we impose is that the limiting ratio $n/m \xrightarrow{a.s.} \vartheta \in (0, 1]$. Thus, our matrices are asymptotically wide; this is not a significant constraint as we can transform $\boldsymbol{X}$ into $\boldsymbol{X}^{\top}$ when we have a tall matrix.

In what follows, we will focus our discussion on controlling FDR for column spaces, i.e., the quantity $\mathrm{FDR}_{\mathbf{l}}$, and we highlight the necessary changes for controlling $\mathrm{FDR}_{\mathbf{r}}$. Under the preceding assumptions, the singular values also experience a BBP transition, and some of the singular values of $\boldsymbol{A}_n$ might get subsumed by the noise. To quantify this, we define a notion of observable rank $r^{\star}$, akin to the case of eigenvalues, which in the present context reduces to

$$r^{\star} := \#\{i \in [r] \mid \theta_i > D_{\mu_{\boldsymbol{E}}}(b^+)^{1/2}\}.$$

The function $D_{\mu_{\boldsymbol{E}}}$ is known as the D-transform of $\mu_{\boldsymbol{E}}$ and it is to the asymmetric case what the Cauchy transform is for the symmetric case. In particular, for any $q \in (0, 1]$ let

$$\varphi_{\mu_{\boldsymbol{E}}}(z; q) := q \cdot \int \frac{z}{z^2 - t^2} d\mu(t) + \frac{1 - q}{z}$$

Then the D-transform of $\mu_{\boldsymbol{E}}$ is given by

$$D_{\mu_{\boldsymbol{E}}}(x) = \varphi_{\mu_{\boldsymbol{E}}}(x; 1) \cdot \varphi_{\mu_{\boldsymbol{E}}}(x; \vartheta)$$

where ϑ is the limiting aspect ratio from Assumption 4. As in the symmetric case, we let $\bar{\mathcal{U}}^{(n)} = \mathrm{span}\{u_1, \ldots, u_{r^{\star}}\}$ be the subspace generated by top $r^{\star}$ components of $\boldsymbol{A}_n$ and we define the upper-bounding FDR

$$\overline{\mathrm{FDR}}_{\mathbf{l}}(\widehat{\mathcal{U}}_k^{(n)}) = \mathbb{E}\left[\frac{\mathrm{tr}\left(P_{\widehat{\mathcal{U}}_k^{(n)}} P_{\bar{\mathcal{U}}^{(n)\perp}}\right)}{k}\right] \quad \text{and} \quad \mathrm{FDR}_{\mathbf{l}}^{\infty}(k) = \lim_{n \to \infty} \overline{\mathrm{FDR}}_{\mathbf{l}}(\widehat{\mathcal{U}}_k^{(n)}). \quad (15)$$

Controlling $\overline{\mathrm{FDR}}_{\mathbf{l}}$ provides control on $\mathrm{FDR}_{\mathbf{l}}$ as $\mathrm{FDR}_{\mathbf{l}} \leq \overline{\mathrm{FDR}}_{\mathbf{l}}$. Moreover, one can obtain an asymptotic characterization of the upper-bounding FDR for any fixed $k \in \mathbb{N}$,

$$\mathrm{FDR}_{\mathbf{l}}^{\infty}(k) = \lim_{n \to \infty} 1 - \frac{2}{k} \sum_{i=1}^{k \wedge r^*} \frac{D_{\mu_{\boldsymbol{E}}}(\sigma_i(\boldsymbol{X}_n)) \cdot \varphi_{\mu_{\boldsymbol{E}}}(\sigma_i(\boldsymbol{X}_n); 1)}{D'_{\mu_{\boldsymbol{E}}}(\sigma_i(\boldsymbol{X}_n))}. \quad (16)$$

The quantity $\mathrm{FDR}_{\mathbf{r}}^{\infty}(k)$ is controlled by an analogous limit where we substitute $\varphi_{\mu_{\boldsymbol{E}}}(\sigma_i(\boldsymbol{X}_n); 1)$ by $\varphi_{\mu_{\boldsymbol{E}}}(\sigma_i(\boldsymbol{X}_n); \vartheta)$. The expression (16) presents the key insight underlying Algorithm 3: the method employs the spectrum of $\boldsymbol{X}_n$ and a rank estimate $\widehat{r}$ to estimate the right-hand-side of (16) and uses this estimate as a proxy for the true FDR.

The following is our main result concerning columnspace FDR control; rowspace FDR control can be obtained with straightforward modifications.

Theorem 4.1 (**FDR control for asymmetric matrices**). *Consider a sequence of growing-size problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \in \mathbb{R}^{n \times m}\}$ that satisfy Assumptions 4 and 5. Further, suppose the rank estimate $\widehat{r}$ required for Algorithm 3 is such that*

$$r^{\star} \leq \widehat{r} \leq r$$

almost surely for all large n. For a generic $\alpha \in (0, 1)$ and for all large n, the output of Algorithm 3 satisfies $\mathrm{FDR}_{\mathbf{l}}\left(\widehat{\mathcal{U}}_k^{(n)}\right) \leq \alpha$, which in turn implies that $\mathrm{FDR}_{\mathbf{l}}\left(\widehat{\mathcal{U}}_k^{(n)}\right) \leq \alpha$. Moreover, if $\alpha \leq \mathrm{FDR}_{\mathbf{l}}^{\infty}(r)$, the output $\widehat{k}$ of Algorithm 3 satisfies

$$\widehat{k} = \max\{k \in [r] \mid \overline{\mathrm{FDR}}_{\mathbf{l}}\left(\widehat{\mathcal{U}}_k^{(n)}\right) \leq \alpha\}.$$

The proof of this theorem is given in Section 6.2.1.

4.2 Rank Estimation

The extension of our rank estimator for the asymmetric case is direct – we substitute eigenvalues with singular values. The method relies on an equivalent version of Assumption 3.

Algorithm 4 RANKESTIMATE for asymmetric matrices

Input: Nonincreasing spectrum $\sigma(\boldsymbol{X})$, *optional $p > 0$* (default to $p = 1$).
Output: Estimate of the identifiable rank $\widehat{r}$.

Step 1 For each $j \leq n - 1$ compute the eigenvalue spacing

$$\Delta_j \leftarrow \sigma_{j-1}(\boldsymbol{X}) - \sigma_j(\boldsymbol{X}).$$

Step 2 Compute

$$\widehat{r} \leftarrow \max\left\{ j \in [n/2] \,\Big|\, \Delta_{j+1} > p \cdot n^{-1/2} \right\}.$$

Assumption 6. The following conditions hold for the sequence $\boldsymbol{E}_n$.

(**Connected support**) The support of the asymptotic distribution $\mu_{\boldsymbol{E}}$ is connected.

(**Singular value spacings**) Let $s_j^{(n)} = \sigma_j(\boldsymbol{E}_n) - \sigma_{j+1}(\boldsymbol{E}_n)$, for all large n and all $i \leq n$,

$$s_i^{(n)} = o\left(n^{-1/2}\right) \qquad \text{almost surely.}$$

Next, we obtain the asymmetric counterpart of Theorem 3.2.

Theorem 4.2 (**Rank estimate**). *Consider a sequence of problem instances $\{\boldsymbol{X}_n = \boldsymbol{A}_n + \boldsymbol{E}_n \mid \boldsymbol{A}_n \in \mathbb{S}_+^n, \ \boldsymbol{E}_n \in \mathbb{S}^n\}$ for which Assumptions 4, 5, and 6 hold. Then, the output of Algorithm 4, RANKESTIMATE($\boldsymbol{X}_n$), almost surely satisfies*

$$r^\star \leq \text{RANKESTIMATE}(\boldsymbol{X}_n) \leq r \qquad \text{for all large enough } n.$$

The proof of this theorem is given in Section 6.2.2.

4.3 Asymmetric Matrix Ensemble satisfying Assumptions 5 and 6

We conclude this section with an example of a matrix ensemble that satisfies our assumptions.

Definition 4.3. A *Wishart-factor ensemble* is a sequence of $\{\boldsymbol{Y} \in \mathbb{R}^{n \times m}\}$ with $n \leq m$ for which the following hold.

1. (**Independence**) The entries $\{\boldsymbol{Y}_{ij} \mid 1 \leq i \leq n, 1 \leq j \leq m\}$ are independent.

2. (**Mean and variance**) For all n and all $i, j \leq n$ we have $\mathbb{E}\,\boldsymbol{Y}_{ij} = 0$ and $\mathbb{E}\,|\boldsymbol{Y}_{ij}|^2 = \frac{1}{\sqrt{m}}$.

3. (**Tail decay**) There exists a constant $c > 0$ such that for all n and all $i, j \leq n$ we have that $\mathbb{P}(|\boldsymbol{Y}_{ij}| \geq t) \leq c \exp\left(-(t\sqrt{m})^c\right)$.

4. (**Limiting size**) The asymptotic ratio of the matrix dimensions $\vartheta := \lim_{n \to \infty} n/m$ satisfies

$$0 < \vartheta \leq 1.$$

These matrices simply correspond to the factors in the definition of a Wishart ensemble. The next proposition shows that Wishart-factor ensembles satisfy the desired assumptions.

Proposition 4.4 (Wishart factors). *Assume that the matrices $\{\boldsymbol{E}_n\}$ form a Wishart-factor ensemble with $\vartheta \neq 1$ and that their singular vector distribution is orthogonally invariant. Then, Assumptions 5 and 6 hold.*

The proof of this result is given in Section 6.3.

5 Numerical Experiments

In this section, we present the results experiments that demonstrate the utility of our methods. The first set of experiments is focused on synthetically-generated data while the second set concerns datasets from RNA sequencing and hyperspectral imaging. All the code used to generate the experimental results and figures in this paper can be found at

<div align="center">

https://github.com/mateodd25/FDRSubspaces.jl.

</div>

In these experiments, we employ a data-driven approach for selecting the parameter p in Algorithm 2 by setting it to $\frac{3}{5} \cdot \texttt{median}\left(\{\Delta_j\}_{j=1}^{n-1}\right) \cdot n$. This choice aims to provide a robust estimate of noise scaling (although any positive constant p is in theory adequate under Assumption 3).

5.1 Synthetic Data

We generate matrices of the form $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{E}$ with a variety of ensembles for the noise $\boldsymbol{E}$ and spectral models for the low-rank signal matrices $\boldsymbol{A}$, which we now summarize.

Noise. We consider seven noise ensembles; see Table 1. These are generated as a product of a matrix drawn from a GOE and another matrix for which we vary the distribution as follows:

$$\boldsymbol{F} = \boldsymbol{GW}. \tag{17}$$

Here $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ has identically distributed independent entries with $\boldsymbol{G}_{ij} \sim N(0, 1/n)$ and $\boldsymbol{W} \in \mathbb{R}^{m \times m}$ has a different distribution depending on the problem, thus leading to matrices with correlated entries. The noise matrix $\boldsymbol{E}$ is derived from $\boldsymbol{F}$ in one of several ways; see Table 1. The noise models in Table 1 are known to satisfy Assumption 2 and 5 in all cases, see for example [24]. However, to the best of our knowledge, it is not known whether the last four models in Table 1 have a rigid spectrum in the sense of Assumptions 3 and 6. Nonetheless, we observe that our methods perform reasonably well for all of them.

Low-rank signal. For each noise model from Table 1, we consider three types of low-rank signal models: `well-separated`, `barely-separated`, and `entangled`. Each of these matrices has rank 20, and their spectrum is given as follows:

$$\theta_i = \texttt{BBP} + \texttt{shift} + 10 \cdot 1.3^{1-i} \quad \text{for all } i \in \{1, \dots, 20\}. \tag{18}$$

The θ_i's correspond to eigenvalues for symmetric problems and to singular values for asymmetric problems. In the preceding expression, `BBP` is an upper estimate of the BBP transition point that changes depending on the noise distribution, `shift` changes depending on the type of low-rank

Distribution	Symmetric	Template for $\boldsymbol{E}$	Distribution of $\boldsymbol{W}$
Wigner	Yes	$(\boldsymbol{F} + \boldsymbol{F}^\top)/2$	Deterministic $\boldsymbol{W} = \boldsymbol{I}$
Wishart	Yes	$\boldsymbol{F}\boldsymbol{F}^\top$	Deterministic $\boldsymbol{W} = \boldsymbol{I}$
WishartFactor	No	$\boldsymbol{F}$	Deterministic $\boldsymbol{W} = \boldsymbol{I}$
Fisher	Yes	$(\boldsymbol{F} + \boldsymbol{F}^\top)/2$	$\boldsymbol{W} \sim$ Wishart
FisherFactor	No	$\boldsymbol{F}$	$\boldsymbol{W} \sim$ Wishart
Uniform	Yes	$(\boldsymbol{F} + \boldsymbol{F}^\top)/2$	Diagonal $\boldsymbol{W}_{ii} = \mathbf{Unif}([0,1])$
UniformFactor	No	$\boldsymbol{F}$	Diagonal $\boldsymbol{W}_{ii} = \mathbf{Unif}([0,1])$

Table 1: Noise distributions: the matrices $\boldsymbol{F}$ is derived from $\boldsymbol{W}$ as described by (17).

signal as follows:

$$
\texttt{shift} := \begin{cases} 1.0 & \text{if generating a } \texttt{well-separated} \text{ instance,} \\ 0.0 & \text{if generating a } \texttt{barely-separated} \text{ instance,} \\ -10 \cdot 1.3^{-9} & \text{if generating a } \texttt{entangled} \text{ instance.} \end{cases} \tag{19}
$$

Intuitively, well-separated instances correspond to low-rank signals having spectra that are easier to distinguish from the noise, barely-separated instances to low-rank signals having spectra that are barely distinguishable from the noise, and entangled instances to low-rank signals having some spectral components that are indistinguishable from the noise. With the above selection of the $\texttt{shift}$ parameter, the observable rank $r^\star$ is equal to 20 (which is the rank r of the low-rank signal) for the first two instances and to 10 for entangled instances.

We report results for two noise models in this section, namely the $\texttt{Wigner}$ (symmetric) and the $\texttt{WishartFactor}$ (asymmetric) ensembles. The conclusions are analogous in the remaining cases and therefore we defer their description to Appendix A. The results for the $\texttt{Wigner}$ and the $\texttt{WishartFactor}$ cases are given in Figures 5 and 6, which plot the true FDR and the estimated FDR against the subspace dimension k. The 'true FDR' is obtained using a Monte Carlo method with 100 repetitions. The 'estimated FDR' is computed using Algorithms 1 and 2 for the symmetric case and Algorithms 3 and 4 for the asymmetric case.

