Title
PCA in High Dimensions: An Orientation

Permalink
https://escholarship.org/uc/item/11v419s3

Journal
Proceedings of the IEEE, 106(8)

ISSN
0018-9219

Authors
Johnstone, Iain M
Paul, Debashis

Publication Date
2018-08-01

DOI
10.1109/jproc.2018.2846730

Peer reviewed

Published in final edited form as:
Proc IEEE Inst Electr Electron Eng. 2018 August ; 106(8): 1277–1292. doi:10.1109/JPROC.2018.2846730.

PCA in High Dimensions: An orientation

Iain M. Johnstone and
Department of Statistics, Stanford University, Stanford CA 94305.

Debashis Paul
Department of Statistics, University of California, Davis.

Abstract

When the data are high dimensional, widely used multivariate statistical methods such as principal component analysis can behave in unexpected ways. In settings where the dimension of the observations is comparable to the sample size, upward bias in sample eigenvalues and inconsistency of sample eigenvectors are among the most notable phenomena that appear. These phenomena, and the limiting behavior of the rescaled extreme sample eigenvalues, have recently been investigated in detail under the spiked covariance model. The behavior of the bulk of the sample eigenvalues under weak distributional assumptions on the observations has been described. These results have been exploited to develop new estimation and hypothesis testing methods for the population covariance matrix. Furthermore, partly in response to these phenomena, alternative classes of estimation procedures have been developed by exploiting sparsity of the eigenvectors or the covariance matrix. This paper gives an orientation to these areas.

Keywords

Marchenko-Pastur distribution; principal component analysis; phase transition phenomena; random matrix theory; spiked covariance model; Tracy-Widom law

I. Introduction

Principal Component Analysis (PCA) is used throughout science and engineering to help summarize, represent and display data measured on many variables in terms of a smaller number of derived variables. The method originated with Karl Pearson in 1901 [1] and Harold Hotelling in 1933 [2]; further historical discussion appears in the book by Joliffe [3]. Its routine use in analysis of data required—and boomed with—the advent of electronic computers: an early classic in meteorology is the use by Lorenz [4] of the Whirlwind general purpose computer at MIT in the early 1950s to summarize air pressure data from $p = 64$ stations across the U.S.

The scale of data collection has exploded in recent decades, and it is no longer uncommon that the number of variables or features collected, p, may be on the order of, or larger than, the number of cases (or sample size), n. In this "high-dimensional" setting, under certain assumptions on the covariance structure of the data, the statistical properties of PCA exhibit phenomena that are perhaps unexpected when viewed from the historically standard perspective of many samples and a fixed number of variables.

This paper seeks to give an orientation to some of these high dimensional phenomena, which we label eigenvalue spreading, Section III, eigenvalue bias, Section IV and eigenvector inconsistency, Section V. The discussion is couched in terms of a particular assumed form for the covariance structure, the “spiked covariance model”, and focuses on a proportional asymptotic growth model in which $p/n \to \gamma \in (0, \infty)$ While certainly somewhat special, it allows clear statements of the phenomena which occur more widely. The results are introduced first in the setting of Gaussian observations, then Section VI reviews some of the many results now available beyond the Gaussian assumption.

Section VII looks at consequences of these high-dimensional phenomena in two estimation settings: covariance matrix estimation in a simple spiked model, and the exploitation of sparsity to estimate leading eigenvectors. Section VIII considers inferential questions in spiked models from the point of view of hypothesis tests.

Section IX departs from the proportional growth asymptotic model to review some non-asymptotic results and bounds, valid for p and n fixed. Finally Section X contains some concluding discussion, including some directions for extension of spiked models both for PCA, and for a variety of other multivariate models.

A. Basics of PCA

We start with observed data $X_1, \ldots, X_n \in \mathbb{R}^p$,, always using n for the number of observations on each of p variables, dimensions, or features. In this paper the observations will be assumed independent, though frequently they are correlated, as in time series. The data for us is real-valued, but everything goes over to the complex valued data that arises in signal processing, and indeed results are sometimes easier to prove for complex valued data.

The $p \times p$ sample covariance matrix is

$$\mathbf{S}_n = n^{-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)'$$

Mean correction by $\bar{X}$ is important in practice, but here for simplicity we assume that $\mathbb{E}X_i = 0$, and work instead with

$$\mathbf{S}_n = n^{-1} \sum_{i=1}^{n} X_i X_i' = n^{-1}\mathbf{X}\mathbf{X}',$$

where the $p \times n$ data matrix X has as its ith column the observation X_i. In the traditional formulation of large sample theory, p is taken as fixed and n is large, and it is then well known that we can estimate the population covariance matrix $\sum$ consistently. For example if the X$_i$ are assumed independent and identically distributed (i.i.d.) with $\mathbb{E}X_iX'_i = \Sigma$, then $\mathbf{S}_n \xrightarrow{\text{a.s.}} \Sigma$ as $n \to \infty$.

Our main focus is1the eigenstructure of covariance matrices. Linguists and information theorists note that the most basic concepts typically have short representations in many languages. So it is with the eigenvalue-eigenvector decomposition of a sample covariance matrix, which has been given acronyms in many fields, curiously always with three letters, for example: PCA: Principal Component Analysis, KLT: Karhunen Loeve Transform, EOF: Empirical Orthogonal Functions, and POD: Proper Orthogonal Decomposition. It is also closely related to the SVD: Singular Value Decomposition of the data matrix X.

It is important to distinguish the *population* and the *sample* versions of the eigendecomposition. For the population covariance matrix, we write the eigenvalue-eigenvector decomposition as

$$\Sigma = \ell_1 \mathbf{u}_1\mathbf{u}'_1 + \ldots + \ell_p \mathbf{u}_p\mathbf{u}'_p = \mathbf{ULU}',$$

where U is a $p \times p$ orthogonal matrix whose columns are the eigenvectors u$_i$ and L is a diagonal matrix, with entries ℓ_i being the eigenvalues of $\sum$, by convention arranged in decreasing order, and assumed here to be distinct. The sample covariance eigendecomposition is

$$\mathbf{S}_n = \lambda_1 \mathbf{v}_1\mathbf{v}'_1 + \ldots + \lambda_p\mathbf{v}_p\mathbf{v}'_p = \mathbf{V}\,\boldsymbol{\Lambda}\,\mathbf{V}',$$

where now the orthogonal matrix **V** has columns which are the sample eigenvectors $\mathbf{v}_i$ and diagonal Λ has entries being the sample eigenvalues λ_i, again in decreasing order. Even with distinct eigenvalues, the *sign* of the eigenvectors is not identified—this is usually handled by specifying a convention for choice of sign. Again, in the traditional setting of p fixed, and large i.i.d. samples n, the sample eigen-quantities converge to their population targets: as $n \to \infty$, we have $\lambda_k \xrightarrow{\text{a.s.}} \ell_k$ and $\mathbf{v}_k \xrightarrow{\text{a.s.}} \mathbf{u}_k$, for each $k = 1,..., p$. Comprehensive references for this setting include classic texts such as [3], [5], [6].

p and n and all that. We noted that the traditional setting has p much smaller than n. Conversely, there are now many applications in which p is much larger than n - for example in genomics, in which there may be hundreds of patients n but millions of SNPs. For definiteness, this paper, and much theory, considers the important “boundary” case, where p is of the same order of magnitude as n, so that it makes sense to do asymptotic approximations in which p and n grow proportionately: $\gamma_\text{n} = p/n \to \gamma \in (0, \infty)$.

B. A low rank “spiked” covariance model

We focus on an idealized model for the population eigende-composition, consisting of a known ‘base’ covariance matrix $\sum_0$ and a low rank perturbation:

$$\Sigma \;=\; \Sigma_0 \;+\; \sum_{k=1}^{K} h_k \mathbf{u}_k \mathbf{u}'_k, \qquad (1)$$

The signal strengths, or “spikes” h_k, as well as the orthonormal eigenvectors u$_k$ are taken as unknown. The rank K is small, (and stays fixed in asymptotic models as p and n grow). In the simplest form of this “spiked” model, $\sum_0$ is assumed equal to the identity matrix, or to a (possibly unknown) scalar multiple $\sum_0 = \sigma^2 I$, [7]. The term *generalized spike model* is used when $\sum_0$ is not necessarily a multiple of the identity [8], [9].

Our study of statistical properties will be primarily asymptotic, and we remark that an essential feature of high dimensionality captured in the (generalized) spiked model is that as p grows, the empirical distribution (defined in Section III) of the population eigenvalues of $\sum$ is approximated by a nontrivial limiting probability distribution—or by a sequence of distributions, one for each p, in the case of approximation by deterministic equivalents, e.g. [10, Ch. 6]) on [0, ∞). We contrast this with, for example, the important domain1of “functional data analysis” and functional PCA, in which the observation vectors X_i have some intrinsic temporal or spatial smoothness, and the eigenvalues of $\sum$ decay at some rapid rate, e.g. [11], [12].

Returning to model (1), regarding the eigenvectors $\mathbf{u}_1,..., \mathbf{u}_K$, we want to contrast two situations. The first makes no assumptions about them, while in the second, we assume that they are sparse in some known orthonormal basis.

For simplicity, we mostly assume distinct population eigenvalues so that eigenvectors are identified. Many of the papers cited in fact include cases when population eigenvalues coalesce (i.e., are degenerate), and some consider estimation of the resulting eigensubspaces, i.e., the linear subspaces spanned by the eigenvectors corresponding to distinct population eigenvalues.

1) Some examples: A (generalized) spike model may be an attractive and plausible idealization for data arising in many domains in science and engineering. An early example [13], [14] is provided by electrocardiogram (ECG) traces, Figure 1. The signal measured in the ith beat might be modeled as

$$X_i = \mu + \sum_{k=1}^{K} \sqrt{h_k} s_{ki} \mathbf{u}_k + \sigma Z_i\,. \qquad (2)$$

The periodic beats vary about the mean beat according to a small number of modes uk with random Gaussian amplitudes, the kth mode having variance hk. Independent Gaussian measurement noise is added. The corresponding covariance matrix Cov(Xi) has the finite rank perturbation form (1) with $\sum_0 = \sigma^2 I$.

Some other examples, described superficially, might include

- Microarrays: X_i might represent the levels of expression of p genes in the ith individual. The eigenvector u_i may be sparse as only a small number of genes may be involved in a given pathway.
- Satellite images: X_i may be suitable sub-images. After a discrete cosine transform, the u_i may be sparse, [15].
- Medical shapes: X_i may be vectors derived from landmarks of body organs. Eigenvectors u_i may be sparse due to localized variation, [16].
- Climate: X_i might be measurements from a global sensor network at time i, the EOFs u_i are often localized.
- Signal detection: X_i are observations at sensors, u_i columns of the steering matrix, with signals s_{ki} (compare model (2)).
- Finance: X_i is a vector of returns of p assets at time i, u_i are factor loadings, often *not* sparse, $f_{ki} = \sqrt{h_k} s_{ki}$ are factors and Z_i idiosyncratic terms.

In summary, the many examples suggest that the spiked model is worthy of theoretical study, especially because it is relatively easy to manage.

II. Statistical framework : Distribution of eigenvalues of the sample covariance matrix

For explicit calculations and proofs it is often helpful to assume that the observations X_i are Gaussian and we will do so in Sections II– V unless stated otherwise. The results often remain true if the vectors X_i are not Gaussian, though still independent, satisfying some structure and moment conditions – these will be discussed in Section VI.

The Wishart Distribution.

Suppose then that the column vectors X_i are independently distributed as $N_p(\mu, \Sigma)$, a p-variate multivariate normal distribution with population mean vector μ and covariance matrix Σ. With mean $\mu = 0$, the unnormalized sample covariance $\mathbf{H} = \mathbf{X}\mathbf{X}'$ is said to have the Wishart $W_p(n, \Sigma)$ distribution with p variables and sample size, or degrees of freedom n. In what follows, the "null" case will have $\Sigma = I$, though one could put in an unknown scale parameter. If $p \leq n$, which we suppose for the rest of this section, then $\mathbf{H}$ is non-singular with probability one. For $p > n$, the singular Wishart distribution is defined in [17].

For completeness, we briefly describe the joint density function of H and its eigenvalues, although they will not be needed in the sequel. Already in 1928, Wishart showed [18] that the density function of $H \sim W_p(n, \Sigma)$ is

$$c_{pn}(\det \Sigma)^{-n/2}(\det\mathbf{H})^{(n-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\mathbf{H}\right)\right\}.$$

For details, including the normalization constant c_{pn}, see e.g. [6, pp. 85, 62]. The joint density of the sample eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ of $\mathbf{S} = n^{-1}\mathbf{H}$ was obtained by James [19], and has the form

$$c'_{pn}(\det\Sigma)^{-n/2}\prod_{i=1}^{p}\lambda_i^{(n-p-1)/2}\prod_{i<j}^{p}\left(\lambda_i-\lambda_j\right)_0F_0^{(p)}\left(-\frac{n}{2}\Lambda,\Sigma^{-1}\right)$$

where the normalizing constant c'_{pn} is given in [6, p. 388], ${}_0F_0^{(p)}(X,Y)$ is a hypergeometric function with two symmetric $p \times p$ matrix arguments X and Y, defined in [6, p. 259], and $\Lambda = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_p\right)$.

These density functions, although explicit, are sufficiently elaborate in form that it is desirable to develop approximations. When dimensionality p and sample size n are large, such approximations may be derived from random matrix theory, as will be reviewed in succeeding sections.

III. High-dimensional phenomena I: Spreading of eigenvalues

Our first high dimensional phenomenon relates to the general fact, perhaps emphasized first by Charles Stein, [20], [21], that the sample eigenvalues are more spread out, or dispersed, than the population eigenvalues.

When p is proportional to n, the effect is very strong. Figure 2 illustrates this: in the left panel, all *population* eigenvalues equal 1, but the $p = 100$ *sample* eigenvalues have a histogram (over repeated sampling) that spreads over an order of magnitude! In the right panel, the population eigenvalues are equally spaced between 5 and 25, and the sample eigenvalues spread over a range from less than 1 to over 50. We see that the effect depends both on the population matrix $\sum$ and also very strongly on the ratio $\gamma_n = p/n$, becoming more pronounced as γ_n increases.

The Quarter Circle Law.

The importance of the spreading phenomenon, and the complexity of the exact joint distributions for fixed n and p makes it natural to look for approximations and limits. Marchenko and Pastur [22], see also [23], [24], gave a celebrated functional equation for the limiting distribution as $p/n \to \gamma$ for general $\sum$. The special case of $\sum = I$, that is, the ‘null’ or ‘white’ case, has a simple and important form that we show here. Suppose that $\mathbf{H} \sim W_p(n, I)$. The empirical distribution function for the p sample eigenvalues of $S_n = \mathbf{H}/n$ is given by $F_p(x) = p^{-1}\#\left\{\lambda_j \le x\right\}$.

