OPTIMAL SHRINKAGE OF EIGENVALUES IN THE SPIKED COVARIANCE MODEL

By

David L. Donoho
Matan Gavish
Iain M. Johnstone

Technical Report No. 2013-10
November 2013

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

OPTIMAL SHRINKAGE OF EIGENVALUES IN THE SPIKED COVARIANCE MODEL

By

David L. Donoho
Matan Gavish
Iain M. Johnstone
Stanford University

Technical Report No. 2013-10
November 2013

This research was supported in part by National Science Foundation grant DMS 0906812.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model

David L. Donoho * Matan Gavish * Iain M. Johnstone *

Abstract

Since the seminal work of Stein (1956) it has been understood that the empirical covariance matrix can be improved by shrinkage of the empirical eigenvalues. In this paper, we consider a proportional-growth asymptotic framework with n observations and p_n variables having limit $p_n/n \to \gamma \in (0,1]$. We assume the population covariance matrix Σ follows the popular spiked covariance model, in which several eigenvalues are significantly larger than all the others, which all equal 1. Factoring the empirical covariance matrix S as $S = V\Lambda V'$ with V orthogonal and Λ diagonal, we consider shrinkers of the form $\hat{\Sigma} = \eta(S) = V\eta(\Lambda)V'$ where $\eta(\Lambda)_{ii} = \eta(\Lambda_{ii})$ is a scalar nonlinearity that operates individually on the diagonal entries of Λ. Many loss functions for covariance estimation have been considered in previous work. We organize and amplify the list, and study 26 loss functions, including Stein, Entropy, Divergence, Fréchet, Bhattacharya/Matusita, Frobenius Norm, Operator Norm, Nuclear Norm and Condition Number losses. For each of these loss functions, and each suitable fixed nonlinearity η, there is a strictly positive asymptotic loss which we evaluate precisely. For each of these 26 loss functions, there is a *unique admissible* shrinker dominating all other shrinkers; it takes the form $\hat{\Sigma}^* = V\eta^*(\Lambda)V'$ for a certain loss-dependent scalar nonlinearity $\eta^* = \eta^*(\,\cdot\,|\gamma, Loss)$, which we characterize. For 17 of these loss functions, we derive a simple analytical expression for the optimal nonlinearity η^*; in all cases we tabulate the optimal nonlinearity and provide software to evaluate it numerically on a computer. We also tabulate the asymptotic slope $\lim_{\lambda\to\infty} \frac{\eta^*(\lambda)}{\lambda}$ and, where relevant, the asymptotic shift $\lim_{\lambda\to\infty}(\eta^*(\lambda) - \lambda)$ of the optimal nonlinearity.

Key Words. Covariance Estimation, Precision Estimation, Optimal Nonlinearity, Stein Loss, Entropy Loss, Divergence Loss, Fréchet Distance, Bhattacharya/Matusita Affinity, Quadratic Loss, Condition Number Loss, High-Dimensional Asymptotics, Spiked Covariance, Principal Component Shrinkage

Acknowledgements. Work partially supported by NSF DMS-0906812 (ARRA). Thanks to Amit Singer, Andrea Montanari, Sourav Chatterjee and Boaz Nadler for helpful discussions. MG was partially supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship.

*Department of Statistics, Stanford University

Contents

1 Introduction

Suppose we observe p-dimensional vectors $X_i \overset{i.i.d}{\sim} N(0, \Sigma_p)$, $i = 1, \ldots, n$, with Σ_p the underlying p-by-p population covariance matrix. To estimate Σ_p, we form the empirical covariance matrix $S \equiv S_{n,p} \equiv n^{-1} \sum_{i=1}^{n} X_i X_i'$; this is the maximum likelihood estimator. Stein [57, 58] observed that the maximum likelihood estimator S ought to be improvable by eigenvalue shrinkage.

Write $S_{n,p} = V \Lambda V'$ for the eigendecomposition of $S_{n,p}$, where V is orthogonal and the diagonal matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ contains the empirical eigenvalues. Stein proposed to shrink the eigenvalues by applying a specific univariate nonlinearity $\eta : [0, \infty) \to \mathbb{R}$ to each eigenvalue of $S_{n,p}$, producing the estimate $\hat{\Sigma}_\eta = V \eta(\Lambda) V'$, where $\eta(\Lambda)$ denotes the application of η entry-wise to the diagonal of Λ. In the ensuing half century, research on eigenvalue shrinkers has flourished, producing an extensive literature; we can point here only to a fraction of this literature, with pointers organized into early decades [29, 21, 26, 27, 7, 25], the middle decades [13, 55, 56, 38, 37, 46, 36, 51, 62, 24, 45], and the last decade [12, 42, 59, 28, 34, 41, 43, 22, 11, 61]. Papers in this literature typically choose some loss function $L_p : S_p^+ \times S_p^+ \to [0, \infty)$, where S_p^+ is the space of positive semidefinite p-by-p matrices, and develop a shrinker η with "favorable" risk $\mathbb{E}\, L_p \left(\Sigma_p \,, \hat{\Sigma}_\eta(S_{n,p}) \right)$.

In high dimensional problems, p and n are often of comparable magnitude. There, the maximum likelihood estimator is no longer a reasonable choice for covariance estimation and the need to shrink becomes acute.

In this paper, we consider a popular large n, large p setting with p comparable to n, and a popular set of assumptions about Σ known as the *Spiked Covariance Model* [30]. We consider a variety of loss functions derived from or inspired by the literature, and show that to each "reasonable" nonlinearity η there corresponds a well-defined asymptotic loss.

In the sibling problem of matrix denoising under a similar setting, it has been shown that there exists a unique asymptotically admissible shrinker [54, 15]. The same phenomenon is shown to exist here: for many different loss functions, we show that there exists a *unique optimal nonlinearity* η^*, which we explicitly provide. Perhaps surprisingly, η^* is the only asymptotically admissible nonlinearity, namely, it offers equal or better asymptotic loss than that of any other choice of η, across all possible spiked covariance models.

1.1 Estimation in the Spiked Covariance Model

Consider a sequence of covariance estimation problems, satisfying the following assumption:

Assumption Asy(γ). The number of observations n and the number of variables p_n in the n-th problem follows the proportional-growth limit $p_n/n \to \gamma$, as $n \to \infty$, for a certain $0 < \gamma \leq 1$.

Denote the population covariance and sample covariance in the n-th problem by Σ_{p_n} and S_{n,p_n}, respectively. Further assume that the eigenvalues of Σ_{p_n} (the population eigenvalues in the n-th problem) satisfy the following assumption:

Assumption Spike($\ell_1, \ldots, \ell_r$). The theoretical (namely, population) eigenvalues in the n-th problem are given by $(\ell_1, \ldots, \ell_r, 1, 1, \ldots, 1)$, where the number of "spikes" r and their amplitudes $\ell_1 \geq \ldots \geq \ell_r \geq 1$ are fixed independently of n and p_n.

Assumptions **[Asy(γ)]** and **[Spike($\ell_1, \ldots, \ell_r$)]** together form the asymptotic model known as the Spiked Covariance Model (or simply the *spiked model*) [30]. Recent results on the spiked model have shown that when $\ell_r > (1+\sqrt{\gamma})$, various regularities among the top r principal components quickly set in as n increases [5, 52, 6, 3, 4]:

(i) Each of the largest r eigenvalues of the sample covariance matrix S_{n,p_n}, call it $\lambda_{i,n}$, tends to a deterministic limit. Specifically, $\lambda_{i,n} \to_P \lambda(\ell_i)$, where

$$\lambda(\ell) \quad = \quad \ell \cdot (1 + \gamma/(\ell - 1)) \,. \tag{1.1}$$

(Here and throughtout the paper, $\to_P$ and $=_P$ denote convergence in probability.)

(ii) Each of the top r eigenvectors of the sample covariance matrix S_{n,p_n}, call it $u_{i,n}$, makes asymptotically deterministic angles with the top r population eigenvectors, call them $v_{1,n}, \ldots, v_{p,n}$. Specifically, $|\langle u_{i,n}, v_{j,n}\rangle| \to_P 0$ if $i \neq j$ and $|\langle u_{i,n}, v_{i,n}\rangle| \to_P c(\ell_i)$, where

$$c(\ell) \quad = \quad \sqrt{\frac{1-\gamma/(\ell-1)^2}{1+\gamma/(\ell-1)}} \,. \tag{1.2}$$

As a result, random variables that depend on the sample covariance matrix S_{n,p_n} only through the top empirical eigenvalues and the angles between corresponding sample and population eigenvectors converge to deterministic quantities.