We make two observations regarding our results. First, our FDR estimates tend to conservative by virtue of being larger than the true FDR; this is a consequence of our rank estimation method yielding smaller values than the actual rank. This is a feature of both the $\texttt{barely-separated}$ and the $\texttt{entangled}$ cases, particularly for the smallest problem dimension (corresponding to $n = 100$). Nonetheless, our approach reliably produces subspace estimates for which the false discovery rate is controlled at the desired level. Second, we observe that the estimated FDR converges rapidly to the true FDR despite the asymptotic nature of our theoretical results, which suggests that our method remains practical even for modest problem dimensions n. In particular, this convergence is notably quicker for smaller values of k, which are the cases of most practical interest in obtaining FDR control below a desired threshold.
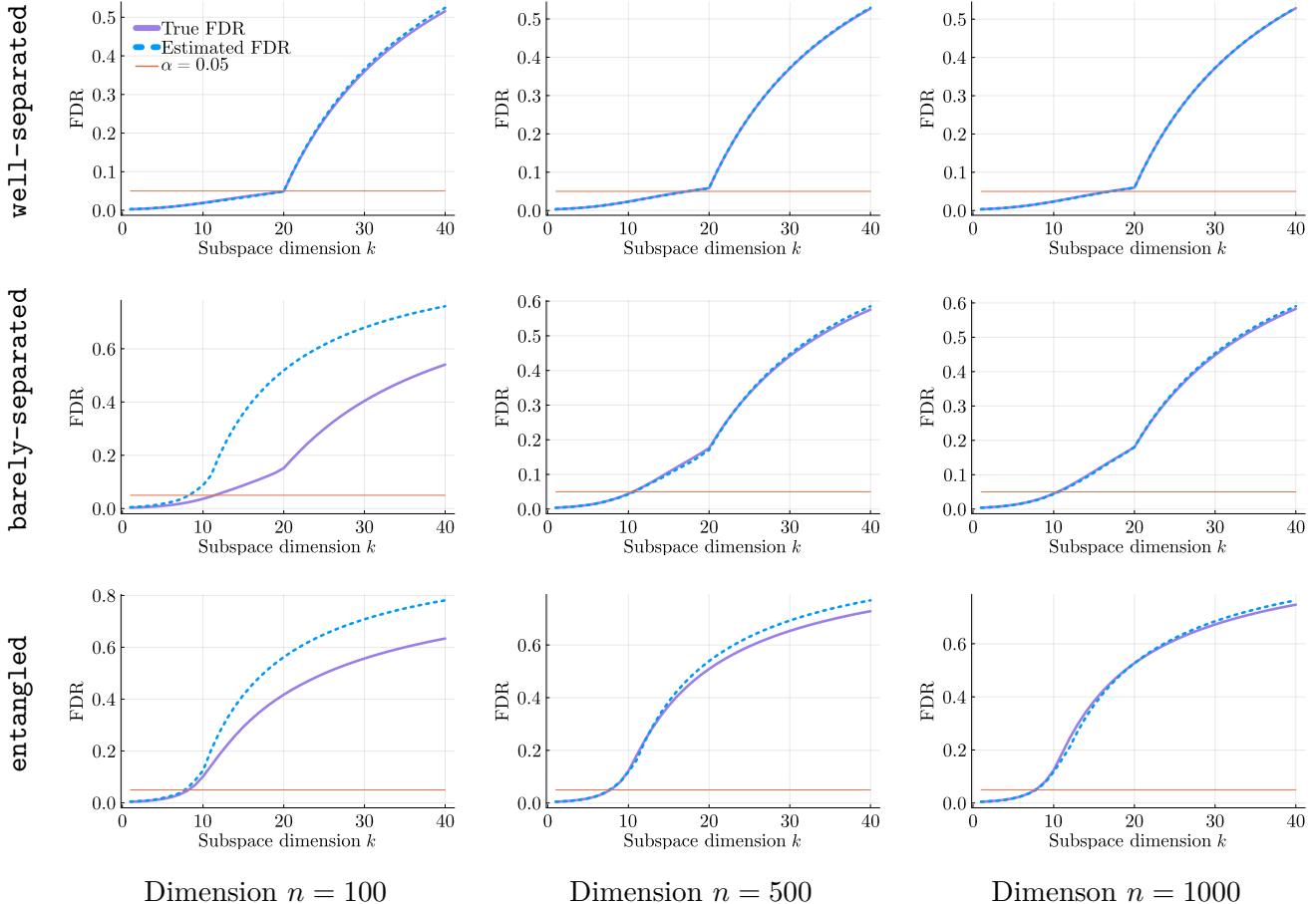
Figure 5: Results on FDR estimation for the `Wigner` ensemble using Algorithms 1 and 2: the columns correspond to problem sizes $n = 100, 500, 1000$ and the rows correspond to the three models for the spectrum of the low-rank signal from (18).

5.2 Experiments with Real Data

Here we present the results obtained by applying our methods to data from single-cell RNA sequencing and from hyperspectral imaging. We compare rank estimates obtained using $p = \frac{3}{5} \cdot \texttt{median}\left(\{\Delta_j\}_{j=1}^{n-1}\right) \cdot n$ — the default value in our implementation of Algorithm 2 — with those obtained using $0.5p$ and $2p$, with lower values yielding larger rank estimates.

Single-cell RNA sequencing. Traditional RNA sequencing technologies could only process mixed populations of cells and as a result they were limited in their ability to identify gene expression profiles of individual cells. The recent development of single-cell sequencing techniques [45] has had a profound impact on both biology and medicine, although it yields massive data matrices of gene expression counts for each cell. Typical workflows employ PCA to identify features that provide a suitable low-dimensional embedding of the gene expressions. We apply our methodology in this context. We consider a publicly available dataset of Peripheral Blood Mononuclear Cells [1] that consists of 2,700 single cells, and we obtain a data matrix $\boldsymbol{X} \in \mathbb{R}^{1872 \times 2700}$ after using a standard pipeline to preprocess the data [59]. The left panel of Figure 7 gives the empirical distribution of the singular values of $\boldsymbol{X}$; the shape of this distribution closely matches the assumptions required for
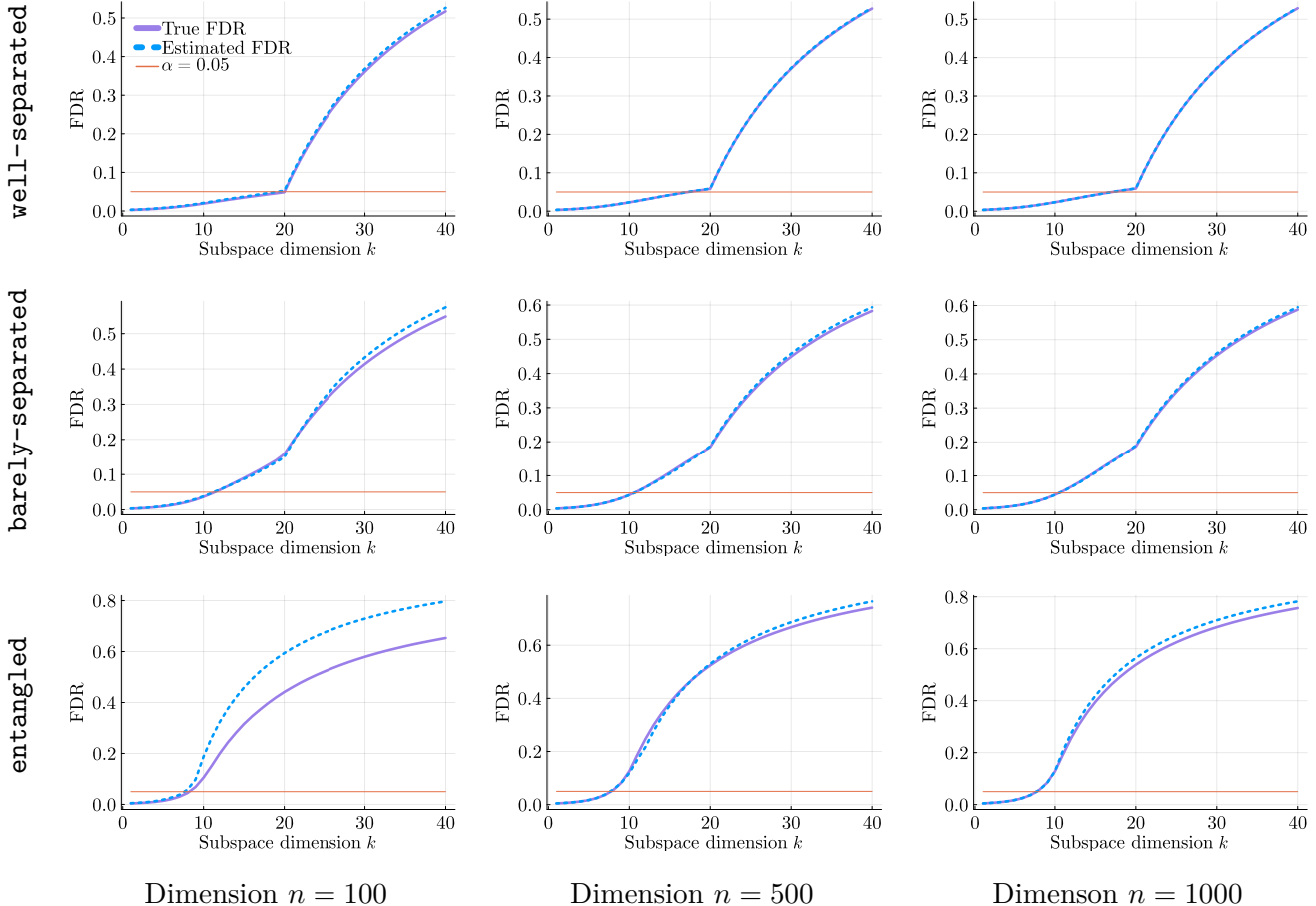
19

Figure 6: Results on FDR estimation for the `WishartFactor` ensemble using Algorithms 1 and 2: the columns correspond to problem sizes $n = 100, 500, 1000$ and the rows correspond to the three models for the spectrum of the low-rank signal from (18). The proportion between dimensions is set to $n/m = 1/2$.

our theory to hold. The right panel of Figure 7 depicts FDR estimates obtained using the three different rank estimates. These FDR estimates change gracefully with different rank estimates, especially exhibiting robustness for smaller subspace dimensions k. For $\alpha = 0.05$, we obtain a subspace estimate of dimension $\widehat{k} = 4$.

Hyperspectral imaging (HSI). HSI is a technique that collects information across a wider spectrum of wavelengths than traditional cameras, and it is employed in application domains such as remote sensing, environmental monitoring, agriculture, and mineralogy for the identification and analysis of different materials and substances based on their unique spectral signatures. HSI data is represented as a third-order tensor $\boldsymbol{W} \in \mathbb{R}^{n \times m \times d}$, where d represents the number of spectral bands. The value of d is 3 for RGB images and it is considerably larger for images obtained via HSI. As spectral bands that are closer in wavelength exhibit high correlation, it is common to reduce dimensionality by applying PCA to the $d \times d$ outer-product matrix $\boldsymbol{X} = \frac{1}{nm} \sum_{ij} \boldsymbol{W}_{ij:} \boldsymbol{W}_{ij:}^{\top}$. For a comprehensive explanation of this process, we direct interested readers to [40]. For this experiment, we consider the Indian Pines dataset [50] in which $\boldsymbol{W} \in \mathbb{R}^{145 \times 145 \times 220}$ and we apply our methodology to obtain estimates with FDR control of the column space of the outer-product matrix $\boldsymbol{X}$. The left
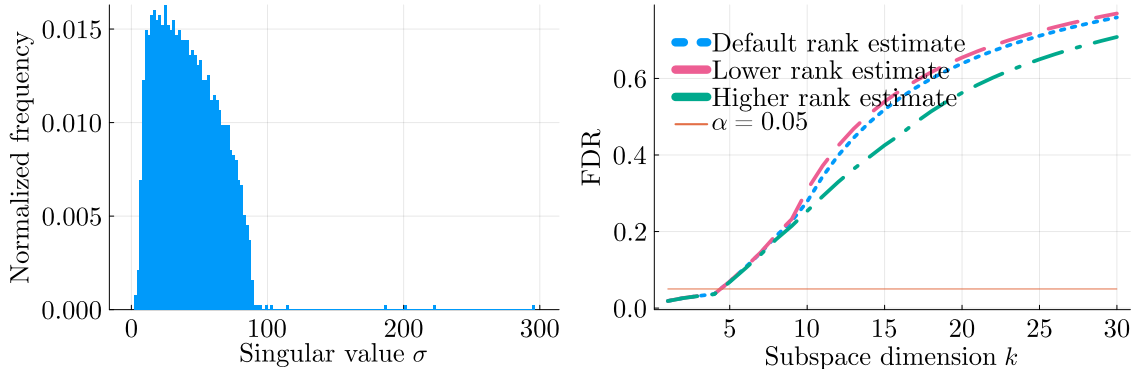
Figure 7: Results for single-cell RNA data: the left plot is a histogram of the singular values of the data matrix, and the right plot shows three FDR estimate using different rank estimators.
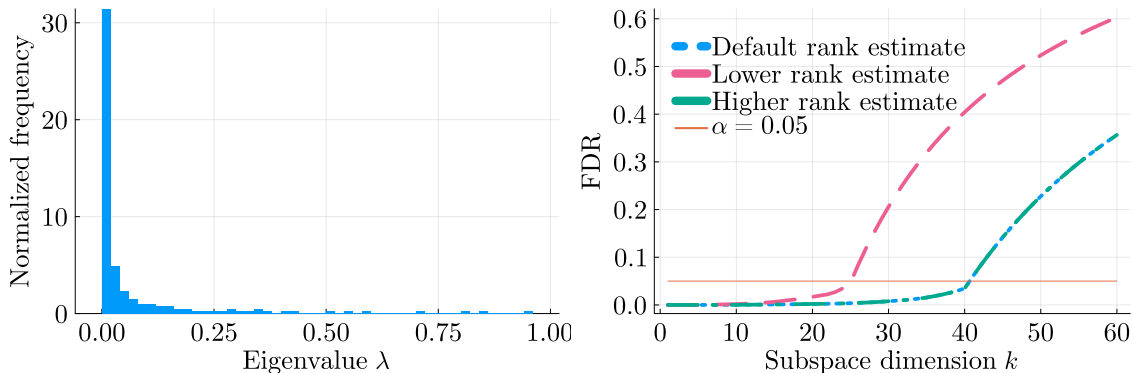

Figure 8: Results for Hyperspectral Imaging: the left plot is a histogram of the singular values of the data, and the right plot shows our FDR estimate using three rank estimators.

panel of Figure 8 displays the empirical distribution of the eigenvalues of $\boldsymbol{X}$, while the right panel displays FDR estimates obtained using our three rank estimation methods. The spectral profile of the HSI data is notably distinct, displaying a less pronounced spectral bulk compared to the RNA sequencing data. Although our FDR estimates are close to each other for smaller values of k, the absence of a well-defined bulk in the HSI spectrum appears to negatively impact the robustness of the FDR estimates for larger k.