If $p/n \to \gamma \le 1$, then the empirical distribution converges $F_p(x) \to F(x)$, with the limit distribution F having a density function of the celebrated quarter-circle form

$$f^{MP}(x) = \frac{1}{2\pi\,\gamma\,x}\sqrt{\left(b_+ - x\right)\left(x - b_-\right)}, \quad (3)$$

for x within the interval defined by the upper and lower *edges* $b_{\pm} = (1 \pm \sqrt{\gamma})^2$. The larger the ratio p/n, and hence γ, the larger the interval, in other words, the larger is the spreading of sample eigenvalues. If p is small relative to n then the distribution concentrates at 1, as occurs for p fixed, and as is to be expected when the data asymptotically allows perfect estimation. Figure 3 shows the examples $p = n/4$ in blue, and the extreme case $p = n$ in green.

If $p > n$, the sample covariance $\mathbf{S}_n$ has only n positive eigenvalues (with probability 1), and the remaining $p - n$ eigenvalues equal 0. If $p/n \to \gamma > 1$, the limit distribution F may be written in differential form as

$$F(dx) = (1 - 1/\gamma)\delta_0(dx) + f^{MP}(x)dx,$$

with δ_0 representing a unit point mass at 0, and f^{MP} as above, supported on the interval $b_{\pm} = (\sqrt{\gamma} \pm 1)^2$.

The name quarter-circle law also recalls the related, and celebrated Wigner semicircle law [25] which describes the limiting eigenvalue distribution for symmetric *square* matrices with i.i.d. entries[1].

IV. High-dimensional phenomena II: Eigenvalue bias

The second important phenomenon to emerge when p is proportional to n is that of bias in the top eigenvalues, combined with a phase transition in behavior that depends on the strength of the spike(s). Known as the Baik-Ben Arous-Péché (BBP) phase transition, it was first established for complex valued data in [27]. We first give an informal description of the bias and limiting distribution of the top sample eigenvalue(s) in the spiked model, and then a fuller set of references to the now large literature.

Consider a single spike, $K = 1$, and suppose that a basis has been chosen so that the population covariance matrix is diagonal: $\Sigma = \mathrm{diag}\left(\ell_1, 1, \ldots, 1\right)$. If $p = \gamma n$, the sample eigenvalues are approximately spread out according to the Marchenko-Pastur distribution, with upper end point of the bulk at $b_+ = (1 + \sqrt{\gamma})^2$, compare Figure 4.

In a "null hypothesis" case, we have also $\ell_1 = 1$. In that case, the largest sample eigenvalue is located near the upper edge b_+ and fluctuates on the (small) scale $n^{-2/3}$ approximately according to the real-valued Tracy Widom distribution: $\lambda_1 \approx \mu(\gamma) + n^{-2/3}\sigma(\gamma)TW_1$, where

$$\mu(\gamma) = b_+ = (1 + \sqrt{\gamma})^2, \ \sigma(\gamma) = (1 + \sqrt{\gamma})^{4/3}\gamma^{-1/6}. \quad (4)$$

[1] An historical note: the quarter-circle law was derived for $\sum = I$ independently, though not published, by Charles Stein [26]. It was presented as a generalization of Wigner's law in a Stanford course in 1966, though "it seemed clear at the time that he had done the derivation somewhat earlier (perhaps in the 1950's)" (Stephen Portnoy, personal communication).

For fixed $\gamma \in (0, \infty)$, this was established for complex-valued data (with limit distribution TW_2) in [28] and for real-valued data in [7]. An extension for $p/n \to 0$ or ∞ (so long as $\min\{p, n\} \to \infty$) was given by [29].

In the non-null cases, with $\ell_1 > 1$, the limiting bulk distribution of all sample eigenvalues is unchanged, essentially since it is unaffected by a single value. More surprisingly however, for $\ell_1 \le 1, + \sqrt{\gamma}$, the largest sample eigenvalue has the *same* limiting Tracy-Widom distribution – the (small) spike in the top population eigenvalue has no limiting effect on the distribution of the sample top eigenvalue. Put another way, asymptotically the largest sample eigenvalue is of no use in detecting a *subcritical* spike in the largest population eigenvalue.

A phase transition occurs at $1 + \sqrt{\gamma}$: for larger values of ℓ_1, the largest sample eigenvalue λ_1 now has a limiting Gaussian distribution, with scale on the usual order of $n^{-1/2}$, compare Figure 5. The mean of this Gaussian distribution shows a significant upward bias, being significantly larger than the true value of ℓ_1. It is perhaps noteworthy that the phase transition point $1 + \sqrt{\gamma}$ for ℓ lies buried within the bulk.

Appendix A gives a heuristic derivation and explanation for the inconsistency and the bias formula above the phase transition.

To summarize, for $1 \le \ell < 1 + \sqrt{\gamma}$, if $p/n = \gamma + o\left(n^{-2/3}\right)$,

$$n^{2/3}\left[\frac{\lambda_1 - \mu(\gamma)}{\sigma(\gamma)}\right] \overset{\mathcal{D}}{\Rightarrow} TW_\beta, \qquad (5)$$

where $\mu(\gamma)$, $\sigma(\gamma)$ in (4) *do not depend on* ℓ. The limiting Tracy-Widom distribution TW_β is more dispersed for real valued data ($\beta = 1$, [30]) than complex data ($\beta = 2$, [31]), but the centering and scaling constants are unaffected. The complex data result is established in [27] via analysis of Fredholm determinants, and the real case in [32] using a triadiagonal representation of XX'.

On the other hand, above the phase transition, $\ell > 1 + \sqrt{\gamma}$,

$$n^{1/2}\left[\frac{\lambda_1 - \lambda(\ell, \gamma)}{\tau_\beta(\ell, \gamma)}\right] \overset{\mathcal{D}}{\Rightarrow} N(0, 1). \qquad (6)$$

There is a simple formula for the limiting upward bias in λ_1,

$$\lambda(\ell, \gamma) - \ell = \gamma \frac{\ell}{\ell - 1}, \qquad (7)$$

which does not vanish as ℓ increases; on the contrary it decreases to the limiting value γ as $\ell \to \infty$. See Figure 6. The bias formula (as an almost sure limit) was established in [33]. The asymptotic variance is

$$\tau_1^2(\ell, \gamma) = 2\ell^2\left(1 - \frac{\gamma}{(\ell - 1)^2}\right).$$

in the real case, the convergence (6) being established by [34] by a perturbative method. An $O(n^{-1/2})$ Edgeworth correction term is given by [35]. In the complex case, $\tau_2^2 = \tau_1^2/2$, as shown in the orignal BBP paper [27], via a Fredholm determinant analysis.

If $\ell - 1 - \sqrt{\gamma} \sim wn^{-1/3}$, corresponding to the *critical regime*, then the limit distribution for λ_1, still on scale $n^{-2/3}$, is given by a modification of TW that depends on w, [27],[32].

In the domain of factor models in economics, [36] shows how these formulas exactly explain formerly puzzling phenomena observed in an influential empirical study [37] of properties of PCA applied to factor models published nearly 20 years earlier.

V. High-dimensional phenomena III: Eigenvector inconsistency

The third new phenomenon of the proportional sample size regime is perhaps the least recognized: namely that the leading eigenvectors in high dimensional PCA can be inconsistent.

Continuing the informal description in Figures 4 and 5, note that even when the population eigenvalue ℓ_1 is well above the phase transition, there is a non-trivial angle between the sample eigenvector $\mathbf{v}_1$ and its population counterpart $\mathbf{u}_1$. As the signal strength of the eigenvalue ℓ_1 decreases, the angle between $\mathbf{v}_1$ and $\mathbf{u}_1$ increases, indeed, when ℓ_1 falls below the phase transition $1 + \sqrt{\gamma}$, $\mathbf{v}_1$ is asymptotically *orthogonal* to $\mathbf{u}_1$, and gives *no* information about $\mathbf{u}_1$. More explicitly, if $p/n \to \gamma > 0$,

$$\langle \mathbf{v}_1, \mathbf{u}_1 \rangle^2 \to \begin{cases} \dfrac{1 - \gamma/(\ell_1 - 1)^2}{1 + \gamma/(\ell_1 - 1)} & \ell_1 > 1 + \sqrt{\gamma} \\ 0 & \ell_1 \in [1, 1 + \sqrt{\gamma}], \end{cases} \qquad (8)$$

Appendix B gives a heuristic derivation of this formula, continuing the arrowhead matrix method set out in Appendix A.

Inconsistency of the largest sample eigenvector in this setting was first established by Lu [38], [39]. The formula (8) appears in various settings in the physics literature [40], [41], and rigorous discussions in high dimensional PCA under various model assumptions have appeared in many articles, including [34], [42]–[46].

Figure 7 depicts the sample PC $\mathbf{v}_1$ as approximately uniformly distributed on a spherical cap at a non-zero angle to the population PC $\mathbf{u}_1$ More precisely, the angle $\langle \mathbf{v}_1, \mathbf{u}_1 \rangle^2$ concentrates near the limiting value given in (8). However, if $\mathbf{v}^{\perp} = \mathbf{v}_1 - \langle \mathbf{v}_1, \mathbf{u}_1 \rangle \mathbf{u}_1$ is the component of $\mathbf{v}_1$ orthogonal to $\mathbf{u}_1$, then $\mathbf{v}^{\perp}/\|\mathbf{v}^{\perp}\|$ is uniformly distributed on a unit sphere $\mathcal{S}^{p-1}$. This picture is exactly correct when the data X_i are i.i.d. Gaussian [34, Theorem 6], and asymptotically accurate more generally – the "delocalization" property to be discussed in Section VI-B below.

The discussion here focuses on the proportional regime $p/n \to \gamma \in (0, \infty)$. The behavior of PCA in settings when dimensionality p is much larger than n has been studied for example in [47]–[50]. Broadly speaking, if the spike eigenvalues remain fixed as p grows, the inconsistency properties can only get worse, but if the spike eigenvalues grow sufficiently fast with p and n, then consistent estimation of eigenvalues and vectors can still be possible.

VI. Universality phenomena

As we have seen, in the context of spectral analysis of the sample covariance matrix, the boundary case, namely when $p/n \to \gamma \in (0, \infty)$, yields many fascinating phenomena associated with the spectral elements of the Wishart matrix $\mathbf{S} = n^{-1}\mathbf{X}\mathbf{X}'$ with far reaching implications for statistical inference and signal processing in high-dimensional problems. However, many of the asymptotic results, especially those related to the behavior of extreme sample eigenvalues, and the sample eigenvectors under a spiked covariance model, were initially derived under the assumption of Gaussianity. In spite of providing valuable insights, these results by themselves are therefore still limited in their scope for practical data analysis. However, during the last decade, a large body of literature has been developed, primarily by analysts and probabilists, under the banner of "universality", that has greatly extended the scope of these results, and therefore enhanced the scope of statistical inference.

Universality in the context of random matrix theory (**RMT**) typically refers to the phenomenon that the limiting behavior of certain eigenvalue/eigenvector statistics does not depend on the distribution of the entries of the random matrix. One of the major threads of contemporary research in **RMT** has been to establish that the asymptotic behavior of eigenvalues both at the bulk and and at the edges essentially remains invariant as long as the first few moments of the distribution of entries match with those of a Gaussian data matrix.

A. Universality of eigenvalue statistics

At the level of convergence of the empirical spectral distribution, the bulk behavior of the eigenvalues of the sample covariance matrix is universal, as long as the entries of that data matrix **X** are standardized independent random variables satisfying a Lindeberg-type condition [51]. The limiting distribution of normalized extreme (or edge) eigenvalues started receiving increased attention with the works of Soshnikov [52] who proved the Tracy-Widom limit of the normalized largest eigenvalues a Wishart matrix. This result, and its extension by[53] required the existence of all moments (in particular, sub-Gaussian tails) and symmetry of the distribution of entries of **X**. Phase transition phenomena for the leading eigenvalues of a sample covariance matrix were established by [54] assuming only the existence of fourth moments of the observations. Bai and Yao ([55]) extended these results and the Gaussian limits for the leading sample eigenvalues when their population counterparts are above the phase transition point. In [56], they extended these results further to the setting of a *generalized spiked model* where the non-leading eigenvalues are slowly varying as opposed to a constant.

Over last few years, radical progress on establishing the universality phenomena have been made by Erdös, Yau and co-authors ([57]–[60]), and Tao and Vu ([61]–[63]) who used analytical techniques to study the question of both bulk and edge universality under much more relaxed assumptions on the entries. The "Four Moments Theorems" of Tao and Vu assert effectively that the limiting behavior of the local eigenvalue statistics of a matrix of the form $\mathbf{X}\mathbf{X}'$ is the same as when the entries of the data matrix $\mathbf{X}$ are i.i.d. standard Gaussian, provided the first four moments of the entries of $\mathbf{X}$ match with those of the standard Gaussian. A prototypical instance of such results is the following.

Theorem *([63]): Let* $\mathbf{X} = ((\mathrm{X}_{\mathrm{ij}}))$ *and* $\widetilde{\mathbf{X}} = \left(\left(\widetilde{X}_{ij}\right)\right)$ *be* $p \times n$ *matrices with* $p,\ n \to \infty$ *such that* $p/n \to y \in (0, 1]$*. The entries* X_{ij} *(respectively,* $\widetilde{X}_{ij}$*) are independently distributed, have mean zero and variance 1, and obey the moment condition* $\sup_{i,j} \mathbb{E}\left[\left|X_{i,j}\right|^{C_0}\right] < C$ *for a sufficiently large constant* $C_0 \geq 2$ *and some C independent of* p, n*. Moreover, all the moments of order up to 4 are identical for* X_{ij} *and* $\widetilde{X}_{ij}$*.. Let* $\mathbf{S}$ *and* $\widetilde{\mathbf{S}}$ *denote the associated covariance matrices. Then the following holds for sufficiently small* c_0 *and for every* $\varepsilon \in (0, 1)$ *and for every* $k \geq 1$:

Let $G: \mathbb{R}^k \to \mathbb{R}$ *be a smooth function obeying the derivative bound*

$$\left\| \nabla^j G(x) \right\|_\infty \leq n^{C_0}$$

(where $\left\| \cdot \right\|_\infty$ *denotes the largest element) for all* $0 \leq j \leq 5$ *and* $x \in \mathbb{R}^k$*. Then for any* $\varepsilon p \leq i_1 < i_2 < \cdots < i_k \leq (1-\varepsilon)p$*, and for sufficiently large n on depending* ε*, k,* c_0*, we have*

$$\left|\mathbb{E}\left[G\left(n\,\lambda_{i_1}(\mathbf{S}), \ldots, n\,\lambda_{i_k}(\mathbf{S})\right)\right] - \mathbb{E}\left[G\left(n\,\lambda_{i_1}(\widetilde{\mathbf{S}}), \ldots, n\,\lambda_{i_k}(\widetilde{\mathbf{S}})\right)\right]\right| \leq n^{-c_0},$$

where λ_j denotes the j-th largest eigenvalue.