Consider a sequence of loss functions $L = \{L_p\}$ and a fixed nonlinearity $\eta : [0,\infty) \to \mathbb{R}$. Define the asymptotic (L-) loss of the shrinkage estimator $\hat{\Sigma}_\eta \,:\, S_{n,p_n} \mapsto V\eta(\Lambda)V'$ in the spiked model satisfying assumption **[Spike($\ell_1, \ldots, \ell_r$)]** by

$$L_\infty(\eta|\ell_1,\ldots,\ell_r) = \lim_{n\to\infty} L_{p_n}\left(\Sigma_{p_n}\,,\, \hat{\Sigma}_\eta(S_{n,p_n})\right)\,, \tag{1.3}$$

assuming such limit exists. If a nonlinearity η^* satisfies

$$L_\infty(\eta^*|\ell_1,\ldots,\ell_r) \leq L_\infty(\eta|\ell_1,\ldots,\ell_r)$$

for any other nonlinearity η, any r and any spikes $\ell_1, \ldots, \ell_r$, and if the inequality is strict at some choice of $\ell_1, \ldots, \ell_r$, then we say that η^* is the *unique asymptotically admissible* nonlinearity (nicknamed "optimal") for the loss sequence L.

1.2 Some Optimal Shrinkers

This paper identifies the optimal nonlinearity for each of 26 loss functions found in, or inspired by, the covariance estimation literature. Most of these nonlinearities have simple closed-form expressions in terms of the functions $\lambda \mapsto \ell(\lambda)$, the inverse map of (1.1), and of $\lambda \mapsto c(\ell(\lambda))$, a composition of $\ell(\lambda)$ with (1.2).

Operator Loss is given by $L(A,B) = \|A-B\|_{op}$ where $\|\cdot\|_{op}$ is the operator norm (namely, the maximal singular value). The unique asymptotically admissible nonlinearity for Operator loss is simply *debiasing*, namely, it shrinks each empirical eigenvalue back to the location of its corresponding population eigenvalue:

$$\eta^*(\lambda;\gamma) = \ell(\lambda), \qquad \lambda > (1+\sqrt{\gamma})^2.$$

Frobenius Loss and Entropy Loss. Frobenius (or Squared Error) loss is given by $L(A,B) = \|A - B\|_F^2$ where $\|\cdot\|_F$ is the Frobenius matrix norm. Entropy loss is given by $L(A,B) = (\mathrm{trace}(B^{-1}A - I) - \log(|A|/|B|))/2$. The unique asymptotically admissible nonlinearity for both Frobenius loss and Entropy loss is

$$\eta^*(\lambda;\gamma) = 1 + (\ell(\lambda) - 1)\cdot c^2(\lambda), \qquad \lambda > (1+\sqrt{\gamma})^2.$$

Frobenius Loss on Precision and Stein Loss. Frobenius loss on precision matrices is given by $L(A,B) = \|A^{-1} - B^{-1}\|_F^2$. Stein loss is given by $L(A,B) = (\mathrm{trace}(A^{-1}B - I) - \log(|B|/|A|))/2$. The unique asymptotically admissible nonlinearity for both Frobenius loss on precision and Stein loss is

$$\eta^*(\lambda;\gamma) = \frac{\ell(\lambda)}{c^2(\lambda) + \ell(\lambda)\cdot(1 - c^2(\lambda))}, \qquad \lambda > (1+\sqrt{\gamma})^2.$$

Fréchet Loss is given by $L(A,B) = \mathrm{trace}(A + B - 2\sqrt{A}\sqrt{B})$. The unique asymptotically admissible nonlinearity for Frechet loss is

$$\eta^*(\lambda;\gamma) = \left(1 - c^2(\lambda) + \sqrt{\ell(\lambda)}\cdot c^2(\lambda)\right)^2, \qquad \lambda > (1+\sqrt{\gamma})^2.$$

All the 26 optimal nonlinearities we study, including the ones above, are depicted in Figure 1.

1.3 Three Key Observations

In what follows, we construct a framework for evaluating the asymptotic loss (1.3). Our framework makes several assumptions about the nonlinearity η and the loss function L, and then exploits three observations flowing from those assumptions. The observations are:

Obs. 1: Block Diagonalization. For certain eigenvalue shrinkage estimators, the population covariance Σ_p and the estimated covariance matrix $\hat{\Sigma}_\eta(S_{n,p})$ are simultaneously block-diagonalizable. Specifically, there is a (random) basis W such that

$$W'\Sigma_p W = (\oplus_i A_i) \oplus I_{p-2r}$$

and

$$W'\hat{\Sigma}_\eta(S_{n,p})W = (\oplus_i B_i) \oplus I_{p-2r}\ ,$$

where A_i and B_i are square blocks of equal size d_i, and $\sum d_i = 2r$.

Obs. 2: Decomposable Loss Functions. Certain matrix loss functions L_p, which are used in the literature to evaluate performance of covariance estimators, are *decomposable* over these blocks. They satisfy either

$$L_p\left(\Sigma_p, \hat{\Sigma}_\eta(S_{n,p})\right) = \sum_i L_{d_i}(A_i, B_i)$$

or

$$L_p\left(\Sigma_p, \hat{\Sigma}_\eta(S_{n,p})\right) = \max_i L_{d_i}(A_i, B_i)\ .$$

Obs. 3. Asymptotically Deterministic Loss. The blocks A_i and B_i only depend on the sample covariance $S_{n,p}$ through its eigenvalues and through the angle between empirical and theoretical eigenvectors. As a result, in the Spiked Covariance asymptotic model, A_i and B_i (and hence the loss) converge to deterministic functions of the spike amplitudes $\ell_1, \ldots, \ell_r$.

Combining these observations, for any loss sequence L consisting of decomposable functions, we obtain an explicit formula for the asymptotic loss (1.3), as a deterministic function of $\ell_1, \ldots, \ell_r$. This allows us to determine the asymptotically optimal eigenvalue shrinkage estimator for covariance estimation in that loss.

This paper is organized as follows. For simplicity of exposition, we assume a single spike ($r = 1$) throughout most of the paper. **[Obs. 1]** is developed in Section 2. Section 3 fleshes out **[Obs. 2]** and introduces our list of 26 decomposable matrix loss functions. Section 4 includes background on the Spiked Covariance model. In Section 5 we discuss **[Obs. 3]** above and derive an explicit formula for the asymptotic loss of a shrinker. In Section 6 we use this formula to characterize the asymptotically unique admissible nonlinearity for any decomposable loss, provide an algorithm for computing the optimal nonlinearity, and provide analytical formulas for some of the losses discussed. In Section 7 we evaluate the large-ℓ asymptotics of the optimal shrinkage estimators, namely their behavior and their performance when the signal $\ell_1 = \ell$ is very strong. In Section 8 we extend these results to the general case where $r > 1$ spikes are present. In Section 9 we show that, at least in the popular Frobenius loss case, our asymptotically optimal univariate shrinkage estimator, which applies the same univariate function to each of the sample principal components, is in fact optimal among equivariant covariance estimators. Our results are discussed in Section 10. Proofs are provided in the appendices. Additional technical details are provided in the supplemental article [17].

2 Simultaneous Block-Diagonalization

Let Σ_p be a population covariance matrix. Assume that Σ_p satisfies assumption **[Spike($\ell_1, \ldots, \ell_r$)]**, and let the empirical covariance matrix $S \equiv S_{n,p} \equiv n^{-1} \sum_{i=1}^{n} X_i X_i'$ have empirical eigenvalues $\lambda_i, i = 1, \ldots, p$, also ordered so $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Let u_i denote the eigenvector (population principal vector) corresponding to the i-th population eigenvalue, and let v_i denote the eigenvector (sample principal vector) corresponding to λ_i.

Specialize now to a population covariance matrix Σ_p with a single "spike", corresponding to $r = 1$ in our notation. Write $S_{n,p} = V \Lambda V'$ for principal component analysis of $S_{n,p}$. Here, V is a p-by-p orthogonal matrix whose i-th column is v_i, and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$.

Let $\eta : [0, \infty) \to \mathbb{R}$ be any scalar nonlinearity and define the eigenvalue shrinkage operator $\hat{\Sigma}_\eta : S_{n,p} \mapsto \hat{\Sigma}_\eta(S_{n,p}) = V \eta(\Lambda) V'$, where by convention η acts entrywise on the diagonal entries of Λ. For a reason that will soon become apparent, we assume that $\eta \geq 1$ and that η and the empirical eigenvalues satisfy $\eta(\lambda_2) = \ldots = \eta(\lambda_p) = 1$. We now develop **[Obs. 1]**, that *while the estimated matrix $\hat{\Sigma}_\eta(S_{n,p})$ and the population covariance Σ_p are not simultaneously diagonalizable in general, under these assumptions, they are simultaneously block-diagonalizable.* Indeed, working in the theoretical eigenbasis (u-basis), we have

$$U' \Sigma_p U = \ell_1 e_1 e_1' \oplus I_{p-1} \,, \tag{2.1}$$

where U is a matrix whose columns are $u_1 \ldots u_p$, and e_i denotes the i-th standard basis vector. Similarly, working in the empirical eigenbasis (v-basis), since by assumption $\eta(\lambda_i) = 1$ ($2 \leq i \leq p$), we have

$$V'\eta(S_{n,p})V = \eta(\lambda_1)e_1e_1' \oplus I_{p-1}. \tag{2.2}$$

Combining the two representations (2.1) and (2.2), we are led to the following "common" basis. Let $w_1, \ldots, w_p$ denote the orthonormal basis constructed by applying the Gram-Schmidt process to the sequence $u_1, v_1, v_2, \ldots, v_{p-1}$. Observe that in the w-basis we have