6 Proofs

In this section, we present all the proofs of the results stated earlier in the paper. The first section concerns the symmetric case and here we give all the relevant details. The second section concerns the asymmetric case, which largely parallels the symmetric case; hence, we only describe the modifications relative to the symmetric case. For notational convenience, we will drop the dependency on $\mathcal{U}_k$ in $\mathrm{FDR}(\mathcal{U}_k)$ and simply write $\mathrm{FDR}(k)$.

6.1 Symmetric Case

6.1.1 Proof of Theorem 3.1

We start by sketching a roadmap for the proof. Recall that $r^\star = \max\{j \in [r] \mid \theta_j > 1/G_{\mu_{\mathbf{E}}}(b^+)\} \le r$ is the number of components that do not get subsumed by the noise. To set the stage, we define and recall a couple of FDR-related sequences that are going to be relevant in the proof.

(**Limiting FDR**)

$$\mathrm{FDR}^\infty(k) = 1 + \frac{1}{k} \sum_{i=1}^{\min\{k,r^\star\}} \frac{1}{\theta_i^2 G'_{\mu_{\mathbf{E}}}(G_{\mu_{\mathbf{E}}}^{-1}(1/\theta_i))} \quad \text{and} \quad k^\infty = \max\left\{j \in [r] \mid \mathrm{FDR}^\infty(j) \le \alpha\right\}. \tag{20}$$

(**True FDR**)

$$\mathrm{FDR}^{(n)}(k) = \mathbb{E}\left[\frac{\mathrm{tr}\left(P_{\widehat{\mathcal{U}}_k^{(n)}} P_{\mathcal{U}^{(n)\perp}}\right)}{k}\right]. \tag{21}$$

(**Upper-bound FDR**) Let $\bar{\mathcal{U}} = \mathrm{span}\{u_1, \ldots, u_{r^\star}\}$ be the subspace generated by top $r^\star$ components of $\boldsymbol{A}$ and define

$$\overline{\mathrm{FDR}}^{(n)}(k) = \mathbb{E}\left[\frac{\mathrm{tr}\left(P_{\widehat{\mathcal{U}}_k^{(n)}} P_{(\bar{\mathcal{U}}^{(n)})^\perp}\right)}{k}\right] \quad \text{and} \quad \bar{k} = \max\{j \in [r] \mid \overline{\mathrm{FDR}}^{(n)}(j) \le \alpha\}. \tag{22}$$

Observe that since $\bar{\mathcal{U}} \subseteq \mathcal{U}$, we have that $\mathrm{FDR}^{(n)}(k) \le \overline{\mathrm{FDR}}^{(n)}(k)$.

(**Distribution-aware FDR Estimate**)

$$\widetilde{\mathrm{FDR}}^{(n)}(k) = 1 + \frac{1}{k} \sum_{j=1}^{\min\{k,r^\star\}} \frac{G_{\mu_{\mathbf{E}}}\left(\lambda_i^{(n)}\right)^2}{G'_{\mu_{\mathbf{E}}}\left(\lambda_j^{(n)}\right)}. \tag{23}$$

Note that unlike the FDR estimate in Algorithm 1, the estimate above uses the true Cauchy transform of the limiting distribution and the true observable rank $r^\star$.

In the remainder of this section, we drop the subscript of n and state $\mathcal{U}, \lambda_i$ instead of $\mathcal{U}^{(n)}, \lambda_i^{(n)}$ for notational brevity whenever the index of n is implicit from the context. Recall that $\widehat{\mathrm{FDR}}^{(n)}(k)$ denotes the FDR estimate given by Algorithm 1, and $\widehat{k}$ is the output of the same algorithm. Theorem 3.1 follows from two simple steps:

Step 1: Convergence of FDR estimate and true FDR. We show that for each fixed $k \in \mathbb{N}$

$$\widehat{\mathrm{FDR}}^{(n)}(k) \to \mathrm{FDR}^\infty(k) \quad \text{a.s.} \quad \text{and} \quad \overline{\mathrm{FDR}}^{(n)}(k) \to \mathrm{FDR}^\infty(k)$$

as n goes to infinity.

Step 2: Equality of maximizers. Using the convergence above, we can conclude that indeed $\bar{k}$ and $\widehat{k}$ match for large n provided that the limiting $\mathrm{FDR}^\infty(k^\infty) \ne \alpha$.

When $r \ne r^\star$ this gives control for the true FDR since we trivially have the upper bound $\mathrm{FDR}^{(n)}(\widehat{k}) \le \overline{\mathrm{FDR}}^{(n)}(\widehat{k}) \le \alpha$, where the second inequality follows since $\widehat{k} = \bar{k}$ and by definition

$\bar{k}$ controls the upper-bound FDR. Further, note that if $r = r^\star$, then $\mathrm{FDR}^{(n)} = \overline{\mathrm{FDR}}^{(n)}$ and the preceding reasoning may be applied to conclude the theorem. Thus, for the remainder of this argument, we focus on bounding $\overline{\mathrm{FDR}}^{(n)}$ instead of $\mathrm{FDR}^{(n)}$. Let us start with the formal proof.

Step 1. First, we establish an intermediate result that ensures that some of the FDR quantities defined above are nonincreasing and takes on values in $[0, 1]$.

Lemma 6.1. *Consider a probability measure over the real line μ with bounded support. Let $b = \sup\{t : t \in \mathrm{supp}(\mu)\}$. Then, the map $\Phi \colon (b, \infty) \to \mathbb{R}$ given by $\Phi(x) = \frac{G_\mu(x)^2}{G'_\mu(x)}$ yields a $[-1, 0]$-valued continuous nonincreasing function.*

Proof. Note that

$$\Phi(y) = -\frac{\left[\mathbb{E}_{\theta \sim \mu}(y - \theta)^{-1}\right]^2}{\mathbb{E}_{\theta \sim \mu}(y - \theta)^{-2}},$$

this quantity is well-defined in the domain (b, ∞). Observe that Φ is defined as a ratio of continuous functions with the denominator bounded away from zero; hence we conclude that Φ is continuous. Moreover, Φ takes nonpositive values and by Jensen's inequality $\Phi(y) \geq -1$. Since both G_μ and G'_μ are integrals of bounded continuous functions over compact sets, they are differentiable. Taking a derivative of Φ reveals

$$\Phi'(y) = \frac{2G_\mu(y)G'_\mu(y)^2 - G_\mu(y)^2 G''_\mu(y)}{G'_\mu(y)^2}.$$

To prove the lemma it suffices to show that the numerator is negative.

$$\begin{aligned}
2G_\mu(y)G'_\mu(y)^2 &= 2\,\mathbb{E}(y - \theta)^{-1}\left[\mathbb{E}(y - \theta)^{-2}\right]^2 \\
&\leq 2\left[\mathbb{E}(y - \theta)^{-1}\right]^2 \mathbb{E}(y - \theta)^{-3} \\
&= G_\mu(y)^2 G''_\mu(y)
\end{aligned}$$

where the inequality uses that for any random variable X we have $\left[\mathbb{E}\,X^2\right]^2 \leq \mathbb{E}\,|X|\,\mathbb{E}\,|X|^3$, which is a simple consequence of the Cauchy-Schwarz and Jensen inequalities. This concludes the proof of the claim. $\square$

Equipped with this lemma, we can tackle the asymptotic convergence of the different FDRs defined at the start of the section.

Proposition 6.2. *Fix any $k \in \mathbb{N}$. Then, the following three assertions hold.*

1. *As $n \to \infty$ we have $\overline{\mathrm{FDR}}^{(n)}(k) \to \mathrm{FDR}^\infty(k)$.*

2. *As $n \to \infty$ we have $\widehat{\widetilde{\mathrm{FDR}}}^{(n)}(k) \to \mathrm{FDR}^\infty(k)$ almost surely.*

3. *As $n \to \infty$ we have $\widehat{\mathrm{FDR}}^{(n)}(k) \to \mathrm{FDR}^\infty(k)$ almost surely.*

Proof. We start by proving the first statement. Recall $\widehat{U}$ and U are matrices whose columns are the eigenvectors of X and A, respectively. For clarity and simplicity, we assume that the eigenvalues of A are all distinct; nonetheless, the same proof strategy follows through with repeated eigenvalues by substituting $\widehat{u}_i \widehat{u}_i^\top$ with the projector onto the eigenvalue subspace associated with the ith eigenvalue.

Expanding the definition of the upper-bound FDR yields

$$\overline{\text{FDR}}^{(n)}(k) = \mathbb{E}\left[\frac{\text{tr}\left(\widehat{\boldsymbol{U}}_{1:k}\widehat{\boldsymbol{U}}_{1:k}^\top(\boldsymbol{I} - \boldsymbol{U}_{1:r^\star}\boldsymbol{U}_{1:r^\star}^\top)\right)}{k}\right]$$

$$= \frac{1}{k}\sum_{j=1}^k \mathbb{E}\left[\text{tr}\left(\widehat{\boldsymbol{u}}_j\widehat{\boldsymbol{u}}_j^\top\left(\boldsymbol{I} - \sum_i^{r^\star}\boldsymbol{u}_i\boldsymbol{u}_i^\top\right)\right)\right]$$

$$= \frac{1}{k}\sum_{j=1}^k\left\{1 - \sum_i^{r^\star}\mathbb{E}\left[\left(\widehat{\boldsymbol{u}}_j^\top\boldsymbol{u}_i\right)^2\right]\right\}$$

$$= 1 - \frac{1}{k}\sum_{j=1}^k\sum_{i=1}^{r^\star}\mathbb{E}\left[\left(\widehat{\boldsymbol{u}}_j^\top\boldsymbol{u}_i\right)^2\right]$$

Thus, it is enough to focus on $\mathbb{E}\left[\left(\widehat{\boldsymbol{u}}_j^\top\boldsymbol{u}_i\right)^2\right]$. Using Theorem B.1 in tandem with Proposition B.3 yields that for all $i, j \subseteq [n]$ the following almost sure convergence holds:

$$\left(\widehat{\boldsymbol{u}}_j^\top\boldsymbol{u}_i\right)^2 \to \begin{cases} -\frac{1}{\theta_i^2 G'_{\mu_{\boldsymbol{E}}}(G_{\mu_{\boldsymbol{E}}}^{-1}(1/\theta_i))} & \text{if } i = j \text{ and } i \leq r^\star, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Recall that we have suppressed the dependence on n, so the above convergence result should be understood in the limit $n \to \infty$. Hence, taking expectations and substituting, we derive that

$$\overline{\text{FDR}}^{(n)}(k) \to 1 + \frac{1}{k}\sum_{i=1}^{\min\{k,r^\star\}}\frac{1}{\theta_i^2 G'_{\mu_{\boldsymbol{E}}}\left(G_{\mu_{\boldsymbol{E}}}^{-1}(1/\theta_i)\right)} = \text{FDR}^\infty(k),$$

as we wanted.

Turning next to the convergence of $\widetilde{\text{FDR}}^{(n)}$, note that $x \to G_{\mu_{\boldsymbol{E}}}(x)^2/G'_{\mu_{\boldsymbol{E}}}(x)$ is a continuous function for $x \notin [a, b]$ from Lemma 6.1. Thus, using the limit $\lambda_i \xrightarrow{a.s.} G_{\mu_{\boldsymbol{E}}}^{-1}(1/\theta_i)$ from Theorem B.1, we get that

$$\frac{G_{\mu_{\boldsymbol{E}}}(\lambda_i)^2}{G'_{\mu_{\boldsymbol{E}}}(\lambda_i)} \xrightarrow{a.s.} \frac{1}{\theta_i^2 G'_{\mu_{\boldsymbol{E}}}\left(G_{\mu_{\boldsymbol{E}}}^{-1}(1/\theta_i)\right)},$$

which leads to the result after a simple substitution.

Finally, let us prove that $\widehat{\text{FDR}}^{(n)}$ also converges to the same limit. Assume that n is large enough to ensure that $r^\star \leq \widehat{r} \leq r$, which happens almost surely by assumption. Expanding the definitions of $\widehat{\text{FDR}}^{(n)}, \widetilde{\text{FDR}}^{(n)}$ and using the triangle inequality, we have for any k that

$$\left|\widehat{\text{FDR}}^{(n)}(k) - \widetilde{\text{FDR}}^{(n)}(k)\right| \leq \frac{1}{k}\left(\underbrace{\sum_{i=1}^{\min\{k,r^\star\}}\left|\frac{G_{\mu_{\boldsymbol{E}}}(\lambda_i)^2}{G'_{\mu_{\boldsymbol{E}}}(\lambda_i)} - \frac{\widehat{G}(\lambda_i)^2}{\widehat{G}'(\lambda_i)}\right|}_{T_1} + \underbrace{\sum_{i=r^\star+1}^{\widehat{r}}\left|\frac{\widehat{G}(\lambda_i)^2}{\widehat{G}'(\lambda_i)}\right|}_{T_2}\right). \tag{25}$$

We will show that each summand goes to zero. Let us start by bounding T_1. Let μ_n be the empirical distribution of $\lambda_{r^\star+1}, \ldots, \lambda_n$. By Assumption 5 and Weyl's interlacing inequalities, we obtain that μ_n converges weakly to $\mu_{\boldsymbol{E}}$. Thus, we have that $\widehat{G}(t) \to G_{\mu_{\boldsymbol{E}}}(t)$ for any t outside the support of the bulk $\text{supp}(\mu_{\boldsymbol{E}}) = (a, b)$. Furthermore, for any $C > b$ that is greater than the right edge of the bulk, there exists $L > 0$ such that the function $G_{\mu_{\boldsymbol{E}}}$ is L-Lipschitz on $[C, \infty)$. The same statement holds almost surely for $\widehat{G}$ once n is large enough thanks to Lemma 6.1. Moreover, by Theorem B.1

24

we have that $\lambda_i^{(n)} \to \rho_i$ for some $\rho_i > b$ as $i \le r^\star$. Fix $C = (b + \min_{i \le r^\star} \rho_i)/2$, then

$$\left| \widehat{G}\left(\lambda_i^{(n)}\right) - G_{\mu_E}\left(\lambda_i^{(n)}\right) \right| \le \left| \widehat{G}\left(\lambda_i^{(n)}\right) - \widehat{G}(\rho_i) \right| + \left| \widehat{G}(\rho_i) - G_{\mu_E}(\rho_i) \right| + \left| G_{\mu_E}(\rho_i) - G_{\mu_E}\left(\lambda_i^{(n)}\right) \right|$$

$$\le 2L \left| \lambda_i^{(n)} - p_i \right| + \left| \widehat{G}(\rho_i) - G_{\mu_E}(\rho_i) \right|.$$

The last expression goes to zero almost surely as n goes to infinity. Thus, we have that $\widehat{G}(\lambda_i^{(n)})$ and $G_{\mu_E}(\lambda_i^{(n)})$ share the same limit. An analogous argument shows the same conclusion for $\widehat{G}'(\lambda_i^{(n)})$ and $G'_{\mu_E}(\lambda_i^{(n)})$, and, by construction, their limit is finite and bounded away from zero. Then, the quotient rule for limits yields that the bound on T_1 in (25) goes to zero.

