The non-parametric bootstrap and spectral analysis in moderate and high-dimension

Noureddine El Karoui
UC Berkeley, Department of Statistics and Criteo AI Lab

Elizabeth Purdom
UC Berkeley, Department of Statistics

Abstract

We consider the properties of the bootstrap as a tool for inference concerning the eigenvalues of a sample covariance matrix computed from an $n \times p$ data matrix X. We focus on the modern framework where p/n is not close to 0 but remains bounded as n and p tend to infinity. Through a mix of numerical and theoretical considerations, we show that the non-parametric bootstrap is not in general a reliable inferential tool in the setting we consider. However, in the case where the population covariance matrix is well-approximated by a finite rank matrix, the non-parametric bootstrap performs as it does in finite dimension.

1 Introduction

The bootstrap (15) is a central tool of applied statistics, enabling inference by assessing the variability of the statistics of interest directly from the data and without explicit appeal to asymptotic theory. The appeal of the bootstrap is especially great when asymptotic theoretical derivations are difficult and/or can be done only under quite restrictive assumptions. For instance, consider the case of Principal Components Analysis (PCA). The classic text of (2; 3) (Chapter 13) gives limit theory for the eigenvalues and eigenvectors of the sample covariance matrix when the data is drawn from a normal population. These limit results are non-trivial to derive, even in the Gaussian case, and depend, for instance, on assumptions regarding the multiplicity of the eigenvalues of the population covariance matrix. Furthermore, it is clear, using approximation arguments from (27), that these limit results are not valid for a broad class of distributions. For instance, they do not apply to populations distributions with kurtosis not equal to 3. The modern theory of PCA which aims for better finite-sample approximations by relaxing the assumption that $p/n \to 0$ is much more difficult technically and relies on very strong assumptions about the geometry of the dataset (see (26; 25; 17), follow-up papers, and Supplementary Section S1 below for a short summary).

Remarkably, from a theoretical standpoint, it has been shown that in many situations the bootstrap estimates the distribution of the statistics of interest accurately, at least with sufficient sample sizes (see (8; 22) for classic references). For the specific example of estimating the eigenvalues of the sample covariance matrix and PCA, numerous papers have been written about the properties of the bootstrap (6; 1; 14; 13; 23). The main results of these papers is that the bootstrap works in an asymptotic regime that assumes that the sample size grows to infinity while the dimension of the data is fixed or grows very slowly, with the additional provision that the population covariance has eigenvalues of multiplicity one. When the assumption of multiplicity equal to one does not hold, subsampling techniques (34) can be used to correctly estimate the distributions of interest by resampling. We note however that these subsampling techniques also require the statistician to have subsamples of size that is infinitely large compared to the dimension of the data.

Given the limitations of existing asymptotic theory and these theoretical results on bootstrapping of eigenvalues, it is not surprising that the bootstrap is a natural tool to use in connection with PCA and inferential questions therein. The bootstrap is mostly used in this context to assess variability of eigenvalues, for instance to come up with principled cutoff selections in PCA and related methods such as factor analysis. For recent examples of an applied nature, we refer the reader to (38; 19; 37; 4). Another application of the bootstrap is of course in bagging (9); a well known instance of bagging related to high-dimensional covariance estimation is in resampled portfolio selection (31).

Our framework: p/n not close to zero The theoretical assumptions that support the use of the bootstrap make the fundamental assumption that the dimension p is much smaller than n (i.e. $p/n \to 0$). The modern asymptotic the-

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

ory of PCA, referenced above, has shown that relaxing that assumption – for example by assuming that $p/n < 1$ but does not tend to 0 – leads to dramatically different theoretical behavior of the eigenvalues and eigenvectors. It is therefore natural to consider the performance of the bootstrap in the situation where p/n is not close to 0. Furthermore, in statistical practice p/n is rarely very close to 0, and hence classical approximations, which rely heavily on that assumption, may lead to theoretical results and interpretations that differ quite drastically from what is observed by practitioners. However, when p/n is not close to 0, developing theoretical results for interesting statistical questions is still quite technically difficult (30; 40; 35; 26; 16; 29), which makes the bootstrap particularly compelling. These considerations motivate our exploration of the bootstrap as an alternative, data-driven way, to perform inferential tasks for spectral properties of large covariance matrices.

Contributions of the paper In Section 2, we study the performance of the bootstrap by simulations in the context of PCA. We assess whether the bootstrap recovers the sampling distribution of various statistics of interest, for instance the largest eigenvalue of a covariance matrix or its bias. Most of our results are negative: only when the largest eigenvalues become quite large compared to the rest does the bootstrap provide accurate inference. Furthermore, the behavior of the bootstrap is unpredictable; in two setups that are nearly similar from a population standpoint, the bootstrap estimate is in one case biased upward and in the other downward.

In Section 3 we provide theoretical results that help explain this behavior. The first results are about the behavior of the bootstrapped empirical distribution of *all* the eigenvalues of the sample covariance matrix $\widehat{\Sigma}$. We show that in the framework we consider, the bootstrapped empirical distribution is biased and asymptotically non-random. We then consider the bootstrap behavior of only the largest eigenvalues of $\widehat{\Sigma}$. We show that when the population covariance Σ has some very large eigenvalues – far separated from the other eigenvalues – the bootstrap distribution of those large eigenvalues does correctly approximate the sampling distribution of the large eigenvalues of $\widehat{\Sigma}$.

The results of this paper confirm that the bootstrap works when the problem is very low-dimensional or can be approximated by a very low-dimensional problem, but is untrustworthy when the problem is genuinely high-dimensional. As such, the current paper complements the findings of (18), which was concerned with problems of the bootstrap for linear regression models in high dimensions.

