Permutation methods for factor analysis and PCA

Edgar Dobriban*

September 16, 2019

Abstract

Researchers often have datasets measuring features x_{ij} of samples, such as test scores of students. In factor analysis and PCA, these features are thought to be influenced by unobserved factors, such as skills. Can we determine how many components affect the data? This is an important problem, because it has a large impact on all downstream data analysis. Consequently, many approaches have been developed to address it. *Parallel Analysis* is a popular permutation method. It works by randomly scrambling each feature of the data. It selects components if their singular values are larger than those of the permuted data. Despite widespread use in leading textbooks and scientific publications, as well as empirical evidence for its accuracy, it currently has no theoretical justification.

In this paper, we show that the parallel analysis permutation method consistently selects the large components in certain high-dimensional factor models. However, it does not select the smaller components. The intuition is that permutations keep the noise invariant, while "destroying" the low-rank signal. This provides justification for permutation methods in PCA and factor models under some conditions. Our work uncovers drawbacks of permutation methods, and paves the way to improvements.

1 Introduction

1.1 Factor Analysis and PCA

Factor Analysis (FA) and Principal Component Analysis (PCA), the unsupervised discovery of components governing variation in the data, is performed routinely by scientists and social scientists in thousands of studies every year. In FA and PCA, we measure a number p of indicators (features, covariates) for a set of n samples. In Spearman (1904)'s original application, this involved p test scores of n students. In another important application to finance, we measure the return for n assets over p (or T) time periods. The goal is to identify the common factors driving variation in the data, such as skills in Spearman's example, or systemic risks in finance. The setup for FA and PCA is similar, while not exactly the same (see e.g., Jolliffe, 2002), and hence we will focus on factor analysis for clarity in most of the paper.

Since Spearman's time factor analysis has found a wide range of applications in a variety of fields, becoming one of the most widely used statistical methods. Applications abound in psychology and education (Fabrigar et al., 1999; Costello and Osborne, 2005; Brown, 2014),

*Department of Statistics, The Wharton School, University of Pennsylvania. E-mail: dobriban@wharton.upenn.edu. We thank Andreas Buja, David Donoho, Alexei Onatski, and Art Owen for stimulating discussions. We are grateful to Jingshu Wang for feedback on the manuscript. The first version of the manuscript had the title "Factor selection by permutation".

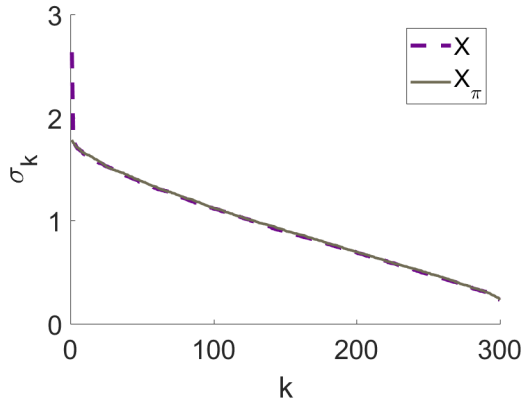

Figure 1: Visual representation of the permutation method Parallel Analysis (PA). We have an $n \times p$ data matrix X measuring p features of n datapoints. We want to determine how many unobserved factors or components affect the data. We examine the *scree plot* of the singular values of X, i.e., the plot of singular values in a decreasing order. Classical methods such as Cattel's scree plot look for the "elbow" in this plot. Instead of such a subjective rule that may be biased by the judgement of the user, we consider a more objective permutation method. We permute the entries within each column of X independently, possibly several times, to get "null" or "fake" data matrices X_π. We plot some fixed percentile of the largest, second largest etc., singular values of these matrices. We select the components of X whose singular values are larger than the permuted ones. Here, only one factor is selected.

public health (Goetz et al., 2008), management/ marketing (Churchill Jr, 1979; Stewart, 1981; Parasuraman et al., 1988), economics/ finance (Bai and Ng, 2008), and genomics (Leek and Storey, 2008; Lin et al., 2016).

The most widely used approach to FA relies on the the common-factor model (e.g., Thurstone, 1947; Anderson, 2003, etc). For each sample i, the j-th indicator x_{ij} is a linear function of one or more common factors η_{ik} and one unique factor (or idiosyncratic noise) ε_{ij}:

$$x_{ij} = \sum_{k=1}^{r} \eta_{ik}\lambda_{jk} + \varepsilon_{ij}. \tag{1}$$

The factor values η_{ik} and the factor loadings λ_{jk} are not observed. In Spearman's example, x_{ij} is the test score of student i on test j, the r factors are interpreted as skills, η_{ik} is student i's level on the k-th skill, and λ_{jk} is the relevance of the k-th skill to the j-th test.

FA is merely one step beyond linear regression. In linear regression, η_{ik} are observed, while in FA they are not. This simplicity is deceiving, however, and FA can be surprisingly difficult. A widely cited tutorial on FA notes: "*Perhaps more than any other commonly used statistical method, FA requires a researcher to make a number of important decisions with respect to how the analysis is performed*" (Fabrigar et al., 1999).

One of the key problems in FA is to select the number of factors. For instance, how many skills control test scores? This is well known to have a large impact on the later steps of data analysis (e.g., Hayton et al., 2004; Brown, 2014). The standard textbook Brown (2014) calls it "*the most crucial decision*" in exploratory FA.

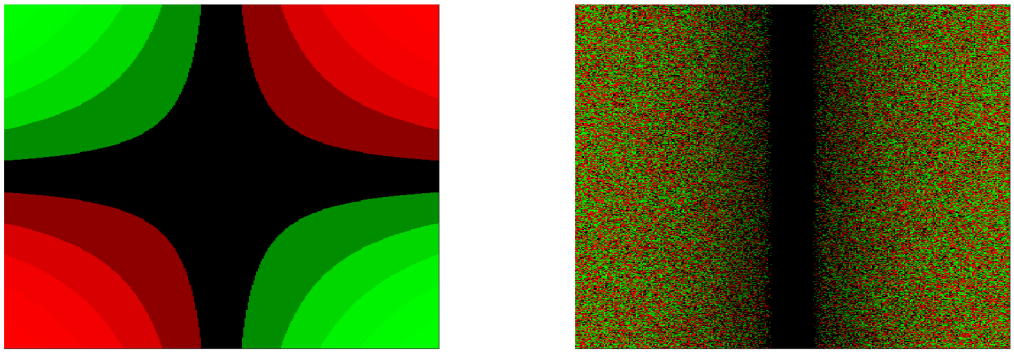

Figure 2: How does PA work? Given a "smooth" signal S of rank one (left), a random permutation transforms it into a "rough", noise-like matrix S_π. The permuted matrix is typically of full rank, and its operator norm is much smaller than that of the signal matrix. Thus, S_π does not perturb the permuted noise matrix N_π significantly, which allows the estimation of the noise level $\|N_\pi\|_{op} =_d \|N\|_{op}$ (equality in distribution). Then, factors above the noise level are selected.

The factor selection problem is also important in principal component analysis (PCA). While PCA and factor analysis are not the same (see e.g., Jolliffe, 2002, for a clear explanation), in practice permutation methods are very popular to select the number of PCs (e.g., Friedman et al., 2009; Zhou et al., 2017). Our work also bears on PCA.

Factor models are also well studied in econometrics, (e.g., Bai and Ng, 2008; Onatski, 2009, 2010; Fan et al., 2011; Bai and Li, 2012, etc). In that setting, the factors are assumed to be so strong that identifying the significant factors is trivial. In contrast, we study settings with weaker, "emergent" factors. These are common in the behavioral and biological sciences.

1.2 Parallel Analysis

Parallel Analysis (PA) (Horn, 1965; Buja and Eyuboglu, 1992) is one of the most popular methods for selecting the number of factors. In the widely used version proposed by Buja and Eyuboglu (1992), we start with the $n \times p$ data matrix X containing the measurements x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$. We generate a matrix X_π by permuting the entries in each column of X separately. Here $\pi = (\pi_1, \pi_2, \ldots, \pi_p)$ is a permutation array, which is a collection of permutations π_j of $\{1, \ldots, n\}$. The permutation π_j permutes the j-th column of X, so X_π has entries $(X_\pi)_{ij} = X_{\pi_j(i),j}$. We repeat this procedure a few times.

We select the first factor if the top singular value of X is larger than a fixed percentile of the top singular values of the permuted matrices. One can use the median, 95%-th, or 100%-th percentile. If the first factor is selected, then we repeat the same procedure for the second largest singular value of X, comparing with the second singular values of permuted matrices, and so on. We stop when a factor is not selected.

Figure 1 illustrates parallel analysis. A data matrix X is randomly permuted, and the singular values of both X and X_π are plotted in decreasing order. Only the first singular value of X is larger than that of X_π, so one factor is selected. In a practical application, one ought to take multiple permutations.

PA has a lot going for it. It is a simple, concrete method. It is easy to code in software,

and it is already implemented in several `R` packages, including `nFactors` (Raiche et al., 2010). In addition, there is a great deal of empirical evidence that it works well, compared to other standard methods. The main other methods are Kaiser's "eigenvalues larger than one" rule (Kaiser, 1960), Bartlett's likelihood ratio test (Bartlett, 1950), the scree plot (Cattell, 1966), and Velicer's minimum-average-partial criterion (MAP) (Velicer, 1976). Empirical evidence favors PA. In an extensive simulation study, Zwick and Velicer (1986) concluded that PA and MAP are consistently accurate. Peres-Neto et al. (2005) compared 20 methods for selecting the number of components in factor analysis. They found the PA and its variants are the best methods. Owen and Wang (2016) proposed a bi-cross-validation method, focusing on estimating the factor component. Even for this new objective function, PA was one of the most accurate methods.

Based on these findings, PA has become a standard textbook method:

1. Brown (2014) notes that PA "*is accurate in the vast majority of cases*"
2. Hayton et al. (2004) review evidence from social science and management that PA is "*one of the most accurate factor retention methods*"
3. Costello and Osborne (2005) write that PA is "*accurate and easy to use*"
4. In the context of PCA, Friedman et al. (2009) use it as the default method for selecting the number of significant components (see Fig 14.24, p. 538).

While PA has not been employed enough by practitioners (Hayton et al., 2004; Gaskin and Happell, 2014), recently it has started to become more widely used by leading researchers in applied statistics, especially in the biological sciences:

1. Leek and Storey (2008) use it in a general framework for multiple testing dependence. It is the default method in the popular `sva` package for "Surrogate Variable Analysis" (Leek and Storey, 2007).
2. Wing H. Wong's group used it to select the number of components when performing dimension reduction while controlling for covariates (Lin et al., 2016).
3. Gerard and Stephens (2017) use it in their general methodology for removing unwanted variation (RUV) based on negative controls.
4. Zhou, Marron and Wright use a block permutation version in eigenvalue significance tests for genetic association (Zhou et al., 2017).

PA is not the end of the story, and there are newer methods (see e.g., Kritchman and Nadler, 2008; Onatski, 2012; Josse and Husson, 2012; Gaskin and Happell, 2014). However PA is by far the most popular method, and thus it is worth studying.

1.3 The lack of theory, and this work

Despite this empirical success, there is essentially no theoretical justification that PA works. For instance, Green et al. (2012) calls PA "*at best a heuristic approach rather than a mathematically rigorous one*". Clearly, this lack of understanding limits the appeal of PA. The perceived lack of rigor prevents practitioners from making the best decision on which method to use.