[64] extended the domain of validity of the bulk and edge universality results even further by only requiring that the first two moments of the entries of $\mathbf{X}$ match that of a standard Gaussian, subject to a sub-exponential tail behavior. Key steps in the derivation of these results are: (i) to derive a strong local Marchenko-Pastur law, which gives a precise estimate

of the local eigenvalue density; (ii) to embed the covariance matrix into a stochastic flow of matrices so that the eigenvalues evolve according to a coupled system of stochastic differential equations, called the Dyson Brownian motion; (iii) to implement a Green function comparison method that establishes closeness between the eigenvalue statistics of the Dyson Browian motion at time $t = O(n^{-1})$ with that of the original matrix, corresponding to flow at time $t = 0$. As a corollary, one obtains that the limiting distribution of the normalized largest eigenvalue of **S** is the Tracy-Widom distribution. The use of Dyson Brownian motion in the work of [64], can be broadly seen as a stochastic interpolation technique that builds a bridge between the data matrix **X** and a matrix of the same dimension with i.i.d. Gaussian entries with identical first two moments.

Above universality results are for the *null Wishart* case. For the more general setting, when $\mathbf{X} = \sum^{1/2}\mathbf{Z}$, where $\sum^{1/2}$ is a symmetric $p \times p$ matrix and **Z** is a $p \times n$ matrix with i.i.d., zero mean, unit variance entries, universality of extreme eigenvalues of $\mathbf{S} = n^{-1}\mathbf{X}\mathbf{X}'$ has been an object of intense study more recently. For a spiked covariance model, universality of extreme eigenvalues was established by [65] under the assumption that $\sum$ is diagonal with all but a finite number of diagonal entries equal to 1, and the entries of **Z** have vanishing odd moments. The latter assumption was relaxed by [66], who also established large deviation bounds on the spiked eigenvalues of **S**. Edge universality, and in particular the phase transition phenomena and Tracy-Widom limit for renormalized spiked sample eigenvalues in the sub-critical regime, for a general non-identity diagonal matrix $\sum$, has been established by [67] and [68]. They have used techniques closely related to those utilized by [64].

B. Universality of sample eigenvectors

Study of the behavior of the eigenvectors of a sample covariance matrix arises in the context of PCA. When **X** has i.i.d. zero mean Gaussian entries, distributional invariance under multiplication of **X** by orthogonal matrices implies that the matrix of eigenvectors of the sample covariance matrix **S** is Haar distributed, that is, the distribution is uniform on the space of orthogonal matrices. In particular, this means that the individual eigenvectors of **S** are uniformly distributed on the unit sphere in $\mathbb{R}^p$. Several results have been derived to describe analogous behavior of the matrix of eigenvectors even when **X** is not Gaussian. A first result of this kind was proved by [69] who showed that if the first four moments of the entries of the data matrix match those of the standard Gaussian, then the matrix of eigenvectors is *asymptotically Haar distributed* as $p/n \to \gamma \in (0,\infty)$.

One of the qualitative features of these results is the observation that entries of individual sample eigenvectors are of similar magnitude, a phenomenon often referred to as a *delocalization property* of eigenvectors. Such delocalization results are typical byproducts, and indeed important ingredients, in the contemporary investigations on universality of sample eigenvalues ([63], [64]).

Eigenvectors associated with the spiked eigenvalues of **S** under a spiked covariance model are of obvious interest. The eigenvector phase transition result (8) suggests that when a population spike is below the phase transition limit $1 + \sqrt{\gamma}$, the corresponding sample eigenvector is orthogonal to the population eigenvector and therefore does not contain any information about the latter. As a generalization, [66] established the large deviation properties of linear functionals of the sample eigenvectors under a spiked population model, with a diagonal population covariance matrix ∑ that is a fixed rank perturbation of the identity, and with distributions of observations having subexponential tails. They showed that, when a population spike eigenvalue ℓ_j is above the phase transition limit $1 + \sqrt{\gamma}$, the corresponding sample eigenvector $\mathbf{v}_j$ concentrates on the intersection of the unit sphere and a cone around the true population eigenvector, as in the Gaussian case (Figure 7). Moreover, the eigenvector $\mathbf{v}_j$ is completely delocalized in any direction orthogonal to the corresponding population eigenvector $\mathbf{u}_j$, while for spikes below and strictly away from the phase transition limit, the corresponding sample eigenvectors are completely delocalized. A surprising finding of [66] is that, when a spiked eigenvalue ℓ_j, say, is in close proximity to the phase transition point, so that, $\left| \ell_j - (1 + \sqrt{\gamma}) \right| \ll 1$, the complete delocalization of the sample eigenvector $\mathbf{v}_j$ in the direction of the corresponding population eigenvector $\mathbf{u}_j$ breaks down.

VII. Estimation in spiked models

Spiked covariance model has a natural interpretation in terms of factor models that are commonly used in econometrics and various branches of sciences. Partly because of this, and partly owing to the well-understood characterization of the asymptotic behavior of the sample eigenvectors and eigenvalues, the spiked covariance model has gained popularity in high-dimensional statistical estimation theory and inference. Investigations have focused on two related problems, one primarily dealing with estimation of the leading eigenvectors of a spiked covariance matrix, and the other focusing on the estimation of the covariance matrix itself.

A. Estimation of leading eigenvectors under sparsity

One branch of this estimation theory assumes some form of sparsity of the eigenvectors associated with the spiked eigenvalues. Specifically, the covariance matrix ∑ is assumed to be of the form $\Sigma = \sum_{k=1}^{K} \tilde{\ell}_k \mathbf{u}_k \mathbf{u}_k^T + \sigma^2 I_p$, where $\tilde{\ell}_1 \geq \cdots \geq \tilde{\ell}_K > 0$, with the orthonormal eigenvectors $\mathbf{u}_1$,..., $\mathbf{u}_K$ having only a few coordinates significantly different from zero. Under this framework, various nonlinear estimation strategies have been proposed for estimating the eigenvectors of ∑. This line of research started with [14] who established consistency of an eigenvector estimator that is obtained by a two-stage procedure. In this method, the first stage involves selection of coordinates based on thresholding the sample variances, which is then followed by a PCA of the selected submatrix of the sample covariance matrix. Improved coordinate selection schemes, together with detailed analyses of the minimax optimality of the proposed estimators have been studied by [70] and [71], while alternative estimation strategies and their asymptotic properties have been investigated by [72] and [73], among others. Under the assumed model, the eigenvector estimators can also be utilized to obtain consistent estimates of ∑ or $\sum^{-1}$ (e.g. [74]).

B. Estimation of the covariance matrix

While the assumption of sparsity of the eigenvectors allows one to solve the problem of covariance estimation in dimensions much larger than the sample size, in the absence of such structural assumptions, there is little hope of obtaining meaningful estimates in the $p \gg n$ setting. However, interesting covariance estimation procedures have been developed by making use of the eigenvalue and eigenvector phase transition phenomena in the "boundary case" $p/n \to \gamma \in (0, \infty)$ under the spiked covariance model. There are alternative estimation strategies (notably, by [75] and [76]) that do not rely on a spiked covariance formulation, but rather restrict attention to rotation-equivariant estimators. These estimators of $\sum$ are of the form $\mathbf{V}_\eta(\Lambda)\mathbf{V}'$, where $\mathbf{V}\,\Lambda\,\mathbf{V}'$ denotes the spectral decomposition of the sample covariance matrix **S**, and $\eta(.)$ denotes an appropriate nonlinear shrinkage applied to the·sample eigenvalues (diagonal of Λ).

To keep the discussion well-connected with the remainder of this paper, below we focus on an estimation strategy [77] specifically designed for a spiked covariance model. It also shows clearly the consequences for estimation of each of the three high-dimensional phenomena discussed in Sections III, IV and V.

The strategy is inspired by early work of Stein, reported in the 1975 IMS Rietz Lecture, partly published in [78]. Suppose that $X_i \xrightarrow{i.i.d.} N_p(0, \Sigma_p)$, for $i = 1,\ldots,n$, with $\sum_p$ having a spiked covariance structure, namely, the eigenvalues of $\sum$ are $\ell_1 \geq \cdots \geq \ell_r > 1 = \cdots = 1$ for some fixed $r \geq 1$. Because of the eigenvalue spreading phenomenon, we want to shrink the sample eigen*values*. Here we propose using a single univariate function η to do the shrinking. With no prior information about the population eigenvectors, we leave the sample eigenvectors alone. This leads to an *orthogonally invariant* estimator of the form

$$\widehat{\Sigma}_\eta(\mathbf{S}_n) = \eta(\lambda_1)\mathbf{v}_1\mathbf{v}'_1 + \ldots + \eta(\lambda_p)\mathbf{v}_p\mathbf{v}'_p$$

While this is a more special form than the general rotation invariant estimator $\mathbf{V}\eta\,\Lambda\,\mathbf{V}'$ mentioned earlier, it turns out that in the present setting, nothing is lost asymptotically by the restriction to scalar shrinkers [77, Sec. 8].

In view of the eigenvalue bias phenomenon, and the explicit upward bias function (7) for the top sample eigenvalue, it is natural to think that one could just undo the bias by choosing the shrinkage function η to be the inverse of $\lambda(\ell)$:

$$\ell(\lambda) = \begin{cases} \dfrac{\lambda + 1 - \gamma + \sqrt{(\lambda + 1 - \gamma)^2 - 4\lambda}}{2} & \lambda > \lambda_+(\gamma) \\ 1 & \lambda \leq \lambda_+(\gamma). \end{cases} \qquad (9)$$

Note that the inversion is to be applied only to sample eigenvalues above the phase transition $\lambda_+(\gamma) = (1 + \sqrt{\gamma})^2$.

However, this idea is complicated by the eigenvector inconsistency phenomenon. In view of (8), the top population and sample eigenvectors $\mathbf{u}_1$ and $\mathbf{v}_1$ span a plane, shown in Figure 8. Inversion of λ_1 to $\ell(\lambda_1)$ still makes an error because the error of $\mathbf{v}_1$ in tracking $\mathbf{u}_1$. Depending on how we measure the error, it seems clear that some other shrinkage value $\eta(\lambda_1)$ might lead to smaller error than simply undoing the bias.

Indeed, Table I shows some many commonly used orthogonally invariant loss functions, such as the Operator loss, Frobenius loss, Entropy loss, Stein's loss and Fréchet loss, and the optimal shrinkage function, available in closed form, that minimizes the limiting loss

$$L_\infty(\eta \mid \ell_1, \ldots, \ell_r) = \lim_{n, p \to \infty} L_p\left(\Sigma_p, \widehat{\Sigma}_\eta(\mathbf{S}_n)\right), \quad (10)$$

in the asymptotic framework $p/n \to \gamma \in (0, 1]$ as $p, n \to \infty$. Indeed, for operator norm, it is best to invert the bias function, but for the other loss functions, a notably larger amount of shrinkage is done, especially for Stein's loss. The key point is that the choice of loss function critically affects which estimator is optimal, and this follows directly from the high dimensional phenomena outlined earlier.

VIII. Inferential questions under the spiked model framework

One of the earliest uses of the distribution of the largest eigenvalue of the sample covariance matrix is in testing the hypothesis $H_0 : \Sigma = I_p$ when i.i.d. samples are drawn from a $N(\mu, \Sigma)$ distribution. This testing problem, typically referred to as testing the hypothesis of sphericity, has a long history. Mauchly [79] first derived the likelihood ratio test for sphericity under the classical fixed p and Gaussian observations regime. The (Gaussian) locally most powerful invariant (under shift, scale and orthogonal transformations) test was obtained by John ([80], [81]) and by [82]. [83] proposed extensions (for the unknown and known scale problems) of John's test, while [84] proposed corrections to Mauchly's likelihood ratio test for the $p/n \to \gamma \in (0, \infty)$ regime. Taking a different approach, Pillai ([85], [86], [87]) utilized the asymptotic behavior of the largest sample eigenvalue to develop tests for sphericity under the fixed p (Gaussian) regime.

The Tracy-Widom law for the largest sample eigenvalue under the null Wishart case, i.e., when the population covariance matrix $\Sigma = I_p$, allows a precise determination of the cut-off value for the largest root test. With a careful calibration of the centering and normalizing sequences, this cut-off value is very accurate in terms of having the correct level of significance even for relatively small p and n ([7], [88], [89]). In addition, the Tracy-Widom law for the largest eigenvalue has been extensively used for signal detection ([90], [91], [92], [93]). Many of these approaches use a sequential hypothesis testing framework whereby the Tracy-Widom law is used to determine the null distribution for testing the presence of an additional signal direction.

In view of various contrasting approaches, a detailed analysis of the behavior of the power function for tests of sphericity requires formulating suitable alternative models. The spiked covariance model provides such a convenient model that has easy interpretability, and at the same time has enabled researchers to carry out precise power analysis. The asymptotic power of various tests for sphericity has been thoroughly investigated by Onatski, Johnstone and coauthors ([94], [95], [96]). We provide here a brief overview of these works.

Onatski et al. [94] studied the asymptotic power of tests of sphericity against perturbations in a single unknown direction as both p and n go to infinity. They established the convergence, under the null hypothesis and contiguous alternatives, of the log ratio of the joint densities of the eigenvalues of the sample covariance under the alternative and the null, to a Gaussian process indexed by the norm of the perturbation. They showed that when the norm of the perturbation is below the phase transition threshold, the limiting log-likelihood ratio process is nondegenerate, and the joint eigenvalue densities under the null and alternative hypotheses are mutually contiguous. Importantly, consistent with formula (5) above, under the contiguous alternative regime, the asymptotic power of the Tracy-Widom-type tests is trivial (i.e., equals the asymptotic size), whereas that of the eigenvalue-based likelihood ratio test is always larger than the size and increases to one as $\ell \nearrow 1 + \sqrt{\gamma}$.