$$W'\Sigma_p W = \begin{bmatrix} \ell_1 & 0 \\ 0 & 1 \end{bmatrix} \oplus I_{p-2} \tag{2.3}$$

$$W'\hat{\Sigma}_\eta(S_{n,p})W = \begin{bmatrix} 1+(\eta(\lambda_1)-1)c_1^2 & (\eta(\lambda_1)-1)c_1s_1 \\ (\eta(\lambda_1)-1)c_1s_1 & 1+(\eta(\lambda_1)-1)s_1^2 \end{bmatrix} \oplus I_{p-2}\,, \tag{2.4}$$

where W is a matrix whose columns are $w_1, \ldots, w_p$, and where [1] $c_1 = \langle u_1, v_1 \rangle$ and $s_1 = \sqrt{1-c_1^2}$. It is convenient to rewrite (2.3) and (2.4) as

$$\begin{aligned} W'\Sigma_p W &= A(\ell_1) \oplus I_{p-2} \\ W'\hat{\Sigma}_\eta(S_{n,p})W &= B\big(\eta(\lambda_1), c_1, s_1\big) \oplus I_{p-2}\,, \end{aligned}$$

where $A(\ell) = diag(\ell, 1)$ and B is the fundamental 2-by-2 matrix

$$B(\lambda, c, s) = \begin{bmatrix} 1+(\lambda-1)c^2 & (\lambda-1)cs \\ (\lambda-1)cs & 1+(\lambda-1)s^2 \end{bmatrix}. \tag{2.5}$$

Below, for a matrix M and a k-dimensional subspace $\mathcal{U}$, we let $M|\mathcal{U}$ denote a k-by-k matrix representing the action of $P_{\mathcal{U}}MP_{\mathcal{U}}$ on vectors in $\mathcal{U}$, where $P_{\mathcal{U}}$ denotes orthoprojection on subspace $\mathcal{U}$. We have proved

Lemma 2.1. Simultaneous Block Diagonalization in the case $r = 1$. *Write $S_{n,p} = V\Lambda V'$ for the diagonalization of the sample covariance matrix. Consider the estimator $\hat{\Sigma}_\eta : V\Lambda V' \mapsto V\eta(\Lambda)V'$ where $\eta : [0,\infty) \to [1,\infty)$ is a nonlinearity that is applied entrywise to the diagonal entries of Λ. Assume that η satisfies $\eta(\Lambda)_i = 1$ for all $i > 1$. Then*

1. *Σ_p and $\hat{\Sigma}_\eta$ are jointly block-diagonalizable, with jointly-invariant subspaces $\mathcal{W}_2$ and $\mathcal{W}_2^\perp$.*
2. *On the first block,*

$$\Sigma_p|\mathcal{W}_2 = A(\ell_1) \equiv diag(\ell_1, 1), \tag{2.6}$$

$$\hat{\Sigma}_\eta|\mathcal{W}_2 = B(\eta(\lambda_1), c_1, s_1)\,. \tag{2.7}$$

3. *On the second block,*

$$\Sigma_p|_{\mathcal{W}_2^\perp} = \hat{\Sigma}_\eta|_{\mathcal{W}_2^\perp} = I|_{\mathcal{W}_2^\perp}\,.$$

[1] In constructing the w-basis by Gram-Schmidt, we choose the orientation of w_2 so $s_1 = +\sqrt{1-c_1^2}$, and not $s_1 = -\sqrt{1-c_1^2}$.

3 Decomposable Loss Functions

Here and below, by *loss function* L_p we mean a function of two p-by-p positive semidefinite matrix arguments obeying $L_p \geq 0$, with $L_p(A, B) = 0$ if and only if $A = B$. **[Obs. 2]** calls out a large class of loss functions which naturally exploit the simultaneously block-diagonalizability property of Lemma 2.1; we now develop this observation.

Definition 3.1. Orthogonal Invariance. *We say the loss function* $L_p(A, B)$ *is* orthogonally invariant *if*

$$L_p(A, B) = L_p(OAO', OBO')$$

for each orthogonal p-by-p matrix O.

We now assume that we have a family of loss functions L_p indexed by $p = 1, 2, \ldots$. Consider all block matrix decompositions of p by p matrices into blocks of size d_i:

$$A = \oplus_i A^i \qquad\qquad B = \oplus_i B^i. \tag{3.1}$$

Definition 3.2. Sum-Decomposability. *We say the loss function* $L_p(A, B)$ *is* sum-decomposable *if, for all decompositions* (3.1),

$$L_p(A, B) = \sum_i L_{d_i}(A^i, B^i).$$

Definition 3.3. Max-Decomposability. *We say the loss function* $L_p(A, B)$ *is* max-decomposable *if, for all decompositions* (3.1),

$$L_p(A, B) = \max_i L_{d_i}(A^i, B^i)\,,$$

Focusing again on the single spike case ($r = 1$), when the representations (2.3) and (2.4) hold, we have for any orthogonally invariant, sum-decomposable or max-decomposable loss function L_p that

$$L_p\left(\Sigma_p, \hat{\Sigma}_\eta(S_{n,p})\right) = L_p\left(A(\ell_1) \oplus I_{p-2}\,,\ B(\eta(\lambda_1), c_1, s_1) \oplus I_{p-2}\right) = L_2\left(A(\ell_1)\,,\ B(\eta(\lambda_1), c_1, s_1)\right)\,.$$

In other words, evaluating L_p reduces to evaluating L_2 on specific matrices A and B, and we may ignore the last $p-2$ coordinates completely. We have proved:

Lemma 3.1. Reduction to Two-Dimensional Problem. *Consider an orthogonally invariant loss function, L_p, which is sum- or max-decomposable. Assume that the population covariance $\Sigma_{n,p}$ and nonlinearity η satisfy the assumptions of Lemma 2.1. When $\eta(\lambda_i) = 1$ for $i > 1$,*

$$L_p\left(\Sigma_p, \hat{\Sigma}_\eta(S_{n,p})\right) = L_2\left(A(\ell_1)\,,\ B(\eta(\lambda_1), c_1, s_1)\right)\,.$$

Many decomposable loss functions that appear in the literature can be built via the following common recipe.

Definition 3.4. Pivots. A matrix pivot *is a matrix-valued function $\Delta(A, B)$ of two real symmetric matrices A, B that is orthogonally invariant and respects block structure:*

$$\Delta(OAO', OBO') = O\Delta(A, B)O', \tag{3.2}$$

$$\Delta(\oplus A^i, \oplus B^i) = \oplus \Delta(A^i, B^i). \tag{3.3}$$

Matrix pivots can be symmetric-matrix valued, for example $\Delta(A,B) = A - B$, but need not be, for example $\Delta(A,B) = A^{-1}B - I$. For symmetric pivots, we consider loss functions of the form

$$L(A,B) = g(\Delta(A,B)), \tag{3.4}$$

where g is orthogonally invariant, $g(O\Delta O') = g(\Delta)$, and so is a function of the eigenvalues $\delta = (\delta_j)$ of Δ, which with slight abuse of notation we write as $g(\delta)$. Such loss functions are orthogonally invariant. If g has either of the forms

$$g(\delta) = \sum_j g_1(\delta_j) \qquad \text{or} \qquad g(\delta) = \max_j g_1(\delta_j),$$

then L is respectively sum- or max-decomposable. For general pivots, we consider loss functions

$$L(A,B) = h(|\Delta|(A,B)), \qquad\qquad |\Delta| = (\Delta'\Delta)^{1/2}. \tag{3.5}$$

Here, h is orthogonally invariant and a function of the *singular values* $\sigma = (\sigma_j)$ of Δ. If $h(\sigma) = \sum h_1(\sigma_j)$ or $\max h_1(\sigma_j)$, we obtain sum- or max-decomposable loss functions. Of course, if Δ is symmetric, then $\sigma_j = |\delta_j|$.

We now discuss some examples in greater detail.

3.1 Sum-Decomposable Losses

Suppose A and B are jointly block diagonalizable with respective blocks A^i, B^i. Then let $\Delta^i = \Delta(A^i, B^i)$ and δ^i_j be the eigenvalues (resp. singular values) of Δ^i. Then, when L has the form (3.4) and g is additive,

$$L(A,B) = \sum_i \sum_j g_1(\delta^i_j).$$

There are several strategies to derive sum-decomposable functions. First, we can use statistical discrepancies between the $N(0,A)$ and the $N(0,B)$ distributions.

1. *Stein Loss* [57, 13, 35]: Let A and B be p by p covariance matrices; Stein's Loss is denoted $L^{st}(A,B) = (\mathrm{trace}(A^{-1}B - I) - \log(|B|/|A|))/2$. This can be understood as the Kullback distance $D_{KL}(N(0,B)|N(0,A))$. Consider the matricial pivot $\Delta = A^{-1/2}BA^{-1/2}$. Then

   $$L^{st}(A,B) = (\mathrm{trace}(\Delta - I) - \log|\Delta|)/2 = g(\Delta)\,.$$

   We may take $g_1(\delta) = (\delta - 1 - \log\delta)/2$.