1.1 Notations and default conventions

If X is an $n \times p$ data matrix, we define $\tilde{X} \triangleq (X - \bar{X})$ and we call $\widehat{\Sigma}$ its associated covariance matrix, i.e. $\widehat{\Sigma} = \frac{1}{n-1}\tilde{X}'\tilde{X}$. We call the *empirical spectral distribution* of a $p \times p$ symmetric matrix M the probability measure such that $dF_p(x) = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(M)}$, where $\lambda_1(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_p(M)$ are the ordered eigenvalues of M. We use the notations $\lambda_{\max}(M)$ or $\lambda_1(M)$ for the largest eigenvalue of the symmetric matrix M. $\lambda_{\min}$ is of course the smallest. For $z \in \mathbb{C}^+$, i.e $z = u + iv$, where $v > 0$, we call $m_p(z)$ the *Stieltjes transform* of the distribution F_p, i.e

$$m_p(z) = \int \frac{1}{x - z} dF_p(x) = \frac{1}{p}\text{trace}\left((\widehat{\Sigma} - z\text{Id}_p)^{-1}\right) .$$

We call the *Gaussian phase transition* the value $1 + \sqrt{p/n}$, which is the value for λ_1 at which point the distribution of $\lambda_1(\widehat{\Sigma}) \triangleq \hat{\lambda}_1$ switches from Gaussian to Tracy-Widom fluctuations for Gaussian designs whose covariance is a rank-1 perturbation of identity (see Supplementary Section S1.2.1 for review of these results). The rows of a Gaussian design matrix are i.i.d Gaussian.

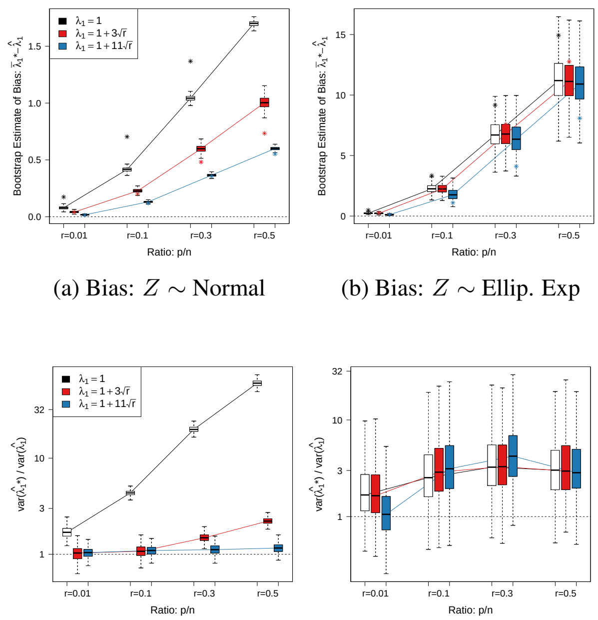
We use the notation $\Longrightarrow$ to denote weak convergence of probability distributions. $|||M|||_2$ is the operator norm of M, i.e its largest singular value. $\|w\|_\infty = \max_{1\leq i\leq p}|w_i|$ is the ℓ_∞-norm of the vector w. We say that the sequence $u_n = \text{polyLog}(n)$ if u_n grows at most like a polynomial in $\log(n)$. In doing asymptotic analysis, we work under the assumptions that $p/n \to r$, $r \in (0, \infty)$. We use the notation o_P to denote a "little-oh" with respect to the data generating distribution. We use $o_{P,w}$ to denote a "little-oh" with respect to the data generating distribution and the bootstrap weights. See (39) Section 2.2 for standard definitions. We use Mult as a shorcut for Multinomial.

2 Simulation Study

We investigate via simulation the behavior of the bootstrap for the top eigenvalue of the standard sample covariance matrix, $\widehat{\Sigma}$. For simplicity, we consider the case where only the top (population) eigenvalue $\lambda_1(\Sigma)$ is allowed to vary and assume the remaining eigenvalues are all equal to 1. The inferential question is to determine whether the top eigenvalue λ_1 differs from 1. In what follows, the data X_i come from either a multivariate normal distribution or an elliptical distribution with exponential weights (see Supplementary Text, Section S2 for details). Many theoretical results on random matrices in high dimensions do not yet extend to the case of elliptical distributions. While still an idealization, it represents a simulation scenario that is more realistic and informative than a straight Gaussian design.

Estimating Bias The eigenvalues of the sample covariance are biased in high dimensions for estimating the true (a.k.a. population) eigenvalues (see Supplementary Text, Section S1 for review of these results), and we consider the question of estimating this bias via the bootstrap. We first note the importance of this bias in understanding the behavior of $\hat{\lambda}_1$. The bias can be substantial unless λ_1 is quite large, and the bias is clearly evident even for low ratios of

$r = p/n$ if λ_1 is close to 1. Furthermore, this bias is more pronounced for elliptical distributions than the normal distribution ((16; 33)). For example, for the null setting of $\lambda_1 = 1$, with a ratio of p/n as low as 0.01, we see a bias in $\hat{\lambda}_1$, overestimating the true λ_1 by 17% for the normal distribution, and 49% for the elliptical distribution with exponential weights (Supplemental Table S1-S4). As the ratio of p/n grows, the bias increases, with $\hat{\lambda}_1$ overestimating λ_1 by 1.88 and 14.93 when $\lambda_1 = 1$ for the normal distribution and the elliptical distribution with exponential weight, respectively when $p/n = 0.5$. The bias declines as λ_1 grows and becomes more separated from the remaining eigenvalues, especially relative to the size of λ_1; Subsection 3.2 provides a theoretical explanation for this phenomenon. But the bias remains for large ratios of p/n even for λ_1 well beyond the Gaussian phase transition $(1 + \sqrt{p/n})$, especially for non-normal distributions; for $p/n = 0.5$, when λ_1 is as large as $1+11\sqrt{p/n} \approx 8.78$, the bias of $\hat{\lambda}_1$ is 8.08 when X follows an elliptical distribution with exponential weights, and 2.01 for an elliptical distribution with normal weights. Any use of $\hat{\lambda}_1$ as an estimate of λ_1 must grapple with the problem of such a highly non-consistent estimator, making bootstrap methods for estimating the bias highly relevant.