IX. Finite sample behavior of principal components

While most of the work within the framework of high-dimensional PCA is asymptotic in nature, with both $p, n \to \infty$ together, there have been notable recent developments in terms of providing finite sample bounds on the discrepancy between the population eigenvalues and eigenvectors and their sample counterparts. One of the first works of this kind is by Nadler [42], who considered a Gaussian observation model with a single spike for the covariance matrix, and established probabilistic bounds for fluctuations of the largest eigenvalue of the sample covariance matrix for arbitrary p and n. He also established a finite sample probabilistic bound for the sine of the angle between the leading sample eigenvector and the corresponding population eigenvector. A feature of the work by Nadler [42] is the use of "small-noise-asymptotics", whereby for fixed p and n, the noise variance (which equals the value of the non-spiked eigenvalues) is allowed to converge to zero. He provided analytic expansions of the leading sample eigenvalue under this asymptotic regime. [97] extends this small-noise approach to other spiked multivariate models.

Understanding the behavior of eigenprojections – orthogonal projection operators onto the eigensubspaces – of the sample covariance matrix, not necessarily under the spiked covariance model, has received attention from multiple communities. Vaswani and coauthors [98], [99] studied signal recovery through PCA in a framework that allows both nonisotropic noise and noise that are correlated with the signal. Specifically, they considered the observation model

$$Y_t = \mathbf{Q}a_t + \mathbf{M}_t\mathbf{Q}a_t + v_t, \quad (11)$$

where $\mathbf{Q}$ is a $p \times m$ matrix with $m \ll p$, with $\mathbf{Q}a_t$ denoting the random signal, while the uncorrelated noise component v_t satisfies $\mathbb{E}\left(\mathbf{Q}a_t v_t^T\right) = 0$. For unknown $p \times p$ matrices $\mathbf{M}_t$, the component $\mathbf{M}_t\mathbf{Q}a_t$ represents the component of noise that is correlated with the data. They discussed various engineering and signal processing applications, including PCA based on missing data. In each case, an estimate of $\mathbf{Q}$ is formed by the matrix consisting of the first r eigenvectors of the sample covariance matrix of $\{Y_1,\ldots, Y_n\}$. Assuming sub-Gaussian signal and noise, [99] established finite sample probabilistic bound for *subspace recovery error*, defined as sine of principal angle between the column spaces of $\mathbf{Q}$ and $\widehat{\mathbf{Q}}$. Under the spiked covariance model, their results are analogous to those by [42].

In related works, Koltchinskii and Lounici [100], [101] studied the behavior of $\left\|\hat{P}_r - \mathrm{P}_r\right\|_2^2$, where $\mathbf{P}_r$ denotes the eigenprojection corresponding to the r-th largest distinct eigenvalue of the population covariance $\sum$, and $\hat{P}_r$ is the corresponding sample eigenprojection based on i.i.d. observations from the population, while the norm is the Hilbert-Schmidt norm. They established the uniform convergence of the standardized version of this quantity to a standard normal distribution. While their work is not within the context of the $p/n \rightarrow c > 0$ setting, they showed that the accuracy of the normal approximation is characterized by the so called "effective rank" $r(\Sigma) := \mathrm{trace}(\Sigma)/\|\Sigma\|$ where $\|\Sigma\|$ denotes the operator norm of $\sum$. They also established finite sample concentration bounds for $\left\|\hat{\mathbf{P}}_r - \mathbf{P}_r\right\|_2^2$ and nonasymptotic bounds its expectation and variance.

The framework adopted by [100] is closely linked with the *functional principal component analysis* framework studied by many researchers. Without delving into the huge literature associated with this topic, we just mention a few works that are most relevant. [102] established nonasymptotic bounds on the L^2 risk of estimating the eigenfunctions of a covariance operator under different regimes, including polynomial and exponential decay of the eigenvalues of the population covariance operator. [103] established nonasymptotic bounds on the expected *excess empirical risk* associated with the projection of the observed data onto the eigensubspaces associated with the leading eigenvalues. The bounds show that the excess risk differs considerably from the subspace distances between the population and sample eigenprojections.

X. Concluding Discussion

We have provided a broad overview of the key phenomena associated with high-dimensional PCA. In this section, we summarize some of the recent trends, and discuss some unresolved questions, in theoretical analyses of PCA and allied methodologies. Strikingly, this literature is characterized by increasingly sophisticated utilization of tools from random matrix theory. The research directions we outline here are broadly categorized into three sub-categories: (i) extensions around high-dimensional PCA in different domains, including time-dependent data, variance components modeling, and hypothesis testing involving the covariance matrix; (ii) exploration of spike phenomena in other multivariate models; and (iii) resampling based inference for principal components.

A. Some extensions (around PCA)

Time dependent data: While traditional multivariate statistical analysis focuses on independently observed samples, much of the data in real world are intrinsically time-dependent. It is notable that PCA is routinely applied for dimension reduction and signal detection in data that can best be characterized as a time series. There is a voluminous econometric literature focusing on static factor models with time-dependent factor loadings when $p/n \to 0$. There is also a growing literature on *dynamic factor models (DFM)* [104]. Until recently, there were little theoretical investigations on statistical properties of estimators under these models when $p/n \to c \in (0, \infty)$. Motivated by the question of determination of the number of dynamic factors in a DFM, Jin *et al.* [105] established the existence of a limiting spectral distribution of the ESD of symmetrized sample autocovariance matrices based on the "null model", i.e., when the observations are assumed to be i.i.d. and isotropic. Liu *et al.* [106] extended these results to a class of linear processes with simultaneously diagonalizable coefficient matrices. A further relaxation on the structure of the linear process was achieved by [107]. These results raise the prospect of extending analyses already carried out for i.i.d. observation, such as establishment of phase transition phenomena, characterization of limiting distribution of extreme eigenvalues, establishment of CLT for linear spectral statistics, estimation of spectra of population covariance, to the setting of time-dependent data. Significant progress related to phase transition phenomena for singular values of sample autocovariances has been made in [108] and [109]. A method for estimating the joint spectrum of coefficient matrices of a class of ARMA processes has been developed in [110]. A different kind of phase transition phenomenon for the largest sample eigenvalues and associate eigenvectors, when the coefficient matrix of an AR(1) process has low rank, has been described in [111]. These results point to the possibility of a rich exploration of phenomena associated with eigen-analysis in the context of high-dimensional time series.

Multivariate variance components: We have been concerned with spectral properties of data relating to a single high-dimensional covariance matrix. In a multivariate variance components model, more than one covariance matrix appears:

$$X = U_1\alpha_1 + \cdots + U_k\alpha_k .$$

Here $\boldsymbol{U}_r$ are fixed $n \times I_r$ design matrices, while the I_r rows of $\boldsymbol{\alpha}_r$ are independently distributed as $N_p(0, \Sigma_r)$. High dimensional settings in which p, n and each I_r grow proportionately are of interest, for example, in quantitative genetics. Spectral properties of quadratic estimators $\widehat{\Sigma}_r = X'B_rX$ of the variance components Σ_r can be investigated: [112] and [113] describe results for the bulk and edge eigenvalues of $\widehat{\Sigma}_r$. Work in progress by the same authors studies analogs of the results of Sections IV and V for spiked models for each Σ_r.

Tests of sphericity beyond the spiked alternative: In a recent work, Dobriban [114] dealt with the question of detection of directionality in high-dimensional data by going beyond the spiked alternatives formulation. His approach addresses the question whether it is possible to detect weak PCs under the general covariance matrix models of [22]. He formulated the hypothesis testing problem within the framework of a non-parametric, non-Gaussian generalization of the spiked model. Specifically, denoting E_p to be the empirical distribution of the eigenvalues of Σ, this formulation boils down to testing

$$H_{0,p}: E_p(1 - h/p)E + (h/p)G_0$$

against

$$H_{1,p}: E_p(1 - h/p)E + (h/p)G_1$$

where E, G_0 and G_1 are pre-specified probability distributions supported on $\mathbb{R}^+$, and $0 < h < p$ is a specified constant. Clearly, by taking $E = G_0 = \delta_1$ (degenerate at 1), the test becomes that of testing sphericity, while at the same time, taking h to be a fixed integer and $G_1 = h^{-1}\sum_{j=1}^{h} \delta_{1+c_j}$ for positive c_j's leads to a spiked alternative. [114] developed new tests based on asymptotic Gaussianity of linear functionals of eigenvalues of the sample covariance matrix to detect weak PCs under this model. A related approach to test of sphericity, involving a correction for the likelihood ratio statistic to compensate for the dimensionality, is discussed in [115].

B. Other multivariate models

PCA is only one of a whole arsenal of methods of multivariate statistics which are based on eigenvalues and eigenvectors of one or two sample covariance matrices. Examples include signal detection, MANOVA and multiple response regression, canonical correlations, discriminant analysis and so on—these form much of the content of textbooks on multivariate statistical analysis such as [5]. James ([116]) organized all these problems into a hierarchy of five different classes (indexed by the classical hypergeometric functions ${}_pF_q$).

The high dimensional phenomena discussed in earlier sections extend to the James hierarchy. For example [95] and [96] consider the spike testing problem. Each of James' five testing problems is related to the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ where $\mathbf{H}$ and $\mathbf{E}$ are independent and proportional to high-dimensional Wishart matrices. Under the null hypothesis, both Wisharts are central with identity covariance. Under the alternative, the non-centrality or the covariance parameter of $\mathbf{H}$ has a single eigenvalue, or a spike, that stands alone. When the spike is larger than a case-specific phase transition threshold, one of the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ separates from the bulk. This makes the alternative easily detectable, so that reasonable statistical tests are consistent, in the sense that their power converges to 1 and a local asymptotic normality theory can be built [96]. In contrast, when the spike lies below the threshold, none of the eigenvalues separates from the bulk, which makes the testing problem more challenging. [95] shows that, the measures corresponding to the joint distributions of the eigenvalues under the alternative and the null hypotheses are mutually contiguous when the magnitude of the spikes are below the phase transition threshold. Furthermore, the log-likelihood ratio processes parametrized by the values of the spikes are asymptotically Gaussian, with logarithmic mean and autocovariance functions. These findings allow computation of the asymptotic power envelopes for the tests for the presence of spikes in the different multivariate models.

C. Bootstrapping high-dimensional PCA

Resampling methods have been very popular in statistics and machine learning due to their distribution-free characteristics and easy applicability. In finite-dimensional problems, under mild regularity conditions, bootstrap techniques provide a useful alternative to (nearly always) asymptotic inference procedures that typically involve quantities requiring costly estimation procedures. However, application of bootstrap techniques to high-dimensional inference, especially in the context of PCA, has had limited success. A succinct explanation of the failure of standard non-parametric bootstrap methods in the $p/n \rightarrow c \in (0, \infty)$ setting has recently been given by El Karoui and Purdom [117]. They also showed that, in the case where the population covariance matrix is well-approximated by a finite rank matrix, which corresponds to a spiked model with much larger spiked eigenvalues compared to the noise eigenvalues, the bootstrap performs as well as it does in the finite-dimensional setting. In a complementary study, Lopes *et al.* [118] developed a consistent method for bootstrapping linear spectral statistics of sample covariance matrices by appropriately modifying the usual parametric bootstrap procedure. This method has the salient feature that it allows the user to circumvent the difficulties of complex asymptotic formulas involved in the description of CLT for linear spectral statistics. Development of provably consistent resampling strategies constitutes an exciting new frontier for high-dimensional PCA and related techniques such as MANOVA and CCA.

Acknowledgment

This paper is in part based on a plenary lecture at ICASSP 2015 in Brisbane, Australia by the first author, who thanks Matthew McKay for encouragement and discussion in preparation for that lecture. Both authors thank the referees and Associate Editor for comments that have improved the exposition of the revision. Johnstone's research is funded in part by NSF DMS 1407813 and NIH R01 EB001988. Paul's research is partially supported by the NSF grants DMS-1407530 and DMS-1713120, and the NIH grant 1R01 EB021707.

Appendix A: Heuristic derivation for bias

We work in the simplest setting, with a single spike. We follow, with modifications, the approach of Nadler [42]. Assume that the observations are Gaussian, and that population covariance matrix is diagonal, with a single signal dimension with variance $\ell_1 > 1$, so that $\Sigma = \mathrm{diag}(\ell_1, 1, \ldots, 1)$. In the spiked model, we can achieve this if necessary by a population-level rotation of the variables.

The data matrix **X**, by assumption, has n independent columns, each with mean zero and covariance $\sum$. Now partition

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \end{bmatrix}$$

with the first $1 \times n$ row containing the “signal” observations with elevated variance ℓ_1, and an $(p-1) \times n$ matrix $\mathbf{X}'_2$ containing the noise variables.

Create modified data

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{V}'_2\mathbf{X}'_2 \end{bmatrix}$$

by rotating the noise variables by an orthogonal matrix V_2 obtained from the eigendecomposition

$$n^{-1}\mathbf{X}'_2\mathbf{X}_2 = \mathbf{V}_2 \Lambda \mathbf{V}'_2, \ \Lambda \ = \mathrm{diag}\left(\lambda_2 , \ldots, \lambda_p\right)$$

The first row $\mathbf{X}'_1$ is left alone. Note that the rotation $\mathbf{V}_2 = \mathbf{V}_2(\mathbf{X}_2)$ is data dependent.

In this new basis, the sample covariance has the form of an *arrowhead* matrix. Indeed, with

$$\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{V}'_2\mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2\mathbf{V}_2 \end{bmatrix},$$

we obtain, on defining the scalar $s = n^{-1}\mathbf{X}'_1\mathbf{X}_1$ and vector $\mathbf{b} = \left(b_2, \ldots, b_p\right)' : = n^{-1}\mathbf{V}'_2\mathbf{X}'_2\mathbf{X}_1 \in \mathbb{R}^{p-1}$,

$$\widetilde{\mathbf{S}} = n^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' = \begin{bmatrix} s & \mathbf{b}' \\ \mathbf{b} & \Lambda \end{bmatrix} = \begin{bmatrix} s & b_2 & \cdots & b_p \\ b_2 & \lambda_2 & & \\ \vdots & & \ddots & \\ b_p & & & \lambda_p \end{bmatrix}$$

The shaft of the arrow consists of the sample noise eigenvalues, which are of order 1 because we have normalized the sample covariance matrix.

The border entries b_j (the “head” of the arrow) are much smaller, as we now show. Since $\sum$ is diagonal and the data Gaussian, the first row $\mathbf{X}'_1$ is independent of the noise matrix X'_2 The entries of $\mathbf{X}_1$ are i.i.d. $N(0, \ell)$, so we calculate

$$\begin{aligned} E\left[\mathbf{b}\middle|\mathbf{X}_2\right] &= 0 \\ E\left[\mathbf{b}\mathbf{b}'\middle|\mathbf{X}_2\right] &= n^{-2}\widetilde{\mathbf{X}}'_2E\left[\mathbf{X}_1\mathbf{X}'_1\right]\widetilde{\mathbf{X}}_2 \\ &= n^{-2}\ \ell\ \widetilde{\mathbf{X}}'_2\widetilde{\mathbf{X}}_2 = n^{-1}\ \ell\ \Lambda \end{aligned}$$

Thus, conditional on $\mathbf{X}_2$, each b_j has mean 0 and variance $n^{-1}\ \ell\ \lambda_j$ and so is $O_p(n^{-1/2})$.