Next, we turn to T_2. Let i be any index in $\{r^\star + 1, \ldots, r\}$. By Lemma 6.1, the ratios $\widehat{G}(\lambda_i^{(n)})^2/\widehat{G}'(\lambda_i^{(n)})$ are bounded between -1 and 0. So it suffices to show that $\widehat{G}'(\lambda_i^{(n)})$ goes to infinity with n to conclude that $T_2 \to 0$. For any fixed $M > 0$, define the truncated function $f_M(z,t) = \min \left\{ M, \frac{1}{(z-t)^2} \right\}$ and the constant $l_M = \lim_{t \downarrow b} \mathbb{E}_{z \sim \mu_E} f_M(z,t)$. Moreover, for any M we have $|\widehat{G}'(\lambda_i^{(n)})| \ge \mathbb{E}_{z \sim \mu_n} f_M(z, \lambda_i^{(n)})$. Taking inferior limits on both sides yields

$$\liminf_{n \to \infty} |\widehat{G}'(\lambda_i^{(n)})| \ge \liminf_{n \to \infty} \mathbb{E}_{z \sim \mu_n} f_M(z, \lambda_i) = l_M \tag{26}$$

where the last relation uses that $\lim_{n \to \infty} f_M(z, \lambda_i^{(n)}) = l_M$, which follows since $f_M(\cdot, \lambda_i^{(n)})$ converges uniformly to $f_M(\cdot, b)$ and μ_n converges weakly almost surely to μ_E. Using Assumption 2, Proposition B.2, and Fatou's Lemma we obtain $\liminf_{M \to \infty} l_M \ge |G'(b^+)| = \infty$. The M we used in (26) was arbitrary, and so we derive that $\liminf_{n \to \infty} |\widehat{G}'(\lambda_i^{(n)})| = \infty$, thus establishing that $T_2 \to 0$ and concluding the proof of the lemma. $\qquad \square$

Step 2. To finish the proof, we need the following auxiliary lemma.

Lemma 6.3. *For all large n, the functions $\widehat{\mathrm{FDR}}^{(n)}(\cdot)$ and $\widetilde{\mathrm{FDR}}^{(n)}(\cdot)$ are nondecreasing and their image is contained in $[0, 1]$. The same properties hold for FDR^∞.*

Proof. We prove the statement for $\widehat{\mathrm{FDR}}^{(n)}$. Define $C = \frac{\lambda_{\hat{r}} + \lambda_{\hat{r}+1}}{2}$ and the function $\Phi \colon [C, \infty) \to \mathbb{R}$ given by $y \mapsto \frac{G_{\hat{\mu}_n}(y)^2}{G'_{\hat{\mu}_n}(y)}$ where $\hat{\mu}_n = \frac{1}{n-\hat{r}} \sum_{i=\hat{r}+1}^n \delta_{\lambda_i}$. The statement follows as direct consequence of Lemma 6.1 by noting that $\widehat{\mathrm{FDR}}^{(n)}(k) = 1 + \frac{1}{k} \sum_{j=1}^{\min\{k,\hat{r}\}} \Phi(\lambda_j)$. The same arguments apply for $\widetilde{\mathrm{FDR}}^{(n)}$ after noting that for large enough n the largest $r^\star$ eigenvalues are outside of the support of the asymptotic distribution μ_E. The limiting FDR inherits both properties since it can be realized as the limit of $\widehat{\mathrm{FDR}}^{(n)}$. $\qquad \square$

Finally, we can show that our estimate $\hat{k}$ matches $\bar{k} = \max\{j \in [r] \mid \overline{\mathrm{FDR}}^{(n)}(j) \le \alpha\}$ for all large n. To prove this, we show that both $\hat{k}$ and $\bar{k}$ almost surely converge to k^∞. Since $\mathrm{FDR}^\infty(k^\infty) \ne \alpha$, we have that

$$\mathrm{FDR}^\infty(k^\infty) < \alpha < \mathrm{FDR}^\infty(k^\infty + 1). \tag{27}$$

Since $\widehat{\mathrm{FDR}}^{(n)}$ converges pointwise to the asymptotic $\mathrm{FDR}^{(\infty)}$ we have that for large n

$$\widehat{\mathrm{FDR}}^{(n)}(k^{(n)}) < \alpha < \widehat{\mathrm{FDR}}^{(n)}(k^{(n)} + 1).$$

Further, by Lemma 6.2 we have that for large enough n, $\widehat{\mathrm{FDR}}^{(n)}$ is nondecreasing in k, and so we derive that $\hat{k} = k^\infty$. Since $\overline{\mathrm{FDR}}^{(n)}$ converges to FDR^∞, for all $k \le r$ and by (27) we also conclude

$\bar{k} = k^\infty$. Thus, we obtain that almost surely $\widehat{k} = \bar{k}$ for large enough n. This finishes the proof of Theorem 3.1. $\qquad\square$

6.1.2 Proof of Theorem 3.2

By Weyl's interlacing inequalities, we have for all $i \geq n$ that

$$\lambda_i(\boldsymbol{E}_n) \leq \lambda_i(\boldsymbol{X}_n) \leq \lambda_{i-r}(\boldsymbol{E}_n) \tag{28}$$

with the convention that $\lambda_k(\boldsymbol{X}) = \infty$ if $k \leq 0$. Fix an index $j > r$; then, using the interlacing inequalities combined with a telescoping sum, we obtain that

$$\Delta_j = \lambda_{j-1}(\boldsymbol{X}_n) - \lambda_j(\boldsymbol{X}_n) \leq \lambda_{j-1-r}(\boldsymbol{E}_n) - \lambda_j(\boldsymbol{E}_n)$$
$$= \sum_{i=0}^{r} \left(\lambda_{j-1-i}(\boldsymbol{E}_n) - \lambda_{j-i}(\boldsymbol{E}_n) \right)$$
$$\leq r \max_{i \in [n-1]} \left\{ \lambda_i(\boldsymbol{E}_n) - \lambda_{i+1}(\boldsymbol{E}_n) \right\}.$$

Recall the labeling $s_i := \lambda_i(\boldsymbol{E}_n) - \lambda_{i+1}(\boldsymbol{E}_n)$. Assumption 3 ensures that for any $\varepsilon > 0$ there exists an $N > 0$ such that for all $n \geq N$, we have that $\max_{i \in [n-1]} \{s_i\} \leq \varepsilon \cdot n^{-1/2}$. Thus, by taking $\varepsilon = p/2r$, we obtain

$$\Delta_j \leq \frac{p}{2} \cdot n^{-1/2} < p \cdot n^{-1/2} \quad \text{for all } j > r. \tag{29}$$

Thus, we conclude that for all large enough n, $\text{RANKESTIMATE}(\boldsymbol{X}_n) \leq r$. Next, let us show the lower bound of $r^\star$. By Theorem B.1 we have that $\lambda_{r^\star}(\boldsymbol{X}) \to \rho_{r^\star} > b$. Furthermore by Theorem B.1 we have that $\lambda_{r^\star+1}(\boldsymbol{X}_n)$ almost surely converges to b. Thus, for large enough n, we obtain that

$$\lambda_{r^\star}(\boldsymbol{X}_n) - \lambda_{r^\star+1}(\boldsymbol{X}_n) \geq \frac{\rho_{r^\star} - b}{2} > p \cdot n^{-1/2}$$

where the last inequality holds for any n large since $n^{-1/2}$ tends to zero. Hence $\text{RANKESTIMATE}(\boldsymbol{X}_n) \geq r^\star$ This concludes the proof of Theorem 3.2. $\qquad\square$

6.1.3 Proof of Proposition 3.4

It is well-known that the empirical distribution of the spectrum of Wigner matrices converges weakly almost surely to the semicircle law, which is given by the density

$$\varrho(t) = \frac{1}{2\pi} \sqrt{(4 - t^2)_+}.$$

The distribution is supported on the interval $[-2, 2]$. Hence, Wigner ensembles satisfy Assumption 2. So to prove that Assumption 3 holds, it suffices to show that $\sup_i |\lambda_i(\boldsymbol{E}_n) - \lambda_{i+1}(\boldsymbol{E}_n)| = o(n^{-1/2})$. To this end, we will use a celebrated result from random matrix theory known as eigenvalue rigidity. Intuitively, rigidity amounts to the concentration of eigenvalue around quantiles of $\mu_{\boldsymbol{E}}$. For any $i \in \{1, \ldots, n\}$ define the quantiles $\gamma_i^{(n)}$ as

$$\int_{\gamma_i^{(n)}}^2 g(t) dt = \frac{i}{n}.$$

As we shall see, $\gamma_i^{(n)}$ gives the typical location of the ith largest eigenvalue $\lambda_i^{(n)}$. To avoid cluttering notation, we will often drop the super index in $\gamma_i^{(n)}$ when it is clear from the context.

Our strategy to bound the spacing will be to control

$$|\lambda_i(\mathbf{H}_n) - \lambda_{i+1}(\mathbf{H}_n)| \leq |\lambda_i(\mathbf{H}_n) - \gamma_i| + |\lambda_{i+1}(\mathbf{H}_n) - \gamma_{i+1}| + |\gamma_i - \gamma_{i+1}|. \tag{30}$$

26

Eigenvalue rigidity will give us a handle on the distance between quantiles and eigenvalues. Before we delve into rigidity, let us bound the deterministic term $|\gamma_i - \gamma_{i+1}|$.

Lemma 6.4. *For large enough n, we have the following bound for all $i \in \{1, \ldots, n\}$:*

$$\left|\gamma_i^{(n)} - \gamma_{i+1}^{(n)}\right| \leq \pi^2 \cdot n^{-2/3}.$$

Proof. For simplicity, we focus on the larger quantiles $i \leq n/2$; the argument follows symmetrically for the other half. Define $\omega_n \in (0, 2)$ to be the largest scalar such that if $t \in \Omega_n = [-\omega_n, \omega_n]$ then $\varrho(t) \geq n^{-1/3}$. We consider a couple of cases.

Case 1. First, we consider the case where quantiles $\gamma_i, \gamma_{i+1} \in \Omega$. By definition, we have that

$$n^{-1/3} \cdot |\gamma_i - \gamma_{i+1}| \leq \int_{\gamma_{i+1}}^{\gamma_i} \varrho(t)dt = n^{-1},$$

and so rearranging terms gives the desired bound with an even better constant.

Case 2. Next, assume that $\gamma_i, \gamma_{i+1} \in (\omega_n, 2]$. Observe that it suffices to prove that $|2 - \omega_n| \leq \pi^2 \cdot n^{-2/3}$. We can write $\omega_n = 2 - h(n^{-1/3})$ where $h(\cdot) = f^{-1}(\cdot)$ is the functional inverse of $f(t) = \varrho(t - 2)$ restricted to the interval $[0, 2]$ so that the inverse is well-defined. Thus, we focus on bounding $|2 - \omega_n| = |h(n^{-1/3})|$. A simple computation reveals that

$$h(t) = 2 - 2\sqrt{1 - \pi \cdot t^2}.$$

Near zero, the second derivative of this function is Q-Lipschitz for some $Q > 0$; to see this just note that its third derivative is bounded. Thus, the Taylor approximation error is bounded by

$$\left|h(t) - \frac{h''(0)}{2}t^2\right| \leq \frac{Q}{6}t^3,$$

here we used that $h(0) = h'(0) = 0$; see [36, Lemma 1.2.4]. Therefore, for t small enough we obtain $\frac{\pi^2}{2}t^2 = \frac{h''(0)t^2}{4} \leq h(t) \leq h''(0)t^2 \leq 2\pi^2 t^2$. For n is large enough, we might substitute $t = n^{-1/3}$ and obtain

$$\frac{1}{2}\pi^2 \cdot n^{-2/3} \leq |2 - \omega_n| \leq \pi^2 \cdot n^{-2/3}, \tag{31}$$

as we wanted.

Case 3. Finally, assume that $\gamma_i \in (-\omega_n, 2)$ and $\gamma_{i+1} \in [-\omega_n, \omega_n]$. The first $\lfloor n/2 \rfloor$ spacings are monotonically decreasing

$$|2 - \gamma_1| > |\gamma_1 - \gamma_2| > \cdots > |\gamma_{\lfloor n/2 \rfloor + 1} - \gamma_{\lfloor n/2 \rfloor}|,$$

which follows since $\int_{\gamma_{i+1}}^{\gamma_i} \varrho(t)dt = n^{-1}$ and ϱ is positive and monotonically decreasing on the positive real line. So, it suffices to prove that $\gamma_1 > \omega_n$. Using the concavity of ϱ, we can lower bound the integral

$$\int_{\omega_n}^{2} \varrho(t)dt \geq \frac{1}{2}(2 - \omega_n) \cdot \varrho(\omega_n) = \frac{1}{2}(2 - \omega_n) \cdot n^{-1/3} \geq \frac{\pi^2}{4}n^{-1} > n^{-1}, \tag{32}$$

where the first equality follows by the definition of ω_n and the second inequality follows by the lower bound in (31). Thus, we derive that $\gamma_1 > \omega_n$. This concludes the proof of the lemma. $\square$

Next, we establish bounds on the remaining terms in (30). To this end, we invoke the following result. This theorem was originally established in [26]; a more accessible proof can be found in [10, Theorem 2.9].