(a) Bias: $Z \sim$ Normal (b) Bias: $Z \sim$ Ellip. Exp

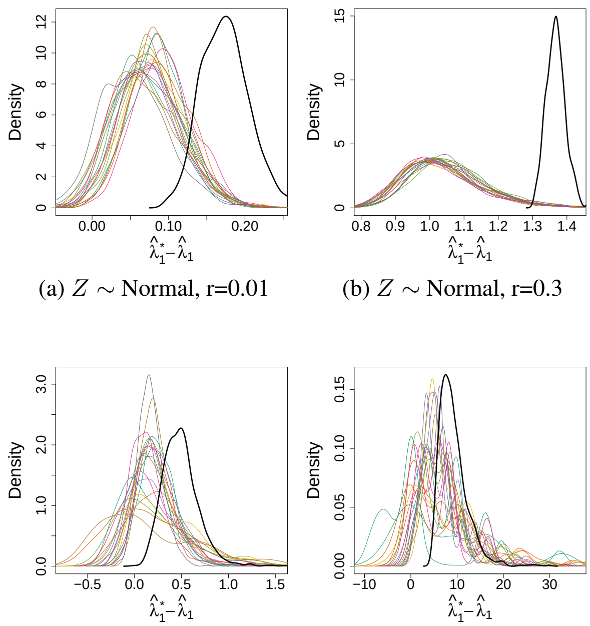
(c) Var: $Z \sim$ Normal (d) Var: $Z \sim$ Ellip. Exp

Figure 1: Boxplots of the bootstrap estimate of (a),(b) bias and (c),(d) variance of $\hat{\lambda}_1$ over 1000 simulations. The bootstrap estimate of variance is shown as a ratio of the true variance while the asterisk (*) in (a),(b) corresponds to the true bias. Each group of boxplots along the x-axis corresponds to a different ratio r of p/n; different colors of the boxplot correspond to different values of the true λ_1. See Supplementary Figures S3 and S4 and Tables S1-S4, and S5 for the median values of these boxplots and for results from larger λ_1 values.

The standard bootstrap estimate of bias is given by, $\bar{\lambda}_1^* - \hat{\lambda}_1$, where $\bar{\lambda}_1^* = \frac{1}{B}\sum_b \hat{\lambda}_1^{*b}$ is the mean of the bootstrap estimates of λ_1. Unfortunately, we see in simulations that this bootstrap estimate of bias is not a reliable estimate of the bias of $\hat{\lambda}_1$ unless λ_1 is quite large relative to the other eigenvalues (Figure 1). For X_i following a normal distribution, the bootstrap estimate of bias in our simulations remains poor for large p/n even for λ_1 past the phase transition, e.g. $\lambda_1 = 1 + 3\sqrt{p/n}$, and only for much larger values of λ_1 does the bootstrap estimate of bias start to approach the true bias in high dimensions (Supplementary Table S1). Further the bootstrap estimate of bias is inconsistent: depending on the true value of λ_1 the bootstrap either under- or over-estimates the bias. Another important feature of the bootstrap shown in our simulations is that when X_i follows an elliptical distribution with exponential weights, while the mean performance of the bootstrap estimate is still poor, it is the extremely high variance of the bootstrap that is even more problematic for the bootstrap estimate of bias.

Estimating the Variance We see similar problems in our simulations for the boostrap estimate of variance (Figure 1). Specifically, the bootstrap dramatically overestimates the variance of $\hat{\lambda}_1$ when λ_1 is close to 1. When the X_i's are normally distributed and $\lambda_1 = 1$, the bootstrap estimates the variance to be four times larger than the true variance for $p/n = 0.1$, and grows to be up to 60 times larger than the true variance when $p/n = 0.5$ (Supplementary Table S5). Even when $\lambda_1 = 1 + 3\sqrt{p/n}$, well beyond the Gaussian phase transition and hence in a relatively easy setup, the bootstrap estimate of variance is inflated to 1.5 to 2.2 times as large as the truth, for $p/n = 0.3$ and 0.5, respectively. Only for large values of λ_1 does the bootstrap inflation of the variance become minimal. Again, when the X_i's follow an elliptical distribution with exponential weights, the behavior of the bootstrap estimate of variance is dominated by the variability in the estimate.

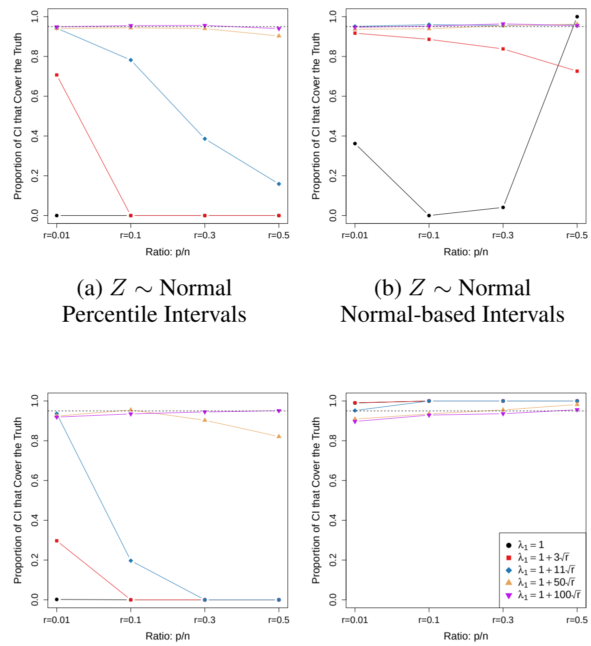
Confidence Intervals for λ_1 Standard techniques for creating confidence intervals for λ_1 are clearly problematic in high dimensions since $\hat{\lambda}_1$ is biased and not a consistent estimator for λ_1. Bootstrap confidence intervals can be created in multiple ways (10). Common techniques include 1) a simple normal confidence interval around $\hat{\lambda}_1$ using the bootstrap estimate of variance, 2) the percentile method, which uses the percentiles of the bootstrap distribution of $\hat{\lambda}_1^{*b}$, or 3) a bias-corrected confidence interval. Examining the actual bootstrap distributions of $\hat{\lambda}_1^* - \hat{\lambda}_1$ from multiple simulations (Figure 2), we see clearly the incorrect bias estimation and the overestimation of variance that results from using the bootstrap. These are features that will also invalidate confidence intervals constructed from the per-