Spectrum of arrowhead matrices For an arrowhead matrix, we can solve more or less directly for the eigenvalues and vectors. Indeed, the equation $\widetilde{\mathbf{X}}\mathbf{v} = x\mathbf{v}$ can be written (if we normalize $\mathbf{v}$ by setting $v_1 = 1$) as

$$\begin{aligned} s + b_2 v_2 + \cdots b_p v_p &= x \\ b_2 + \lambda_2 v_2 &= x v_2 \\ b_3 + \lambda_3 v_3 &= x v_3 \\ &\vdots \\ b_p + \lambda_p v_p &= x v_p \end{aligned}$$

From the last $p - 1$ equations, it is immediate that

$$\mathbf{v} \propto \left(1, \frac{b_2}{x - \lambda_2}, \ldots, \frac{b_p}{x - \lambda_p}\right), \quad (12)$$

while the first equation reduces to the secular or characteristic equation[2]

$$f(x) = x - s - \sum_{j=2}^{p} \frac{b_j^2}{x - \lambda_j} = 0. \quad (13)$$

Since the noise eigenvalues λ_j are distinct with probability one, a graph of $f(x)$ against x shows that the sample covariance eigenvalues x_i interleave the λ_j:

$$\lambda_p < x_p < \lambda_{p-1} < \cdots < \lambda_3 < x_2 < \lambda_2 < x_1 .$$

We can now read off the behavior of the top sample eigenvalue from the eigenvalue equation, rewritten in the form

$$x_1 = s + \sum_{j=2}^{p} \frac{b_j^2}{x_1 - \lambda_j} \quad (14)$$

[2] Here is another route to the eigenvalue equation. We can write $\tilde{S} - xI$ as a rank two perturbation of the diagonal matrix of the diagonal matrix $\mathbf{D} = \mathrm{diag}(s - x, \lambda_2 - x, \ldots, \lambda_p - x)$:

$$\tilde{S} - xI = \mathbf{D} + \begin{bmatrix} \mathbf{e}_1 & \mathbf{b} \end{bmatrix} \begin{bmatrix} \mathbf{b}' \\ \mathbf{e}'_1 \end{bmatrix}$$

where, with slight abuse of earlier notation, now $\mathrm{b}' = (0\ b_2 \ldots b_\mathrm{p})$ and $\mathrm{e}'_1 = (1\ 0 \ldots 0)$. Apply the matrix determinant lemma $|\mathbf{D} + \mathbf{U}\mathbf{V}'| = |\mathbf{D}| \left| I + \mathbf{V}'\mathbf{D}^{-1}\mathbf{U} \right|$: the diagonal entries of $\mathbf{V}'\mathbf{D}^{-1}\mathbf{U}$ vanish because of the zero pattern in b and e_1, hence

$$\left| I + \mathbf{V}'\mathbf{D}^{-1}\mathbf{U} \right| = \begin{vmatrix} 1 & b'\mathbf{D}^{-1}\mathrm{b} \\ \mathrm{e}_1'\mathbf{D}^{-1}\mathrm{e}_1 & 1 \end{vmatrix} = 1 - \frac{1}{s - x} \sum_{j=2}^{p} \frac{b_j^2}{\lambda_j - x}.$$

Equivalently the eigenvalues x_i solve the secular equation (13). together with the fact that $s = \ell_1 + O_p\left(n^{-1/2}\right)$ and the b_j have expected square of order $1/n$.

If p is *fixed*, the contribution of the sum is negligible and the leading eigenvalue converges to ℓ and so is consistent:

$$x_1 = \ell_1 + O_p\left(n^{-1/2}\right) + O_p\left(n^{-1}\right)$$

However everything changes if $p/n \to \gamma > 0$. Recalling the behavior when we condition on the noise variables X_2, we have

$$E\left(s \middle| \mathbf{X}_2\right) = \ell\, E\left(b_j^2 \middle| \mathbf{X}_2\right) = \frac{\ell}{n} \lambda_j$$

Proceeding heuristically, the sum on j in (14) now looks like an empirical average of a function of the noise eigenvalues λ_i:

$$x_1 \approx E\left[x_1 \middle| \mathbf{X}_2\right] = \ell + \ell \frac{p}{n} \cdot \frac{1}{p} \sum_{j=2}^{p} \frac{\lambda_j}{x_1 - \lambda_j}$$

Since the empirical distribution of the sample λ_j converges to Marchenko-Pastur, it is plausible and can be shown that as $p/n \to \gamma$, the largest eigenvalue x_1 converges to a limit $\lambda(\ell)$ which satisfies the equation

$$\lambda(\ell) = \ell + \ell\gamma \int \frac{\lambda}{\lambda(\ell) - \lambda} dF_\gamma^{MP}(\lambda) \qquad (15)$$

While the integral can certainly be evaluated directly, it is instructive to consider an alternative indirect approach. The Stieltjes transform of a probability measure, in this case the Marchenko-Pastur law, is defined by

$$m(z) = \int \frac{dF_\gamma^{MP}(\lambda)}{\lambda - z}, \quad z \in \mathbb{C}^+,$$

and is known [51] to satsify the quadratic equation

$$\gamma z m^2(z) + (z + \gamma - 1) m(z) + 1 = 0. \qquad (16)$$

Equation (15) above can be rewritten using the Stieltjes transform as

$$\ell \gamma \lambda m(\lambda) = \ell - \ell \gamma - \lambda, \quad (17)$$

where we abbreviate $\lambda = \lambda(\ell)$. Evidently, we may substitute the latter equation into the former, evaluated at $z = \lambda(\ell)$. When the resulting equation is viewed as a (quadratic) polynomial in λ, it turns out that the constant term vanishes, and so one arrives at the evaluation

$$\lambda(\ell) = \ell + \frac{\ell\gamma}{\ell - 1}. \quad (18)$$

Thus, above the phase transition at $1 + \sqrt{\gamma}$, the bias is given by $\gamma\ell/(\ell - 1)$. So the bias is always at least γ, no matter how large the top population eigenvalue is.

Appendix B: Heuristic derivation for eigenvector inconsistency

Recall that by assumption the top population eigenvectur $\mathbf{u}_1 = \mathbf{e}_1$. Using the explicit form found in (12), one easily calculates the cosine between population and sample as

$$\cos^2\alpha = \frac{\langle \mathrm{v}, \mathrm{e}_1 \rangle^2}{\left\| \mathrm{v} \right\|^2} = \frac{1}{1 + T^2}, \qquad T^2 = \sum_{j=2}^{p} \frac{b_j^2}{(x_1 - \lambda_j)^2} \quad (19)$$

When p is fixed there are a finite number of terms each of order $1/n$, so $\cos^2 \alpha \rightarrow 1$ and the sample eigenvector is consistent.

However, when p is large and proportional to γn, then T^2 converges to a positive constant:

$$T^2 \rightarrow \ell\gamma \int \frac{\lambda}{[\lambda(\ell) - \lambda]^2} dF_\gamma(\lambda) > 0$$

An easy way to evaluate this integral is to observe from (15) and (18) that

$$\int \frac{\lambda}{[\lambda(\ell) - \lambda]} dF_\gamma(t) = \frac{1}{\ell - 1}.$$

Differentiating w.r.t. ℓ, we obtain

$$\lambda'(\ell) \int \frac{\lambda}{[\lambda(\ell) - \lambda]^2} dF_\gamma(\lambda) = \frac{1}{(\ell - 1)^2}.$$

From (18), $\lambda'(\ell) = 1 - \gamma/(\ell - 1)^2$.Substituting into (19), we find that

$$\cos^2\alpha \to \begin{cases} \dfrac{1-\gamma/(\ell-1)^2}{1+\gamma/(\ell-1)} & \ell > 1+\sqrt{\gamma} \\ 0 & \ell \le 1+\sqrt{\gamma}. \end{cases}$$

Biographies

Iain Johnstone received the B.Sc. (Hons) degree in Pure Mathematics and Statistics and M.Sc. degree in Statistics from the Australian National University in 1977 and 1978 respectively, and the Ph.D. degree in Statistics from Cornell University in 1981. Since 1981 he has been Assistant, Associate and then Full Professor in the Department of Statistics at Stanford University. Since 1989, he has also held a 50% time appointment in Biostatistics in the Stanford University School of Medicine. His research in theoretical statistics has used ideas from harmonic analysis, such as wavelets, to understand noise-reduction methods in signal and image processing. More recently, he has applied random matrix theory to the study of high-dimensional multivariate statistical methods, such as principal components and canonical correlation analysis. In biostatistics, he has collaborated with investigators in cardiology and prostate cancer. He is a member of the U.S. National Academy of Sciences and the American Academy of Arts and Sciences and a former president of the Institute of Mathematical Statistics.

Debashis Paul finished his B.Stat. and M.Stat. from Indian Statistical Institute, Kolkata, in 1997 and 1999, respectively, and received his Ph.D. degree in Statistics from Stanford University in 2005 under the supervision of Professor Iain Johnstone. In 2005, he joined the Department of Statistics at the University of California, Davis as Assistant Professor. One of his main research interests is in high-dimensional inference and random matrix theory. He has served on the editorial boards of several journals including Annals of Statistics, Bernoulli and Statistica Sinica.

References

[1]. Pearson K, “On lines and planes of closest fit to systems of points in space,” Philosophical Magazine, vol. 2, no. 6, pp. 559–572, 1901.

[2]. Hotelling H, “Analysis of a complex of statistical variables into principal components,” J. Educational Psychology, vol. 24, pp. 417–441,498–520, 1933.

[3]. Jolliffe IT, Principal Component Analysis, 2nd ed. Springer, 2002.

[4]. Lorenz EN, “Empirical orthogonal functions and statistical weather prediction,” 1956.

[5]. Anderson TW, An Introduction to Multivariate Statistical Analysis, 3rd ed. Wiley, 2003.

[6]. Muirhead RJ, Aspects of Multivariate Statistical Theory. Wiley, 1982.

[7]. Johnstone IM, “On the distribution of the largest eigenvalue in principal components analysis,” Annals of Statistics, vol. 29, pp. 295–327, 2001.

[8]. Bai Z and Yao J, “On sample eigenvalues in a generalized spiked population model,” J. Multivariate Anal, vol. 106, pp. 167–177, 2012.

[9]. Yao J, Zheng S, and Bai Z, Large sample covariance matrices and high-dimensional data analysis, ser Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, 2015, vol. 39.

[10]. Couillet R and Debbah M, Random matrix methods for wireless communications Cambridge University Press, Cambridge, 2011.

[11]. Ramsay JO and Silverman BW, Functional data analysis, 2nd ed., ser. Springer Series in Statistics. Springer, New York, 2005.

[12]. Horváth L and Kokoszka P, Inference for functional data with applications, ser Springer Series in Statistics. Springer, New York, 2012.

[13]. Lu AY, “Sparse principal components analysis for functional data,” Ph.D. dissertation, Department of Statistics, Stanford University, 2002.

[14]. Johnstone IM and Lu AY, “On consistency and sparsity for principal components analysis in high dimensions,” Journal of the American Statistical Association, vol. 104, pp. 682–693, 2009. [PubMed: 20617121]

[15]. Schizas ID and Giannakis GB, “Covariance eigenvector sparsity for compression and denoising,” IEEE Transactions on Signal Processing, vol. 60, no. 5, pp. 2408–2421, 2012.

[16]. Sjöstrand K, Stegmann MB, and Larsen R, “Sparse principal component analysis in medical shape modeling,” in Medical Imaging, Reinhardt JM and Pluim JPW, Eds. SPIE, Mar. 2006, pp. 61 444X–61 444X–12.

[17]. Uhlig H, “On singular Wishart and singular multivariate beta distributions,” The Annals of Statistics, pp. 395–405, 1994.

[18]. Wishart J, “The generalised product moment distribution in samples from a normal multivariate population,” Biometrika, vol. 20A, no. 1/2, pp. 32–52, 1928.

[19]. James AT, “The distribution of the latent roots of the covariance matrix,” Annals of Mathematical Statistics, vol. 31, pp. 151–158, 1960.

[20]. Stein CM, “Some problems in multivariate analysis, part I,” Department of Statistics, Stanford University, Tech. Rep., 1956, Technical Report CHE ONR 6. [Online]. Available: https://statistics.stanford.edu/resources/technical-reports

[21]. James W and Stein C, “Estimation with quadratic loss,” in Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory. University of California Press, 1961, pp. 361–380.

[22]. Marchenko VA and Pastur LA, “Distributions of eigenvalues of some sets of random matrices,” Mathematics of the USSR-Sbornik, vol. 1, pp. 507–536, 1967.

[23]. Silverstein JW and Bai ZD, “On the empirical distribution of eigenvalues of a class of large dimensional random matrices,” Journal of Multivariate Analysis, vol. 54, pp. 175–192, 1995.

[24]. Silverstein JW, “Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices,” J. Multivariate Anal, vol. 55, no. 2, pp. 331–339, 1995 [Online]. Available: 10.1006/jmva.1995.1083.

[25]. Wigner EP, “Characteristic vectors of bordered matrices of infinite dimensions,” Annals of Mathematics, vol. 62, pp. 548–564, 1955.

[26]. Stein CM, “Multivariate analysis I,” Department of Statistics, Stanford University, Tech. Rep., 1969, Technical Report OLK NSF 42. [Online]. Available: https://statistics.stanford.edu/resources/technical-reports

[27]. Baik J, Ben Arous G, and Péché S, “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,” Ann. Probab, vol. 33, no. 5, pp. 1643–1697, 2005.

[28]. Johansson K, “Shape fluctuations and random matrices,” Communications in Mathematical Physics, vol. 209, pp. 437–476, 2000.

[29]. El Karoui N, “On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity,” 2003, arXiv:math.ST/0309355.

[30]. Tracy CA and Widom H, “Level-spacing distributions and the Airy kernel,” Communications in Mathematical Physics, vol. 159, pp. 151–174, 1994.

[31]. —, “On orthogonal and symplectic matrix ensembles,” Communications in Mathematical Physics, vol. 177, pp. 727–754, 1996.

[32]. Bloemendal A and Virág B, “Limits of spiked random matrices I,” Probab. Theory Related Fields, vol. 156, no. 3–4, pp. 795–825, 2013.

[33]. Baik J and Silverstein JW, “Eigenvalues of large sample covariance matrices of spiked population models,” Journal of Multivariate Analysis, vol. 97, pp. 1382–1408, 2006.