Theorem 6.5 (**Wigner eigenvalue rigidity**). *Assume that $\mathbf{H}_n \in \mathbb{R}^{n \times n}$ is a Wigner ensemble. Then, for all small $\varepsilon > 0$ and large M we have that*

$$\sup_i \mathbb{P}\left(|\lambda_i(\mathbf{H}_n) - \gamma_i| \geq n^\varepsilon \cdot n^{-2/3} \left(\min\{i, (n+1-i)\}\right)^{-1/3}\right) \leq n^{-M} \tag{33}$$

for all large enough n.

To complete the proof of Proposition 3.4, take $\varepsilon < 1/6$ and $D \geq 3$. Define the events

$$\mathcal{E}_n = \left\{ \text{For all } i \in [n], \ |\lambda_i(\mathbf{H}_n) - \gamma_i| \geq n^\varepsilon \cdot n^{-2/3} \left(\min\{i, (n+1-i)\} \right)^{-1/3} \right\}.$$

Using the union bound in tandem with Theorem 6.5 we derive that $\mathbb{P}(\mathcal{E}_n) \leq n^{1-M}$. By our choice of M, these probabilities are summable. So, the Borel-Cantelli Lemma ensures that, with probability one, only finitely many $\mathcal{E}_n$ occur. Thus, almost surely, there exists an N such that for all $n \geq N$ we have

$$|\lambda_i(\mathbf{H}_n) - \gamma_i| \leq n^\varepsilon \cdot n^{-2/3} \left(\min\{i, (n+1-i)\} \right)^{-1/3} \leq n^{\varepsilon - 2/3} \quad \text{for all } i \in [n]. \tag{34}$$

By our choice of ε, we obtain that $|\lambda_i(\mathbf{H}_n) - \gamma_i| = o(n^{-1/2})$.

Combining (30), Lemma 6.4, and (34) we conclude that for all large enough n, we have that

$$|\lambda_i(\mathbf{H}_n) - \lambda_{i+1}(\mathbf{H}_n)| \leq 2n^{\varepsilon - 2/3} + \pi^2 n^{-2/3} = o(n^{-1/2}) \quad \text{for all } i \in [n].$$

This finishes the proof of the proposition. $\qquad\square$

6.1.4 Proof of Proposition 3.6

The proof follows the exact same reasoning as that of Proposition 3.4. Again, for each $i \in [n]$, define the typical location $\gamma_i^{(n)}$ as

$$\int_{\gamma_i^{(n)}}^{c_+} \nu(t)dt = \frac{i}{n},$$

where ν denotes the Marchenko-Pastur distribution (12). In order for the arguments to go through, we need an equivalent of Lemma 6.4 and Theorem 6.7. Let us state their counterparts for Wishart matrices.

Lemma 6.6. *There exists a constant $C > 0$ depending only on ϑ such that for large enough n,*

$$|\gamma_i^{(n)} - \gamma_{i+1}^{(n)}| \leq C \cdot n^{-2/3}.$$

The proof parallels that of Lemma 6.4, and so we omit the details. The eigenvalue rigidity of Wishart matrices was established in [39].

Theorem 6.7 (**Wishart eigenvalue rigidity**)**.** *Assume that $\mathbf{W}_n \in \mathbb{R}^{n \times n}$ is a Wishart ensemble satisfying $\vartheta \neq 1$. Then, for all small $\varepsilon > 0$ and large M we have that*

$$\sup_{i \in [\min\{n_1, n_2\}]} \mathbb{P}\left(|\lambda_i(\mathbf{H}_n) - \gamma_i| \geq n_1^\varepsilon \cdot n_1^{-2/3} \left(\min\{i, (\min\{n_1, n_2\} + 1 - i)\} \right)^{-1/3} \right) \leq n_1^{-M} \tag{35}$$

for all large enough n.

The statement we transcribed here is a little weaker than that of [39, Theorem 3.3], but it suffices for our purposes. Armed with these two results, the reasoning follows exactly the same logic as the proof of Proposition 3.4. We omit the details.

6.2 Asymmetric Case

6.2.1 Proof of Theorem 4.1

The proof is analogous to that of Theorem 3.1. Let us summarize the main changes one must implement for the same proof to follow. First, define the analogous FDR quantities (20), (21), (22), and (23) accordingly to account for the asymmetry. Then, use Theorem B.6, Proposition B.7, and Proposition B.8 in place of Theorem B.1, Proposition B.2, and Proposition B.3. With these changes

in place, almost all the same arguments follow, with the only exception being the equivalent of Lemma 6.1, which we now state and prove. We omit all other details since they are equivalent to those in the proof of Theorem 3.1.

Lemma 6.8. *Consider a probability measure over the real line μ with bounded support such that $\operatorname{supp}(\mu) \subset [0, \infty)$ and let $b = \sup\{t : t \in \operatorname{supp}(\mu)\}$. Then, the map $\Phi \colon (b, \infty) \to \mathbb{R}$ given by $\Phi(x) = 2\frac{D_\mu(x)\varphi_\mu(x;1)}{D'_\mu(x)}$ yields a $[-1,0]$-valued continuous nonincreasing function.*

Proof. Recall that

$$\varphi_{\mu_E}(x; s) := s \cdot \int \frac{x}{x^2 - t^2} d\mu(t) + \frac{1-s}{x} \quad \text{and} \quad \varphi'_\mu(x; s) := -\left(s \cdot \int \frac{x^2 + t^2}{(x^2 - t^2)^2} d\mu(t) + \frac{1-s}{x^2} \right),$$

for any $s \in (0, 1]$ and, for a fixed $q \in (0, 1]$ the D-transform and its derivative are given by

$$D_\mu(x) = \varphi_\mu(x; 1) \cdot \varphi_\mu(x; q). \quad \text{and} \quad D'_\mu(x) = \varphi'_\mu(x; 1) \cdot \varphi_\mu(x; q) + \varphi_\mu(x; 1) \cdot \varphi'_\mu(x; q).$$

Immediately we see that $\Phi(x) < 0$ because $\varphi_\mu(x, s) > 0$ and $\varphi'_\mu(x; s) < 0$ for any $s \in (0, 1]$ and $x > b \geq 0$. Next, we show $\Phi(x) \geq -1$. Let $\tilde{\mu} = q\mu + (1 - q)\delta_0$, then $\varphi_\mu(x; q) = \varphi_{\tilde{\mu}}(x, 1)$. Expanding reveals that the following relations hold

$$2\,\mathbb{E}_{t \sim \mu}\left[\frac{x^2}{(x^2 - t^2)^2} \right] = \frac{\varphi_\mu(x, 1)}{x} - \varphi'_\mu(x, 1), \tag{36}$$

$$2\,\mathbb{E}_{t \sim \tilde{\mu}}\left[\frac{t^2}{(x^2 - t^2)^2} \right] = -\frac{\varphi_\mu(x, q)}{x} - \varphi'_\mu(x, q). \tag{37}$$

Then,

$$
\begin{aligned}
2D_\mu(x)\varphi_\mu(x; 1) &= 2\varphi_\mu(x; q)\varphi_\mu(x; 1)^2 \\
&= 2\varphi_\mu(x; q)\,\mathbb{E}_{t \sim \mu}\left[\frac{x}{x^2 - t^2} \right]^2 \\
&\leq 2\varphi_\mu(x; q)\,\mathbb{E}_{t \sim \mu}\left[\frac{x^2}{(x^2 - t^2)^2} \right] \\
&= \frac{\varphi_\mu(x; q)\varphi_\mu(x; 1)}{x} - \varphi_\mu(x; q)\varphi'(x; 1) \\
&= -\varphi_\mu(x; 1)\left[2\,\mathbb{E}_{t \sim \tilde{\mu}}\left[\frac{t^2}{(x^2 - t^2)^2} \right] + \varphi'_\mu(x, q) \right] - \varphi_\mu(x; q)\varphi'(x; 1) \\
&\leq -\varphi_\mu(x; 1)\varphi'_\mu(x, q) - \varphi_\mu(x; q)\varphi'(x; 1) = -D'_\mu(x)
\end{aligned}
$$

where the third line uses Jensen's inequality, the fourth line uses (36), and the second to last line uses (37). Multiplying by the negative number $1/D'_\mu(x)$ yields $\Phi(x) \geq -1$.

Finally, we prove that $\Phi(x)$ is nonincreasing. Since $\Phi(x) < 0$, the function $\Phi(\cdot)$ is nonincreasing if, and only if, $\Theta(\cdot) = 1/\Phi(\cdot)$ is nondecreasing. Let's show that $\Theta(\cdot)$ is nondecreasing. Expanding,

$$\Theta(x) = \frac{1}{2}\Big(\underbrace{\frac{\varphi'_\mu(x; 1)}{\varphi_\mu(x; 1)^2}}_{T_1(x):=} + \underbrace{\frac{\varphi'_\mu(x; q)}{\varphi_\mu(x; q)\varphi(x; 1)}}_{T_2(x;q):=} \Big).$$

In turn, the first term T_1 is strictly increasing, while the second term T_2 is increasing for every x when q is close to one and decreasing for every x when q gets closer to zero. However, we shall see T_1 increases fast enough to guarantee that the whole function is nondecreasing for every x for any fixed q. The second term has the fastest decrease when q goes to zero, as the next claim shows.

29

Claim 6.9. *For any fixed x, we have that*

$$T_2'(x; q) \geq T_2'(x; 0)$$

for all $q \in [0, 1]$.

We defer the proof of this claim to the end of this subsection and focus on the extreme case $q = 0$. In this case, the function Θ reduces to

$$\Theta(x) = \frac{1}{2}\left(\frac{x\varphi_\mu'(x; 1) - \varphi_\mu(x; 1)}{x\varphi_\mu(x; 1)^2}\right).$$

Taking a derivative yields

$$\Theta'(x) = \left(\frac{x^2\left(\varphi_\mu''(x; 1)\varphi_\mu(x; 1)^2 - 2\varphi_\mu(x; 1)\varphi_\mu'(x; 1)^2\right) + \varphi_\mu(x; 1)^3 + x\varphi_\mu(x; 1)^2\varphi_\mu'(x; 1)}{2x^2\varphi_\mu(x; 1)^4}\right)$$

$$= \left(\frac{x^2\left(\varphi_\mu''(x; 1)\varphi_\mu(x; 1)^2 - 2\varphi_\mu(x; 1)\varphi_\mu'(x; 1)^2\right) - 2x\varphi_\mu(x; 1)^2\,\mathbb{E}\left(\frac{t^2}{(x^2 - t^2)^2}\right)}{2x^2\varphi_\mu(x; 1)^4}\right)$$

$$= \left(\frac{x^2\varphi_\mu(x; 1)\left(\left(\varphi_\mu''(x; 1) - 2\,\mathbb{E}\left(\frac{t^2/x}{(x^2 - t^2)^2}\right)\right)\varphi_\mu(x; 1) - 2\varphi_\mu'(x; 1)^2\right)}{2x^2\varphi_\mu(x; 1)^4}\right)$$

where the second equality follows by (37). Thus, to show that Θ is nondecreasing, it suffices to show that the numerator is positive. We upper bound

$$2\varphi_\mu'(x; 1)^2 = 2\,\mathbb{E}\left[\frac{x^{1/2}}{(x^2 - t^2)^{1/2}}\frac{\left(x^{3/2} + \frac{t^2}{x^{1/2}}\right)}{(x^2 - t^2)^{3/2}}\right]^2 \leq 2\,\mathbb{E}\left[\frac{x}{x^2 - t^2}\right]\mathbb{E}\left[\frac{\left(x^{3/2} + \frac{t^2}{x^{1/2}}\right)^2}{(x^2 - t^2)^3}\right]$$

$$= 2\,\mathbb{E}\left[\frac{x}{x^2 - t^2}\right]\mathbb{E}\left[\frac{x\left(x + \frac{t^2}{x}\right)^2}{(x^2 - t^2)^3}\right]$$

$$= 2\,\mathbb{E}\left[\frac{x}{x^2 - t^2}\right]\mathbb{E}\left[\frac{x\left(x^2 + 2t^2 + \frac{t^4}{x^2}\right)}{(x^2 - t^2)^3}\right]$$

$$= 2\,\mathbb{E}\left[\frac{x}{(x^2 - t^2)}\right]\left(\mathbb{E}\left[\frac{x(x^2 + 3t^2)}{(x^2 - t^2)^3}\right] - \mathbb{E}\left[\frac{x(t^2 - t^4/x^2)}{(x^2 - t^2)^3}\right]\right)$$

$$= \varphi_\mu(x; 1)\left(\varphi_\mu''(x; 1) - 2\,\mathbb{E}\left[\frac{t^2/x}{(x^2 - t^2)^2}\right]\right)$$

where the first inequality follows by Cauchy-Schwarz. This establishes that Θ is nondecreasing, and so Φ is nonincreasing, thus proving the Lemma. $\qquad\square$

Proof of Claim 6.9. Taking derivatives

$$T_2'(x; q) = \frac{\varphi''(x; q)\varphi(x; q)\varphi(x; 1) - \varphi'(x; q)\left(\varphi'(x; q)\varphi(x; 1) + \varphi(x; q)\varphi'(x; 1)\right)}{\varphi(x; q)^2\varphi(x; 1)^2}.$$

Evaluating at $q = 0$ yields

$$T_2'(x; 0) = \frac{\varphi(x; 1) + x\varphi'(x; 1)}{x^2\varphi(x; 1)^2}.$$

To simplify the notation, we drop the dependency on x via $\varphi_q := \varphi(x;q)$. Thus,

$$T_2'(x;q) - T_2'(x;0) = \frac{x^2\varphi_q''\varphi_q\varphi_1 - x^2\varphi_q'\left(\varphi_q'\varphi_1 + \varphi_q\varphi_1'\right) - \varphi_q^2\left(\varphi_1 + x\varphi_1'\right)}{x^2\varphi_q^2\varphi_1^2}.$$