2. *Entropy/Divergence Losses:* Because the Kullback discrepancy is not symmetric in its arguments, we may consider two other losses: reversing the arguments we get *Entropy* loss $L^{ent}(A,B) = L^{st}(B,A)$ [56, 36] and summing the Stein and Entropy losses gives *divergence* loss:

   $$L^{div}(A,B) = L^{st}(A,B) + L^{st}(B,A) = \tfrac{1}{2}[\mathrm{trace}(A^{-1}B - I) + \mathrm{trace}(B^{-1}A - I)]\,,$$

   see [39, 24]. Each can be shown sum-decomposable, following the same argument as above.

3. *Bhattarcharya/Matusita Affinity* [32, 48]: Let $L^{aff} = \frac{1}{2}\log\frac{|A+B|/2}{|A|^{1/2}|B|^{1/2}}$. This measures the statistical distinguishability of $N(0,A)$ and $N(0,B)$ based on independent observations, since $L^{aff} = \frac{1}{2}\log(\int\sqrt{\phi_A}\sqrt{\phi_B})$ with ϕ_A and ϕ_B the densities of $N(0,A)$ and $N(0,B)$. Hence convergence of affinity loss to zero is equivalent to convergence of the underlying densities in Hellinger or Variation distance. Using the pivot $\Delta = A^{-1/2}BA^{-1/2}$, we have

$$L^{\text{aff}} = \tfrac{1}{4}\log(|2I + \Delta + \Delta^{-1}|/4),$$

as is seen by setting $C = A^{-1/2}(A+B)B^{-1/2}$ and noting that $C'C = (2I + \Delta + \Delta^{-1})$. Then we have $g_1(\delta) = \frac{1}{4}\log(2 + \delta + \delta^{-1})/4$.

4. *Fréchet Discrepancy* [50, 19]: Let $L^{fre}(A,B) = \mathrm{trace}(A+B-2A^{1/2}B^{1/2})$. This measures the minimum possible mean-squared difference between zero-mean random vectors with covariances A and B respectively. The pivot $\Delta = A^{1/2} - B^{1/2}$, and $L^{fre} = \mathrm{trace}(\Delta^2)$, so that $g_1(\delta) = \delta^2$.

We may also obtain sum-decomposable losses by applying certain standard matrix norms to pivot matrices.

1. *Squared Error Loss* [29, 11, 41, 43]: Let $L^{F,1}(A,B) = \|A-B\|_F^2$, using pivot $\Delta = A - B$, so that $g(\Delta) = \mathrm{trace}\Delta'\Delta$ and $g_1(\delta) = \delta^2$.

2. *Squared Error Loss on Precision* [25]: Let $L^{F,2}(A,B) = \|A^{-1} - B^{-1}\|_F^2$. The pivot $\Delta = A^{-1} - B^{-1}$ and again $g(\Delta) = \mathrm{trace}\Delta'\Delta$.

3. *Nuclear Norm Loss.* Let $L^{N,1}(A,B) = \|A-B\|_*$ where $\|A-B\|_*$ denotes the nuclear norm of matrix Δ, i.e. the sum of singular values. The symmetric pivot $\Delta = A - B$, and so if blocks Δ^i have singular values σ^i_j and eigenvalues δ^i_j, (say), then

$$L^{N,1} = \sum_{i,j}\sigma^i_j = \sum_{i,j}|\delta^i_j|.$$

Many other norm-based loss functions offer additive decomposability. The previous examples all involved the use of the r-th power of an ℓ_r norm, applied to the eigenvalues or singular values of a pivot $\Delta(A,B)$ having the property $\Delta(A,A) = 0$. We now adopt the systematic naming scheme $L^{\text{norm,pivot}}$ where norm $\in \{F, O, N\}$, and pivot $\in \{1, \ldots, 7\}$. The resulting 21 different possible combinations are all studied in this article. Under this naming scheme, the three examples immediately above are called $L^{F,1}$, $L^{F,2}$ and $L^{N,1}$, respectively.

Remarks.

- The squared Frobenius norm of the pivot $A^{-1}B - I$ is an invariant loss function – here denoted $L^{F,3}$; it was studied in [53, 27, 55] and later work.
- The pivot $\log(A^{-1/2}BA^{-1/2})$, where $\log()$ denotes the matrix logarithm [23, 44], also yields invariant losses. The matrix logarithm transfers the matrices from the symmetric space (Riemannian manifold) of symmetric positive definite (SPD) matrices to its tangent space at A. Applying the squared Frobenius norm to this logarithm, we obtain an invariant loss function, here denoted $L^{F,7}$. It is simply the squared geodesic distance in the manifold of positive

semidefinite matrices. This metric between covariances has attracted attention in diffusion tensor MRI [44, 20].[2]

3.2 Max-Decomposable Losses

Max-decomposable losses arise by applying an operator norm to a suitable matrix pivot, so that when A and B are jointly block diagonalizable, with respective blocks A^i, B^i and if σ^i_j and δ^i_j are the singular values and eigenvalues of $\Delta^i = \Delta(A^i, B^i)$, then

$$L(A,B) = \max_{i,j} \sigma^i_j = \max_{i,j} |\delta^i_j|,$$

the latter holding for symmetric pivots.

Here are a few examples.

1. *Operator Norm Loss* [33]: Let $L^{op}(A,B) = \|A - B\|_{op}$. The pivot $\Delta = A - B$.
2. *Operator Norm Loss on Precision:* Let $L^{op}(A,B) = \|A^{-1} - B^{-1}\|_{op}$. Then the pivot $\Delta = A^{-1} - B^{-1}$.
3. *Condition Number Loss*: Let $L^{cond}(A,B) = \|\log(A^{-1/2}BA^{-1/2})\|_{op}$. The pivot $\Delta = \log(A^{-1/2}BA^{-1/2})$. (Related to [61].)

Under our naming scheme, the preceding 3 examples are $L^{O,1}$, $L^{O,2}$ and $L^{O,7}$. We note that, in the spiked model, $L^{O,7}$ effectively measures the condition number of $A^{-1/2}BA^{-1/2}$.

The systematic notation we adopt for naming these 26 losses is summarized in Table 1.

	MatrixNorm		
Pivot	Frobenius	Operator	Nuclear
$A - B$	$L^{F,1}$	$L^{O,1}$	$L^{N,1}$
$A^{-1} - B^{-1}$	$L^{F,2}$	$L^{O,2}$	$L^{N,2}$
$A^{-1}B - I$	$L^{F,3}$	$L^{O,3}$	$L^{N,3}$
$B^{-1}A - I$	$L^{F,4}$	$L^{O,4}$	$L^{N,4}$
$A^{-1}B + B^{-1}A - 2I$	$L^{F,5}$	$L^{O,5}$	$L^{N,5}$
$A^{-1/2}BA^{-1/2} - I$	$L^{F,6}$	$L^{O,6}$	$L^{N,6}$
$\log(A^{-1/2}BA^{-1/2})$	$L^{F,7}$	$L^{O,7}$	$L^{N,7}$
	Statistical Measures		
	St	Ent	Div
Stein	L^{st}	L^{ent}	L^{div}
Affinity	L^{aff}		
Fréchet	L^{fre}		

Table 1: Systematic naming of 26 loss functions

[2]We view our ability to derive the optimal shrinker for such losses as a contribution.

4 The Spiked Covariance Model

So far, under certain conditions, we have reduced the evaluation of a decomposable loss L_p for $p > 2$ to evaluation of a corresponding loss L_2 on specific 2-by-2 matrices $A(\ell)$ and $B(\eta(\lambda_1), c_1, s_1)$. We now develop **[Obs. 3]**, which asserts that, in the spiked model [30, 52, 5], these matrices, and hence the limiting loss, are in fact *deterministic* functions of the population spike amplitude ℓ.

To set notation, let us work under assumptions **[Asy(γ)]** and **[Spike($\ell_1, \ldots, \ell_r$)]**. We denote the i-th empirical eigenvalue of S_{n,p_n} by $\lambda_{i,n}$ and its corresponding eigenvector by $v_{i,n}$. Similarly, we denote the i-th theoretical eigenvector by $u_{i,n}$. The symbol $\to_P$ denotes convergence in probability.

The Bulk. In what we might call the "unspiked" or "null" special case of this model, all eigenvalues are equal to one: $\ell_1 = \cdots = \ell_r = 1$. Although here $\Sigma_{p_n} = I_{p_n}$ is the standard identity covariance, many empirical eigenvalues will be far from the underlying theoretical eigenvalue 1. In fact, in this case the empirical eigenvalues are concentrated in an interval extending roughly from $\lambda_-(\gamma)$ to $\lambda_+(\gamma)$, where

$$\textbf{[RMT 1]} \qquad \lambda_\pm(\gamma) = (1 \pm \sqrt{\gamma})^2 \,, \tag{4.1}$$

see [1]. This compact scatter forms the so-called Marčenko-Pastur "bulk" (or "sea") and $\lambda_\pm$ are known as the limiting "bulk edges" [3].