[34]. Paul D, “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model,” Statistica Sinica, vol. 17, pp. 1617–1642, 2007.

[35]. Yang J and Johnstone I, “Edgeworth correction for the largest eigenvalue,” Statistica Sinica, 2018, to appear.

[36]. Harding MC, “Explaining the single factor bias of arbitrage pricing models in finite samples,” Economics Letters, vol. 99, no. 1, pp. 85–88, 2008 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165176507002169

[37]. Brown SJ, “The number of factors in security returns,” The Journal of Finance, vol. XLIV, no. 5, pp. 1247–1261, 1989.

[38]. Lu AY, “Sparse principal components analysis for functional data,” Ph.D. dissertation, Stanford University, 2002.

[39]. Johnstone IM and Lu AY, “Sparse principal components analysis,” 2009, arXiv:0901.4392, written 1/1/94.

[40]. Biehl M and Mietzner A, “Statistical mechanics of unsupervised structure recognition,” Journal of Physics A: Mathematical and General, vol. 27, no. 6, pp. 1885–1897, 1994.

[41]. Hoyle DC and Rattray M, “Principal-component-analysis eigenvalue spectra from data with symmetry breaking structure,” Physical Review E, vol. 69, p. 026124, 2004.

[42]. Nadler B, “Finite sample approximation results for principal component analysis: A matrix perturbation approach,” Annals of Statistics, vol. 36, pp. 2791–2817, 2008.

[43]. Mestre X, “On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices,” IEEE Transactions on Signal Processing, vol. 56, no. 11, pp. 5353–5368, 11 2008.

[44]. Benaych-Georges F and Nadakuditi RR, “The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices,” Advances in Mathematics, vol. 227, pp. 494–521, 2011.

[45]. Benaych-Georges F and Nadakuditi RR, “The singular values and vectors of low rank perturbations of large rectangular random matrices,” Journal of Multivariate Analysis, vol. 111, pp. 120–135, 2012.

[46]. Benaych-Georges F, Guionnet A, and Maida M, “Large deviations of the extreme eigenvalues of random deformations of matrices,” Probability Theory and Related Fields, vol. 154, pp. 703–751, 2012.

[47]. Jung S and Marron JS, “PCA consistency in high dimension, low sample size context,” Ann. Statist, vol. 37, no. 6B, pp. 4104–4130, 2009.

[48]. Shen D, Shen H, Zhu H, and Marron JS, “The statistics and mathematics of high dimension low sample size asymptotics,” Statist. Sinica, vol. 26, no. 4, pp. 1747–1770, 2016.

[49]. Fan J, Liao Y, and Mincheva M, “Large covariance estimation by thresholding principal orthogonal complements,” J. R. Stat. Soc. Ser. B. Stat. Methodol, vol. 75, no. 4, pp. 603–680, 2013, with 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva.

[50]. Wang W and Fan J, “Asymptotics of empirical eigenstructure for high dimensional spiked covariance,” Ann. Statist, vol. 45, no. 3, pp. 1342–1374, 2017.

[51]. Bai ZD and Silverstein JW, Spectral Analysis of Large Dimensional Random Matrices. Springer, 2010.

[52]. Soshnikov A, “A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices,” Journal of Statistical Physics, vol. 108, pp. 1033–1056, 2002.

[53]. Péché S, "Universality results for the largest eigenvalues of some sample covariance matrix ensembles," Probability Theory and Related Fields, vol. 143, pp. 481–516, 2009.

[54]. Baik J and Silverstein JW, "Eigenvalues of large sample covariance matrices of spiked population models," Journal of Multivariate Analysis, vol. 97, pp. 1382–1408, 2006.

[55]. Bai ZD and Yao J, "Central limit theorems for eigenvalues in a spiked population model," Annales de l'Institut Henri Poincaré - Probabilités et Statstiques, vol. 44, pp. 447–474, 2008.

[56]. Bai ZD and Yao J, "On sample eigenvalues in a generalized spiked population model," Journal of Multivariate Analysis, vol. 106, pp. 167–177, 2012.

[57]. Erdös L, Schlein B, and Yau H-T, "Local semicircle law and complete delocalization for Wigner random matrices," Communications in Mathematical Physics, vol. 287, pp. 641–655, 2009.

[58]. Erdös L and Yau H-T, "Universality of local spectral statistics of random matrices," Bulletin of the American Mathematical Society, vol. 49, pp. 377–414, 2012.

[59]. Erdös L, Yau H-T, and Yin J, "Rigidity of eigenvalues of generalized Wigner matrices," Advances in Mathematics, vol. 229, pp. 1435–1515, 2012.

[60]. Erdös L, Schlein B, Yau H-T, and Yin J, "The local relaxation flow approach to universality of the local statistics for random matrices," Annales Institute H. Poincaré (B) Probability and Statistics, vol. 48, pp. 1–46, 2012.

[61]. Tao T and Vu V, "Random matrices: universality of local eigenvalue statistics up to the edge," Communications in Mathematical Physics, vol. 298, pp. 549–572, 2010.

[62]. Tao T and Vu V, "Random matrices: universality of local eigenvalue statistics," Acta Mathematica, vol. 206, pp. 127–204, 2011.

[63]. Tao T and Vu V, "Random covariance matrices: universality of local statistics of eigenvalues," Annals of Probability, vol. 40, pp. 1285–1315, 2012.

[64]. Pillai NS and Yin J, "Universality of covariance matrices," Annals of Applied Probability, vol. 24, pp. 935–1001, 2014.

[65]. Féral D and Péché S, "The largest eigenvalues of sample covariance matrices for a spiked population: diagonal case," Journal of Mathematical Physics, vol. 50, p. 073302, 2009.

[66]. Bloemendal A, Knowles A, Yau H-T, and Yin J, "On the principal components of sample covariance matrices," Probability Theory and Related Fields, vol. 164, pp. 459–552, 2016.

[67]. Bao Z, Pan GM, and Zhou W, "Universality for the largest eigenvalue of sample covariance matrices with general population," Annals of Statistics, vol. 43, pp. 382–421, 2015.

[68]. Lee JO and Schnelli K, "Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population," Annals of Applied Probability, vol. 26, pp. 3786–3839, 2016.

[69]. Silverstein JW, "Some limit theorems on the eigenvectors of large dimensional sample covariance matrices," Journal of Multivariate Analysis, vol. 15, pp. 295–324, 1984.

[70]. Birnbaum A, Johnstone IM, Nadler B, and Paul D, "Minimax bounds for sparse PCA with noisy high-dimensional data," Annals of Statistics, vol. 41, pp. 1055–1084, 2013. [PubMed: 25324581]

[71]. Cai TT, Ma Z, and Wu Y, "Sparse PCA: optimal rates and adaptive estimation," Annals of Statistics, vol. 41, pp. 3074–3110, 2013.

[72]. Ma Z, "Sparse principal component analysis and iterative thresholding," Annals of Statistics, vol. 41, pp. 772–801, 2013.

[73]. Vu V and Lei J, "Minimax sparse principal subspace estimation in high dimensions," Annals of Statistics, vol. 41, pp. 2905–2947, 2013.

[74]. Cai T, Ma Z, and Wu Y, "Optimal estimation and rank detection for sparse spiked covariance matrices," Probab. Theory Related Fields, vol. 161, no. 3–4, pp. 781–815, 2015.

[75]. El Karoui N, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," Annals of Statistics, vol. 36, pp. 2757–2790, 2008.

[76]. Ledoit O and Wolf M, "Nonlinear shrinkage estimation of large-dimensional covariance matrices," Annals of Statistics, vol. 40, pp. 1024–1060, 2012.

[77]. Donoho DL, Gavish M, and Johnstone IM, "Optimal shrinkage of eigenvalues in the spiked covariance model," Annals of Statistics, 2018, to appear.

[78]. Stein C, “Lectures on the theory of estimation of many parameters,” Journal of Mathematical Sciences, vol. 34, no. 1, pp. 1373–1403, 1986.

[79]. Mauchly JW, “Significance test for sphericity of a normal n-variate distribution,” Annals of Mathematical Statistics, vol. 11, pp. 204–209, 1940.

[80]. John S, “Some optimal multivariate tests,” Biometrika, vol. 58, pp. 123–127, 1971.

[81]. John S, “The distribution of a statistic used for testing sphericity of normal distributions,” Biometrika, vol. 59, pp. 169–173, 1972.

[82]. Sugiura N, “Locally best invariant test for sphericity and the limiting distributions,” Annals of Mathematical Statistics, vol. 43, pp. 1312–1316, 1972.

[83]. Ledoit O and Wolf M, “Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size,” Annals of Statistics, vol. 30, pp. 1081–1102, 2002.

[84]. Bai ZD, Jiang D, Yao J, and Zheng S, “Corrections to LRT on large-dimensional covariance matrix by RMT,” Annals of Statistics, vol. 37, pp. 3822–3840, 2009.

[85]. Pillai KCS, “Some new test criteria in multivariate analysis,” Annals of Mathematical Statistics, vol. 26, pp. 117–121, 1955.

[86]. Pillai KCS, “On the distribution of the largest or smallest root of a matrix in multivariate analysis,” Biometrika, vol. 43, pp. 122–127, 1956.

[87]. Pillai KCS, “On the distribution of the largest characteristic root of a matrix in multivariate analysis,” Biometrika, vol. 52, pp. 405–414, 1965.

[88]. Johnstone IM, “Approximate null distribution of the largest root in multivariate analysis,” Annals of Applied Statistics, vol. 3, pp. 1616–1633, 2009. [PubMed: 20526465]

[89]. Ma Z, “Accuracy of the Tracy-Widom limits for the extreme eigenvalues in white Wishart matrices,” Bernoulli, vol. 18, no. 1, pp. 322–359, 2012.

[90]. Bianchi P, Debbah M, Maïda M, and Najim J, “Performance of statistical tests for single source detection using random matrix theory,” IEEE Transactions on Information Theory, vol. 57, pp. 2400–2419, 2011.

[91]. Kritchman S and Nadler B, “Determining the number of components in a factor model from limited noisy data,” Chemometrics and Intelligent Laboratory Systems, vol. 94, pp. 19–32, 2008.

[92]. Nadler B, Penna F, and Garello R, “Performance of eigenvalue-based signal detectors with known and unknown noise level,” in IEEE International Conference on Communications ICC2011, 2011.

[93]. Onatski A, “Testing hypotheses about the number of factors in large factor models,” Econometrica, vol. 77, pp. 1447–1479, 2009.

[94]. Onatski A, Moreira MJ, and Hallin M, “Asymptotic power of sphericity tests for high-dimensional data,” Annals of Statistics, vol. 41, pp. 1204–1231, 2013.

[95]. Johnstone IM and Onatski A, “Testing in high-dimensional spiked models,” 2015, arXiv: 1509.07269.

[96]. Dharmawansa P, Johnstone IM, and Onatski A, “Local asymptotic normality of the spectrum of high-dimensional spiked F-ratios,” 2014, arXiv:1411.3875.

[97]. Johnstone IM and Nadler B, “Roys largest root test under rank-one alternatives,” Biometrika, vol. 104, no. 1, p. 181, 2017. [PubMed: 29430030]

[98]. Vaswani N and Guo H, “Correlated-PCA: principal components analysis when data and noise are correlated,” Advances in Neural Information Processing Systems, pp. 1768–1776, 2016.

[99]. Vaswani N and Narayanamurthy P, “Finite sample guarantees for PCA in non-isotropic and data-dependent noise,” 2017, arXiv:1709.06255.

[100]. Koltchinskii V and Lounici K, “Normal approximation and concentration of spectral projectors of sample covariance,” Annals of Statistics, vol. 45, pp. 121–157, 2017.

[101]. Koltchinskii V and Lounici K, “New asymptotic results in principal component analysis,” Sankhya A, vol. 79, pp. 254–297, 2017.

[102]. Mas A and Ruymgaart F, “High-dimensional principal projections,” Complex Analysis and Operator Theory, vol. 9, pp. 35–63, 2015.

[103]. Reiss M and Wahl M, “Non-asymptotic upper bounds for the reconstruction error of PCA,” 2016, arXiv:1609.03779v2.

[104]. Forni M, Hallin M, Lippi M, and Reichlin L, “The generalized dynamic-factor model: Identification and estimation,” Review of Economics and Statistics, vol. 82, pp. 540–554, 2000.

[105]. Jin B, Wang C, Bai ZD, Nair KK, and Harding M, “Limiting spectral distribution of a symmetrized auto-cross covariance matrix,” Annals of Applied Probability, vol. 24, pp. 1199–1225, 2014.

[106]. Liu H, Aue A, and Paul D, “On the Marčenko?Pastur law for linear time series,” Annals of Statistics, vol. 43, pp. 675–712, 2015.

[107]. Bhattacharjee M and Bose A, “Large sample behaviour of high dimensional autocovariance matrices,” Annals of Statistics, vol. 44, pp. 598–628, 2016.

[108]. Li Z, Pan G, and Yao J, “On singular value distribution of large-dimensional autocovariance matrices,” Journal of Multivariate Analysis, vol. 137, pp. 119–140, 2015.

[109]. Li Z, Wang Q, and Yao J, “Identifying the number of factors from singular values of a large sample auto-covariance matrix,” Annals of Statistics, vol. 45, pp. 257–288, 2017.

[110]. Namdari J, “Estimation of spectral distributions of a class of high-dimensional linear processes,” Ph.D. dissertation, University of California, Davis, 2018.

[111]. Paul D and Wang L, “Discussion of “Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation”,” Electronic Journal of Statistics, vol. 10, pp. 74–80, 2016.

[112]. Fan Z and Johnstone IM, “Eigenvalue distributions of variance components estimators in high-dimensional random effects models,” 2016, arXiv:1607.02201.

[113]. Fan Z and Johnstone IM, “Tracy-Widom at each edge of real covariance estimators,” 2017, arXiv:1707.02352.

[114]. Dobriban E, “Sharp detection in PCA under correlations: all eigenvalues matter,” Annals of Statistics, vol. 45, pp. 1810–1833, 2017.

[115]. Zheng S, Bai ZD, and Yao J, “Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing,” Annals of Statistics, vol. 43, pp. 546–591, 2015.

[116]. James AT, “Distributions of matrix variates and latent roots derived from normal samples,” Annals of Mathematical Statistics, vol. 35, pp. 475–501, 1964.

[117]. El Karoui N and Purdom E, “The bootstrap, covariance matrices and PCA in moderate and high-dimensions,” 2016, arXiv:1608.00948.

[118]. Lopes M, Blandino M, and Aue A, “Bootstrapping spectral statistics in high dimensions,” 2017, arXiv:1709.08251.