In this paper, we will develop the first fully rigorous understanding of PA. We will show that PA selects the large factors in a broad range of factor models. Importantly, PA selects only the factors whose size is above a certain well-specified threshold. The key requirements are that (1) the dimension p is large compared to the sample size n, and (2) each factor loads on more than just a few variables. These are quantified into precise mathematical statements. See Thm, 2.1 for a clean result.

The basic idea is simple: The factor model can be written in a signal-plus-noise form $X = S + N$, where S is of low rank. A random permutation "destroys" the signal S, transforming it into a noise-like matrix (see Fig. 2); while keeping the noise distribution invariant. This allows the identification of the factors above noise level.

We hope that our results will de-mystify PA, and help practitioners understand when it is the most suitable method in factor analysis and PCA. We also hope that the mathematical approach developed in this paper will become useful to improve PA, as discussed at the end of the paper. In a follow-up work, we have been able to address several of the limitations of PA (Dobriban and Owen, 2017).

Roughly speaking, our contributions are as follows:

1. We provide an asymptotic analysis of PA in "low-rank-signal plus noise" models (Sec. 2.2). We prove a basic Consistency Lemma (Lem. 2.2) giving general conditions on the signal and the noise when PA selects the large factors, which we call *perceptible*.
2. We then provide concrete assumptions under which the general conditions for the signal (Sec. 3) and the noise (Sec. 4) hold. This involves new bounds on operator norms of permutation random matrices (Thm. 3.1).
3. We apply these results to show that PA selects the perceptible factors in factor models (Sec. 2.1). For pedagogical reasons, this is presented *before* the general signal-plus-noise approach.
4. We discuss the application of PA in PCA (Sec. 5). We provide numerical evidence supporting our claims (Sec. 6), which are all reproducible with software available at `github.com/dobriban/PA`. We close with a discussion of future work (Sec. 7).

2 Consistency of permutation methods (PA)

2.1 A simple result

We start by presenting a simple consistency result for PA. Recall that we have observations x_{ij} following the standard factor model (1). The vector $x_i = (x_{i1}, \dots, x_{ip})^\top$ of observations for the i-th sample can be expressed as

$$x_i = \Lambda \eta_i + \varepsilon_i,$$

where Λ is the $p \times r$ factor loading matrix with entries λ_{jk}, η_i is the r-vector of factor values for the i-th sample, and ε_i is the p-vector of unique factors. The factors η_i are random variables, while the loadings Λ are fixed parameters. The $n \times p$ matrix $X = (x_1, \dots, x_n)^\top$ can be written as

$$X = H\Lambda^\top + \mathcal{E}.$$

Here H is the $n \times r$ matrix containing the factor values η_{ij}, and $\mathcal{E}$ is the $n \times p$ matrix containing the noise ε_{ij}. The covariance matrix of one sample x_i is

$$\Sigma = \Lambda \Psi \Lambda^\top + \Phi,$$

where $\Phi = \mathrm{diag}(\Phi_i)$ is the diagonal matrix of idiosyncratic variances.

It is well known that the parameters are not uniquely identified in this model (Anderson, 2003). It turns out, however, that the number of *large factors* is asymptotically well defined. The key is to quantify *the size of the noise* via the operator norm of the noise matrix. For this, we consider a sequence of factor models where the sample size n or the dimension p grows. In this asymptotic setting, we suppose that we can re-normalize the data so that

$$c_{n,p}^{-1} \|\mathcal{E}\| \to b > 0$$

almost surely (a.s.), or in probability. Here $\|M\|$ denotes the operator norm, the largest singular value of a matrix M, and $c_{n,p}$ are some constants. We define b as *the size of the noise*. As we will see, there are many settings where $c_{n,p}$ exists and thus b is well defined. Convergence a.s. and in probability are both considered, and allow for "parallel" theories.

We define the *above-noise factors* as those whose contribution to variation in the data is larger than the size of the noise. We measure the contribution of the k-th factor by the k-th largest singular value $\sigma_k(X)$. We say that the k-th factor is above-noise if $c_{n,p}^{-1}\sigma_k(X) > b$ a.s, (or in probability). It may seem surprising that this definition depends on the entire asymptotic setting, and not just on finite values of n, p. However, since our entire approach is asymptotic, this is reasonable. In practice, a factor is above-noise if its effect on variation is larger than the noise.

In addition, it will be useful to define *perceptible factors*, whose effect on variation differs from the size of the noise by some definite value. We define *perceptible factors* as those indices k for which $c_{n,p}^{-1}\sigma_k(X) > b + \varepsilon$ a.s (or in probability) for some $\varepsilon > 0$. Similarly, we define *imperceptible factors* as those indices k for which $c_{n,p}^{-1}\sigma_k(X) < b - \varepsilon$ for some $\varepsilon > 0$.

Our main result is that PA selects the perceptible factors. To state this we will need the *distribution function* of Φ, the discrete probability mass function placing equal mass on all Φ_i. For a bounded probability distribution H, we will also need the quantity $s(H) = \sup \mathrm{supp}(H)$, the supremum of the support of H. For a discrete distribution H taking values on $h_1 < h_2 < \ldots < h_l$, we have $s(H) = \max_i h_i = h_l$. Let $\Psi^{1/2}$ be the symmetric square root of Ψ, and let us define the scaled factor loading matrix $\Lambda\Psi^{1/2} = [f_1, \ldots, f_r]$.

Theorem 2.1 (Parallel analysis selects the perceptible factors)**.** *Suppose we observe n independent samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. Assume the following conditions:*

1. **Factors***: The number r of factors is fixed. The factors η_i have the form $\eta_i = \Psi^{1/2}U_i$, where U_i have r independent entries with mean zero and variance 1.*
2. **Idiosyncratic terms***: The idiosyncratic terms are $\varepsilon_i = \Phi^{1/2}Z_i$, where $\Phi^{1/2}$ is a diagonal matrix, and Z_i have p independent entries with mean zero, variance one, and bounded fourth moment.*
3. **Asymptotics***: $n, p \to \infty$ such that one of the following conditions holds:*
   (a) *$p/n \to \gamma > 0$, while the distribution function of Φ converges weakly to H, and $\max \Phi_j \to s(H)$. The entries of Z_i have bounded $6+\varepsilon$-th moment.*
   (b) *$p/n \to \infty$, while the entries $\Phi_j \le C\,\mathrm{tr}[\Phi]/p$ for all j.*
4. **Factor loadings***: The r vectors of scaled loadings f_k are each bounded, in the sense that $|f_k|_2 \le Cn^{1/4-\delta/2}$. They are also delocalized, in the sense that $|f_k|_4/|f_k|_2 \to 0$.*

Then with probability tending to one, parallel analysis selects all perceptible factors*, and* no separated imperceptible factors*.*

The proof of the theorem follows from the more general approach developed below, and is given later in Sec. 8.1.

Importantly, PA selects only the sufficiently large factors, whose size is above a certain well-specified threshold. While in general there we are not aware of a simple description of how large the factors must be to be selected, in the special case of spiked models, the thresholds become much more explicit, see Corollary 5.1.

The theory allows growing factors, but only at the rate $|f_k|_2^2 \sim n^{1/2-\delta}$. In econometics, (e.g., Bai and Ng, 2008; Onatski, 2009, etc), the factors grow linearly at rate n, so the current theory is weaker. However, PA is actually used more in computational genomics and social science, where the factors are typically much weaker. So, we think that the current theoretical

results are a good first step. In future work it would be important to examine the issue of strong factors in more detail.

Assumption 3(a) is somewhat technical. We assume that the distribution function of Φ converges weakly to a certain limit probability distribution H. This means that there is a certain "regularity" in the overall distribution of the variances. Mathematically, it is a standard assumption in random matrix theory (Bai and Silverstein, 2009; Yao et al., 2015). This and $\max \Phi_j \to s(H)$ are technical assumptions needed to guarantee that the *size of the noise* $b > 0$ is well defined.

The conclusion is that under reasonable conditions, PA selects the perceptible factors with high probability. A key feature of the theorem is that it allows both the sample size n and the dimension p to be large. Both $p/n \to \gamma > 0$ and $p/n \to \infty$ are handled, so that p can be much larger than n. However $p/n \to 0$ is not handled, and we will see in simulations that PA does not always work in this regime. This is one of the main conclusions of the current paper: *PA works well when the dimension is large.* This can be interpreted as a blessing of dimensionality.

The intuitive explanation is that when p is small, the factor loadings increase the effective variance of the features of the data. Thus, when the features are permuted, the variances are overestimated. Hence, the true noise level is overestimated, and some smaller factors may not be detected. However, this heuristic argument at least indicates that PA will likely be *conservative* in this case.

A second key feature is that the factor loading vectors λ_j need to load on more than just a few variables. Suppose for simplicity that we have an *orthogonal factor model*, $\Psi = I_r$, so $f_k = \lambda_k$. The formal requirement is that the ratio of the ℓ_4 and ℓ_2 norms vanishes: $|\lambda_j|_4/|\lambda_j|_2 \to 0$. For instance, $\lambda_j = (1, 0, 0, \ldots)$ does not satisfy this, but $\lambda_j = p^{-1/2}(1, 1, \ldots)$ does. In fact, λ_j can have nonzero loadings on a vanishing fraction δ of the entries, as long as $n\delta \to \infty$, and the entry sizes are lower bounded. An interesting example is a *clustering* pattern, where the $\lambda_{jk} = |S_j|^{-1/2} I(k \in S_j)$, and S_j are mutually disjoint clusters with sizes $|S_j| \to \infty$[1].

Thus, our assumptions are not restrictive. In practice, they mean that the loadings cannot concentrate on a small number of variables. This assumption is similar to non-sparsity, delocalization, or incoherence conditions seen in other works. This is the second main conclusion of the current paper: *PA works well when the factors load on more than just a few variables.*

In summary, our main conclusion is that PA works well when:

1. the dimension of the data is large, and
2. the factors load on more than just a few variables

Finally, this result concerns only separated factors, for which $c_{n,p}^{-1}\sigma_k(X) > b+\varepsilon$ or $c_{n,p}^{-1}\sigma_k(X) < b - \varepsilon$, but not factors near the noise level. Intuitively, these latter are "hard to distinguish" from the noise. This is related to the difficulty of understanding the critical regime of spiked models (e.g., Yao et al., 2015). At the moment, we do not have a clear understanding of PA in the critical regime.

In addition, we note that the new theoretical guarantees for PA cover roughly the same regime as when some of the existing methods based on eigenvalues are known to work (Kritchman and Nadler, 2008; Onatski, 2012). However, PA is a very popular method, used widely in science, and thus it is important to understand what it does.

2.2 The general approach: signal-plus-noise matrices

We now shift to a more general approach, which will be used in the rest of the paper. We will work with *signal-plus-noise* matrices $X = S + N$, where S is an $n \times p$ signal matrix of low rank,

[1] We thank Jingshu Wang for this example.

and N is an $n \times p$ noise matrix. In the standard factor model, $X = H\Lambda^\top + \mathcal{E}$. The first term is the signal due to the factor component, whose rank is at most r. Thus the factor model falls into the signal-plus-noise framework.