Emergence from the Bulk. When $\ell_1 \geq \cdots \geq \ell_r > 1$ (the "actually spiked" or "non-null" case) and ℓ_r is large enough, the top empirical eigenvalues typically "emerge from the bulk", namely, their asymptotic limits in probability exist and exceed $\lambda_+(\gamma)$ (while all the other empirical eigenvalues have cluster points $\leq \lambda_+$). This emergence from the bulk happens when ℓ_r crosses the so-called Baik-Ben Arous-Péché phase transition

$$\textbf{[RMT 2]} \qquad \ell_+(\gamma) = (1 + \sqrt{\gamma}) \,, \tag{4.2}$$

see [4]. Gathering together results from several papers [5, 52, 6, 3] we have the following general picture:

Theorem 4.1. *(**Spiked Covariance Model**) Let the theoretical covariance have leading eigenvalues ℓ_i, $i = 1, \ldots, r$ obeying $\ell_i > \ell_+(\gamma)$. Then, as $n \to \infty$, the leading empirical eigenvalues obey*

$$\lambda_{i,n} \to_P \lambda(\ell_i), \qquad i = 1, \ldots, r \,, \tag{4.3}$$

where $\lambda(\ell)$ is defined in (1.1)*; See [5, 3, 6]. Moreover, let $v_{i,n}$ denote the empirical eigenvector corresponding to $\lambda_{i,n}$ and let $u_{i,n}$ denote the theoretical eigenvector corresponding to ℓ_i. Suppose that the principal theoretical eigenvalues are distinct, namely $\ell_i \neq \ell_j$, $1 \leq i, j \leq r$. Then, as $n \to \infty$, we have*

$$|\langle u_{i,n}, v_{j,n} \rangle| \to_P \delta_{i,j} \cdot c(\ell_i) \qquad 1 \leq i, j \leq r \,,$$

where $c(\ell)$ is defined in (1.2)*; See [52, 6].*

[3] The marginal distribution of the eigenvalues falling in the bulk is the famous Marčenko-Pastur distribution [47]

In words, the empirical eigenvalues are shifted upwards from their theoretical counterparts by a displacement of asymptotically predictable size; and the empirical eigenvectors are rotated away from their theoretical positions, through a noticeable angle of asymptotically predictable size. These two effects drive our main results, as the fundamental 2-by-2 matrix B only depends on ℓ through these quantities.

In all the expressions below we freely use that, for $\lambda > \lambda_+(\gamma)$ and $\ell > \ell_+(\gamma)$, there is a one-one relationship between λ and ℓ such that any expression may be written in terms of either, without chance of ambiguity. A simple calculation shows that the one-to-one map $\lambda(\ell)$ and $\ell(\lambda)$ is given by (1.1) and

$$\ell(\lambda) = \frac{(\lambda + 1 - \gamma) + \sqrt{(\lambda + 1 - \gamma)^2 - 4\lambda}}{2}, \qquad \lambda > \lambda_+(\gamma). \tag{4.4}$$

5 Asymptotic Loss in the Spiked Covariance Model

Consider the spiked model with a single spike, $r = 1$, namely, make assumptions **[Asy(γ)]** and **[Spike(ℓ_1)]**. In accord with **[Obs. 3]**, we now show that the asymptotic loss (1.3) is a deterministic, explicit function of the population spike ℓ_1.

For $\varepsilon > 0$, consider the event

$$\Omega'_{n,\varepsilon} = \{\lambda_{1,n} > \lambda_+(\gamma) + \varepsilon, \lambda_{2,n} < \lambda_+(\gamma) + \varepsilon\}.$$

Under the single-spiked model and the assumption $\ell_1 > \ell_+(\gamma)$ we have, for sufficiently small $\varepsilon > 0$,

$$P(\Omega'_{n,\varepsilon}) \to 1 \text{ as } n \to \infty.$$

Definition 5.1.

- *We say that a scalar nonlinearity $\eta : [0, \infty) \to [0, \infty)$* collapses the bulk to 1 *if $\eta(\lambda) = 1$ whenever $\lambda \in (\lambda_-(\gamma), \lambda_+(\gamma))$.*
- *We say that a scalar nonlinearity $\eta : [0, \infty) \to [0, \infty)$* collapses the vicinity of the bulk to 1 *if, for some $\varepsilon > 0$, we have $\eta(\lambda) = 1$ whenever $\lambda \in (\lambda_-(\gamma) - \varepsilon\,,\ \lambda_+(\gamma) + \varepsilon)$.*

Observe that, if η is a nonlinearity collapsing the vicinity of the bulk to 1, then, on the event $\Omega'_{n,\varepsilon}$ for an appropriate ε, Lemma 2.1 holds: for each n there is a random p_n-by-p_n orthogonal matrix W_n that simultaneously block-diagonalizes Σ_{p_n} and $\hat{\Sigma}_\eta(S_{n,p_n})$.

We now turn to investigate the convergence of the 2-by-2 matrix B. By Theorem 4.1 we have $\lambda_{1,n} \to_P \lambda(\ell)$ and $c_{1,n} \to_P c(\ell)$. Define $s(\ell) = \sqrt{1 - c^2(\ell)}$. Suppose that $\eta(\lambda)$ is continuous. It follows that the 2-by-2 matrix $B(\eta(\lambda_{1,n}), c_{1,n}, s_{1,n})$ in (2.4) converges in probability to a limit 2-by-2 matrix $B(\ell_1) = B(\ell_1, \eta)$, where

$$B(\ell, \eta) := B\Big(\eta(\lambda(\ell)), c(\ell), s(\ell)\Big) = \begin{bmatrix} 1 + (\eta(\lambda(\ell)) - 1)c(\ell)^2 & (\eta(\lambda(\ell)) - 1)c(\ell)s(\ell) \\ (\eta(\lambda(\ell)) - 1)c(\ell)s(\ell) & 1 + (\eta(\lambda(\ell)) - 1)s(\ell)^2 \end{bmatrix}. \tag{5.1}$$

We have proved:

Lemma 5.1. Convergence of the principal block. *Let the scalar nonlinearity η collapse the vicinity of the bulk to 1 and assume that it is continuous. Then for every $\varepsilon > 0$ small enough there is a sequence of events $\Omega_{n,\varepsilon}$ with $P(\Omega_{n,\varepsilon}) \to 1$ as $n \to \infty$ such that on $\Omega_{n,\varepsilon}$:*

1. *Σ_{p_n} and $\hat{\Sigma}_\eta(S_{n,p_n})$ are simultaneously block-diagonalizable, and*

2. $\|B\left(\eta(\lambda_{1,n}), c(\lambda_{1,n}), s(\lambda_{1,n})\right) - B(\ell_1)\|_F < \varepsilon$.

We now specify a family $L = \{L_p\}$ of loss functions and consider the asymptotic loss

$$\lim_{n\to\infty} L_{p_n}(\Sigma_{p_n}, \hat{\Sigma}_\eta(S_{n,p_n}))$$

of a shrinkage rule η.

Lemma 5.2. A deterministic formula for the asymptotic loss. *Let $L = \{L_p\}$ be a family of loss functions which is orthogonally invariant in the sense of Definition 3.1, and also sum-decomposabile or max-decomposabile in the sense of Definitions 3.2 and 3.3. Assume that for some exponent $\alpha > 0$, and any $\ell > 0$, L_2^α is Lipschitz continuous on the set*

$$\left\{(A,B)\,\middle|\,\lambda_{min}(A), \lambda_{min}(B) \geq \ell\right\} \subset S_2^+ \times S_2^+$$

with respect to the matrix Frobenius distance, with Lipschitz constant C_ℓ. Consider a problem sequence with $n, p \to \infty$, with $p = p_n$ obeying $p_n/n \to \gamma$, with theoretical covariance matrices Σ_{p_n} following the single-spike model with fixed population spike $\ell > \ell_+(\gamma)$. Suppose the scalar nonlinearity η collapses the vicinity of the bulk to 1 and is continuous. Then

$$L_{p_n}\left(\Sigma_{p_n}\,,\,\hat{\Sigma}_\eta(S_{n,p_n})\right) \to_P L_\infty(\ell|\eta) \equiv L_2\left(A(\ell)\,,\,B\left(\eta(\lambda(\ell), c(\ell), s(\ell)\right)\right)\,, \tag{5.2}$$

where $\lambda(\ell)$, $c(\ell)$ and $s(\ell)$ are defined above, $A(\ell) = diag(\ell, 1)$, $B(\eta, c, s)$ is defined in (2.5)*, and where the convergence is in probability.*

Proof. Fix the spike amplitude ℓ and let C_ℓ be the corresponding Lipschitz constant of L_2^α. WLOG assume $\alpha = 1$. We will show that

$$P\left\{\left|L_{p_n}\left(\Sigma_{p_n}\,,\,\hat{\Sigma}_\eta(S_{n,p_n})\right) - L_2\left(A(\ell)\,,\,B\left(\eta(\lambda(\ell), c(\ell), s(\ell)\right)\right)\right| \leq C_\ell\varepsilon\right\} \to 1$$

as $n \to \infty$. Let $\Omega_{n,\varepsilon}$ denote the sequence of events in Lemma 5.1. For $\varepsilon > 0$ is small enough, $P(\Omega_{n,\varepsilon}) \to 1$ and on each event $\Omega_{n,\varepsilon}$, Σ_{p_n} and $\hat{\Sigma}_\eta(S_{n,p_n})$ are simultaneously block-diagonalizable. By Lemma 3.1, on $\Omega_{n,\varepsilon}$ we have

$$L_p\left(\Sigma_p, \hat{\Sigma}_\eta(S_{n,p})\right) = L_2\left(A(\ell)\,,\,B(\eta(\lambda_1), c_1, s_1)\right)\,.$$

By Lemma 3.1, on $\Omega_{n,\varepsilon}$ we also have $\|B\left(\eta(\lambda_{1,n}), c(\lambda_{1,n}), s(\lambda_{1,n})\right) - B(\ell_1)\|_F < \varepsilon$. Since L_2 is Lipschitz continuous,

$$\left|L_2(A(\ell), B\left(\eta(\lambda_{1,n}), c(\lambda_{1,n}), s(\lambda_{1,n})\right)) - L_2(A(\ell), B(\ell_1))\right| \leq C_\ell\varepsilon\,,$$

as required.