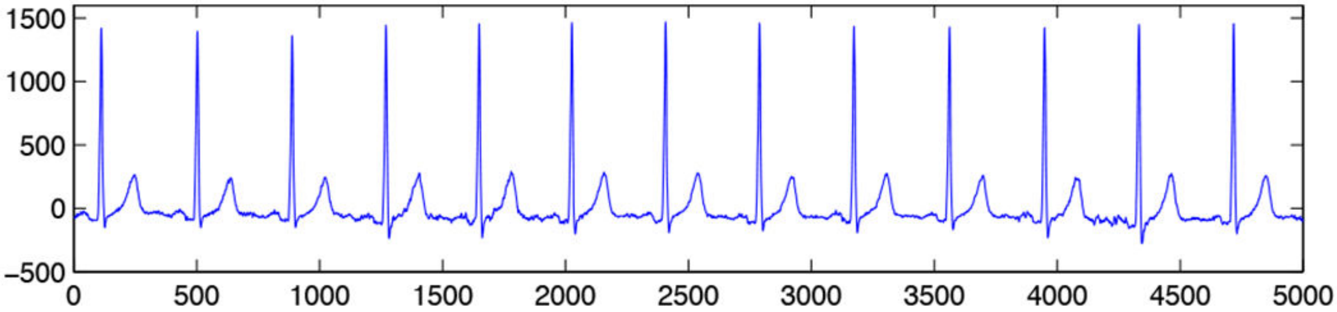

Fig. 1.

sample from an ECG trace sampled at 500 Hz, via Jeffrey Froning and Victor Froelicher, then at cardiology group at Palo Alto Veterans Affairs Hospital.

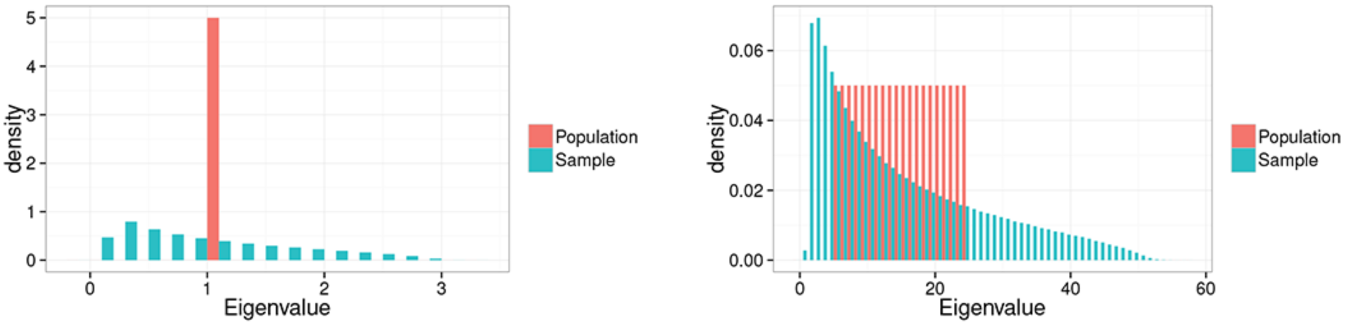

Fig. 2.

Eigenvalue spreading, for $p = 100$, $n = 200$. Population eigenvalues shown as histograms in red: left all at 1, $\sum = I$, right equally spaced on [5, 25]: $\sum = \text{diag}(25,..., 5)$. Corresponding histograms of sample eigenvalues shown in blue. Figure credit: Brett Naul.

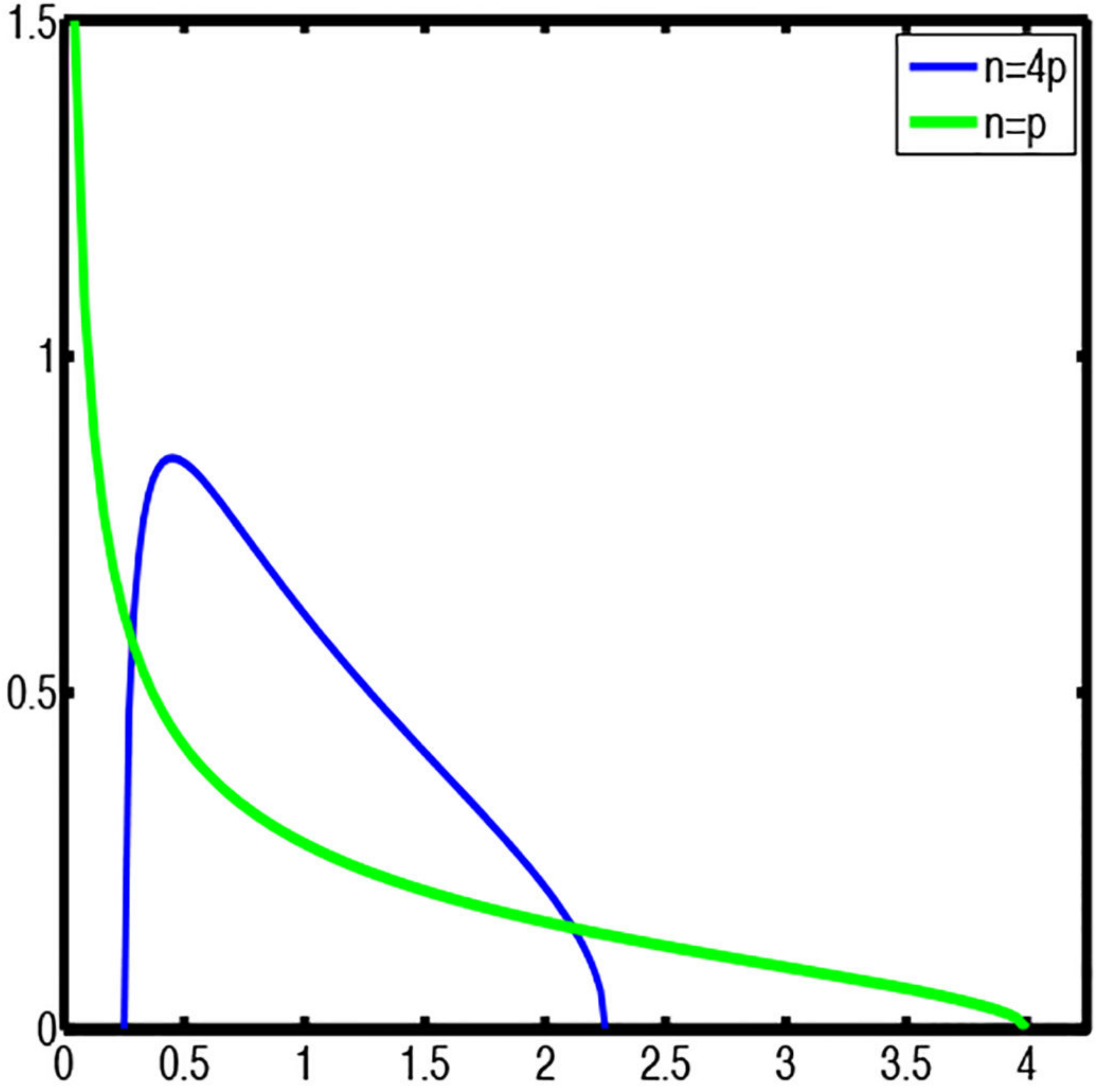

Fig. 3.

Two instances of the Marchenko-Pastur quarter circle law $f^{MP}(x)$ from (3): blue for $\gamma = p/n = 1/4$, green for $\gamma = 1$.

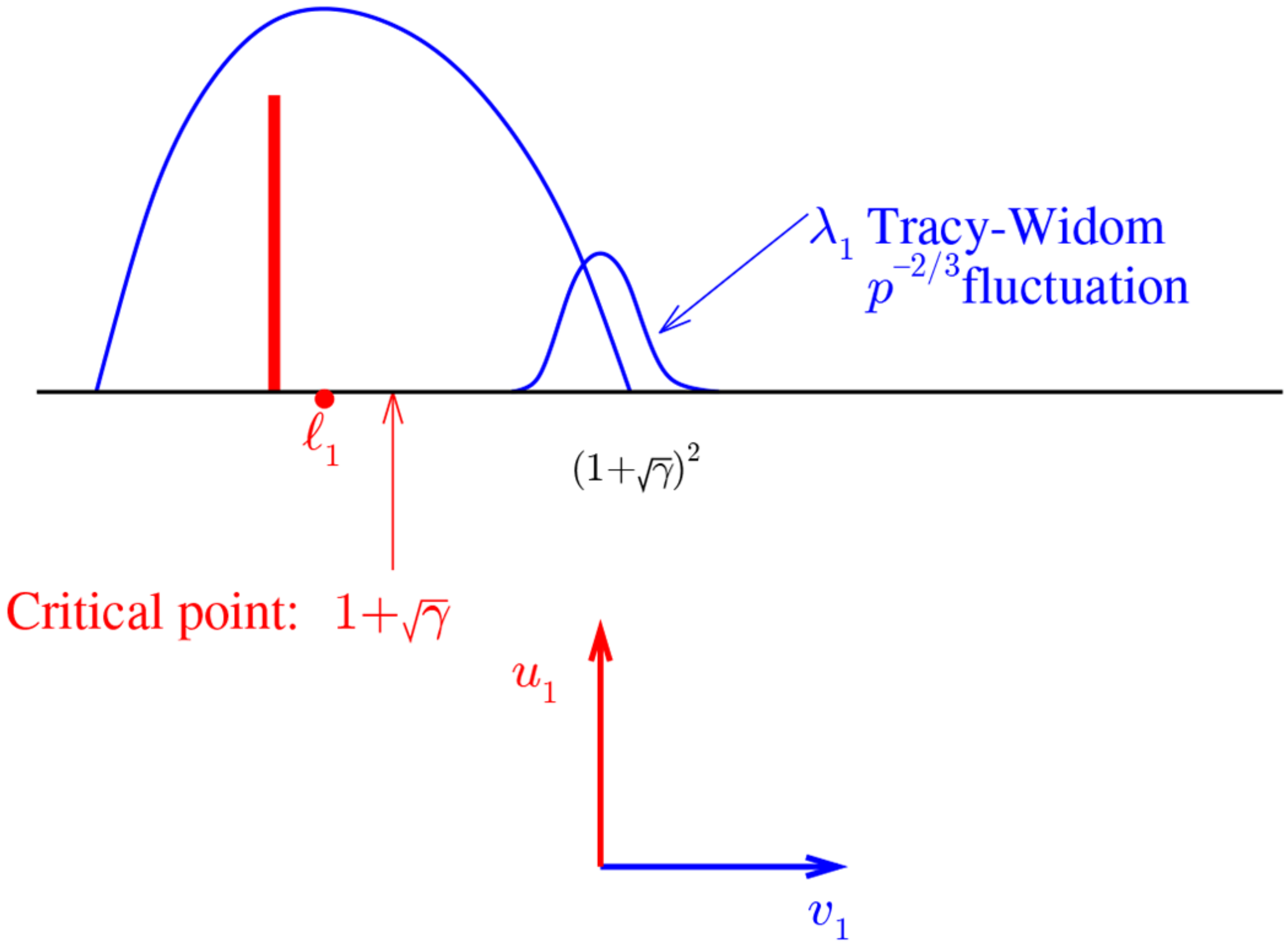

Fig. 4.

Below phase transition: Population quantities are in red, sample ones in blue. All population eigenvalues equal 1 except perhaps for the top one, ℓ_1. Below the critical value $1 + \sqrt{\gamma}$,, the value of ℓ_1 has no effect on the limiting distribution of λ_1, the Tracy-Widom distribution as in (5).

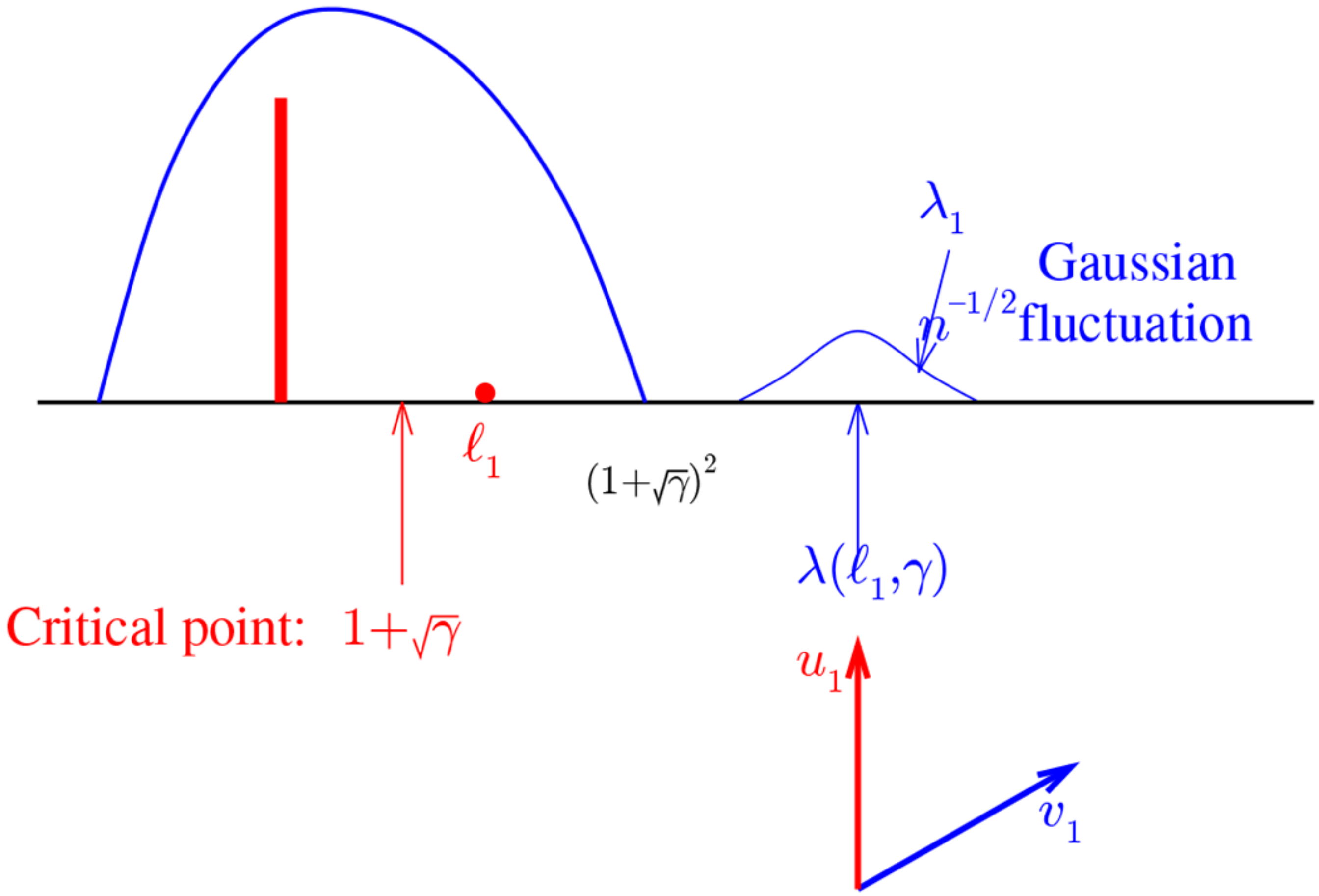

Fig. 5.
Above the phase transition: for $\ell_1 > 1 + \sqrt{\gamma}$, the limiting distribution of λ_1 is now Gaussian, as in (6).