PA works with the permuted matrices X_π. By linearity, π acts separately on S and N, so $X_\pi = S_\pi + N_\pi$. Intuitively, permutations keep the noise distribution unchanged, while "destroying" the signal. Think of S as a "smooth" matrix, which can be achieved by reordering rows and columns. A typical permutation π transforms this into a "rough", "noise-like" matrix S_π. See Figure 2. This has the same entries as S, but is typically of full rank. While the Frobenius norm (sum of squared entries) is preserved, the operator norm can be dramatically reduced. Symbolically:

$$X_\pi = S_\pi + N_\pi \approx N_\pi.$$

Therefore, X_π behaves like the noise N_π. If the noise is sufficiently "invariant under permutations", then it may be possible to estimate its "size". Write $X =_d Y$ if the random variables X, Y have the same distribution, and suppose that $N_\pi =_d N$. Then from the previous two equations,

$$\|X_\pi\| \approx_d \|N\|.$$

Thus, the operator norms of the permuted matrix X_π and the noise matrix are roughly the same. Selecting factors whose singular values are larger than $|X_\pi|$ is roughly the same as comparing to the operator norm of the noise. This provides an intuitive justification that PA selects the perceptible factors, as defined above. The rest of the paper makes this intuition precise.

2.2.1 The consistency lemma

The first step is to formalize the above argument into a rigorous *consistency lemma*. This is a general result that gives broad conditions for the *signal* and the *noise* under which PA is consistent. We will then give examples when the two sets of conditions hold.

In the signal-plus-noise model $X = S + N$, suppose S is deterministic and N is random; this can be achieved by conditioning on S. We want to provide a result that holds under a variety of asymptotic settings. Thus, consider an asymptotic setting $\mathcal{A}$, for instance

1. Classical asymptotics, where $n \to \infty$ and p is fixed
2. Proportional-limit asymptotics, where $n, p \to \infty$, such that $p/n \to \gamma > 0$. This is also known as high-dimensional asymptotics, random matrix asymptotics, or the thermodynamic limit (e.g., Paul and Aue, 2014; Yao et al., 2015; Dobriban and Wager, 2015).
3. "Big n, bigger p" asymptotics, where $n, p \to \infty$, such that $p/n \to \infty$,

Our consistency lemma does not depend on the specific type of asymptotics. It applies to all of the above settings.

Next, fix a mode of stochastic convergence, either convergence almost surely (a.s.), or in probability. Below we will use a.s., but the equivalent results hold for in probability, *mutatis mutandis*. We will use both in the application to factor models.

In the asymptotic setting $\mathcal{A}$, suppose that the signal matrix belongs to some parameter space $S \in \Theta$, and the noise has some distribution $N \sim P_N$. Suppose that we have re-normalized the data such that under $\mathcal{A}$,

$$\|N\| \to b > 0$$

a.s. As in the special case of factor models discussed above, we define the *above-noise factors* as those indices k for which $\sigma_k(X) > b$ a.s.

Consider also a distribution of permutation arrays Π, defined for all n, p of interest; for instance each permutation π_j sampled independently and uniformly from the set of all permutations of $[n] = \{1, \ldots, n\}$.

Before turning to usual PA, it is pedagogically helpful to define *asymptotic PA* as a theoretical version of PA leading to a particularly simple analysis. Consider a finite set of permutation arrays Π_0 sampled independently, each according to Π. Asymptotic PA selects those factors for which $\sigma_k(X) > \max_{\pi\in\Pi_0} \|X_\pi\|$ a.s. This definition depends on the entire asymptotic setting $\mathcal{A}$, not just on finite values of n, p. In finite samples, we instead select the factors for which $\sigma_k(X) > \max_{\pi\in\Pi_0} \|X_\pi\|$. Asymptotic PA is an "oracle method", but it leads to very elegant results. The second definition is practically feasible, and we will see that the results are still nice.

As we will see, in our setting selecting the factors above the 95th percentile of $\{\|X_\pi\| : \pi \in \Pi_0\}$ leads to an entirely equivalent method. This is because the values $\|X_\pi\|, \pi \in \Pi_0$ all converge a.s. to the same value. The difference is only important for the properties of PA as a hypothesis testing method, specifically its control of the type I error. Thus, we focus on asymptotic PA as defined above:

Lemma 2.2 (Consistency lemma)**.** *Suppose the following*

1. **Noise invariance***: The distribution of the noise is invariant under permutations, so $N =_d N_\pi$, where the equality in distribution is taken with respect to the joint randomness of the noise matrix $N \sim P_N$ and the independently chosen permutation $\pi \sim \Pi$.*
2. **Signal destruction***: Under the asymptotics $\mathcal{A}$, we have $\|S_\pi\| \to 0$ a.s., for all $S \in \Theta$, where the randomness is induced by the random permutation $\pi \sim \Pi$.*

Then, asymptotic parallel analysis is consistent for selecting the above-noise factors.

Proof. Since $X_\pi = S_\pi + N_\pi$, by the triangle inequality we have $|\|X_\pi\| - \|N_\pi\|| \leq \|S_\pi\| \to 0$. Now, by invariance $N =_d N_\pi$, and by the convergence $\|N\| \to b$ to the noise level, we have that the operator norms of the permuted matrices also converge: $\|N_\pi\| =_d \|N\| \to b$. Hence, it follows that $\|X_\pi\| \to b$.

Asymptotic parallel analysis selects the factors for which $\sigma_k(X) > \max_{\pi\in\Pi_0} \|X_\pi\|$ a.s. Since Π_0 is of fixed size, based on the above argument, this is the same as those factors for which $\sigma_k(X) > b$ a.s., which are exactly the above-noise factors. This finishes the proof.

□

This result is a very elegant form of the statement that PA selects the number of above-noise factors. However, it deals with asymptotic PA, which is an oracle method only defined asymptotically. Can we remove the asymptotics from the definition of the method?

Recall that we consider the version of non-asymptotic parallel analysis which selects the factors for which $\sigma_k(X) > \max_{\pi\in\Pi_0} \|X_\pi\|$. Since above-noise factors are defined asymptotically by comparing $\sigma_k(X)$ to b, and non-asymptotic PA depends only on finite n, p, it is not clear how to show that PA selects all above-noise factors. However, this becomes clear if we focus on *separated above-noise factors*, called *perceptible factors*, as indicated previously. Thus, we define *perceptible factors* as the k for which $\sigma_k(X) > b + \varepsilon$ a.s. for some $\varepsilon > 0$. We also define *imperceptible factors* as the k for which $\sigma_k(X) < b - \varepsilon$ a.s. for some $\varepsilon > 0$. We then have the following analogue of the previous lemma:

Lemma 2.3 (Consistency lemma for non-asymptotic PA)**.** *Under the conditions of Lemma 2.2, PA selects all perceptible factors, and no imperceptible factors, a.s.*

Proof. Non-asymptotic parallel analysis selects the factors for which $\sigma_k(X) > \max_{\pi\in\Pi_0} \|X_\pi\|$. Since $\max_{\pi\in\Pi_0} \|X_\pi\| \to b$ a.s., it is immediate that this includes all perceptible factors, and no imperceptible factors, a.s.

□

These results give broad conditions for the signal and the noise under which PA is consistent. The real work is always to show that these conditions hold in particular cases of interest, such as for factor models.

2.2.2 Conditions on the signal and the noise

When do our assumptions hold? We start here with a brief discussion.

For the noise, we need two assumptions.

1. The existence of a well-defined asymptotic noise level $b > 0$ such that $\|N\| \to b > 0$. This imposes a restriction on the noise models. For this condition, it will be helpful that operator norms of random matrices have been studied thoroughly, and thus there are broad conditions under which such convergence is known.
2. The invariance of the distribution of noise to permutations: $N =_d N_\pi$. There is a tradeoff: the more general the noise distribution, the smaller the set of permutations that keeps it invariant. Thus, this also imposes a restriction, because we may need a large set of permutations to cancel out the signal terms, as we see next.

For the signal, we need one assumption:

1. The operator norm of the permuted signal matrices vanishes, $\|S_\pi\| \to 0$ a.s. for all $S \in \Theta$. There is a tradeoff here too: The larger the parameter space Θ, the harder this is, and the more permutations are needed to get enough "averaging" for this to hold. For certain signals, e.g., the all ones matrix with $S_{ij} = 1$, this is entirely impossible, because every permutation keeps the matrix unchanged.

In the next two sections, we examine the two sets of conditions in more detail.

3 Signal models

When do our assumptions on the signal hold? We need that permutations "destroy" the signal structure, so that $\|S_\pi\| \to 0$ a.s., for all $S \in \Theta$. Consider a rank one signal matrix $S = uv^\top$. Then, acting on this by a permutation array π we get (denoting by $\odot$ elementwise product of matrices):

$$S_\pi = (uv^\top)_\pi = (u1^\top)_\pi \odot 1v^\top = [\pi_1(u); \pi_2(u); \ldots; \pi_p(u)] \odot 1v^\top.$$

Each permutation π_j permutes the corresponding column j of S. This column equals $v_j u$, so π_j permutes the entries of u. Since π_j is a uniformly random permutation, the distribution of this column is uniform on all permutations of u, and is "modulated" by v_j. If the entries of u sum to zero, this is effectively "noise" of variance approximately v_j^2/n. The n entries of column j are exchangeable random variables, which are almost independent for large n. Hence, $(uv^\top)_\pi$ is a random matrix whose columns are independent, and within each column the entries are nearly independent, with variance approximately v_j^2/n. If the entries of the matrix were independent, we could use well-known results controlling its operator norm (e.g., Vershynin, 2010). However, since the entries are dependent, we need to establish these results here from first principles.

A first simplification is that we can separate the component corresponding to $u \in span(1)$, where $1 = (1, 1, \ldots)^\top$ is the all ones vector, and its orthocomplement. The first part is kept invariant by the permutation. So we just need to assume that it goes to zero. Let $\theta \cdot n^{-1/2} 1 \cdot v^\top$ be this component, where $|u|_2 = |v|_2 = 1$. Then, we need to assume that $\theta \to 0$, because we need

$$\|\pi(\theta \cdot n^{-1/2} 1 \cdot v^\top)\| = \|\theta \cdot n^{-1/2} 1 \cdot v^\top\| = \theta |v|_2 = \theta \to 0.$$

In our application to factor models, we will separate this component, and show that $\theta \to 0$ holds.

On the orthocomplement, we will use the moment method to show $\|S_\pi\| \to 0$. We have that $\|A\|^k \leq \operatorname{tr}(A^\top A)^k$ for all k. Hence,

$$P(\|A\| > \varepsilon) = P(\|A\|^k > \varepsilon^k) \leq P(\operatorname{tr}(A^\top A)^k > \varepsilon^k) \leq \varepsilon^{-k} \mathbb{E} \operatorname{tr}(A^\top A)^k$$

Thus, to show that $\|S_\pi\| \to 0$ in probability, it is enough to argue that $\mathbb{E} \operatorname{tr}(A^\top A)^k \to 0$ for some $k > 0$. To show a.s. convergence, by the Borel-Cantelli lemma we need that $\mathbb{E} \operatorname{tr}(A^\top A)^k$ is summable for some $k > 0$. After the appropriate moment calculations, we obtain the following result:

Theorem 3.1 (Requirements on the signal)**.** *Consider signals of the form $S = n^{-1/2}\theta \cdot 1 v_0^\top + T$, where $T = \sum_{i=1}^r \theta_i u_i v_i^\top$, with $|u_i|_2 = |v_i|_2 = 1$, and $u_i^\top 1 = 0$ for all i. Here r can be fixed or change with the dimensions n, p. Suppose that $\theta \to 0$. Define the constants A_{nk} for $k = 2, 3, 4$ as*

$$A_{nk} = \sum_{i=1}^r \theta_i \cdot C_k(v_i)^{1/(2k)}$$

where $C_k(v)$ are defined as

1. $C_2(v) = 1/(n-1) + |v|_4^4$
2. $C_3(v) = 1/(n-1)^2 + 9n^{-1}|v|_4^4 + |v|_6^6$
3. $C_4(v) = 1/(n-1)^3 + 4/(n-1)^2|v|_4^4 + 12n^{-1}[|v|_4^8 + |v|_6^6] + |v|_8^8$

Then,

$$[\mathbb{E} \operatorname{tr}(S_\pi^\top S_\pi)^k]^{1/(2k)} \leq A_{nk}.$$

Therefore,

1. *If $A_{nk} \to 0$, then $\|S_\pi\| \to 0$ in probability.*
2. *If A_{nk}^{2k} are summable, then $\|S_\pi\| \to 0$ almost surely.*

The proof is provided in Sec. 8.2. Note that the second condition can only hold for $k \geq 3$, because $n^{-1} \lesssim A_{n2}^4$.