$\square$

6 Optimal Shrinkage for Decomposable Losses

6.1 Formally Optimal Shrinker

Formula (5.2) for the asymptotic loss has only been shown to hold under **[Asy(γ)]** and **[Spike(ℓ_1)]**, for certain nonlinearities η. In fact, the loss converges to $L_\infty(\ell|\eta)$ under **[Asy(γ)]** and **[Spike($\ell_1,\ldots,\ell_r$)]**, and for a much broader class of nonlinearities η. To preserve the narrative flow of the paper, we defer the proof, which is largely technical, to Section 8. Instead, we proceed by simply assuming that (5.2) holds, namely that $L_\infty(\ell|\eta)$ is the correct asymptotic loss, and drawing conclusions on the optimal shape of the shrinker η.

Definition 6.1. *Let $L = \{L_p\}$ be a given loss family and let $L_\infty(\ell|\eta)$ be the asymptotic loss corresponding to a nonlinearity η, as defined in* (5.2)*. If a nonlinearity η^* satisfies*

$$L_\infty(\ell|\eta^*) = \min_\eta L_\infty(\ell|\eta), \qquad \forall \ell \geq 1\,, \tag{6.1}$$

we say that η^* is a formally optimal, or simply optimal, *shrinker.*

By definition, the corresponding rule $\eta^*(\lambda;\gamma,L)$ is the unique admissible rule, in the asymptotic sense, among rules of the form $\hat{\Sigma}_\eta(S_{n,p}) = V\eta(\Lambda)V'$ in the single-spike model. In view of the Lemma 5.2, solving the problem (6.1) is simpler than one might expect, as it only involves optimization over 2-by-2 matrices:

Definition 6.2. *(***Optimal Shrinker.***) Let $\ell > \ell_+(\gamma)$ and correspondingly $\lambda = \lambda(\ell) > \lambda_+(\gamma)$. Given a component loss function $L_2(A,B)$, consider the optimization problem*

$$\min_{\eta \geq 1} F(\eta,\ell) \tag{6.2}$$

where

$$F(\eta,\ell) = L_2\Big(\begin{bmatrix} \ell & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1+(\eta-1)c^2 & (\eta-1)cs \\ (\eta-1)cs & 1+(\eta-1)s^2 \end{bmatrix} \Big)\,. \tag{6.3}$$

Here, $c = c(\ell)$ and $s = s(\ell)$ satisfy $c^2(\ell) = \frac{1-\gamma/(\ell-1)^2}{1+\gamma/(\ell-1)}$ and $s^2(\ell) = 1 - c^2(\ell)$. Suppose this optimization problem has a unique solution $\eta^*(\ell)$ for each $\ell > \ell_+$ and write the solution as a function of λ using the relation (4.4) *to obtain $\eta^*(\lambda) \equiv \eta^*(\ell(\lambda))$. We call the resulting function $\eta^*(\lambda) = \eta^*(\lambda;\gamma,L)$, the* optimal shrinker *for the family of loss functions L.*

Alternatively suppose $1 < \ell < \ell_+(\gamma)$. Consider the minimization $\min_{\eta\geq 1} G(\eta)$*, where*

$$G(\eta,\ell) = L_2\Big(\begin{bmatrix} \ell & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \eta \end{bmatrix} \Big); \tag{6.4}$$

if, for every $\ell \in [1,\ell_+)$, this minimum is achieved at $\eta = 1$, then we say that the optimal shrinker collapses the bulk to 1: i.e. the optimal shrinker satisfies $\eta^*(\lambda) = 1$ for $\lambda < \lambda_+(\gamma)$.

Many of the 26 loss families discussed in Section 3 admit a closed form expression for the optimal shrinker; see Table 2. For others, we computed the optimal shrinkage numerically, by implementing the optimization problem of Definition 6.2 in software. Figure 1 portrays the optimal shrinkers for 26 loss functions. For readers interested in using specific individual shrinkers, we recommend to read our reproducibility advisory at the bottom of this paper, and explore the data and code supplement [16], consisting of online resources and code we offer.

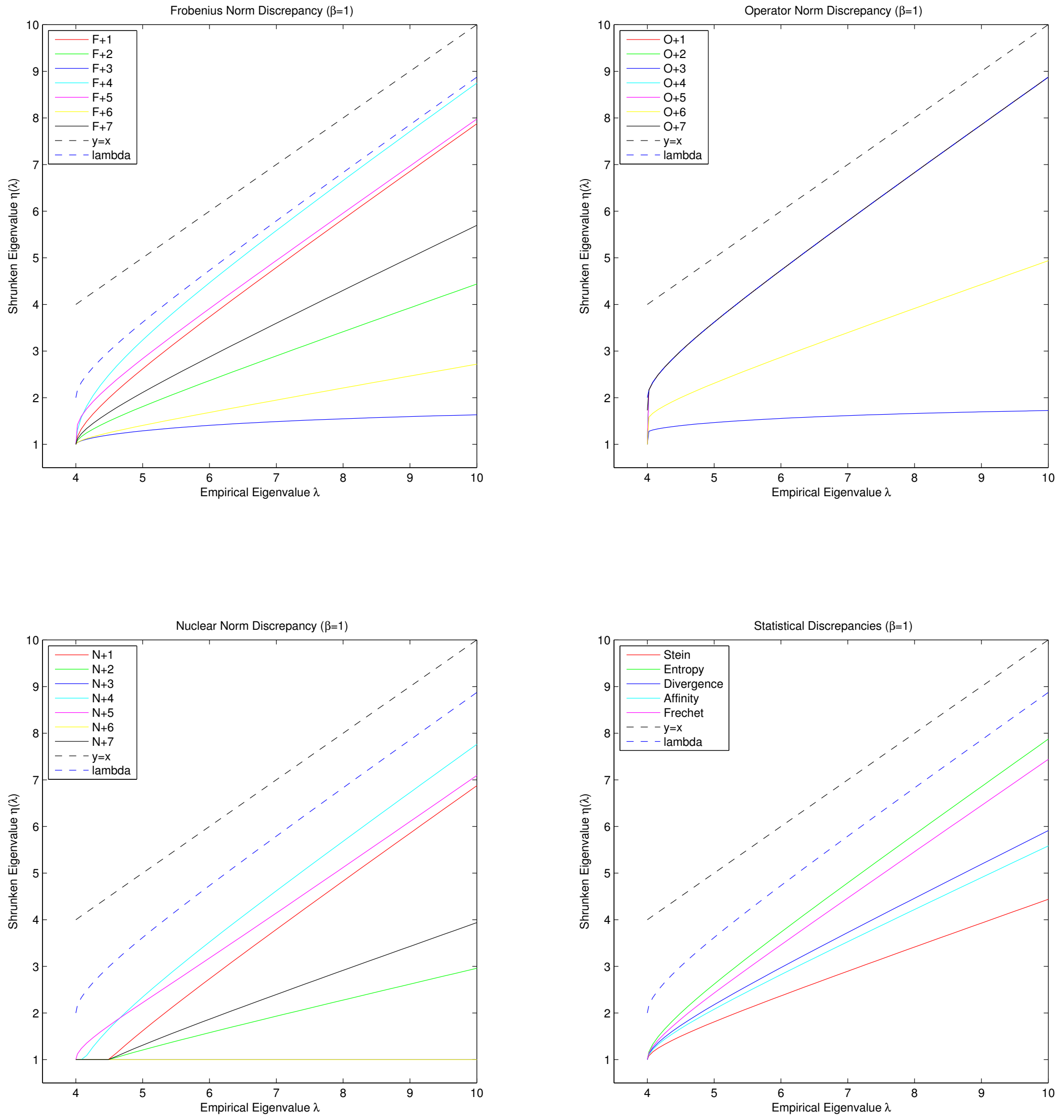

Figure 1: Optimal Shrinkers for 26 Component Loss Functions. $\gamma = 1$. $4 < \lambda < 10$. Upper Left: Frobenius-based losses; Lower Left: Nuclear-Norm based losses; Upper Right: Operator-norm-based losses; Lower Right: Statistical Discrepancies. *Reproducibility advisory:* The data and code supplement [16] includes a script to generate any one of these individual curves.