(a) $Z \sim$ Normal, r=0.01 (b) $Z \sim$ Normal, r=0.3

(c) $Z \sim$ Ellip. Exp, r=0.01 (d) $Z \sim$ Ellip. Exp, r=0.3

Figure 2: Estimated bootstrap density of $\hat{\lambda}_1^{*b} - \hat{\lambda}_1$ from twenty simulations (picked at random from 1,000 simulations). The solid black line line represents the true distribution of $\hat{\lambda}_1 - \lambda_1$ (over 1,000 simulations). For larger value of λ_1, see Supplementary Figure S6.



(a) $Z \sim$ Normal Percentile Intervals (b) $Z \sim$ Normal Normal-based Intervals

(c) $Z \sim$ Ellip. Exp Percentile Intervals (d) $Z \sim$ Ellip. Exp Normal-based Intervals

Figure 3: Plotted are the percentage of the bootstrap confidence intervals that cover the true λ_1 (out of 1,000 simulations), for different values of $r = p/n$ and for different true values of λ_1. See Supplementary Tables S6 and S7 for more precise numerical values and Supplemental Figure S5 for additional distributions.

centiles of the distribution of $\hat{\lambda}_1^*$. We also see that when the X_i's follow an elliptical distribution with exponential weights, the bootstrap distributions do not appear to be converging to a limit for small values of λ_1, at least for $n = 1000$ – an even greater problem in using the bootstrap in these settings.

As expected, the resulting bootstrap confidence intervals are not useful in inference on the true value of λ_1. Bootstrap confidence intervals based on the percentile estimates do not cover the true value with any kind of reasonable probability until λ_1 becomes quite large due to the bias in the estimate of λ_1 (Figure 3 and Supplementary Table S6). Bootstrap confidence intervals based on normal intervals around $\hat{\lambda}_1$ using the bootstrap estimate of variance *do* cover the true value of the λ_1 with high probability. However, this coverage is due to the fact that the bootstrap estimate of variance is much larger than the true variance of $\hat{\lambda}_1$, as seen above, and thus results in overly large confidence intervals. As a result, the normal-based bootstrap confidence intervals also incorrectly cover the putative null hypothesis ($\lambda_1 = 1$) with high probability when the alternative is true, particularly for X_i following an elliptical distribution (Supplementary Table S7). In short, such bootstrap confidence intervals that use the bootstrap estimate of variance suffer from lack of power once the distribution of X_i deviates from strictly normal because of the large size of the confidence interval.

2.1 Detecting Gaps in the Eigenvalue Spectrum

The behavior of the top eigenvalues is often studied in theoretical work, but in practice, examination of the eigenvalues of the sample covariance matrix is largely done to find gaps in the eigenvalue spectrum which indicate a logical point at which to reduce the dimension of the data. Again, we focus for simplicity on detecting the separation of just the top eigenvalue from the remainder. Then a natural statistic is the gap statistic, $\hat{\lambda}_1 - \hat{\lambda}_2$, and large values suggest a true difference between the first and second eigenvalue. These statistics are difficult to understand theoretically, with limit distributions that are even less standard than the Tracy-Widom distribution (see (36; 12; 16)), which again makes them good candidates for using the bootstrap for inference.

The gap statistic $\hat{\lambda}_1 - \hat{\lambda}_2$ is also a biased estimate for the true population value. Our simulations show that unlike $\hat{\lambda}_1$, the direction of the bias for the gap statistic differs depending on the value of λ_1. For $\lambda_1 = 1$, the gap statistic overestimates the true difference, while for $\lambda_1 > 1$, the gap statistic underestimates the true difference (Supplementary Tables S8-S11); how large λ_1 needs to be before the bias becomes negative depends on the distribution of X.

As in the case of the top eigenvalue, the bootstrap estimate of bias does not accurately estimate this bias (Supplementary Figure S7). The bootstrap under-estimates the absolute size of the bias, and for elliptical distributions can misspecify the direction of the bias (Supplementary Tables S8-S11, Supplementary Figure S7). As with the top eigenvalue, the bootstrap estimate of bias improves as the top eigenvalue becomes more separated from the bulk. Estimating the variance of the gap statistic with the bootstrap shows similar problems, with the bootstrap widely over-estimating the variance of $(\lambda_1(\widehat{\Sigma}) - \lambda_2(\widehat{\Sigma}))$ in high dimensions (Supplementary Figure S8). Bootstrap confidence intervals also suffer from the same problem as those of the top eigenvalue: percentile CIs have low coverage of the truth in high dimensions and normal-based CI being much wider than necessary because of the over estimation of the variance.

Gap Ratio Statistic Another alternative that tries to normalize the gap statistic is the gap ratio, $(\lambda_1 - \lambda_2)/(\lambda_2 - \lambda_3)$ (32). In the scenario we are evaluating, this population quantity is not well defined ($\lambda_2 - \lambda_3 = 0$), but the estimate and its distribution are well defined, and the statistic is still a tool for deciding whether λ_1 and λ_2 are well separated. Again, the bootstrap estimate gives poor estimates of various features of the actual distribution of the gap ratio statistic, and the bias of the estimates depends on the true value of λ_1 (Supplementary Figures S11 and S12). The bootstrap distribution does not appear to be converging even in the case of normal X(Supplementary Figure S13).