An interesting consequence of this result is that PA works in certain cases even when the *number* of signals as well as the *strength* of signals grows to infinity simultaneously. Indeed, suppose v_i are all maximally delocalized, so that $|v_i|_\infty \leq C p^{-1/2}$ for some $C > 0$. Then, we have $|v_i|_4^4 \leq C^4 p^{-1}$, and $|v_i|_6^6 \leq C^6 p^{-2}$, therefore $C_3(v_i) \leq C'(n^{-2} + p^{-2})$ and

$$A_{n3}^6 \leq C' \left[\sum_{i=1}^r \theta_i\right]^6 \cdot \left(\frac{1}{p^2} + \frac{1}{n^2}\right)$$

So we need to find conditions under which this goes to zero or is summable. When $n, p \to \infty$, for $A_{n3} \to 0$ it is enough that $\sum_{i=1}^r \theta_i = O(\min(n,p)^{1/3-\varepsilon})$ for some $\varepsilon > 0$. For instance, when $p/n \to \gamma > 0$, is enough that $m|\theta|_\infty = O(n^{1/3-\varepsilon})$. We can take $n^{1/6}$ spikes (signals) of size $n^{1/6-\varepsilon}$ each, and parallel analysis will work. Alternatively, we can take one spike of strength $\theta = n^{1/3-\varepsilon}$.

This is important, because it allows us to handle two seemingly very different regimes simultaneously: the "explosive", i.e., growing, spikes of the type that are common in econometrics (Bai and Ng, 2008), while also handling the constant-sized spikes that are common in the literature in random matrix theory and applications to statistics (e.g., Paul and Aue, 2014; Yao et al., 2015).

3.1 Optimality considerations

3.1.1 Signal strength

The above results and discussion show that PA selects the perceptible factors in models of the form $\theta uv^\top + N$ as long as the signal strengh θ is not too large. For instance, we saw that $\theta = \min(n,p)^{1/3-\varepsilon}$ suffices for delocalized signals. It may seem counterintuitive that a strong signal can decrease the performance of PA. Is this a weakness of our theoretical analysis, or a weakness of the method?

To understand this issue, we recall that PA "transforms the signal into noise". Thus, a large signal is transformed into large noise, which can lead to the overestimation of the true noise level. In turn, this may prevent the selection of the above-noise factors which are *not* above the estimated noise level. So the problem is with PA, not with our result.

More precisely, the permuted matrix $S_\pi = (\theta uv^\top)_\pi$ is a matrix with independent columns, and within each column, with approximately independent (in truth, exchangeable) bounded entries. If the entries were independent with variance σ^2/n, the operator norm would be of order $\sigma \cdot [1 + (p/n)^{1/2}]$ (Bai and Silverstein, 2009; Vershynin, 2010). In our case, $\operatorname{tr} S_\pi^\top S_\pi = \operatorname{tr} S^\top S = \theta^2$, so heuristically, $p\sigma^2 \approx \theta^2$. Thus, heuristically

$$\|S_\pi\| \approx \theta \cdot [n^{-1/2} + p^{-1/2}].$$

In our consistency lemma, we showed that PA will select the above-noise factors if $\|S_\pi\| \to 0$, which amounts to $\theta \cdot [n^{-1/2} + p^{-1/2}] \to 0$. In particular, under high-dimensional asymptotics when $p/n \to \gamma > 0$, this holds if $\theta/n^{1/2} \to 0$. This suggests that our result $\theta = n^{1/3-\varepsilon}$ is *not* optimal, and PA works more broadly. We note that a k-th moment bound in Theorem 3.1 will allow $\theta = o(p^{1/2-1/(2k)})$. However, much more work is needed to show such a bound. In principle, the current moment calculations should work, but this is much beyond the scope of this work, as they become too hard for large k.

Thus it appears that very strong factors lead to problems with PA. This is counter-intuitive, because strong factors should be easy to detect. However, this apparent paradox can be resolved. The noise level estimated by PA is of the order of

$$f_{est} \approx \max(b, \sum_{k=1}^{r} \theta_k \cdot [n^{-1/2} + p^{-1/2}]).$$

A factors is not selected if $\sigma_k(X) < f_{est}$. From the analysis of spiked covariance matrix models, when the noise is of the form $n^{-1/2}X$ for X with iid entries, we expect the empirical singular values to behave (very roughly speaking) like $\sigma_k(X) \approx \theta_k + (p/n)^{1/2}$. From these two approximations, a factor k is not selected only if $\theta_k / \sum_{k=1}^r \theta_k \lesssim n^{-1/2} + p^{-1/2}$. This shows that only the *relatively unimportant* factors are not selected by PA, in the sense that the relative strength of the factor k, $\theta_k / \sum_{k=1}^r \theta_k$, must be small. In this sense, PA still selects the "relatively large" factors.

3.1.2 Delocalization

What is the precise condition needed on v? In Theorem 3.1 we gave several conditions depending on the norms $|v|_k$, for $k = 4, 6, 8$, which all amount to some form of delocalization, in the sense that v is non-sparse. Some non-sparsity condition is needed. Indeed, when $v = (1, 0, \ldots, 0)$, then a permutation only acts on the first column of $uv^\top$, thus the operator norm is unchanged. Some form of delocalization is needed, but finding the precise condition may need a new theoretical approach, which is beyond our current scope.

4 Noise models

When do our assumptions on the noise hold? We need two assumptions: invariance to permutations, and operator norm convergence.

4.1 Invariance

We need the that noise is invariant in distribution to permutations $N =_d N_\pi$, where the permutation π is also random. We will study when invariance holds for any *fixed* permutation π; then it will also hold for random permutations $\pi \sim \Pi_0$ chosen indepenendently from N. This is a non-asymptotic condition, so the findings will apply to any asymptotic setting we consider.

For $N =_d N_\pi$ it is enough if columns of N are independent, and each column has exchangeable entries. But a weaker condition is enough. Suppose for instance that $(N_{ij})_{ij}$ are an equicorrelated Gaussian random vector, in matrix form. Then clearly they are not independent, but are still invariant under any permutation.

Following this logic, if we vectorize the matrix N into an np-length vector, whose blocks of size n are the different features indexed from 1 to p, the condition $N =_d N_\pi$ means that the distribution is invariant under permutations within the blocks. For a Gaussian random vector, in terms of the covariance matrix, this means that it has the following block structure:

- $\mathrm{Var}\,[N_{ij}] = \sigma_j^2$: Within any column j, the entries N_{ij} are exchangeable random variables. Thus, they must have the same variance σ_j^2.
- $\mathrm{Cov}\,[N_{ij}, N_{i'j}] = \tau_j^2$ for $i \neq i'$: Similarly, distinct entries $N_{ij}, N_{i'j}$ in the same column must have the same covariance.
- $\mathrm{Cov}\,[N_{ij}, N_{kj'}] = \eta_{jj'}^2$ for $j \neq j'$: Consider two distinct columns j, j'. Since the entries within each of them can be permuted independently of each other while preserving the distribution, the covariance between any two entries $N_{ij}, N_{kj'}$ must be the same.

Equivalently, one has the explicit representation:

$$N = \mathcal{E} D^{1/2} + 1 z^\top \Sigma^{1/2},$$

where $\mathcal{E}$ is $n \times p$ matrix of iid Gaussians, D is diagonal, $z \sim \mathcal{N}(0, I_p)$, and Σ is $p \times p$ PSD. Here Σ induces the correlations between the different columns, and is not necessarily diagonal. Thus each sample has the form $N_i = D^{1/2}\varepsilon_i + \Sigma^{1/2} z \sim \mathcal{N}(0, D + \Sigma)$, which is a sum of a sample-specific independent diagonal normal random vector $D^{1/2}\varepsilon_i$, and the same normal random vector $\Sigma^{1/2} z$ added to each sample.

A bit more generally, we have the following representation for noise models invariant under permutations. The proof is immediate, and thus it is omitted.

Proposition 4.1 (Requirements for noise invariance)**.** *Suppose that the noise matrix N has rows of the form $N_i = D^{1/2}\varepsilon_i + z$, where ε_i are iid across i, and z is any random vector independent of all ε_i. Suppose that ε_i have independent standardized entries. and $D^{1/2}$ is diagonal. Then, the distribution of N is invariant under column permutations, i.e., $N =_d N_\pi$ for any fixed permutation array π.*

This result covers factor models, where the noise is of the form $N_i = D^{1/2}\varepsilon_i$. The term z is allowed by the theory, but it is usually not of practical interest. This common term can be viewed as the mean of N_i. Even if the mean is zero, this can be viewed as a common perturbation affecting all samples. However, z increases the overall noise level, because the operator norm $\|N\|$ is of order $(np)^{1/2}$. Thus, the presence of z renders many previously

above-noise factors to sink below the noise. The practically interesting scenarios usually have $z = 0$, which can be achieved by de-meaning the data.

In addition, there are other noise models that can be reduced to the model from Prop. 4.1. A prime example is correlated noise models where taking the Fourier transform, or some other known orthogonal transform in space-time, leads to independent coordinates.

For instance, in time series analysis (e.g., Brockwell and Davis, 2009), stationary processes can be transformed into having approximately independent coordinates by the Fourier transform. By the spectral representation theorem, every zero-mean stationary process has the representation $X_t = \int_{(-\pi,\pi]} \exp(it\nu) dZ(\nu)$, where Z is an orthogonal-increment process.

The autocovariance function can be written as $\gamma(h) = \int_{(-\pi,\pi]} \exp(it\nu)\ dF(\nu)$, for a distribution function F. If γ is absolutely summable and real-valued, then the process has asymptotically uncorrelated Fourier components (e.g., Brockwell and Davis, 2009, Prop. 4.5.2.). In particular, permutation methods are heuristically reasonable. However, making this rigorous would require us to understand what happens to the permutation distribution when we have only an approximate invariance of the noise. This is beyond our scope, but is interesting for future work (see Sec. 7).

4.2 Operator norm convergence

The second condition that we need for the noise is the convergence of the operator norm: $\|N\| \to b > 0$. Operator norms of random matrices have been studied for a long time, see e.g., Bai and Silverstein (2009); Vershynin (2010). We are fortunate that we can leverage some of these results. For instance, Bai, Yin, Silverstein and others have showed convergence of the operator norm of matrices of the form $N = n^{-1/2} X T^{1/2}$, where the entries of X are iid standardized random variables, and where $p, n \to \infty$ such that $p/n \to \gamma > 0$. We state this result together with another one for the case $p/n \to \infty$.