6.2 Collapse of the Bulk

We first observe that, for any of the 26 losses considered, the optimal shrinker collapses the bulk to 1. The following lemma is proved in Appendix A.1:

Lemma 6.1. *Let L be any of the 26 losses mentioned in Table 1. Then the rule $\eta^{**}(\ell) = 1$ is unique asymptotically admissible on $[1, \ell_+(\gamma))$, namely, for every $\ell \in [1, \ell_+(\gamma))$ we have $\mathbb{E}L(\ell, \eta) \geq L(\ell, \eta^{**})$, with* *strict inequality for at least one point in* $[1, \ell_+(\gamma))$.

To determine the optimal shrinker η^* for each of our loss functions, it therefore remains to determine the map $\ell \mapsto \eta(\ell)$ or $\lambda \mapsto \eta(\lambda)$ only for $\ell > \ell_+(\gamma)$. This is our next task.

6.3 Optimal Shrinkers by Computer

The optimization problem (6.2) is easy to solve numerically, so that one can always compute the optimal shrinker at any desired value λ. In the data and code supplement [16] we provide Matlab code to compute the optimal nonlinearity for each of the 26 loss families discussed, as well as tabulated values of each.

6.4 Optimal Shrinkers in Closed Form

The rest of this section presents analytic formulas for the optimal shrinker η^* in each of 17 loss families from Section 3. While the optimal nonlinearities provided are of course functions of the empirical eigenvalue λ, in the interest of space, we provide the formulas in terms of the quantities ℓ, c and s. To calculate any of the nonlinearities below for a specific empirical eigenvalue λ, use the following procedure:

1. Calculate $\ell(\lambda)$ using (4.4)
2. Calculate $c(\lambda) = c(\ell(\lambda))$ using (1.2) and (4.4).
3. Calculate $s(\lambda) = s(\ell(\lambda))$ using $s(\ell) = \sqrt{1 - c^2(\ell)}$.
4. Substitute $\ell(\lambda)$, $c(\lambda)$ and $s(\lambda)$ into the formula provided[4].

The formulas we provide are summarized in Table 2. These formulas are derived in the following sequence of Lemmas, which are proved in Appendix A.1.

Lemma 6.2. (**Operator Norms.**) *For the direct operator norm loss* $L^{O,1}$ *and the operator norm loss on precision matrices* $L^{O,2}$*, we have*

$$\eta^*(\lambda; \gamma, L^{O,1}) = \eta^*(\lambda; \gamma, L^{O,2}) = \begin{cases} \ell, & \ell > \ell_+(\gamma) \\ 1, & \ell \le \ell_+(\gamma) \end{cases} . \tag{6.5}$$

This asymptotic relationship reflects the classical fact that in finite samples the top empirical eigenvalue is always biased upwards of the underlying empirical eigenvalue [60, 9]. Formally defining the (asymptotic) bias as

$$bias(\eta, \ell) = \eta(\lambda(\ell)) - \ell \,,$$

we have $bias(\lambda(\ell), \ell) > 0$. The formula $\eta^*(\lambda) = \ell$ shows that the optimal nonlinearity for operator norm loss is what we might simply call a *debiasing* transformation, mapping each empirical eigenvalue back to the value of its "original" population eigenvalue, and the corresponding shrinkage estimator $\hat{\Sigma}_\eta$ uses each sample eigenvectors with its corresponding population eigenvalue. In

[4]We stress that these formulas describe the optimal shrinkers only for $\lambda > \lambda_+(\gamma)$; all optimal shrinkers discussed set to 1 any $\lambda < \lambda_+(\gamma)$.

	MatrixNorm		
Pivot	Frobenius	Operator	Nuclear
$A-B$	$\ell c^2 + s^2$	ℓ	$\max\left(1+(\ell-1)(1-2s^2)\,,\,1\right)$
$A^{-1}-B^{-1}$	$\frac{\ell}{c^2+\ell s^2}$	ℓ	$\max\left(\frac{\ell}{c^2+(2\ell-1)s^2}\,,\,1\right)$
$A^{-1}B-I$	$\frac{\ell c^2+\ell^2 s^2}{c^2+\ell^2 s^2}$	N/A	$\max\left(\frac{\ell}{c^2+\ell^2 s^2}\,,\,1\right)$
$B^{-1}A-I$	$\frac{\ell^2 c^2+s^2}{\ell c^2+s^2}$	N/A	$\max\left(\frac{\ell^2 c^2+s^2}{\ell}\,,\,1\right)$
$A^{-1/2}BA^{-1/2}-I$	$\frac{(\ell-1)c^2}{(c^2+\ell s^2)^2}$	$\frac{\ell-1}{c^2+\ell s^2}$	$\max\left(\frac{\ell-(\ell-1)^2c^2s^2}{(c^2+\ell s^2)^2}\,,\,1\right)$
	Statistical Measures		
	St	Ent	Div
Stein	$\frac{\ell}{c^2+\ell s^2}$	ℓc^2+s^2	$\sqrt{\frac{\ell^2c^2+\ell s^2}{c^2+\ell s^2}}$
Fréchet	$\left(\sqrt{\ell}c^2+s^2\right)^2$		

Table 2: Optimal shrinkers $\eta^*(\lambda;\gamma,L)$ for 17 of the loss families L discussed. Values shown are shrinkers for $\lambda > \lambda_+(\gamma)$. All shrinkers obey $\eta^*(\lambda;\gamma,L)=1$ for $\lambda \le \lambda_+(\gamma)$. Here, ℓ, c and s depend on λ according to (4.4), (1.2) and $s=\sqrt{1-c^2}$. In the cases marked "N/A" we were not able to obtain the optimal shrinker in a simple, appealing form.

words, within the top branch of (6.5), the effect of *operator-norm optimal shrinkage is to debias the top eigenvalue*:

$$bias(\eta^*(\cdot;\gamma,L^{O,1}),\ell) = bias(\eta^*(\cdot;\gamma,L^{O,2}),\ell) = 0, \qquad \forall \ell > \ell_+(\gamma).$$

On the other hand, within the bottom branch, the effect is to *collapse the bulk to 1*. Comparing this with Definition 5.1, we see that η^* does not collapse the vicinity of the bulk to 1, only the bulk itself.

One might expect asymptotic debiasing from *every* loss function, but, perhaps surprisingly, precise asymptotic debiasing is exceptional. In fact, none of the following optimal nonlinearities is precisely debiasing.

Lemma 6.3. *(**Frobenius Matrix Norms.**) For the squared error losses $L_2^{F,1}$, $L_2^{F,2}$, $L_2^{F,3}$, $L_2^{F,4}$ and $L_2^{F,6}$, the optimal nonlinearity collapses the bulk to 1, and outside the bulk, obeys the formulas*

$$\eta^*(\lambda;\gamma,L_2^{F,1}) = \ell c^2 + s^2 \tag{6.6}$$

$$\eta^*(\lambda;\gamma,L_2^{F,2}) = \frac{\ell}{c^2+\ell s^2} \tag{6.7}$$

$$\eta^*(\lambda;\gamma,L_2^{F,3}) = \frac{\ell c^2+\ell^2 s^2}{c^2+\ell^2 s^2} \tag{6.8}$$

$$\eta^*(\lambda;\gamma,L_2^{F,4}) = \frac{\ell^2 c^2+s^2}{\ell c^2+s^2} \tag{6.9}$$

$$\eta^*(\lambda;\gamma,L_2^{F,6}) = \frac{(\ell-1)c^2}{(c^2+\ell s^2)^2}\,. \tag{6.10}$$

Lemma 6.4. *(**Nuclear Matrix Norms.**) For the squared error losses $L_2^{N,1}$, $L_2^{N,2}$, $L_2^{N,3}$, $L_2^{N,4}$ and $L_2^{N,6}$, the optimal nonlinearity collapses the bulk to 1, and outside the bulk, obeys the formulas*

$$\eta^*(\lambda;\gamma,L_2^{N,1}) = \max\big(1+(\ell-1)(1-2s^2)\,,\,1\big) \tag{6.11}$$

$$\eta^*(\lambda;\gamma,L_2^{N,2}) = \max\left(\frac{\ell}{c^2+(2\ell-1)s^2}\,,\,1\right) \tag{6.12}$$

$$\eta^*(\lambda;\gamma,L_2^{N,3}) = \max\left(\frac{\ell}{c^2+\ell^2 s^2}\,,\,1\right) \tag{6.13}$$

$$\eta^*(\lambda;\gamma,L_2^{N,4}) = \max\left(\frac{\ell^2 c^2+s^2}{\ell}\,,\,1\right) \tag{6.14}$$

$$\eta^*(\lambda;\gamma,L_2^{N,6}) = \max\left(\frac{\ell-(\ell-1)^2c^2s^2}{(c^2+\ell s^2)^2}\,,\,1\right)\,. \tag{6.15}$$

Lemma 6.5. *(**Stein, Entropy and Divergence Losses.**) For the Stein, Entropy and Divergence losses L_2^{st}, L_2^{ent} and L_2^{div}, the optimal nonlinearity collapses the bulk to 1, and outside the bulk, obeys the formula*

$$\eta^*(\lambda;\gamma,L_2^{st}) = \frac{\ell}{c^2+\ell s^2} \tag{6.16}$$

$$\eta^*(\lambda;\gamma,L_2^{ent}) = \ell c^2+s^2 \tag{6.17}$$

$$\eta^*(\lambda;\gamma,L_2^{div}) = \sqrt{\frac{\ell^2c^2+\ell s^2}{c^2+\ell s^2}}\,. \tag{6.18}$$

Lemma 6.6. *(**Fréchet Loss**) For the Fréchet loss $L_2^{fre}(A,B)$, the optimal nonlinearity collapses the bulk to 1, and outside the bulk, obeys the formula*

$$\eta^*(\lambda;\gamma,L_2^{fre}) = \left(\sqrt{\ell}c^2+s^2\right)^2\,. \tag{6.19}$$

7 Large-λ Asymptotics of the Optimal Shrinker

As the top theoretical eigenvalue $\ell \to \infty$, Theorem 4.1 shows that the top empirical eigenvalue is asymptotic to the top theoretical eigenvalue $\lambda(\ell)/\ell \to 1$, and that the corresponding empirical eigendirection becomes an increasingly accurate estimate of the underlying theoretical eigendirection. These observations might suggest that also $\eta^*(\lambda)/\lambda \to 1$ for $\lambda \to \infty$.