It suffices to show that the numerator is nonnegative. Reorganizing yields that the numerator is equal to

$$\left(x^2\varphi_q''\varphi_q - x^2(\varphi_q')^2 - \varphi_q^2\right)\varphi_1 - \left(x\varphi_q^2 + x^2\varphi_q'\varphi_q\right)\varphi_1'$$

$$\geq \left(x^2\varphi_q''\varphi_q/2 - \varphi_q^2\right)\varphi_1 - \left(x\varphi_q + x^2\varphi_q'\right)\varphi_q\varphi_1'$$

$$= \left[\left(x^2\varphi_q''/2 - \varphi_q\right)\varphi_1 + 2\,\mathbb{E}_{\tilde\mu}\left(\frac{x^2t^2}{(x^2-t^2)^2}\right)\varphi_1'\right]\varphi_q$$

$$= \left[\mathbb{E}_{\tilde\mu}\left(\frac{xt^2(5x^2 - t^2)}{(x^2-t^2)^3}\right)\varphi_1 + 2\,\mathbb{E}_{\tilde\mu}\left(\frac{xt^2}{(x^2-t^2)^2}\right)x\varphi_1'\right]\varphi_q$$

$$= \left[\mathbb{E}_{\tilde\mu}\left(\frac{xt^2(5x^2 - t^2)}{(x^2-t^2)^3}\right)\varphi_1 - 2\,\mathbb{E}_{\tilde\mu}\left(\frac{xt^2}{(x^2-t^2)^2}\right)\left(\varphi_1 + 2\,\mathbb{E}_\mu\left(\frac{xt^2}{(x^2-t^2)^2}\right)\right)\right]\varphi_q$$

$$= \left[\mathbb{E}_{\tilde\mu}\left(\frac{xt^2(3x^2 + t^2)}{(x^2-t^2)^3}\right)\varphi_1 - 4\,\mathbb{E}_{\tilde\mu}\left(\frac{xt^2}{(x^2-t^2)^2}\right)\left(\mathbb{E}_\mu\left(\frac{xt^2}{(x^2-t^2)^2}\right)\right)\right]\varphi_q$$

$$= q\left[\mathbb{E}_\mu\left(\frac{xt^2(3x^2 + t^2)}{(x^2-t^2)^3}\right)\varphi_1 - 4x^2\,\mathbb{E}_\mu\left(\frac{t^2}{(x^2-t^2)^2}\right)^2\right]\varphi_q$$

$$\geq q\left[\mathbb{E}_\mu\left(\frac{xt^2(3x^2 + t^2)}{(x^2-t^2)^3}\right)\varphi_1 - 4x^2\,\mathbb{E}_\mu\left(\frac{t^3}{(x^2-t^2)^3}\right)\mathbb{E}_\mu\left(\frac{t}{(x^2-t^2)}\right)\right]\varphi_q$$

$$\geq q\left[\mathbb{E}_\mu\left(\frac{xt^2(3x^2 + t^2)}{(x^2-t^2)^3}\right)\varphi_1 - 4x^2\,\mathbb{E}_\mu\left(\frac{t^3}{(x^2-t^2)^3}\right)\varphi_1\right]\varphi_q$$

$$= q\left[\mathbb{E}_\mu\left(\frac{xt^2(3x^2 + t^2 - 4xt)}{(x^2-t^2)^3}\right)\varphi_1\right]\varphi_q$$

$$= q\,\mathbb{E}_\mu\left(\frac{xt^2(3x - t)(x - t)}{(x^2-t^2)^3}\right)\varphi_1\varphi_q$$

$$\geq 0$$

This concludes the proof.

6.2.2 Proof of Theorem 4.2

By singular-value interlacing [41, Theorem 3.1.2], we have that for all $i \geq n$

$$\sigma_{i+r}(\boldsymbol{E}_n) \leq \sigma_i(\boldsymbol{X}_n) \leq \sigma_{i-r}(\boldsymbol{E}_n) \tag{38}$$

with the convention that $\sigma_k(\boldsymbol{X}) = \infty$ if $k \leq 0$. Fix an index $\lceil n/2 \rceil \geq j > r$; then, using the interlacing inequalities combined with a telescoping sum, we obtain that

$$\Delta_j = \sigma_{j-1}(\boldsymbol{X}_n) - \sigma_j(\boldsymbol{X}_n) \leq \sigma_{j-1-r}(\boldsymbol{E}_n) - \sigma_{j+r}(\boldsymbol{E}_n)$$

$$= \sum_{i=-r}^{r} (\sigma_{j-1-i}(\boldsymbol{E}_n) - \sigma_{j-i}(\boldsymbol{E}_n))$$

$$\leq 2r \max_{i \in [n-1]} \{\sigma_i(\boldsymbol{E}_n) - \sigma_{i+1}(\boldsymbol{E}_n)\}.$$

Recall the labeling $s_i := \sigma_i(\boldsymbol{E}_n) - \sigma_{i+1}(\boldsymbol{E}_n)$. Assumption 3 ensures that for any $\varepsilon > 0$ there exists an $N > 0$ such that for all $n \geq N$, we have that $\max_{i \in [n-1]} \{s_i\} \leq \varepsilon \cdot n^{-1/2}$. Thus, by taking $\varepsilon = p/4r$, we obtain

$$\Delta_j \leq \frac{p}{2} \cdot n^{-1/2} < p \cdot n^{-1/2} \quad \text{for all } j > r. \tag{39}$$

Thus, we conclude that for all large enough n, $\text{RANKESTIMATE}(\boldsymbol{X}_n) \leq r$. Next, let us show the lower bound of $r^\star$. By Theorem B.1 we have that $\lambda_{r^\star}(\boldsymbol{X}) \to \rho_{r^\star} > b$. Furthermore by Theorem B.6 we have that $\sigma_{r^\star+1}(\boldsymbol{X}_n)$ almost surely converges to b. Thus, for large enough n, we obtain that

$$\sigma_{r^\star}(\boldsymbol{X}_n) - \sigma_{r^\star+1}(\boldsymbol{X}_n) \geq \frac{\rho_{r^\star} - b}{2} > p \cdot n^{-1/2}$$

where the last inequality holds for any n large since $n^{-1/2}$ tends to zero. Hence $\text{RANKESTIMATE}(\boldsymbol{X}_n) \geq r^\star$ This concludes the proof of Theorem 4.2. $\qquad \square$

6.3 Proof of Proposition 4.4

By definition, the product $\boldsymbol{W}_n = \boldsymbol{E}_n \boldsymbol{E}_n^\top$ is a Wishart ensemble; moreover, the left singular vectors of $\boldsymbol{E}_n$ match the eigenvectors of $\boldsymbol{W}_n$. Then, invoking Proposition 3.6 we obtain that $\boldsymbol{W}_n$ satisfies Assumptions 2, and 3. Combining this with the fact that $\sigma_i^2(\boldsymbol{E}_n) = \lambda_i(\boldsymbol{W}_n)$ shows that $\{\boldsymbol{E}_n\}$ satisfies Assumption 4 with a limiting singular value distribution $\mu_{\boldsymbol{E}}$ that is characterized by the density

$$f_{\boldsymbol{E}}(t) = 2t \cdot \nu(t^2) = \frac{1}{\pi \vartheta t} \sqrt{((c_+ - t^2)(t^2 - c_-))_+} \quad \text{with} \quad c_\pm = \left(1 \pm \sqrt{\vartheta}\right)^2,$$

which is supported on the interval $(\sqrt{c_-}, \sqrt{c_+})$. Thus, the first part of Assumption 6 also holds true. To prove that the second part of Assumption 6 holds, recall that $\vartheta < 1$. Therefore, for all large n and $i \leq n$ we have $\sigma_i(\boldsymbol{X}_n) \in \left(\frac{\sqrt{c_-}}{2}, 2\sqrt{c_+}\right)$. Within that interval, the function $x \mapsto \sqrt{x}$ is L-Lipschitz for some $L > 0$. Therefore,

$$|\sigma_i(\boldsymbol{E}_n) - \sigma_{i+1}(\boldsymbol{E}_n)| \leq L \cdot |\sigma_i(\boldsymbol{E}_n)^2 - \sigma_{i+1}(\boldsymbol{E}_n)^2| = L \cdot |\lambda_i(\boldsymbol{W}_n) - \lambda_{i+1}(\boldsymbol{W}_n)| = o(n^{-1/2}),$$

where the last relation follows since Assumption 3 holds for $\boldsymbol{W}_n$; see Proposition 3.6. This completes the proof. $\qquad \square$

7 Future Directions

We presented methodology for subspace selection in PCA in which the objective is to obtain subspace estimates that come with FDR control guarantees. Although subspace estimation has a long history, the perspective in our paper is grounded in multiple testing and this is not common in the literature. Our methods are free of tuning parameters and although our analysis is asymptotic in nature, the empirical results showcase strong performance even for modest-sized problem instances.

There are a number of research directions that are suggested by our work. It is of interest to evaluate the extent to which some of our method yields non-asymptotic FDR control; progress in this

direction likely require new advances in random matrix theory. More generally, subspace estimation arises in many applications in which one may only have partial observations of a data matrix (e.g., matrix completion and collaborative filtering), and generalizing our methods to such settings would broaden their applicability. Finally, in recent work [44] natural notions of FDR, generalizing those for multiple testing and subspace selection, were obtained for a variety of model selection problems such as clustering, ranking, and causal inference. Developing methodology for controlling FDR in these contexts would substantially expand the scope of the FDR control paradigm.

Acknowledgements

References

[1] 10x Genomics. Pbmc 3k filtered gene-barcode matrices. URL https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz. Accessed: January 28, 2024.

[2] E. Abbe, J. Fan, and K. Wang. An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

[3] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584 – 653, 2006. doi: 10.1214/009053606000000074. URL https://doi.org/10.1214/009053606000000074.

[4] J. Agterberg, Z. Lubberts, and C. E. Priebe. Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Transactions on Information Theory*, 68(7):4618–4650, 2022.

[5] B. Au. Bbp phenomena for deformed random band matrices. *arXiv preprint arXiv:2304.13047*, 2023.

[6] Z. Bai and J. Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

[7] J. Baik, G. Ben Arous, and S. Peche. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of probability*, 33(5):1643–1697, 2005.

[8] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, pages 2055–2085, 2015.

[9] C. Baumgartner, C. Plant, K. Railing, H.-P. Kriegel, and P. Kroger. Subspace selection for clustering high-dimensional data. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 11–18. IEEE, 2004.

[10] F. Benaych-Georges and A. Knowles. Lectures on the local semicircle law for wigner matrices. *arXiv preprint arXiv:1601.04055*, 2016.

[11] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[12] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111: 120–135, 2012.

[13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[14] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[15] G. Blanchard, S. Delattre, and E. Roquain. Testing over a continuum of null hypotheses with false discovery rate control. *Bernoulli*, 20(1), 2014.

[16] P. Bourgade, L. Erdős, and H.-T. Yau. Bulk universality of general β-ensembles with non-convex potential. *Journal of mathematical physics*, 53(9), 2012.

[17] P. Bourgade, H.-T. Yau, and J. Yin. Random band matrices in the delocalized phase i: Quantum unique ergodicity and universality. *Communications on Pure and Applied Mathematics*, 73(7), 2020.

[18] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen. Subspace estimation from unbalanced and incomplete data matrices: $l_{2,\infty}$ statistical guarantees. *Annals of Statistics*, 49(2):944–967, 2021.

[19] J. Cape, M. Tang, and C. E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019.

[20] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.

[21] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy. Composite optimization for robust rank one bilinear sensing. *Information and Inference: A Journal of the IMA*, 10(2):333–396, 2021.

[22] Y. Choi, J. Taylor, and R. Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, pages 2590–2617, 2017.

[23] M. Díaz. The nonsmooth landscape of blind deconvolution. *Workshop on Optimization for Machine Learning*, 2019.

[24] D. Donoho, M. Gavish, and E. Romanov. Screenot: Exact mse-optimal singular value thresholding in correlated noise. *The Annals of Statistics*, 51(1):122–148, 2023.

[25] L. Erdős and H.-T. Yau. Gap universality of generalized wigner and β-ensembles. *Journal of the European Mathematical Society*, 17(8):1927–2036, 2015.

[26] L. Erdős, H.-T. Yau, and J. Yin. Rigidity of eigenvalues of generalized wigner matrices. *Advances in Mathematics*, 229(3):1435–1515, 2012.

[27] D. P. Foster and R. A. Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2): 429–444, 2008.

[28] J. J. Goeman and U. Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24(4):537–544, 2008.

[29] H. Z. James X. Hu and H. H. Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, 2010. doi: 10.1198/jasa.2010.tm09329.

[30] A. Javanmard and A. Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics*, 46(2):526–554, 2018.

[31] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

[32] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[33] M. L. Mehta. *Random matrices*. Elsevier, 2004.

[34] H. L. Montgomery. The pair correlation of zeros of the zeta function. In *Proc. Symp. Pure Math*, volume 24, pages 181–193, 1973.

[35] R. R. Nadakuditi. Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60 (5):3002–3018, 2014.

[36] Y. Nesterov. *Lectures on convex optimization*. Springer, 2nd edition, 2018.

[37] H. Nguyen, T. Tao, and V. Vu. Random matrices: tail bounds for gaps between eigenvalues. *Probability Theory and Related Fields*, 167:777–816, 2017.

[38] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

[39] N. S. Pillai and J. Yin. Universality of covariance matrices. *The Annals of Applied Probability*, 24(3):935–1001, 2014.

[40] C. Rodarmel and J. Shan. Principal component analysis for hyperspectral image classification. *Surveying and Land Information Science*, 62(2):115–122, 2002.

[41] H. Roger and R. J. Charles. Topics in matrix analysis, 1994.

[42] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

[43] A. Taeb, P. Shah, and V. Chandrasekaran. False discovery and its control in low rank estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):997–1027, 2020.

[44] A. Taeb, P. Bühlmann, and V. Chandrasekaran. Model selection over partially ordered sets. *Proceedings of the National Academy of Sciences*, 121(8):e2314228121, 2024.

[45] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[46] T. Tao. The asymptotic distribution of a single eigenvalue gap of a wigner matrix. *Probability Theory and Related Fields*, 157(1-2):81–106, 2013.

[47] T. Tao and V. Vu. Random matrices: Universality of local eigenvalue statistics up to the edge. *Communications in Mathematical Physics*, 298:549–572, 2010.

[48] T. Tao and V. Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 1(206):127–204, 2011.

[49] S. Torquato. Hyperuniform states of matter. *Physics Reports*, 745:1–95, 2018.

[50] University of the Basque Country. Indian Pines Dataset. URL https://www.ehu.eus/ccwintco/uploads/6/67/Indian_pines_corrected.mat. Accessed: January 28, 2024.

[51] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

[52] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

[53] R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

[54] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognition*, 48(1):10–19, 2015.

[55] E. P. Wigner. Results and theory of resonance absorption. In *Conference on neutron physics by time-of-flight*, pages 1–2, 1956.

[56] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1958.

[57] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions i. *The Collected Works of Eugene Paul Wigner: Part A: The Scientific Papers*, pages 524–540, 1993.

[58] N. Wolchover. A Bird's-Eye View of Nature's Hidden Order. In *The Prime Number Conspiracy: The Biggest Ideas in Math from Quanta*. The MIT Press, 11 2018. ISBN 9780262350549. doi: 10.7551/mitpress/11910.003.0012. URL https://doi.org/10.7551/mitpress/11910.003.0012.