3 Theoretical results

The problems with the bootstrap can be explained in part by the difference between the spectral behavior of weighted and unweighted covariance matrices when p/n is not small. This is because bootstrapping the observations (rows) of X is equivalent to randomly re-weighting the observations which changes the spectral distribution of $\widehat{\Sigma}$. Specifically, suppose that $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$. When bootstrapping, the bootstrap estimate, $\widehat{\Sigma}^*$, is given by $\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'$, where w_i is the proportion of times index i is picked in our resample. To understand the effect of this weighting consider the case where the X_i's are normally distributed; randomly weighting the X_i's effectively transforms, for spectral purposes, X_i into $\mathcal{X}_i = \sqrt{w_i}X_i$, and hence transforms the data distribution to an elliptical distribution. This leads to a very different spectral distribution of eigenvalues for $\widehat{\Sigma}^*$ when p/n is not close to 0 (a technical review of the spectral distribution of eigenvalues for these distributions is given in Section S1.1 in the Supplementary Text). Similarly, the distribution of the largest eigenvalues are dramatically affected by reweighting; the one exception to this rule is the situation where the largest eigenvalues of Σ are very separated from the rest.

In what follows we provide theoretical results that help explain the results of our numerical simulations and also complete them. The first set of results concerns the impact of bootstrapping on the spectral distribution of a sample covariance matrix. We explain that this creates bias in the setting we consider and it helps explain some of the misbehavior of the bootstrap we observed in the numerical study. We then consider the case of extreme eigenvalues, in the case where the largest population eigenvalues are well-separated from the bulk of the eigenvalues. We show that then the bootstrap works asymptotically under certain conditions. This helps explain why the performance of the bootstrap improves in our numerical study when we increase the largest population eigenvalue.

3.1 Bootstrapped empirical distribution

3.1.1 Spectral distribution of bootstrapped covariance matrix

The result that follows is in fact more general than the bootstrap setting, and applies to a broad class of randomly reweighted estimates $S_w = \frac{1}{n}\sum_{i=1}^n w_i X_i X_i'$, where w_i are random variables and X_i's are considered as non-random. The case of $w \sim Mult(n, 1/n)$ corresponds to the standard bootstrap estimate $\widehat{\Sigma}^*$. We recall that it is well-known that pointwise convergence of the Stieltjes transform (see Subsection 1.1) implies weak convergence of the corresponding spectral distributions (see (5; 21)).

Lemma 3.1. *Suppose $\{X_i\}_{i=1}^n$ are fixed vectors in $\mathbb{R}^p$. Consider $S_w = \frac{1}{n}\sum_{i=1}^n w_i X_i X_i'$ and $m_p(z)$ the Stieltjes transform of S_w. (Recall that $v = Im[z]$.) Suppose that w_i's are independent random variables. Then*

$$P(|m_p(z)) - \mathbf{E}(m_p(z))| > t) \leq C\exp(-cp^2v^2t^2/n) ,$$

with $C = 4$ and $c = 1/16$ for instance. Furthermore, the same result holds when w_i's have $Mult(n, 1/n)$ distribution.

Naturally, since the X_i's are considered as non-random in the previous lemma, the result applies directly to $\tilde{S}_w = \frac{1}{n}\sum_{i=1}^n w_i(X_i - \bar{X}_n)(X_i - \bar{X}_n)'$, by simply considering $\tilde{X}_i = X_i - \bar{X}_n$ in the previous lemma. See Section S3.1 for the proof.

Corollary 3.1. *When $p/n \to r \in (0, \infty)$, the Stieltjes transform $m_p(z)$ of the independently-weighted bootstrapped covariance matrix, S_w, is asymptotically deterministic. The same is true with the standard bootstrap, where the weights have a Multinomial(n,1/n) distribution. In particular, if f is a bounded continuous function, as n*

and p tend to infinity, while $p/n \to r$,

$$\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i^*) - \frac{1}{p}\sum_{i=1}^{p} \mathbf{E}^*\left(f(\lambda_i^*)\right) \to 0 \text{ in probability },$$

where λ_i^* are the decreasingly ordered bootstrapped eigenvalues and $\mathbf{E}^*(\cdot)$ refers to expectation under the bootstrap distribution.

The corollary follows from our lemma by simply using the fact that pointwise convergence of the Stieltjes transform implies weak convergence of the corresponding spectral distributions. It says that bootstrap average of "nice" functions of the spectral distribution are asymptotically non-random. We apply the result below for shrunken estimates of the covariance matrix.

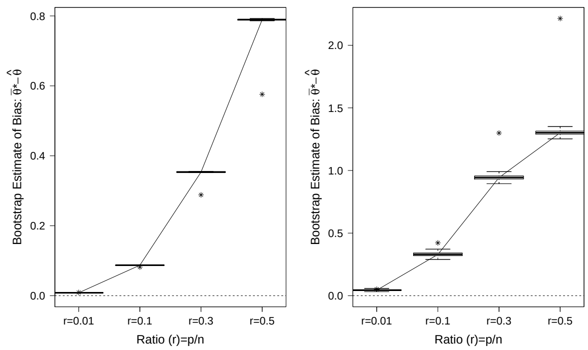
3.1.2 Bias of the bootstrap spectral distribution when p/n is not small: the case of Gaussian data

Lemma 3.1 above shows that the spectral distribution of the bootstrap eigenvalues has a non-random limit distribution, like that of the sample eigenvalues. The important question then remains as to whether the relationship between bootstrap and sample eigenvalues is the same as the relationship between sample and population eigenvalues, which would allow us to use the bootstrap for eigenvalues.