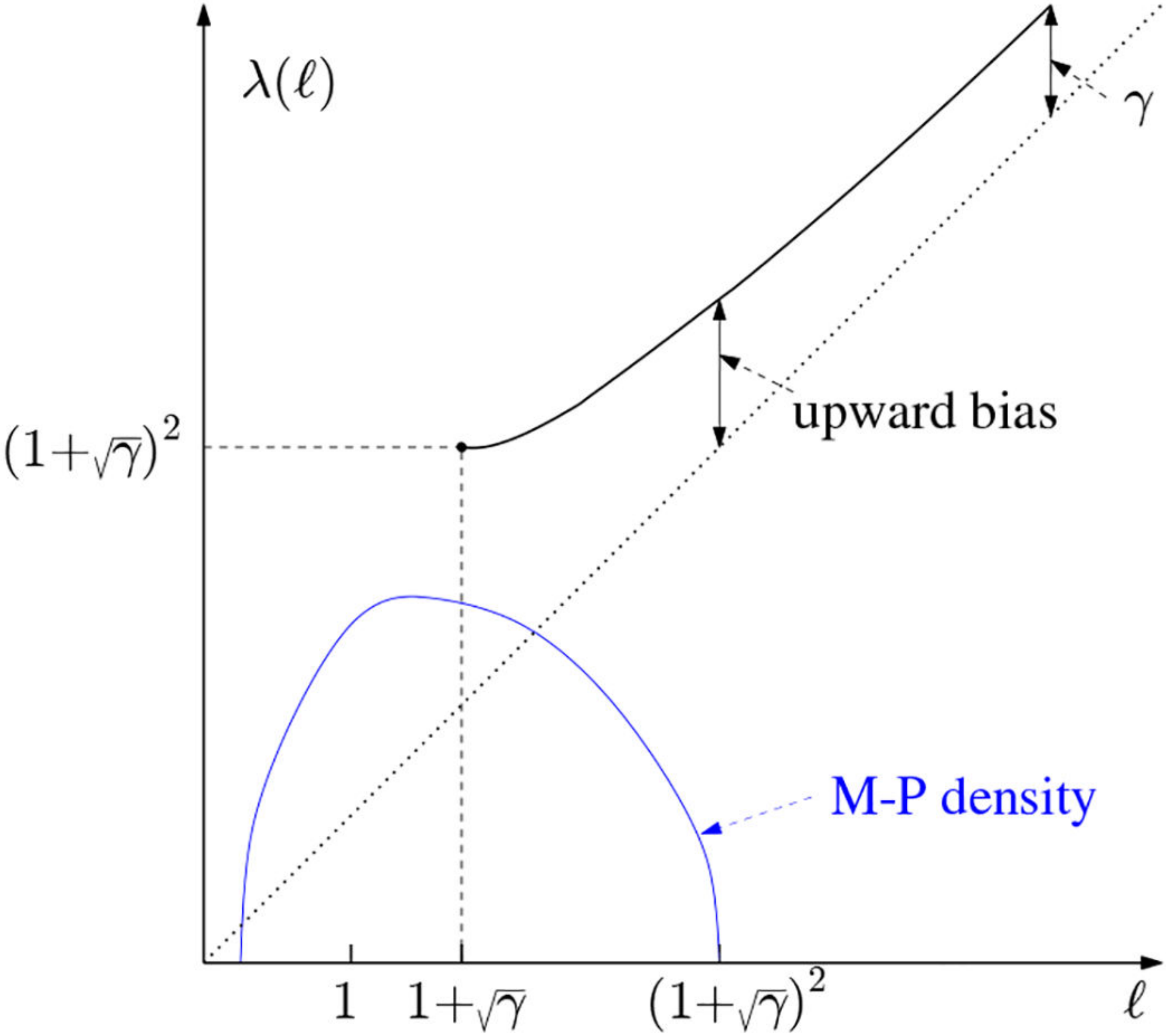

Fig. 6.

Upward bias of (asymptotic) mean of λ_1, denoted by $\lambda(\ell) = \lambda(\ell, \gamma)$ in (7) above, for $\ell > 1 + \sqrt{\gamma}$. Marchenko-Pastur density shown for reference in blue – the phase transition point is well inside the bulk.

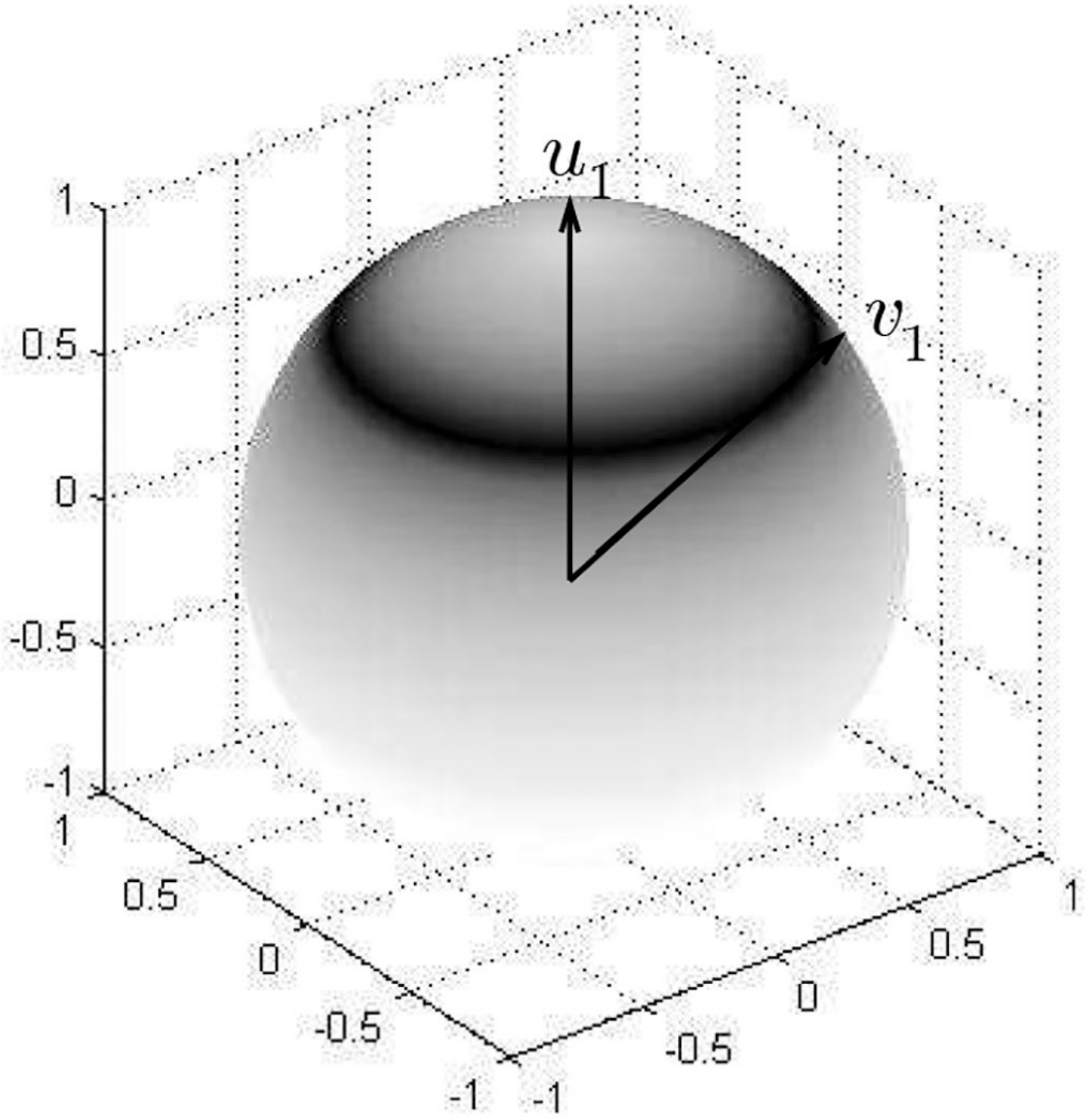

Fig. 7.

Inconsistency of the top sample eigenvector $\mathbf{v}_1$ for estimating the top population eigenvector $\mathbf{u}_1$ in the spiked covariance model when $p \propto n$. See (8).

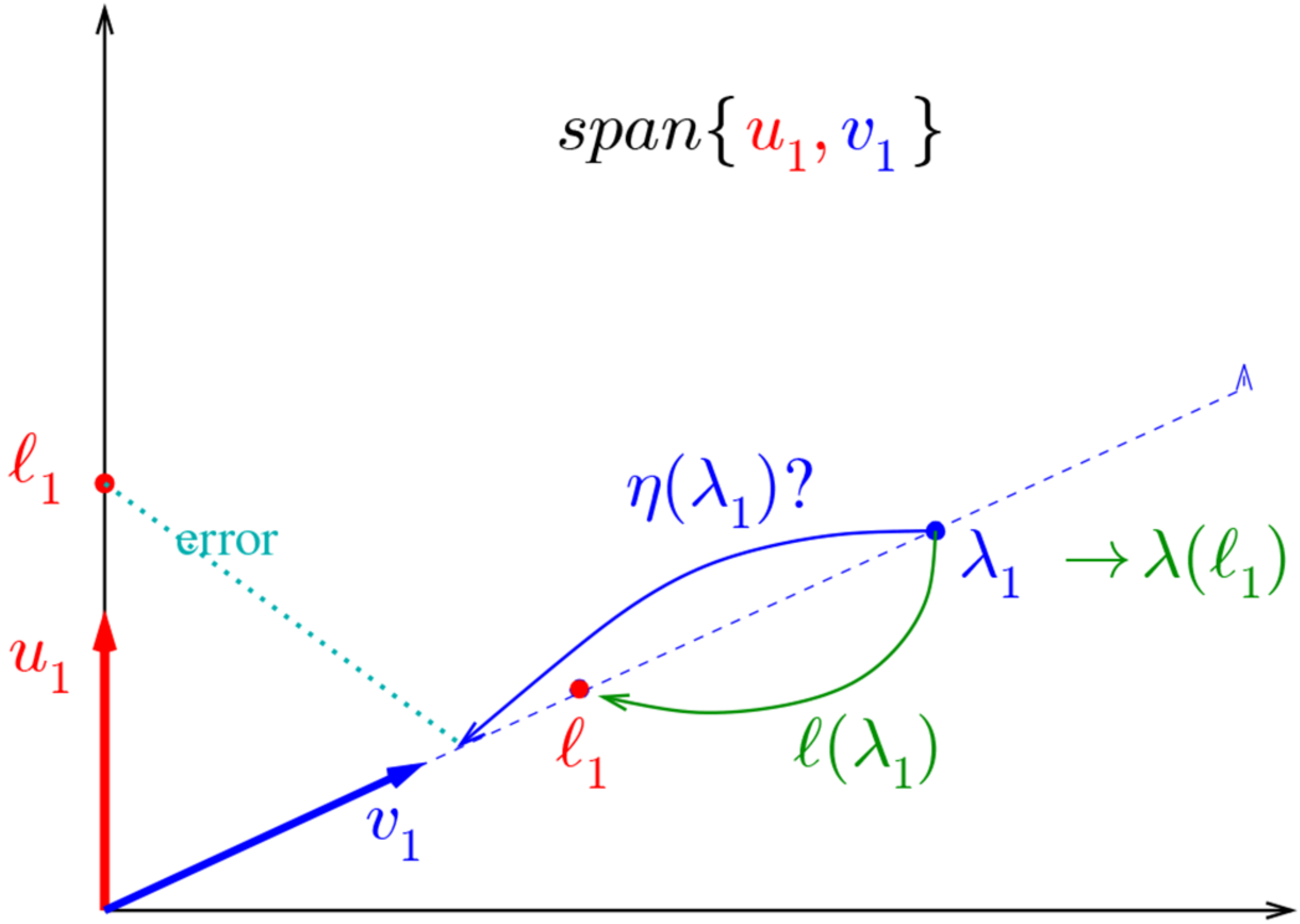

Fig. 8.

Schematic to motivate dependence of optimal shrinkage on the error measure. The observed sample eigenvalue λ_1 is shrunk by $\eta(\lambda_1)$ along sample eigenvector $\mathbf{v}_1$. Since $\mathbf{v}_1$ is necessarily mis-aligned with the truth $\mathbf{u}_1$, the error incurred depends on the metric used.

TABLE I

Optimal shrinkage functions $\eta^*(\lambda; \gamma)$ for a selelection of loss functions under a spiked covariance model [77]. We set $c = c(\ell) = \sqrt{(1 - \gamma/(\ell - 1)^2)/(1 + \gamma/(\ell - 1))}$ and $s = \sqrt{1 - c^2}$.. Here $\ell = \ell(\lambda), c = c(\ell(\lambda))$ and $s = s(\ell(\lambda))$ depend on λ through (9) and implicitly also on γ. Values shown are shrinkers for $\lambda > \lambda_+(\gamma)$; all shrinkers satisfy $\eta^*(\lambda) = 1$ for $\lambda \le \lambda_+(\gamma)$.

Loss function	**Form of $L_p(A,B)$**	**Optimal shrinker $\eta^*(\lambda; \gamma)$**
Operator loss	$\|A - B\|_{op}$	ℓ
Frobenius loss	$\|A - B\|_F^2$	$\ell c^2 + s^2$
Entropy loss	$\frac{1}{2}(\mathrm{trace}(B^{-1}A - I) - \log(\|A\|/\|B\|))$	$\ell c^2 + s^2$
Frobenius loss on precision	$\|A^{-1} - B^{-1}\|_F^2$	$\ell/(c^2 + \ell s^2)$
Stein loss	$\frac{1}{2}(\mathrm{trace}(A^{-1}B - I) - \log(\|B\|/\|A\|))$	$\ell/(c^2 + \ell s^2)$
Fréchet loss	$\mathrm{trace}(A + B) - 2(\sqrt{A}\sqrt{B})$	$(s^2 + \sqrt{\ell c^2})^2$