[59] F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[60] D. Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008. doi: 10.1198/016214507000001373.
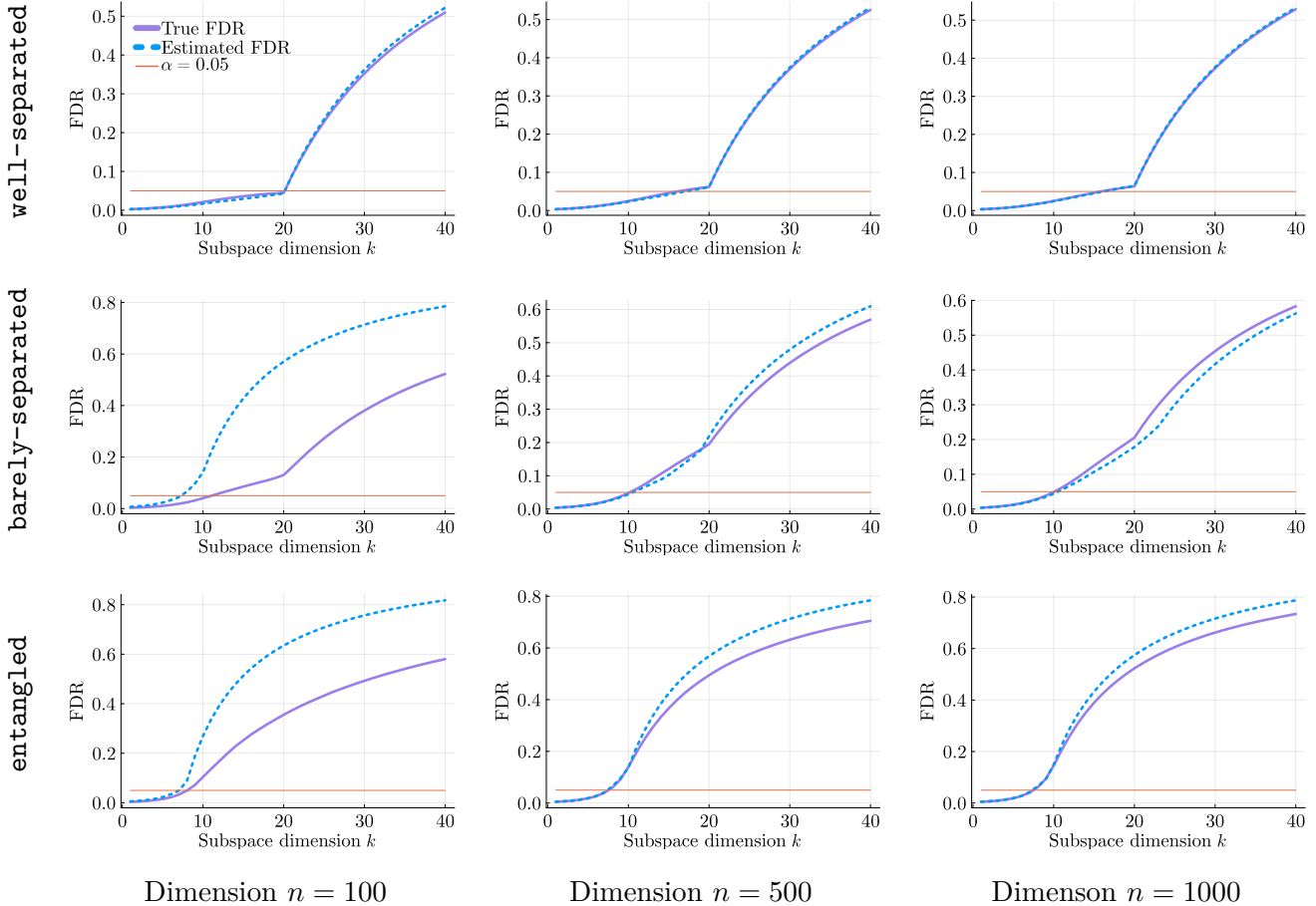
Figure 9: Results for estimating the FDR of the `Wishart` ensemble using Algorithms 1 and 2.

A   Additional Numerical Experiments

In this Appendix, we include the additional numerical experiments missing from Section 5. In particular, we include plots for all the ensembles in Table 1. The setup here is exactly the same as the one described in Section 5.1. The proportions of all the experiments with rectangular matrices are set $n/m = 1/2$.

B   Supplementary Results

This section reviews the necessary background to establish theoretical guarantees for our methods. We highlight that the results in this section are known.

B.1   Symmetric Case

Let us start with an asymptotic characterization of the eigenvalues and eigenvectors of $\boldsymbol{X}$.

Theorem B.1 (Theorems 2.1 and 2.2 in [11])**.** *Suppose that Assumption 1 and 2 hold. For each n, let $\lambda_i^{(n)} = \lambda_i(\boldsymbol{X}_n)$ and let $\tilde{u}$ be a unit-norm eigenvector associated with $\lambda_i^{(n)}$. Then, we have that as n grows to infinity,*

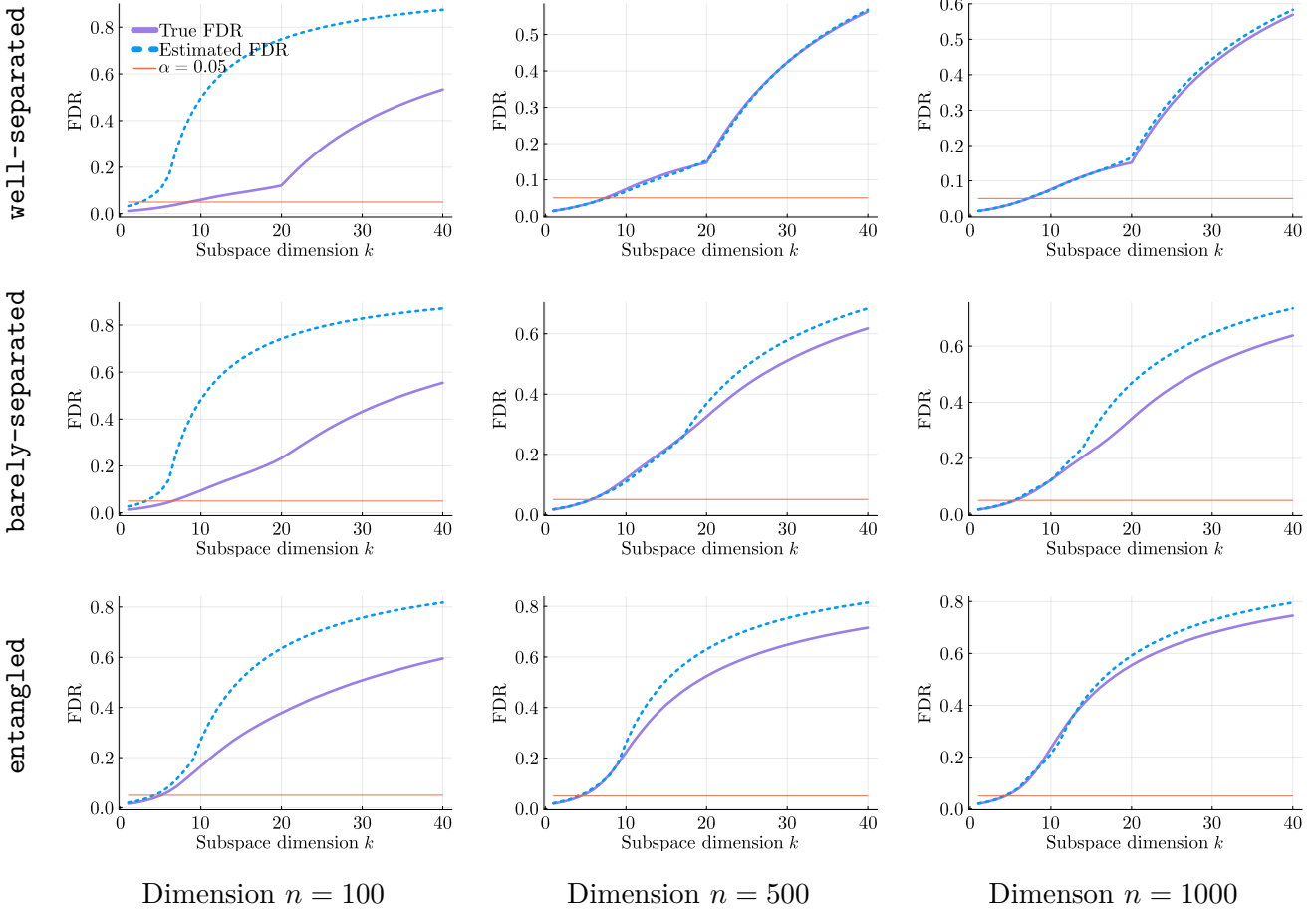Figure 10: Results for estimating the FDR of the `Uniform` ensemble using Algorithms 1 and 2.

(**Eigenvalue limit**) *Fix* $j \in \{1, \ldots, r\}$, *then,*

$$\lambda_i^{(n)} \xrightarrow{a.s.} \begin{cases} \rho_i := G_{\mu_E}^{-1}(1/\theta_i) & \text{if } \theta_j > 1/G_{\mu_E}(b^+), \\ b & \text{otherwise.} \end{cases} \tag{40}$$

(**Eigenvector correlation limit**) *Fix* $j \in \{1, \ldots, r\}$ *such that* $\frac{1}{\theta_j} \in \left(G_{\mu_E}(a^-), G_{\mu_E}(b^+)\right)$, *let* $\mathcal{W}_j = \ker(\theta_j \boldsymbol{I}_n - \boldsymbol{A}_n)$ *be the subspace of eigenvectors associated with* θ_i, *and let* $\widehat{u}$ *be a unit norm eigenvector associated with* λ_j. *Then,*

$$\|\mathcal{P}_{\mathcal{W}_j}(\widehat{u})\|^2 \xrightarrow{a.s.} -\frac{1}{\theta_j^2 G'_{\mu_E}(\rho_j)} = -\frac{G_{\mu_E}(\rho_j)^2}{G'_{\mu_E}(\rho_j)}. \tag{41}$$

Furthermore, let $\mathcal{T}_j = \bigoplus_{i \neq j}^r \ker(\theta_i \boldsymbol{I}_n - \boldsymbol{A}_n)$ *be the eigenspace corresponding to the first* r *eigenvectors that are not associated with* θ_i. *Then,*

$$\|\mathcal{P}_{\mathcal{T}_j}(\widehat{u})\|^2 \xrightarrow{a.s.} 0. \tag{42}$$

The following proposition unveils the behavior of the Cauchy transform and its derivative at the edge of the spectrum. It relies crucially on the square-root decay assumption. The previous Theorem holds without the need for the square-root decay condition.
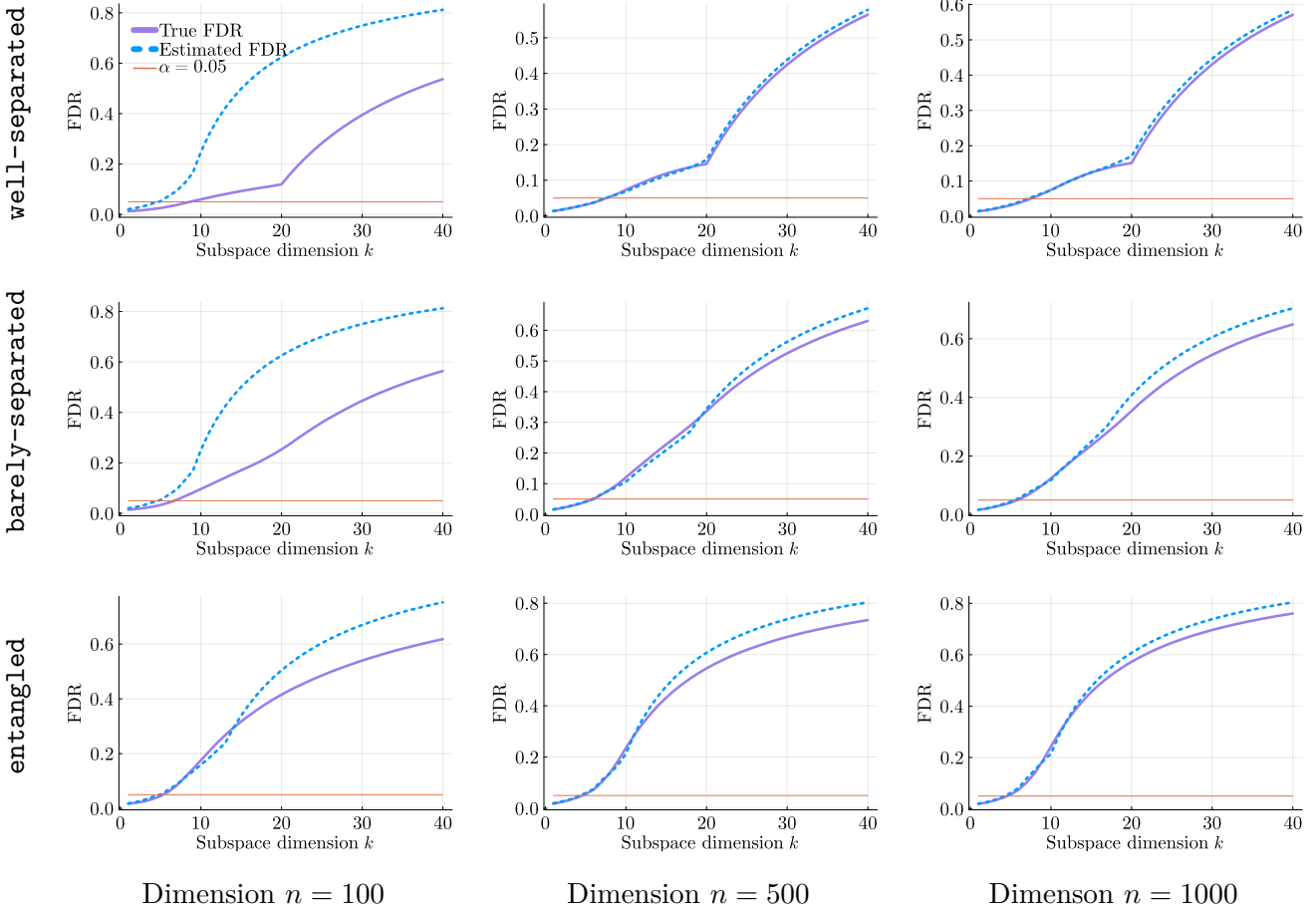
38

Figure 11: Results for estimating the FDR of the `UniformFactor` ensemble using Algorithms 3 and 4.