We first consider for concreteness and simplicity the case of a Gaussian design matrix, i.e. $X_i \overset{iid}{\backsim} \mathcal{N}(0, \Sigma)$. In doing so, we can clearly demonstrate (see Supplementary material S1.1.1) that the spectral distribution of $\widehat{\Sigma}^*$ is a biased estimate of that of $\widehat{\Sigma}$ and that furthermore the spectral relationship between $\widehat{\Sigma}^*$ and $\widehat{\Sigma}$ is different from that between $\widehat{\Sigma}$ and Σ.

Lemma 3.1 and Corollary 3.1 apply without restrictions on the design, which is one of their main strengths. However, current results in high-dimensional random matrix theory are still somewhat limited from a design standpoint (see Section S1.1 in Supplementary Text for a review). Hence we cannot make precise quantitative statements about bootstrapping empirical spectral distributions that are very general when it comes to the design matrix. It suffices to say that the impact of the distribution of w_i's spectral distribution is highly non-linear. Hence picking bootstrap weights according to the $Mult(n, 1/n)$ will generally not work.

A geometric interpretation Intuitively, one can think that bootstrapping moves the data in the setting considered here from a Gaussian setting to an elliptical one. It is well-known ((11), (24), (17) and Supplementary Material) that in moderate and high-dimension elliptical distributions have completely different geometric properties than Gaussian ones and that this impacts strongly the statistical behavior of many spectral estimators ((17)). As such, it is not that surprising that the bootstrap does not perform well: from an eigenvalue point of view, it is as if the bootstrap changed "the geometry of the dataset" and this geometry has an important impact on their behavior. *Bootstrapping is therefore not a good way to mimic the data generating process in this context.*



(a) $Z \sim$ Normal (b) $Z \sim$ Ellip. Exp.

Figure 4: Boxplots of the bootstrap estimate of the bias of the plug-in estimate $\hat{\theta} = \frac{1}{p}\text{trace}\left((\widehat{\Sigma} + s\text{Id}_p)^{-1}\right)$ over 1000 simulations. The true bias is shown with an asterisk. The regularization parameter s is set to $s = 0.1$.

Non-Gaussian Design Matrices Much work has been done in random matrix theory to extend the domain of validity of the Marchenko-Pastur equation, which holds beyond the case of Gaussian data. The bootstrap bias problem remains the same, because the limiting properties of the matrices of interest are unaffected by the move from Gaussian to these more general models (see (17)).

Example: Shrunken Inverse covariance matrix The bulk behavior of the bootstrapped eigenvalues is important for inference regarding various properties of the inverse covariance matrix. In particular, a non-trivial question is to understand the bias in shrunken versions of $\widehat{\Sigma}^{-1}$, i.e. $(\widehat{\Sigma} + s\text{Id}_p)^{-1}$. These questions are natural in the context of portfolio optimization (see (28)) or the study of regularized discriminant analysis (20). Furthermore, analytic characterization of this bias is hard (or currently impossible for general designs), making this example a prime candidate for the use of the bootstrap in applications. Corollary 3.1 imply that the bootstrap estimate of bias of $\frac{1}{p}\text{trace}\left((\widehat{\Sigma} + s\text{Id}_p)^{-1}\right)$ is asymptotically non-random. Figure 4 shows its performance in simulations: it is itself biased, and can either over-estimate or under-estimate the true bias, depending on the distribution of the underlying design matrix X. This confirms the conclusion of our earlier simulation results for extreme eigenvalues, namely that it is very hard to predict how the bootstrap fails.

3.2 Extreme Eigenvalues: well-separated case

We first consider the case where the top eigenvalues are much larger than the rest of the eigenvalues. In the context of PCA, this corresponds to the case where the data lives predominantly in a smaller subspace. We show that in this subsection that the bootstrap eigenvalues are consistent in probability.

Notations Call S_n the sample covariance matrix of the data and Σ_n the population covariance. We use the block notation

$$S_n = \begin{pmatrix} T_n & U_n \\ U_n' & n^{-\alpha} V_n \end{pmatrix} , \Sigma_n = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & n^{-\alpha}\Sigma_{22} \end{pmatrix} .$$

T_n and Σ_{11} are both assumed to be $q \times q$.

Assumptions: **A1** We assume that $|||\Sigma_{22}|||_2 = \mathrm{O}(1)$ and that $\lambda_{min}(\Sigma_{11}) > \eta > 0$. We assume that Σ_{11} is $q \times q$ with q fixed. **A2** X_i's are i.i.d with $X_i = r_i Z_i$, where $Z_i \sim \mathcal{N}(0, \Sigma_n)$, and $0 < \delta_0 < r_i < \gamma_0$ is a bounded random variable independent of Z_i, with $\mathbf{E}\left(r_i^2\right) = 1$.**A3** The bootstrap weights w_i have infinitely many moments, $\|w\|_\infty = \mathrm{O}(\mathrm{polyLog}(n))$ and $\mathbf{E}\left(w_i\right) = 1$. These weights can either be independent or Multinomial$(n, 1/n)$. **A4** p/n remains bounded as n and p tend to infinity.

Theorem 3.1. *Under our assumptions A1-A4, if $\alpha > 1/2 + \epsilon$ for any $\epsilon > 0$,*

$$\sup_{1\leq i \leq q} \sqrt{n}(\lambda_i(S_n) - \lambda_i(T_n)) = \mathrm{o}_P(1) . \quad (1)$$

Furthermore, if w denotes the vector of weights used in the bootstrap and the corresponding bootstrapped matrices are S_n^* and T_n^*, we have

$$\sup_{1\leq i \leq q} \sqrt{n}(\lambda_i(S_n^*) - \lambda_i(T_n^*)) = \mathrm{o}_{P,w}(1) . \quad (2)$$

The proof of this theorem is given in the Supplementary Text, Section S3.2. As noted in the proof, our arguments rely heavily on the very nice results of (14).

3.2.1 Consistency of the bootstrap

The definition of bootstrap consistency is recalled in the Supplementary material, Definition 1, p.8 there.