Proposition 4.2 (Requirements for noise operator norm, partly a corollary of Cor. 6.6 in (Bai and Silverstein, 2009))**.** *Suppose that the noise matrices have the form $N = c_p^{-1/2} X T^{1/2}$ with $c_p = \operatorname{tr} T$, where the entries of X are independent standardized random variables with bounded fourth moment, and T are diagonal positive semi-definite matrices. Suppose that $p \to \infty$, and one of the following two sets of assumptions holds:*

1. *$p/n \to \gamma > 0$, while the distribution function of the entries of T converges weakly to a limit distribution H, $F_T \Rightarrow H$. Moreover, the operator norm of T converges to the supremum of the support of H, $\|T\| \to \operatorname{supp}\sup(H)$, and the entries of X have bounded $6+\varepsilon$-th moment.*
2. *The entries of T are bounded as $t_j \leq C \operatorname{tr}[T]/p$ for all j, while (A) $p/n \to \infty$ or (B) $n^{2+\varepsilon} \leq p$ for some $\varepsilon > 0$.*

Then, we have $\|N\| \to b$ for some $b > 0$, in probability under (2A), and almost surely under (1) or (2B).

The second statement allows n fixed while $p \to \infty$, which is the "transpose" of classical asymptotics where p is fixed and $p \to \infty$. The proof is provided later in Sec. 8.3. Combined with the conditions on noise invariance, and with the conditions on the signal, this result provides a broad set of concrete scenarios when PA selects the perceptible factors.

5 PCA and spiked models

Should we select the number of components in PCA using PA? As Jolliffe (2002) clearly explains, there is a substantial difference between PCA and FA, and "*it is usually the case that*

the number of components needed to achieve the objectives of PCA is greater than the number of factors in a FA of the same data".

However, we can understand the behavior of PA in PCA within a certain class of popular *spiked models*. Spiked models have served as a theoretical tool to understand PCA in high dimensions. There are several versions, some of them mutually exclusive, see for instance Johnstone (2001); Paul (2007); Nadler (2008); Bai and Ding (2012); Benaych-Georges and Nadakuditi (2012); Onatski et al. (2013); Nadakuditi (2014), and Paul and Aue (2014); Yao et al. (2015) for more references.

An important class of signal-plus-noise spiked models was studied in Benaych-Georges and Nadakuditi (2012). Here $X = S + N$, where $S = \sum_{i=1}^{r} \theta_i u_i v_i^\top$, and $n^{1/2}u_i, p^{1/2}v_i$ are each iid vectors with iid entries from a distribution that satisfies a log-Sobolev inequality. It is assumed that $n, p \to \infty$ such that $p/n \to \gamma > 0$, the spectral distribution of N converges to a compactly supported distribution, and the top and bottom singular values converge to the respective edges $a < b$ of the distribution. The rank r and the spike strengths θ_i are fixed constants. Under these conditions, Benaych-Georges and Nadakuditi (2012) derive the asymptotic limits of the empirical singular values of X. They establish the *BBP phase transition* phenomenon discovered earlier by Baik et al. (2005) in a special case. For θ_i above a critical value, the corresponding empirical spike $\sigma_i(X)$ will converge to a definite value larger than b. In this case θ_i is said to be *above the phase transition*. These correspond to the perceptible factors in our terminology. For θ_i below the critical value, $\sigma_i(X) \to b$.

Our assumptions are neither more general, nor more specific. Indeed, we allow $p/n \to \infty$ and diverging spikes, while they allow a general converging spectral distribution, without requiring permutation-invariance.

However, our assumptions do have a nontrivial intersection. We can state the conclusion as a corollary. This justifies the use of permutation methods in PCA:

Corollary 5.1. *(PA in spiked models) Suppose we observe a signal-plus-noise spiked model $X = S + N$, where $S = \sum_{k=1}^{r} \theta_k u_k v_k^\top$, and $n^{1/2}u_k, p^{1/2}v_k$ are each iid vectors with iid entries from a distribution that satisfies a log-Sobolev inequality. Suppose that $n, p \to \infty$ such that $p/n \to \gamma > 0$. Suppose that the noise matrix is of the form $N = n^{-1/2} Y T^{1/2}$, where the entries of Y are independent standardized random variables with bounded 6+ε-th moments, and T are diagonal positive semi-definite matrices such the distribution function of the entries of T converges weakly to a limit distribution H. Suppose that the operator norm of T converges to the supremum of the support of H, $\|T\| \to \operatorname{supp}\sup(H)$.*

According to Benaych-Georges and Nadakuditi (2012), Thm. 2.9, for θ_k above the phase transition*, when $\theta_k > \bar{\theta}$ for a certain $\bar{\theta}$, the empirical singular values $\sigma_k(X)$ converge, $\sigma_k(X) \to \lambda_k$ a.s., for some $\lambda_k > b$, where $b > 0$ is the limit $\|N\| \to b$, as guaranteed by Prop. 4.2.*

Then, parallel analysis selects all spikes above the phase transition.

The analysis for the spikes below the transition is more delicate, and our results do not address it.

We also emphasize that, the threshold $\bar{\theta}$ above which the factors are selected becomes much more explicit. In particular, when the covariance of the noise is identity, $\bar{\theta} = \sqrt{\gamma}$, which is completely explicit as a function of n and p.

6 Numerical simulations

We perform numerical simulations to understand the behavior of PA. We wish to understand the effect of key parameters of the factor model, including signal strength and delocalization of loadings, on the accuracy of PA.

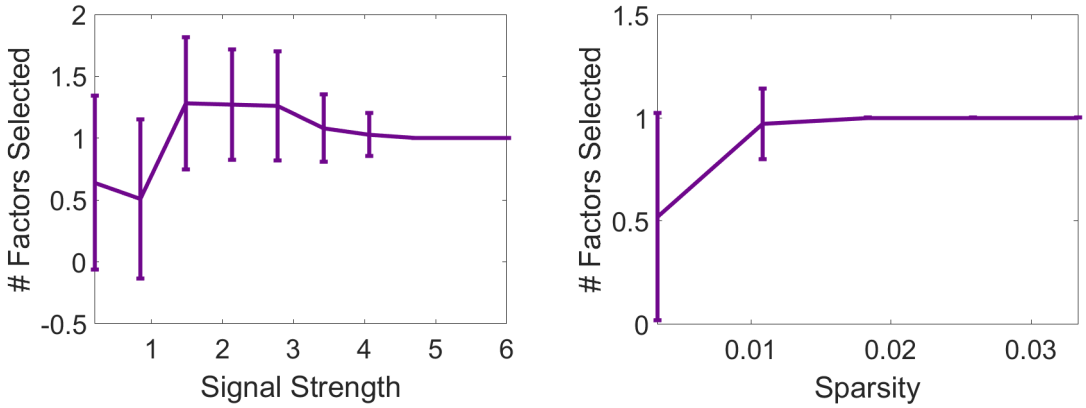

Figure 3: Mean and SD of number of factors selected by PA as a function of signal strength (left), and sparsity (right).

6.1 Effect of signal strength

We simulate from the factor model $x_i = \Lambda\eta_i + \varepsilon_i$. We generate the noise $\varepsilon_i \sim \mathcal{N}(0, I_p)$, and the factor loadings as $\Lambda = \theta\tilde{Z}$, where $\theta > 0$ is a scalar corresponding to factor strength, and $\tilde{Z}$ is generated by normalizing the columns of a random matrix $Z \sim \mathcal{N}(0, I_{p\times m})$.

We use a one-factor model, so $m = 1$, and work with sample size $n = 500$ and dimension $p = 300$. It is well known that the critical regime for the signal strength θ is of the order of $\gamma^{1/2}$. We vary θ on a grid of the form $\gamma^{1/2} \cdot s$, for s on a linear grid between 0.2 and 6.

We use PA to select the number of factors. We perform 10 Monte Carlo iterations for each parameter. Motivated by our theoretical understanding, for each Monte Carlo realization of X, we generate only one permutation X_π. We select the first factor if $\|X\| > \|X_\pi\|$. The results in Fig. 3 (a) show that PA is selects the right number of factors as soon as the signal strength s is larger than ~ 4. This agrees with our theoretical predictions, since it shows that PA selects the perceptible factors.

It may seem "wrong" that PA selects a factor even when the signal strength is nearly 0. However, this result is in agreement with our theoretical predictions. Indeed, such a factor is below-noise, but *non-separated*. In line with the discussion in Sec. 5, the singular value σ_k corresponding to a spike below the phase transition converges to the noise level b. Thus, the empirical singular value does not separate from the noise level, hence PA cannot identify it as below-noise.

6.2 Effect of delocalization

We provide numerical evidence for our claim that "PA works when the factors load on more than just a few variables." We use the same model as above. To change the delocalization of the factor scores, we define the sparsity parameter c, and generate c-sparse factors, by setting the first $\lfloor c \cdot p \rfloor$ coordinates of Z to be iid Gaussians, and the remaining coordinates to be zero. One can verify that for every vector λ of factor scores, the expected "localization" parameter $L = |\lambda|_4/|\lambda|_2$ is approximately $L = (9/cp)^{1/4}$, which decreases with c. Our theoretical results suggest that PA should select the right number of factors for "delocalized" or "non-sparse" vectors, when c is large and L is small.

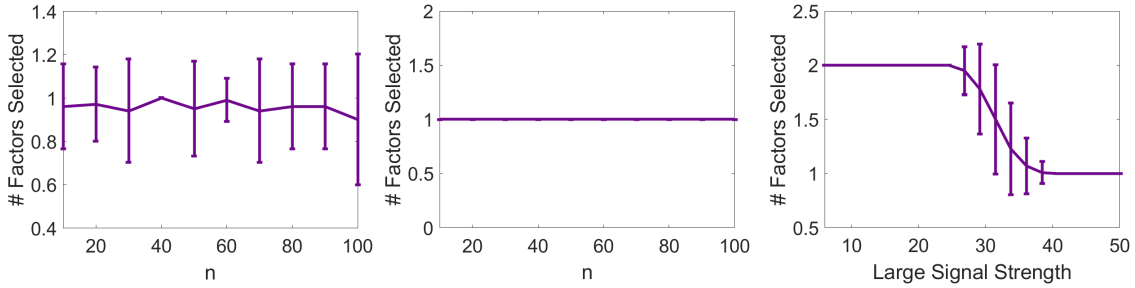

Figure 4: Mean and SD of number of factors selected by PA as a function of sample size for $p = 3$ (left) and $p = 1000$ (middle). Same quantity in a 2-factor model as a function of stronger factor value (right).

We set $\theta = 2$ to place ourselves in a critical regime where the effect of delocalization is visible. This choice was made empirically. We vary c on a grid from $1/p$ to $10/p$. We perform 100 Monte Carlo iterations for each setting of the parameters.

The results in Fig. 3 (b) show that PA tends to select the right number of factors for non-sparse, delocalized factor loadings (large c). This agrees with our theoretical predictions.

It is remarkable that PA already works when the sparsity is 2% ($c = 0.02$). That is, if the factor loads on *at least 6 out of 300 variables*, PA selects the right number of factors! This surprising result suggests that PA is likely to perform well in many realistic settings, and that delocalization is not a stringent requirement.

6.3 Effect of dimension

We provide numerical evidence for our claim that "PA works when the dimension of the data is large, even when the di." Using the same model as in the first simulation, we compare the accuracy of PA for $p = 3$ and $p = 1000$. We set the signal strength to $\theta = 6\gamma^{1/2}$, which is a perceptible factor. This corresponds to the same signal strength for all p. Thus, the two problems are equally hard statistically; or put it differently, the SNR is the same for the two values of p. We vary the sample size from $n = 10$ to $n = 100$ in steps of 10.