Proposition B.2 (Proposition 2.4 in [11])**.** *Suppose that Assumption 2 holds. Then, we have that*

$$G_{\mu_{\boldsymbol{E}}}(b^+) < \infty \quad and \quad G'_{\mu_{\boldsymbol{E}}}(b^+) = \infty.$$

Finally, the following proposition shows that eigenvectors of $\boldsymbol{X}$ and $\boldsymbol{A}$ that are not associated with the same eigenvalue are decorrelated.

Proposition B.3. *Suppose that Assumption 1 and 2 hold. Fix $j \in \{1, \dots, r\}$ and $i \neq j$. Let $\boldsymbol{u}_j$ be a unit-norm eigenvector of $\boldsymbol{A}_n$ associated with θ_j and let $\widehat{\boldsymbol{u}}_i$ be a unit-norm eigenvector of $\boldsymbol{X}_n$ associated with $\lambda_i(\boldsymbol{X})$. Then,*

$$\langle \widehat{u}_\ell, u_j \rangle \xrightarrow{a.s.} 0 \qquad almost\ surely. \tag{43}$$

The conclusion of this guarantee is similar to (46), but it does not assume that $i \leq r$.

Proof. The proof uses an analogous result for singular vectors; see Proposition B.8 below. We will leverage the following folklore result.

Fact B.4. *Let $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \in \mathbb{R}^{d \times d}$ be two independent orthogonal matrices with uniform distribution over the orthogonal group. Then, the product $\boldsymbol{UW}$ and $\boldsymbol{VW}$ are independent.*
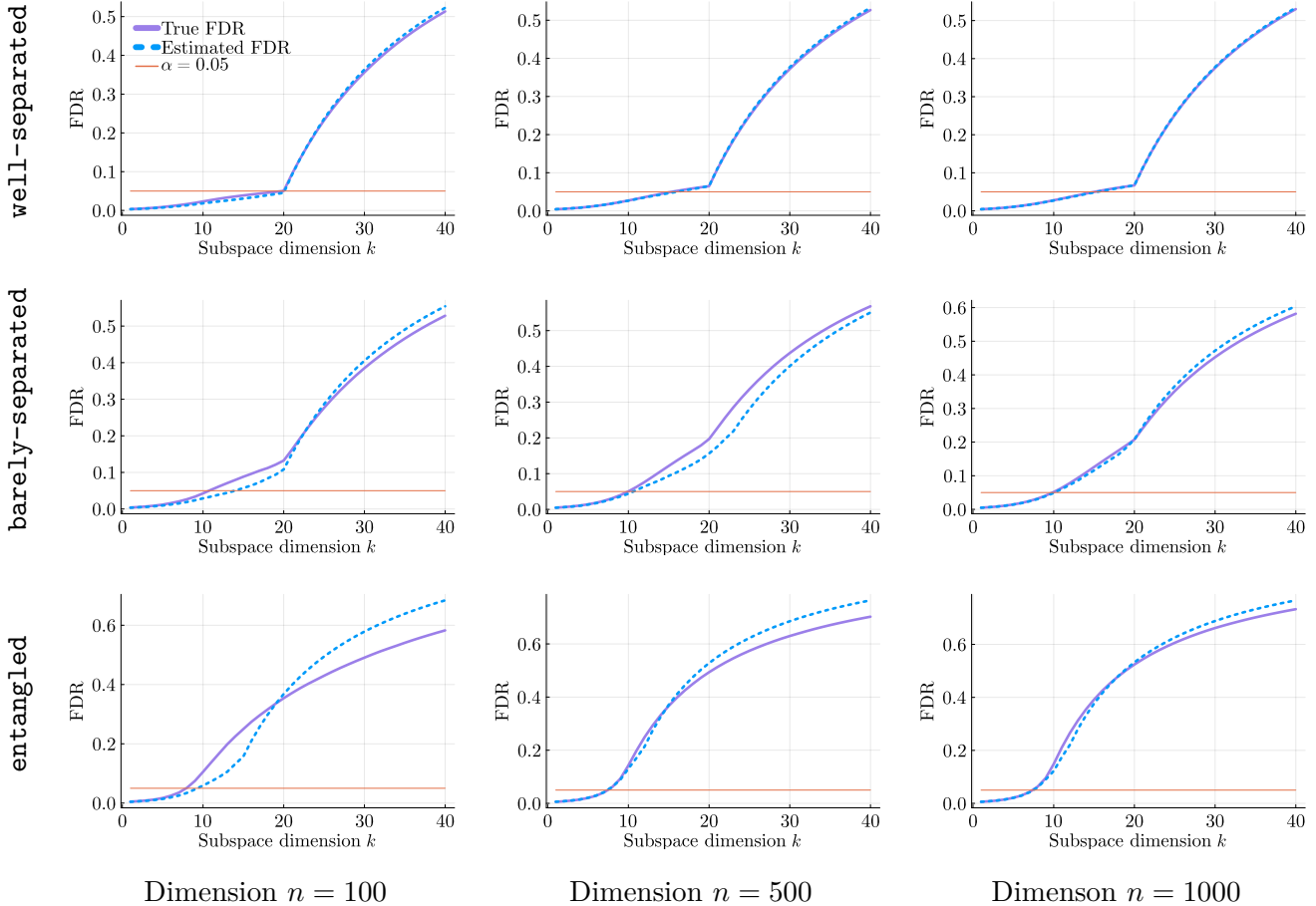
Figure 12: Results for estimating the FDR of the `Fisher` ensemble using Algorithms 1 and 2.

Fact B.5. Let $\boldsymbol{U}, \boldsymbol{W} \in \mathbb{R}^{d \times d}$ be two independent orthogonal matrices with uniform distribution over the orthogonal group. Then, the product $\boldsymbol{U}\boldsymbol{W}$ is also uniform over the orthogonal orthogonal group, and $\boldsymbol{U}\boldsymbol{W}$ is independent of both $\boldsymbol{U}$ and $\boldsymbol{W}$.

Thus, without loss of generality, we can assume that the distributions of the eigenvectors of both $\boldsymbol{A}$ and $\boldsymbol{E}$ are orthogonally invariant. Moreover, we can post-multiply the matrix $\boldsymbol{X}$ by a uniform at random orthogonal matrix $\boldsymbol{W}$, then $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}} + \tilde{\boldsymbol{E}}$ where $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{W}$ and $\tilde{\boldsymbol{E}} = \boldsymbol{E}\boldsymbol{W}$. The left singular vectors of $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{A}}$ match the eigenvectors of $\boldsymbol{X}$, $\boldsymbol{E}$ and $\boldsymbol{A}$, respectively. Moreover, thanks to Facts B.4 and B.5, the singular vectors of $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{E}}$ are independent, and so are the left and right singular vectors of $\boldsymbol{E}$. Notice that this implies that $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{E}}$ satisfy Assumptions 4 and 5. Thus, the result follows after invoking Proposition B.8.

$\square$

B.2 Asymmetric Case

Theorem B.6 (Theorems 2.8 and 2.9 in [12])**.** *Suppose that Assumption 4 and 5 hold. For each n, let $\sigma_i^{(n)} = \sigma_i(\boldsymbol{X}_n)$ and let $\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{v}}$ be unit-norm left and right singular vectors associated with $\sigma_i^{(n)}$. Then, we have that as n grows to infinity,*
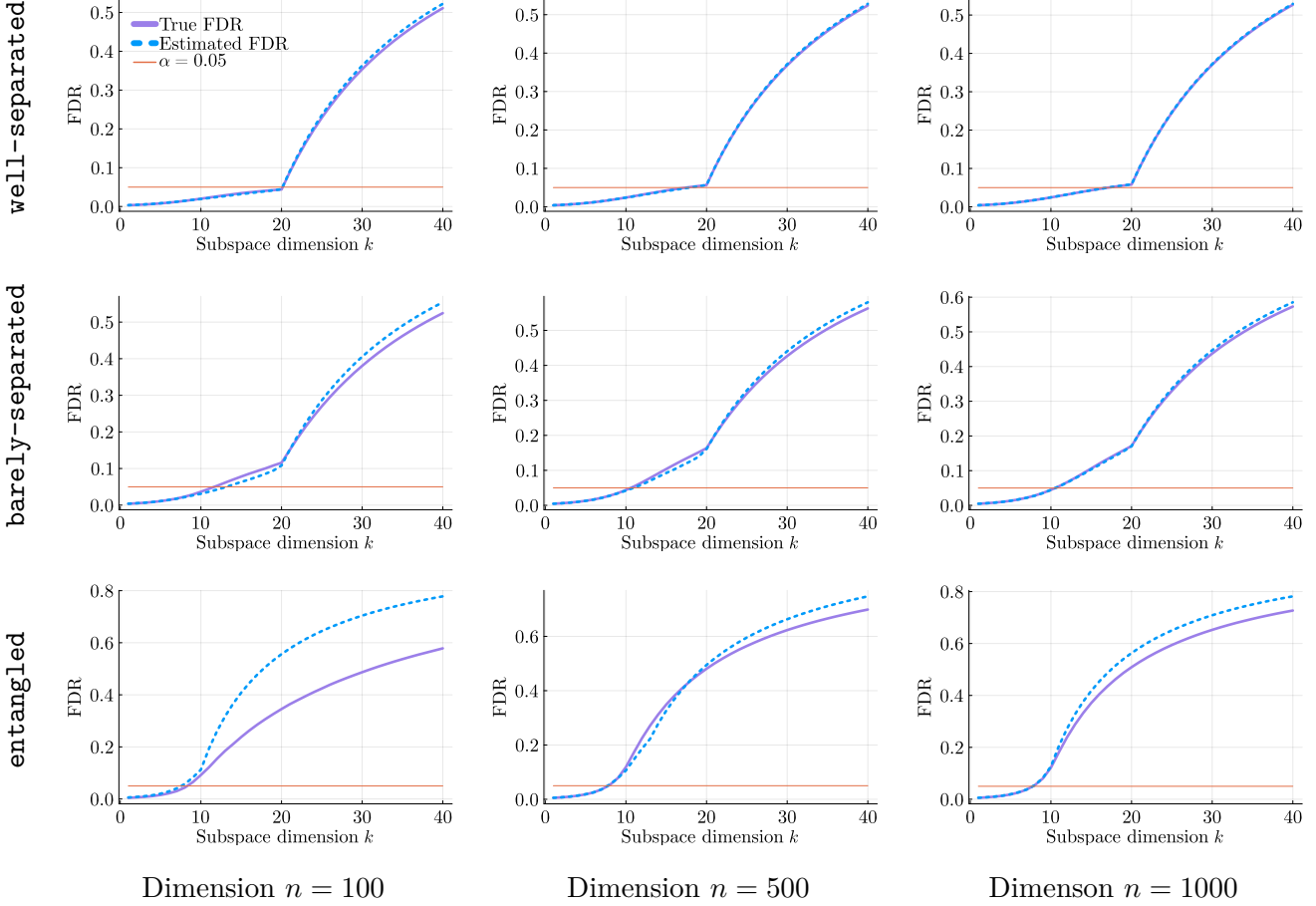
Figure 13: Results for estimating the FDR of the `FisherFactor` ensemble using Algorithms 3 and 4.

(Eigenvalue limit) *Fix $j \in \{1, \ldots, r\}$, then,*

$$\sigma_i^{(n)} \xrightarrow{a.s.} \begin{cases} \rho_i := D_{\mu_E}^{-1}(1/\theta_i^2) & \text{if } \theta_j^2 > 1/D_{\mu_E}(b^+), \\ b & \text{otherwise.} \end{cases} \tag{44}$$

(Eigenvector correlation limit) *Fix $j \in \{1, \ldots, r\}$ such that $\theta_j^2 \geq 1/D_{\mu_E}(b^+)$, let $\mathcal{W}_j = \ker(\theta_j \boldsymbol{I}_n - \boldsymbol{A}_n \boldsymbol{A}_n^\top)$ be the subspace of singula vectors associated with θ_i, and let $\widehat{u}$ be a unit norm singular-vector of $\boldsymbol{X}_n$ associated with σ_j. Then,*

$$\|\mathcal{P}_{\mathcal{W}_j}(\widehat{u})\|^2 \xrightarrow{a.s.} -\frac{2\varphi_\mu(\rho_j; 1)}{\theta_j^2 D_{\mu_E}'(\rho_j)} = -2\frac{D_{\mu_E}(\rho_j)\varphi_\mu(x; 1)}{D_{\mu_E}'(\rho_j)}. \tag{45}$$

Furthermore, let $\mathcal{T}_j = \bigoplus_{i \neq j}^r \ker(\theta_i \boldsymbol{I}_n - \boldsymbol{A}_n \boldsymbol{A}_n^\top)$ be the eigenspace corresponding to the first r eigenvectors that are not associated with θ_i. Then,

$$\|\mathcal{P}_{\mathcal{T}_j}(\widehat{u})\|^2 \xrightarrow{a.s.} 0. \tag{46}$$

Similarly to the one in the symmetric case, the following proposition unveils the behavior of the D-transform and its derivative at the edge of the spectrum. It relies crucially on the square-root decay assumption.

Proposition B.7 (Proposition 2.11 in [12]). *Suppose that Assumption 5 holds. Then, we have that*

$$D_{\mu_E}(b^+)^{-1/2} > 0 \quad and \quad \varphi'_{\mu_E}(b^+; 1) = \varphi'_{\mu_E}(b^+; q) = -\infty.$$

Finally, we close with the asymptotic decorrelation of singular values associated with different singular values.

Proposition B.8 (Proposition 2 in [24]). *Suppose that Assumption 4 and 5 hold. Fix $j \in \{1, \dots, r\}$ and $i \neq j$. Let $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ be unit-norm left and right singular value of $\boldsymbol{A}_n$ associated with θ_j and let $\widehat{\boldsymbol{u}}_i$ and $\widehat{\boldsymbol{v}}_i$ be the left and right unit-norm eigenvectors of $\boldsymbol{X}_n$ associated with $\sigma_i(\boldsymbol{X})$. Then,*

$$\langle \widehat{u}_i, u_j \rangle \xrightarrow{a.s.} 0 \qquad and \quad \langle \widehat{v}_i, v_j \rangle \xrightarrow{a.s.} 0 \qquad almost\ surely. \tag{47}$$

We highlighted that although the statement of the proposition in [24] states the conclusion for the product of the two inner products, a quick examination of the proof reveals that the conclusion holds for both terms separately.