Theorem 3.2. *Suppose the eigenvalues of Σ_{11} are simple and the assumptions A1-A4 of Theorem 3.1 hold. Then the bootstrap distribution of the q largest eigenvalues of S_n is consistent in probability, provided $\alpha > 1/2 + \epsilon$, with $\epsilon > 0$.*

See Section S3.2 for the proof of this theorem, which uses the nice classical results of (6; 7) and (14). While we do not discuss bootstrap eigenvector results in this paper because they are much less used by practitioners and because of space limitations, the techniques we use here could be extended to these questions.

3.2.2 Discussion of some assumptions and remarks

$\boldsymbol{\alpha > 1/2 + \epsilon}$ This assumption is not terribly restrictive in a PCA context: in fact under our assumptions, if $\Sigma_{22} = \mathrm{Id}_{p-q}$, the fraction of variance explained by the top q eigenvalues is, if $C = \mathrm{trace}\left(\Sigma_{11}\right)/q$

$$\frac{\mathrm{trace}\left(\Sigma_{11}\right)}{\mathrm{trace}\left(\Sigma_n\right)} = \frac{qC}{qC + (p-q)n^{-\alpha}} .$$

So if $1/2 + \epsilon < \alpha < 1$, in our asymptotics this fraction of variance is asymptotically 0 ($pn^{-\alpha} \to \infty$ and qC is bounded). On the other hand, if $\alpha > 1$, the fraction of variance explained by the top q eigenvalues is approximately 1, which is the standard setting where PCA is used. If $\alpha = 1$, the fraction of variance varies between 0 and 1, depending on C. Of course, a similar analysis is possible and actually easy to carry out if Σ_{22} is not a multiple of the identity and corresponding details are left to the interested reader.

The m-out-of-n bootstrap As noted in (7; 14), subsampling approaches fix the problem of bootstrap inconsistency in the setting where Σ_{11} has eigenvalues of multiplicity higher than 1. Our approximation results for the pair (S_n^*, T_n^*) can be extended to subsampling approaches, and hence our results could be extended to cover these ideas. However, since this question is a bit distant from our main motivations, we do not treat it in detail here.

Discussion of other assumptions Assumptions on X_i's, Z_i's, and Σ could be relaxed significantly. In particular, the block representation assumptions of Σ_n are made for analytic convenience and can be easily dispensed of. We discuss these issues that are a bit secondary for this paper in the Supplementary material, Subsubsection S3.3.

3.3 When the extreme eigenvalues are not well-separated from the bulk

Precise results regarding the fluctuation behavior of the extreme eigenvalues of the bootstrapped covariance matrix, i.e. $\lambda_1(\widehat{\Sigma}^*)$, when the extreme eigenvalues of Σ are not well separated from the bulk, appear difficult to obtain in full generality. Indeed, our simulations in Section 2, particularly Figure 2, indicate that a great variety of behaviors seem to be possible; see also the discussion below. However, our results and discussion regarding the bias of the bulk distribution in the case of Gaussian designs induced by the reweighting of the X_i's give some insight into the behavior of the extreme eigenvalues in this setting. While a general theory is out of the reach of this paper, we give some concrete results concerning the bias of the bootstrap estimate of bias and discuss further this issue in the Supplementary material, Subsubsection S3.3.

A great many mathematically interesting questions remain concerning the fluctuation behavior of bootstrapped eigenvalues when the largest population eigenvalues are not well-separated from the bulk. For instance, one could ask whether it is the case that the bootstrap distribution of the largest eigenvalue of a sample covariance matrix is Tracy-Widom when $\lambda_1(\widehat{\Sigma})$ has this distribution? The simulation study in Section 2 suggests however that the answer to this question may be of limited practical statistical interest. For instance Figure 1 indicates that the bootstrap estimate of bias of $\lambda_1(\widehat{\Sigma})$ is itself very biased. Figure 2 and in particular Subfigure 1b suggests that the bootstrap distribution of our statistics is a poor approximation of the sampling distribution. These simulations also suggest that the characteristics of the bootstrap distribution and its relationship to the sampling distribution depend strongly on the characteristics of the design matrix, a problematic feature for a black box method such as the bootstrap. With current tools, mathematical characterization of the bootstrap distribution would be intractable outside of "simple situations", such as Gaussian or elliptical designs and slight generalizations. The lack of robustness to alternative distributions seen in simulations suggests that such mathematical characterization would be of limited practical statistical interest: the simulations show that in the high-dimensional setting the bootstrap distribution is very sensitive to characteristics of the data generating process that would be unknown in statistical practice. For this reason, we postpone these mathematically interesting and delicate questions to possible future work.

4 Conclusion

We have investigated in this paper the properties of the non-parametric bootstrap for spectral analysis of high-dimensional covariance matrices, namely when p/n is not close to 0, a realistic setup in current statistical practice. Our theoretical results concern two different aspects of the eigenvalue distribution. The first concerns the bulk of the eigenvalues. We show that the spectral distribution of the bootstrapped covariance matrix has a non-random limiting distribution. We go on to demonstrate that in the case of the Gaussian design, where this limiting distribution can be explicitly compared to that of the sample covariance matrix, the two do not have the same limiting distributions and the bootstrap version is thus a biased estimate of the sample spectral distribution. Furthermore, this bias is different from the bias that the sample spectral distribution exhibits when compared to the population spectral distribution. This set of results is of interest for instance when dealing with functions of the inverse covariance matrix for which all eigenvalues contribute.

The other aspect of the spectral analysis that we consider concerns inference for only the extreme eigenvalues. This is the setting most applicable for dimensionality reduction via PCA, for example, where inference on the top eigenvalues helps pick the subspace into which the data is projected. We show that when the top eigenvalues are well-separated from the rest and of multiplicity 1, the bootstrap is consistent (Theorem 3.1). We show that because the data are effectively low-dimensional in this setting, existing results derived for the case of $n \to \infty$ and fixed p can be extended to the case of $p/n \to r \in (0, \infty)$.