The results in Fig. 4 show that PA tends to select the right number of factors almost without error for $p = 1000$, but not for $p = 3$. This holds already for $p = 10$ (data not shown). This agrees with our theoretical predictions. Moreover, this also suggests that the requirement on the sample size is not stringent.

6.4 Effect of strong signals on detectability of weak signals: Shadowing

We provide numerical evidence for the claim that "PA selects the relatively important factors." Using the same model as in the first simulation, we evaluate the accuracy of PA in a two-factor model. We set the smaller signal strength to $\theta_1 = 6\gamma^{1/2}$, which is a perceptible factor. We vary the larger signal strength as $\theta_2 = c_2\gamma^{1/2}$ on a grid between $c_2 = 6$ and 50.

The results in Fig. 4 (c) show that PA tends to select the right number of factors almost without error for $c_2 < 25$, but it starts making errors above that value. Above $c_2 > 35$, PA consistently selects only one perceptible factor. Qualitatively, these agree with our theoretical predictions. A strong factor is transformed into noise by PA, thus "shadowing" the weaker

factor. Quantitatively, in this example the ratio of small-to-large signal strength where PA breaks down is $\theta_1/(\theta_1+\theta_2) \approx 6/35 \approx 0.17$. According to our theory, this should be on the order of $n^{-1/2}+p^{-1/2} = 0.1$. Thus, our predictions seem quite accurate.

7 Discussion

In this paper we provided a theoretical analysis of parallel analysis (PA). We established precise conditions under which PA consistently selects the perceptible factors for large datasets. We argued that PA works when the dimension of the data is large, and when the factors load on more than just a few variables.

There are numerous important directions for future research. First, there are variants of PA developed in applied research (see e.g., Peres-Neto et al., 2005; Brown, 2014; Crawford et al., 2010; Gaskin and Happell, 2014). When are they useful? These methods differ in:

1. The test statistic, for instance: singular value gap, fraction of variance explained, robust correlations, loadings (Buja and Eyuboglu, 1992).
2. The number of permutations, and percentile used: mean of eigenvalues (Horn, 1965), other percentiles (Buja and Eyuboglu, 1992; Glorfeld, 1995).
3. Using stepwise testing (Horn, 1965).
4. Using the correlation matrix.

Can we understand when they help, and possibly develop improvements? This is especially interesting for tasks *other* than selecting the number of factors, such as estimating the factor loadings.

Second, what should one do when the noise is correlated? Independent permutations on the original columns do not generate the correct null distribution. In Sec. 4.1 we saw that taking the Fourier transform may help for stationary time series. However, this will need a much more careful analysis.

Third, an important issue that we have not discussed is the computational cost of PA. The cost of permutations and SVDs for PA can become a problem for "big data". Another important issue is the randomness introduced by PA, which can lead to arbitrary decisions. Can we speed up PA, or remove the randomness? Zhou et al. (2017) developed such a method, by employing Dobriban (2015)'s Spectrode algorithm to approximate the noise level. Can we develop a theoretical understanding of this method, with suitable improvements?

8 Proofs

8.1 Proof of Thm. 2.1

We will check that the conditions of the Consistency Lemma 2.3 hold with probability tending to one. In matrix form, the factor model reads $X = U\Psi^{1/2}\Lambda^\top + Z\Phi^{1/2}$. We first normalize it to have operator norm of unit order: $n^{-1/2}X = n^{-1/2}U\Psi^{1/2}\Lambda^\top + n^{-1/2}Z\Phi^{1/2}$.

Let us verify the required conditions:

1. Signal: Let $\Lambda\Psi^{1/2} = [f_1, \dots, f_r]$.
   The signal component is $S = n^{-1/2}U\Psi^{1/2}\Lambda^\top = n^{-1/2}\sum_{k=1}^r u_k f_k^\top$
   Since there are only a fixed number of factors, it is enough to analyze one term $n^{-1/2}uf^\top = n^{-1/2}|u|_2|f|_2 \cdot \tilde{u}\tilde{f}^\top$, where $\tilde{\cdot}$ denotes normalized vectors. Let $\mathrm{e} = n^{-1/2}1$.

(a) Mean term: The term in the span of 1 is $n^{-1/2}|u|_2|f|_2 \cdot \tilde{u}^\top \mathrm{e} \cdot \mathrm{e}\tilde{f}^\top$. We need that $n^{-1/2}|u|_2|f|_2 \cdot |\tilde{u}^\top \mathrm{e}| \to 0$.
Now, $|f|_2 \le Cn^{1/4-\delta/2}$ by assumption. Moreover, if u_i has iid entries with mean 0 and variance 1, then by the LLN $n^{-1/2}|u|_2 \to 1$. Thus it is enough that $n^{1/4-\delta/2}|\tilde{u}^\top \mathrm{e}| = |n^{-1/4-\delta/2}(\sum u_i)| \to 0$. This holds by the CLT.

(b) Zero-mean term: We need to analyze the term in the orthocomplement of 1. Let $P = I - \mathrm{e}\mathrm{e}^\top$ be the de-meaning projection operator. Our term is $n^{-1/2}Puf^\top = n^{-1/2}|f|_2|Pu|_2 \cdot \widetilde{Pu}\tilde{f}^\top$
From Thm 3.1, where the low rank part has form $\theta uv^\top$, we need a bound on $\theta(2n^{-1} + |v|_4^4)^{1/4}$. But note that

$$\theta/[Cn^{1/4-\delta/2}] = n^{-1/2}|f|_2/[Cn^{1/4-\delta/2}]|Pu|_2 \le n^{-1/2}|u|_2 \to 1$$

a.s., so we need only

$$n^{1/4-\delta/2}(2n^{-1} + |v|_4^4)^{1/4} \to 0.$$

i.e., $n^{1/4-\delta/2}|\tilde{f}|_4 \to 0$, or also $n^{1/4-\delta/2}|f|_4/|f|_2 \to 0$.

2. Noise: We need first that $n^{-1/2}Z\Phi^{1/2}$ has a distribution that is invariant under permutations, as discussed in Sec. 4.1. This holds by inspection.
We need second that $\|n^{-1/2}Z\Phi^{1/2}\|_2 \to b$. Conditions for this are given in Prop. 4.2, and one can verify that the conditions given in the theorem match these. Thus, PA selects all perceptible factors, and no imperceptible factors.

8.2 Proof of Thm. 3.1

We need to show that $\|S_\pi\| \to 0$, where $S = n^{-1/2}\theta \cdot 1v^\top + \sum_{i=1}^r \theta_i u_i v_i^\top$. The term $n^{-1/2}\theta \cdot 1v^\top$ is handled by the assumption $\theta \to 0$, so we can focus on the rest, and assume $\theta = 0$ from now on. Note that $[\mathrm{tr}(A^\top A)^k]^{1/(2k)} = \|A\|_{2k}$ is the Schatten $2k$-norm of A. By the triangle inequality for the Schatten norm, $\|S_\pi\|_{2k} \le \sum_{i=1}^r \theta_i |(u_i v_i^\top)_\pi|_{2k}$. Hence,

$$\|S_\pi\|_{2k}^{2k} \le \left[\sum_{i=1}^r \theta_i \|(u_i v_i^\top)_\pi\|_{2k}\right]^{2k},$$

therefore

$$[\mathbb{E}\,\mathrm{tr}(S_\pi^\top S_\pi)^{2k}]^{1/(2k)} \le \left[\mathbb{E}(\sum_{i=1}^r \theta_i D_i^{1/(2k)})^{2k}\right]^{1/(2k)},$$

where $D_i = \mathbb{E}\,\mathrm{tr}[(u_i v_i^\top)_\pi^\top (u_i v_i^\top)_\pi]^{2k}$. Next, by the triangle inequality for the ℓ_{2k} norm $X \to [\mathbb{E}\|X\|^{2k}]^{1/(2k)}$,

$$[\mathbb{E}(\sum_{i=1}^r \theta_i D_i^{1/(2k)})^{2k}]^{1/(2k)} \le \sum_{i=1}^r \theta_i [\mathbb{E}D_i]^{1/(2k)}.$$

Let us focus on bounding one such term $\mathbb{E}D_i$, and denote $u = u_i$, $v = v_i$ for simplicity. Let us write $A = (uv^\top)_\pi$.

The simplest calculation is the first moment bound $\|A\|^2 \le \mathrm{tr}(A^\top A)$. However, this is not effective, as $\mathrm{tr}(A^\top A) = \sum_{ij}[\pi(uv^\top)_{ij}]^2 = \sum_{ij}[(uv^\top)_{ij}]^2 = \|uv^\top\|^2 = |u|_2^2|v|_2^2 = 1$, because the permutation of the entries does not change the sum of squares.

8.2.1 Second moment

We turn to the second simplest calculation, the second moment bound. This starts with the identity

$$\operatorname{tr}(A^\top A)^2 = \operatorname{tr} A^\top A A^\top A = \sum_{ijkl} A_{ij} A_{jk}^\top A_{kl} A_{li}^\top = \sum_{ijkl} A_{ij} A_{il} A_{kj} A_{kl}.$$

We have $A_{ab} = u_{\pi_b(a)} v_b$. The vectors u, v are fixed, while π_b are random and independent across b. Thus, if $j \neq l$, then $A_{\cdot j}$ and $A_{\cdot l}$ are independent. Moreover, the joint distribution of (A_{ij}, A_{kj}), for $i \neq k$, is equal to that of $v_j \cdot (u_{\tau_1}, u_{\tau_2})$, where τ is a permutation chosen uniformly at random. Thus,

$$(u_{\tau_1}, u_{\tau_2}) \sim Unif\{(u_i, u_j) : i \neq j\}.$$

With this observation, we can make the following moment calculations:

1. $\mathbb{E} A_{ij} = v_j \cdot \mathbb{E} u_{\tau_1} = v_j \cdot \sum_i u_i / n = 0$.
2. $\mathbb{E} A_{ij}^2 = v_j^2 \cdot \mathbb{E} u_{\tau_1}^2 = v_j^2 \cdot \sum_i u_i^2 / n = v_j^2 / n$.
3. $\mathbb{E} A_{ij} A_{kj} = v_j^2 \cdot \mathbb{E} u_{\tau_1} u_{\tau_2} = v_j^2 \cdot \sum_{i \neq j} u_i u_j / [n(n-1)] = v_j^2 \cdot [(\sum u_i)^2 - 1]/[n(n-1)] = -v_j^2/[n(n-1)]$.

Therefore, we conclude that

$$\begin{aligned}
\sum_{ijkl} \mathbb{E} A_{ij} A_{kj} A_{il} A_{kl} &= \sum_{ik, j \neq l} \mathbb{E} A_{ij} A_{kj} \cdot \mathbb{E} A_{il} A_{kl} + \sum_{ik, j} \mathbb{E} (A_{ij} A_{kj})^2 \\
&= n(n-1) \sum_{j \neq l} \mathbb{E} A_{1j} A_{2j} \cdot \mathbb{E} A_{1l} A_{2l} + n \sum_{j \neq l} \mathbb{E} A_{1j}^2 \cdot \mathbb{E} A_{1l}^2 \\
&\quad + \sum_j (\mathbb{E} \sum_i A_{ij}^2)^2 \\
&= n(n-1) \cdot I + n \cdot II + III.
\end{aligned}$$

Then, we have the following bounds for I, II, and III:

1. $$\begin{aligned} I &= \sum_{j \neq l} \mathbb{E} A_{1j} A_{2j} \cdot \mathbb{E} A_{1l} A_{2l} = \sum_{j \neq l} -v_j^2/[n(n-1)] \cdot (-v_l^2/[n(n-1)]) \\ &= 1/[n(n-1)]^2 \sum_{j \neq l} v_j^2 v_l^2 = (1 - \sum_j v_j^4)/[n(n-1)]^2 \leq 1/[n(n-1)]^2 \end{aligned}$$

   Note that $I \geq 0$, so $|I| \leq 1/[n(n-1)]^2$
2. $II = \sum_{j \neq l} \mathbb{E} A_{1j}^2 \cdot \mathbb{E} A_{1l}^2 = 1/n^2 \sum_{j \neq l} v_j^2 v_l^2 = (1 - \sum_j v_j^4)/n^2 \leq 1/n^2$
3. $III = \sum_j (\mathbb{E} \sum_i A_{ij}^2)^2 = \sum_j (n \mathbb{E} A_{1j}^2)^2 = \sum_j v_j^4$
   Above we used $\mathbb{E} \sum_i A_{ij}^2 = n \mathbb{E} A_{1j}^2 = v_j^2$.

Combining the bounds for I, II, and III, we get the upper bound

$$\mathbb{E} \operatorname{tr}(A^\top A)^2 \leq 1/[n(n-1)] + 1/n + \sum_j v_j^4 = 1/(n-1) + |v|_4^4$$

In conclusion $\mathbb{E} D_i \leq 2/n + |v_i|_4^4$, and $[\mathbb{E} \operatorname{tr}(S_\pi^\top S_\pi)^4]^{1/4} \leq \sum_{i=1}^r \theta_i [1/(n-1) + |v_i|_4^4]^{1/4}$. This finishes the second moment bound.

The overall rate at which $\operatorname{tr}(S_\pi^\top S_\pi)^2$ decays is $1/n$. For a.s convergence, we need the bounds to be a summable sequence; thus one can not to prove a.s convergence using only a second moment argument. This motivates us to look at the third moment.

8.2.2 Third moment

The third moment bounds proceeds similarly. We start with the identity

$$\operatorname{tr}(A^\top A)^3 = \operatorname{tr} A^\top A A^\top A A^\top A = \sum_{ijklmq} A_{ij}A_{kj}A_{kl}A_{ml}A_{mq}A_{iq}.$$

In this expression, the random variables with this the same second index (j, l or q) are dependent, thus there are at most three groups of independent random variables. There are three cases, depending on how many distinct indices there are among j, l or q:

Three distinct indices: $j \neq l \neq q$. In this case, we can write the sum over all $j \neq l \neq q$ as $A_1 = \sum_{ijklmq} \mathbb{E}A_{ij}A_{kj} \cdot \mathbb{E}A_{kl}A_{ml} \cdot \mathbb{E}A_{mq}A_{iq}$. We already calculated that $\mathbb{E}A_{ij}A_{kj} = v_j^2/n \cdot [\tau + \delta_{ik}\eta]$, where $\tau = -1/(n-1)$, $\eta = 1 - \tau = n/(n-1)$. Thus,

$$A_1 = n^{-3} \sum_{j\neq l\neq q} v_j^2 v_l^2 v_q^2 \cdot \sum_{ikm} [\tau + \delta_{ik}\eta] \cdot [\tau + \delta_{km}\eta] \cdot [\tau + \delta_{mi}\eta].$$

Now, we need to evaluate

$$A_2 = \sum_{ikm} [\tau + \delta_{ik}\eta] \cdot [\tau + \delta_{km}\eta] \cdot [\tau + \delta_{mi}\eta].$$

For the formal calculation, we can factor out τ^3, even though τ may be 0, and so this may technically not be allowed. However, the formal calculation still leads to the correct answer. If $\tau = 0$, the result is $A_2 = n\eta^3$, which agrees with what we get below. Let thus $\zeta = \eta/\tau$, and we want

$$\begin{aligned}
\tau^{-3}A_2 &= \sum_{ikm} [1 + \delta_{ik}\zeta] \cdot [1 + \delta_{km}\zeta] \cdot [1 + \delta_{mi}\zeta] \\
&= \sum_{ik} [1 + \delta_{ik}\zeta] \cdot \sum_{m} [1 + \delta_{km}\zeta] \cdot [1 + \delta_{mi}\zeta] \\
&= \sum_{ik} [1 + \delta_{ik}\zeta] \cdot [n + 2\zeta + \delta_{ki}\zeta^2] \\
&= [n + 2\zeta] \sum_{ik} [1 + \delta_{ik}\zeta] + \zeta^2 \sum_{ik} [1 + \delta_{ik}\zeta]\delta_{ik} \\
&= [n + 2\zeta] \cdot [n^2 + n\zeta] + n\zeta^2[1 + \zeta] \\
&= n^3 + 3n^2\zeta + 3n\zeta^2 + n\zeta^3.
\end{aligned}$$

Above we used that $\sum_m [1 + \delta_{km}\zeta] \cdot [1 + \delta_{mi}\zeta] = n + 2\zeta + \delta_{ki}\zeta^2$. Hence

$$A_2 = n^3 \cdot \tau^3 + 3n^2 \cdot \tau^2\eta + 3n \cdot \tau\eta^2 + n \cdot \eta^3 = (n\tau + \eta)^3 + (n-1)\eta^3.$$

However, we also have $n\tau + \eta = n\tau + 1 - \tau = (n-1)\tau + 1 = 0$, and $(n-1)\eta^3 = n^3/(n-1)^2$. Therefore, we conclude that $A_2 = n^3/(n-1)^2$. Going back to the definition of A_1, we thus see $A_1 = 1/(n-1)^2 \cdot \sum_{j\neq l\neq q} v_j^2 v_l^2 v_q^2$.

Now $\sum_{j\neq l\neq q} v_j^2 v_l^2 v_q^2 \leq \sum_{jlq} v_j^2 v_l^2 v_q^2 = |v|_2^6 = 1$, so we conclude that $A_1 \leq 1/(n-1)^2$.

Two distinct indices: $j = l \neq q$ **and the other two symmetric cases.** In this case, we can write the sum as $B_1 = \sum_{j\neq q, ikm} \mathbb{E}A_{ij}A_{kj}^2A_{mj} \cdot \mathbb{E}A_{mq}A_{iq}$. Now, it is easy to see that the v_j terms contribute a factor of at most $(\sum_j v_j^4)(\sum_j v_j^2) = \sum_j v_j^4$. In the remainder, it is enough to work with the u-part. This equals $B_2 = \sum_{ikm} \mathbb{E}\tau_i\tau_k^2\tau_m \cdot \mathbb{E}\tau_i\tau_m$, where τ is a random permutation of the set of values $u_1, \ldots, u_n$. Now we can sum over k first to get

$$B_2 = \sum_{im} \mathbb{E}\tau_i\tau_m \cdot \sum_k \mathbb{E}\tau_i\tau_k^2\tau_m = \sum_{im} \mathbb{E}\tau_i\tau_m \cdot \mathbb{E}[\tau_i\tau_m(\sum_k \tau_k^2)]$$

However, $\sum_k \tau_k^2 = \sum_k u_k^2 = 1$ is a deterministic quantity, so we obtain $B_2 = \sum_{im} (\mathbb{E}\tau_i\tau_m)^2$. Now, recall that $\mathbb{E}\tau_i\tau_k = 1/n \cdot [\tau + \delta_{ik}\eta]$, where $\tau = -1/(n-1)$, $\eta = 1 - \tau$. Therefore,

$$n^2 B_2 = n(n-1) \cdot 1/(n-1)^2 + n \cdot n^2/(n-1)^2.$$

So $B_2 \le 3/n$ and $B_1 \le 3n^{-1}|v|_4^4$

One unique index: $j = l = q$**.** In this case, we can write the sum as

$$C_1 = \sum_{j,ikm} \mathbb{E}A_{ij}^2 A_{kj}^2 A_{mj}^2 = \sum_j (\sum_i \mathbb{E}A_{ij}^2)^3 = \sum_j (v_j^2)^3 = \sum_j v_j^6.$$

Putting together the results from the three cases, we obtain

$$\mathbb{E}\,\mathrm{tr}(A^\top A)^3 \le 1/(n-1)^2 + 9n^{-1}|v|_4^4 + |v|_6^6.$$

This finishes the proof.

8.2.3 Fourth moment

The fourth moment bounds proceeds similarly, except the calculation is more complicated. We start with the identity

$$\mathrm{tr}(A^\top A)^4 = \sum_{i_1 i_2 i_3 i_4 j_1 j_2 j_3 j_4} A_{i_1 j_1} A_{i_2 j_1} A_{i_2 j_2} A_{i_3 j_2} A_{i_3 j_3} A_{i_4 j_3} A_{i_4 j_4} A_{i_1 j_4}.$$

As before, in this expression, the random variables with this the same second index ($j.$) are dependent, thus there are at most four groups of independent random variables. There are now four cases, depending on how many distinct indices there are among them.

Four distinct indices. In this case, we can write the sum over all j_s as

$$A_1 = \sum_{i_1 i_2 i_3 i_4 j_1 j_2 j_3 j_4} \mathbb{E}A_{i_1 j_1} A_{i_2 j_1} \mathbb{E}A_{i_2 j_2} A_{i_3 j_2} \mathbb{E}A_{i_3 j_3} A_{i_4 j_3} \mathbb{E}A_{i_4 j_4} A_{i_1 j_4}.$$

We already calculated that $\mathbb{E}A_{ij}A_{kj} = v_j^2/n \cdot [\tau + \delta_{ik}\eta]$, where $\tau = -1/(n-1)$, $\eta = 1 - \tau = n/(n-1)$. Thus, denoting $I = (i_1, i_2, i_3, i_4)$, $J = (j_1, j_2, j_3, j_4)$, $\tilde{v}_J = v_{j_1} v_{j_2} v_{j_3} v_{j_4}$, $i \in I$ summation over all i_s, and $j \in J$ summation over distinct j_s: $A_1 = n^{-4} \sum_{J \in S_J} \tilde{v}_J \cdot A_2(J)$, where, with $\zeta = \eta/\tau$,

$$\tau^{-3} A_2(J) = \sum_{I \in S_I} \prod_{l=1}^4 [1 + \delta_{i_l i_{l+1}} \zeta].$$

This equals

$$\begin{aligned}
&\sum_{i_1, i_3} \sum_{i_2} [1 + \delta_{i_1 i_2}\zeta][1 + \delta_{i_3 i_2}\zeta] \cdot \sum_{i_4} [1 + \delta_{i_1 i_4}\zeta][1 + \delta_{i_3 i_4}\zeta] \\
&= \sum_{i_1, i_3} \left(\sum_a [1 + \delta_{i_1 a}\zeta][1 + \delta_{i_3 a}\zeta] \right)^2 \\
&= \sum_{ik} [n + 2\zeta + \delta_{ik}\zeta^2]^2 \\
&= n^4 + 4n^3\zeta + 6n^2\zeta^2 + 4n\zeta^3 + n\zeta^4.
\end{aligned}$$

Here we used identities established in the previous section. This also equals $(n+\zeta)^4 + (n-1)\zeta^4$. Hence $A_2 = (n\tau + \eta)^4 + (n-1)\eta^3 = n^4/(n-1)^3$. Therefore, $A_1 = 1/(n-1)^3 \sum_{J \in S_J} \tilde{v}_J \le 1/(n-1)^3$.