While this is a positive theoretical result for bootstrapping eigenvalues, its practical performance is less encouraging. Our simulation results for the simple question of detecting a single separated eigenvalue show that the separation of the top eigenvalue must be quite large to have bootstrap consistency. Moreover, the required size of the separation in eigenvalues for bootstrap consistency depends heavily on the distribution of the X_i's. When the separation is not sufficient, our simulation results show that the bootstrap distribution of the top eigenvalue is a highly biased estimate of the sampling distribution of the top eigenvalues. Hence, a naive application of the bootstrap method for these practically important eigenvalues will give very bad results. While this contradicts intuition based on low-dimensional analysis, this is not completely surprising, since the sample eigenvalue $\lambda_1(\widehat{\Sigma})$ is itself known to be a biased estimator of $\lambda_1(\Sigma)$ when $p/n \to r \in (0, \infty)$. However, even for other uses of the bootstrap that are less prone to bias-induced problems, for example to estimate the bias of the largest sample eigenvalues or to provide inference regarding the gap between eigenvalues, the bootstrap fails dramatically for values of p/n not near zero unless the top eigenvalue is enormously separated from the bulk.

This is particularly unfortunate, because it is exactly when the extreme eigenvalues are not well-separated from the bulk that theoretical results are hardest to obtain, most limited and require the most assumptions about the distribution of the design matrix. When the eigenvalues are well-separated, the low-dimensional nature of the problem essentially means that classical theoretical results regarding the distribution of the eigenvalues still hold.

Ultimately, in the very cases where the bootstrap would be most helpful, it fails as an inferential method. Moreover, this happens when one would expect - based on classical intuition and results - that the bootstrap "should" work. This echoes the findings of (18) in the case of bootstrapping in linear regression: in high-dimension, bootstrapping does not mimic the data generating process. Hence, standard bootstraps appear to work only for problems that are effectively low-dimensional.

Acknowledgements

We gratefully acknowledge support from the NSF grant DMS 1510172 and the ENS-CFM Data Science Chair.

References

[1] Demissie Alemayehu. Bootstrapping the latent roots of certain random matrices. *Comm. Statist. Simula-*

tion Comput., 17(3):857–869, 1988.

[2] T. W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34:122–148, 1963.

[3] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[4] Hamid Babamoradi, Frans van den Berg, and Åsmund Rinnan. Bootstrap based confidence limits in principal component analysis – a case study. *Chemometrics and Intelligent Laboratory Systems*, 120:97–105, January 2013.

[5] Z. D. Bai. Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica*, 9(3):611–677, 1999. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

[6] Rudolf Beran and Muni S. Srivastava. Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.*, 13(1):95–115, 1985.

[7] Rudolf Beran and Muni S. Srivastava. Correction: "Bootstrap tests and confidence regions for functions of a covariance matrix" [Ann. Statist. **13** (1985), no. 1, 95–115; MR0773155 (86g:62054)]. *Ann. Statist.*, 15(1):470–471, 1987.

[8] Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6):1196–1217, 1981.

[9] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[10] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997.

[11] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12(3):793–815, 1984.

[12] Momar Dieng. Distribution functions for edge eigenvalues in orthogonal and symplectic ensembles: Painlevé representations. *Int. Math. Res. Not.*, 37(37):2263–2287, 2005.

[13] Lutz Dümbgen. On nondifferentiable functions and the bootstrap. *Probab. Theory Related Fields*, 95(1):125–140, 1993.

[14] Morris L. Eaton and David E. Tyler. On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.*, 19(1):260–271, 1991.

[15] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[16] Noureddine El Karoui. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, March 2007.

[17] Noureddine El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362–2405, December 2009.

[18] Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimension? *Journal of Machine Learning Research*, 19(5), 2018.

[19] Aaron Fisher, Brian Caffo, Brian Schwartz, and Vadim Zipunnikov. Fast, exact bootstrap principal component analysis for $p > 1$ million. *Journal of the American Statistical Association*, 2015.

[20] Jerome H. Friedman. Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, 84(405):165–175, 1989.

[21] Jeffrey S. Geronimo and Theodore P. Hill. Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform. *J. Approx. Theory*, 121(1):54–60, 2003.

[22] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[23] Peter Hall, Young K. Lee, Byeong U. Park, and Debashis Paul. Tie-respecting bootstrap methods for estimating distributions of sets and functions of eigenvalues. *Bernoulli*, 15(2):380–401, 2009.

[24] Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(3):427–444, 2005.

[25] Iain M. Johnstone. High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I*, pages 307–333. Eur. Math. Soc., Zürich, 2007.

[26] I.M. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.*, 29(2):295–327, 2001.

[27] Tosio Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[28] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411, 2004.

[29] Ji oon Lee and Kevin Schnelli. Tracy-widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *arxiv:1409.4979*, 2014.

[30] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967.

[31] Richard O. Michaud. *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Oxford University Press, USA, June 1998.

[32] A. Onatski. Talk at mit conference, July 2006.

[33] Alexei Onatski. The Tracy-Widom limit for the largest eigenvalues of singular complex Wishart matrices. *Ann. Appl. Probab.*, 18(2):470–490, 2008.

[34] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.

[35] Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.*, 55(2):331–339, 1995.

[36] Alexander Soshnikov. Universality at the edge of the spectrum in Wigner random matrices. *Comm. Math. Phys.*, 207(3):697–733, 1999.

[37] Stephen K. Mitchell Sungjin. Hong and Richard A. Harshman. Bootstrap scree tests: A monte carlo simulation and applications to published data. *British Journal of Mathematical and Statistical Psychology*, 59(1):35–57, May 2006.

[38] M. E. Timmerman, H. A. Kiers, and A. K. Smilde. Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, 60(2):295–314, 2007.

[39] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[40] Kenneth W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability*, 6(1):1–18, 1978.