Three distinct indices: $j_1 = j_2$, other j_s different, and the other three symmetric cases. In this case, we can write the sum as

$$B_1 = \sum_{i_1 i_2 i_3 i_4; j_1 j_3 j_4} \mathbb{E} A_{i_1 j_1} A_{i_2 j_1}^2 A_{i_3 j_1} \cdot \mathbb{E} A_{i_3 j_3} A_{i_4 j_3} \mathbb{E} A_{i_4 j_4} A_{i_1 j_4}.$$

As before, it is easy to see that the v_j terms contribute a factor of at most $(\sum_j v_j^4)(\sum_j v_j^2)^2 = \sum_j v_j^4$. In the remainder, it is enough to work with the u-part. This equals

$$B_2 = \sum_{ikml} \mathbb{E} \tau_i \tau_k^2 \tau_m \cdot \mathbb{E} \tau_m \tau_l \cdot \mathbb{E} \tau_l \tau_i = \sum_{iml} \mathbb{E} \tau_i \tau_m \cdot \mathbb{E} \tau_m \tau_l \cdot \mathbb{E} \tau_l \tau_i,$$

using that $\sum_k \tau_k^2 = 1$. However, this expression is precisely the one that came up in the calculation of the third moment bound for three distinct indices. There we saw that it equals $1/(n-1)^2$. Hence, $B_1 \leq |v|_4^4/(n-1)^2$, and the overall contribution of the terms with three distinct indices is four times this.

Two distinct indices: $j_1 = j_2 \neq j_3 = j_4$, and the other three symmetric cases. Here we need

$$B_1 = \sum_{i_1 i_2 i_3 i_4; j_1 j_3} \mathbb{E} A_{i_1 j_1} A_{i_2 j_1}^2 A_{i_3 j_1} \cdot \mathbb{E} A_{i_3 j_3} A_{i_4 j_3}^2 A_{i_1 j_3}.$$

The v_j terms contribute a factor of at most $(\sum_j v_j^4)^2$. The u-part contributes

$$B_2 = \sum_{ikml} \mathbb{E} \tau_i \tau_k^2 \tau_m \cdot \mathbb{E} \tau_m \tau_l^2 \tau_i = \sum_{im} (\mathbb{E} \tau_i \tau_m)^2,$$

using that $\sum_k \tau_k^2 = 1$. In the calculation of the third moment bound for two distinct indices we saw that $B_2 \leq 3/n$. Hence, $B_1 \leq 3|v|_4^8/n$. The overall contribution is four times this.

Two distinct indices: $j_1 = j_2 = j_3 \neq j_4$, and the other three symmetric cases. In this case we need

$$B_1 = \sum_{i_1 i_2 i_3 i_4; j_1 j_4} \mathbb{E} A_{i_1 j_1} A_{i_2 j_1}^2 A_{i_3 j_1}^2 A_{i_4 j_1} \cdot \mathbb{E} A_{i_4 j_4} A_{i_1 j_4}.$$

The v_j terms contribute a factor of at most $(\sum_j v_j^6)(\sum_j v_j^2) = \sum_j v_j^6$. The u-part contributes

$$B_2 = \sum_{ikml} \mathbb{E} \tau_i \tau_k^2 \tau_m^2 \tau_l \cdot \mathbb{E} \tau_l \tau_i = \sum_{im} (\mathbb{E} \tau_i \tau_m)^2,$$

using that $\sum_k \tau_k^2 = 1$. In the third moment bound for two distinct indices we saw that $B_2 \leq 3/n$. Hence, $B_1 \leq 3|v|_6^6/n$, and the overall bound is four times this.

One distinct index: $j_1 = j_2 = j_3 = j_4$, and the other three symmetric cases. In this case, we can write the sum as

$$B_1 = \sum_{i_1 i_2 i_3 i_4; j_1 j_4} \mathbb{E} A_{i_1 j_1}^2 A_{i_2 j_1}^2 A_{i_3 j_1}^2 A_{i_4 j_1}^2.$$

The v_j contribute a factor of at most $\sum_j v_j^8$, while the u-part contributes $\sum_{ikml} \mathbb{E} \tau_i^2 \tau_k^2 \tau_m^2 \tau_l^2 = 1$, using that $\sum_k \tau_k^2 = 1$. Hence, $B_1 \leq |v|_8^8$.

In conclusion we obtain the desired bound

$$\mathbb{E} \operatorname{tr}(A^\top A)^4 \leq 1/(n-1)^3 + 4/(n-1)^2 |v|_4^4 + 12 n^{-1} [|v|_4^8 + |v|_6^6] + |v|_8^8.$$

8.3 Proof of Prop. 4.2

The first part essentially follows from (Bai and Silverstein, 2009, Cor 6.6). A small modification is needed to deal with the non-iid-ness, as explained in Dobriban et al. (2017).

For the second part, we will show that $|[\operatorname{tr} T]^{-1/2} X T^{1/2}| \to 1$. For this, it suffices to show that $|[\operatorname{tr} T]^{-1} X T X^\top - I_n| \to 0$ in probability (or a.s.). For the convergence in probability, it is in turn enough to show that $\mathbb{E} \operatorname{tr}[X\Sigma X^\top - I_n]^2 \to 0$, where $\Sigma = [\operatorname{tr} T]^{-1} T$. We calculate

$$A = \mathbb{E} \operatorname{tr}[X\Sigma X^\top - I_n]^2 = \mathbb{E} \operatorname{tr}[X\Sigma X^\top]^2 - 2\mathbb{E} \operatorname{tr} X\Sigma X^\top + n.$$

Now $X\Sigma X^\top = \sum_{j=1}^p \sigma_j x_j x_j^\top$, where the x_j are independent $n \times 1$ random vectors whose entries are iid random variables (whose distribution may depend on j). They collect the j-th coordinates of the observed data. So, $\mathbb{E} \operatorname{tr} X\Sigma X^\top = \sum_{j=1}^p \sigma_j \mathbb{E}|x_j|^2 = n\sum_{j=1}^p \sigma_j = n$. Also,

$$\mathbb{E} \operatorname{tr}[X\Sigma X^\top]^2 = \mathbb{E} \operatorname{tr}[\sum_{j=1}^p \sigma_j x_j x_j^\top][\sum_{k=1}^p \sigma_k x_k x_k^\top] = \sum_{j,k=1}^p \sigma_j \sigma_k \mathbb{E}[x_j^\top x_k]^2.$$

To evaluate this expression, we need to find $\mathbb{E}[x_j^\top x_k]^2$. If $j \neq k$, then x_j and x_k are independent, and we can take expectation over j first, to get $\mathbb{E}[x_j^\top x_k]^2 = \mathbb{E} \operatorname{tr}[x_j x_j^\top x_k x_k^\top] = \mathbb{E} \operatorname{tr}[x_k x_k^\top] = n$. This leads to

$$\mathbb{E} \operatorname{tr}[X\Sigma X^\top]^2 = n \sum_{j,k=1}^p \sigma_j \sigma_k + \sum_{j=1}^p \sigma_j^2 [\mathbb{E}|x_j|^4 - n].$$

Therefore we find that $A = \sum_{j=1}^p \sigma_j^2 [\mathbb{E}|x_j|^4 - n]$. Thus, we need to show that $A \to 0$. Since $\sigma_j \leq Cp^{-1}$ for all j,

$$A = p^{-2} \sum_{i=1}^n \sum_{j=1}^p (\mathbb{E} x_{ij}^4 - 1) \leq p^{-2} \cdot Cnp = Cn/p \to 0$$

if $n/p \to 0$, since the 4-th moments are bounded. This shows that $n/p \to 0$ guarantees convergence in probability, and finishes the proof of (2A). If in addition $n/p \leq 1/n^{1+\varepsilon}$, then by the Borel-Cantelli lemma we conclude that $|[\operatorname{tr} T]^{-1} X T X^\top - I_n| \to 0$ a.s., as needed. This finishes the proof of (2B). Therefore, the proof of the proposition is complete.

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley New York, 2003.

J. Bai and K. Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465, 2012.

J. Bai and S. Ng. *Large dimensional factor analysis.* Now Publishers Inc, 2008.

Z. Bai and X. Ding. Estimation of spiked eigenvalues in spiked models. *Random Matrices: Theory and Applications*, 1(02):1150011, 2012.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices.* Springer Series in Statistics. Springer, 2009.

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

M. S. Bartlett. Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, 3(2):77–85, 1950.

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.

P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer, 2009.

T. A. Brown. *Confirmatory factor analysis for applied research*. Guilford Publications, 2014.

A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27 (4):509–540, 1992.

R. B. Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2): 245–276, 1966.

G. A. Churchill Jr. A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, pages 64–73, 1979.

A. B. Costello and J. W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7):1–9, 2005.

A. V. Crawford, S. B. Green, R. Levy, W.-J. Lo, L. Scott, D. Svetina, and M. S. Thompson. Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, 70(6):885–901, 2010.

E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.

E. Dobriban and A. B. Owen. Deterministic parallel analysis: An improved method for selecting the number of factors and principal components. *arXiv preprint arXiv:1711.04155. To appear in JRSS-B*, 2017.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *arXiv preprint arXiv:1507.03003, to appear in the Annals of Statistics*, 2015.

E. Dobriban, W. Leeb, and A. Singer. Optimal prediction in the linearly transformed spiked model. *arXiv preprint arXiv:1709.03393*, 2017.

L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

J. Fan, Y. Liao, and M. Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39(6):3320, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics, 2009.

C. J. Gaskin and B. Happell. On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International journal of nursing studies*, 51(3):511–521, 2014.

D. Gerard and M. Stephens. Unifying and generalizing methods for removing unwanted variation based on negative controls. *arXiv preprint arXiv:1705.08393*, 2017.

L. W. Glorfeld. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, 55(3):377–393, 1995.

C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. *Movement disorders*, 23(15):2129–2170, 2008.

S. B. Green, R. Levy, M. S. Thompson, M. Lu, and W.-J. Lo. A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, 72(3):357–374, 2012.

J. C. Hayton, D. G. Allen, and V. Scarpello. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7(2):191–205, 2004.

J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

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

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

J. Josse and F. Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6): 1869–1879, 2012.

H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151, 1960.

S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

J. T. Leek and J. D. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

Z. Lin, C. Yang, Y. Zhu, J. Duchi, Y. Fu, Y. Wang, B. Jiang, M. Zamanighomi, X. Xu, M. Li, et al. Simultaneous dimension reduction and adjustment for confounding variation. *Proceedings of the National Academy of Sciences*, 113(51):14662–14667, 2016.

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.

B. Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

A. Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.

A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

A. Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258, 2012.

A. Onatski, M. J. Moreira, and M. Hallin. Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231, 2013.

A. B. Owen and J. Wang. Bi-cross-validation for factor analysis. *Statistical Science*, 31(1): 119–139, 2016.

A. Parasuraman, V. A. Zeithaml, and L. L. Berry. Servqual: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of retailing*, 64(1):12, 1988.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

P. R. Peres-Neto, D. A. Jackson, and K. M. Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997, 2005.

G. Raiche, D. Magis, and M. G. Raiche. Package nfactors. 2010.

C. Spearman. "General Intelligence", objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

D. W. Stewart. The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, pages 51–62, 1981.

L. L. Thurstone. *Multiple-factor analysis*. University of Chicago Press, 1947.

W. F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327, 1976.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

Y.-H. Zhou, J. Marron, and F. A. Wright. Eigenvalue significance testing for genetic association. *Biometrics*, 2017.

W. R. Zwick and W. F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432, 1986.