Figure 2 dashes this expectation, for several of the loss functions we are considering. In effect, there is a nontrivial asymptotic proportional shrinkage, $\eta(\lambda)/\lambda \to b < 1$ as $\lambda \to \infty$. We investigate this in more detail in subsection 7.1 below.

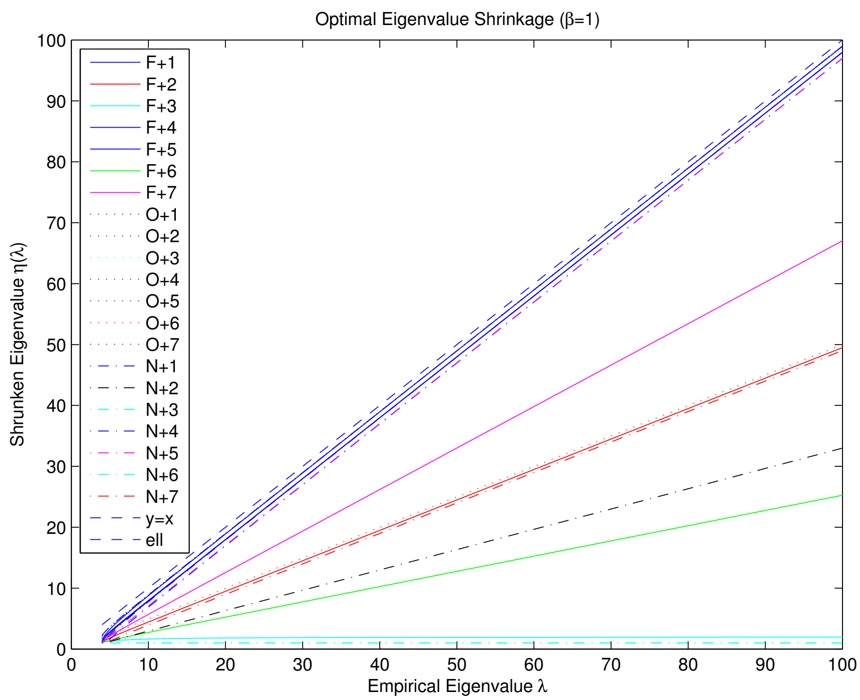

Figure 2: Optimal Shrinkers for 21 Component Loss Functions ($\gamma = 1$) – expanded range $4 < \lambda < 100$. Coloring classifies shrinkers by their asymptotic slope. Blue: 1. Magenta: roughly .7. Red: $1/2$, Black dashdot $1/3$, Green: $1/4$. Several loss functions (indicated with cyan), demand extreme shrinkage. *Reproducibility advisory:* The data and code supplement [16] includes a script that generates any of these individual curves.

For other loss functions, we may still have $\eta(\lambda)/\lambda \to 1$ as $\lambda \to \infty$, but nevertheless there is a nontrivial asymptotic shift $\eta(\lambda) - \lambda \to a$ as $\lambda \to \infty$, with $a \neq 0$. We study this in detail in subsection 7.2 below.

Finally, one might imagine that because there is asymptotically stronger 'signal' as ℓ grows, that $Loss(\ell, \eta^*) \to 0$ as $\ell \to \infty$, or at least that the optimal shrinker behaves 'much better' than the non-shrinker $\eta(\lambda) = \lambda$. We study this in detail in subsection 7.3 below.

7.1 Asymptotic Slopes

Definition 7.1. *The* asymptotic slope *of a shrinker η is*

$$asySlope(\eta) = \lim_{\lambda\to\infty} \frac{\eta(\lambda)}{\lambda}.$$

The bias *of a shrinker is defined by*

$$bias(\eta, \ell) = \eta(\lambda(\ell)) - \ell\,.$$

Note that the asymptotic slope of a shrinker η is connected to its biased by

$$\lim_{\ell\to\infty} \frac{bias(\eta,\ell)}{\ell} = asySlope(\eta) - 1\,.$$

Lemma 7.1. *(***Asymptotic Slopes for Shrinkers with Closed Forms***) We have the asymptotic slopes in Table 3.*

	MatrixNorm		
Pivot	Frobenius	Operator	Nuclear
$A - B$	1	1	1
$A^{-1} - B^{-1}$	$\frac{1}{1+\gamma}$	1	$\frac{1}{1+2\gamma}$
$A^{-1}B - I$	0	N/A	0
$B^{-1}A - I$	1	N/A	1
$A^{-1/2}BA^{-1/2} - I$	$\frac{1}{(1+\gamma)^2}$	$\frac{1}{1+\gamma}$	$\frac{1-\gamma}{(1+\gamma)^2}$
	Statistical Measures		
	St	Ent	Div
Stein	$\frac{1}{1+\gamma}$	1	$\frac{1}{\sqrt{1+\gamma}}$
Fréchet	1		

Table 3: Asymptotic Slopes; See Lemma 7.1

Proof. First observe that $asySlope(\eta) = 1 + \lim_{\ell\to\infty} bias(\eta,\ell)/\ell = \lim_{\ell\to\infty} \eta(\lambda(\ell))/\ell$. Using the identities $\lim_{\ell\to\infty} c^2(\ell) = 1$, $\lim_{\ell\to\infty} s^2(\ell) = 0$ and $\lim_{\ell\to\infty} \ell s^2(\ell) = \gamma$, calculating the limit $\lim_{\ell\to\infty} \eta(\lambda(\ell))/\ell$ for each of the shrinkers in Table 2 is a simple calculus exercise. $\square$

All our shrinkers with $asySlope(\eta) = 1$ are approximately unbiased for large λ; while shrinkers with $asySlope(\eta) < 1$ are noticeably biased for large λ.

For example, the optimal shrinker for $L^{F,2}$ is noticeably biased as $\lambda \to \infty$. For example when $\gamma = 1$, so $n \sim p$, it performs 50% asymptotic shrinkage. The same is true for the optimal shrinker for L^{st}. The divergence loss shrinker is less biased than either of these: it sits between those for L^{st} and L^{ent}, as $1 > \frac{1}{\sqrt{1+\gamma}} > \frac{1}{1+\gamma}$.

In two cases where we did not obtain closed-form expressions for the optimal shrinker, we nevertheless were able to obtain the asymptotic slopes:

Lemma 7.2. (**Asymptotic Slopes for $L^{F,7}$ and L^{aff}**) *For the Bhattacharya/Matusita Affinity, the asymptotic slope is the unique solution $b^* \in (0,1)$ to*

$$b^{3/2} = \frac{2}{\gamma}(1-b).$$

For $L^{F,7}$ the asymptotic slope is the value $b^* \in (0,1)$ minimizing

$$J(b;\gamma) = \sum_{\pm} \log^2(\lambda_{\pm}(b;\gamma))$$

where

$$\lambda_{\pm}(b;\gamma) = \Big((1+b(1+\gamma)) \pm \sqrt{(1+b(1+\gamma))^2 - 4b}\Big)/2$$

In cases where we could not determine asymptotic slopes analytically, we resorted to numerical studies; we simply define $\widehat{asySlope}(\eta) = \frac{\eta(\lambda)}{\lambda}|_{\lambda=100}$ (Figure 3). Table 4 compiles results obtainable by either analytical or numerical means, in case $\gamma = 1$.

	MatrixNorm		
Pivot	Frobenius	Operator	Nuclear
$A - B$	1	1	1
$A^{-1} - B^{-1}$	1/2	1	1/3
$A^{-1}B - I$	0	0	0
$B^{-1}A - I$	1	1	1
$A^{-1}B + B^{-1}A - 2I$	1	1	0.7
$A^{-1/2}BA^{-1/2} - I$	1/4	1/2	0
$\log(A^{-1/2}BA^{-1/2})$	$\approx 2/3$	1	1/2
	Statistical Measures		
	St	Ent	Div
Stein	1/2	1	$1/\sqrt{2}$
Affinity	$\approx .66$		
Fréchet	1		

Table 4: Approximate Asymptotic Slopes; $\gamma = 1$

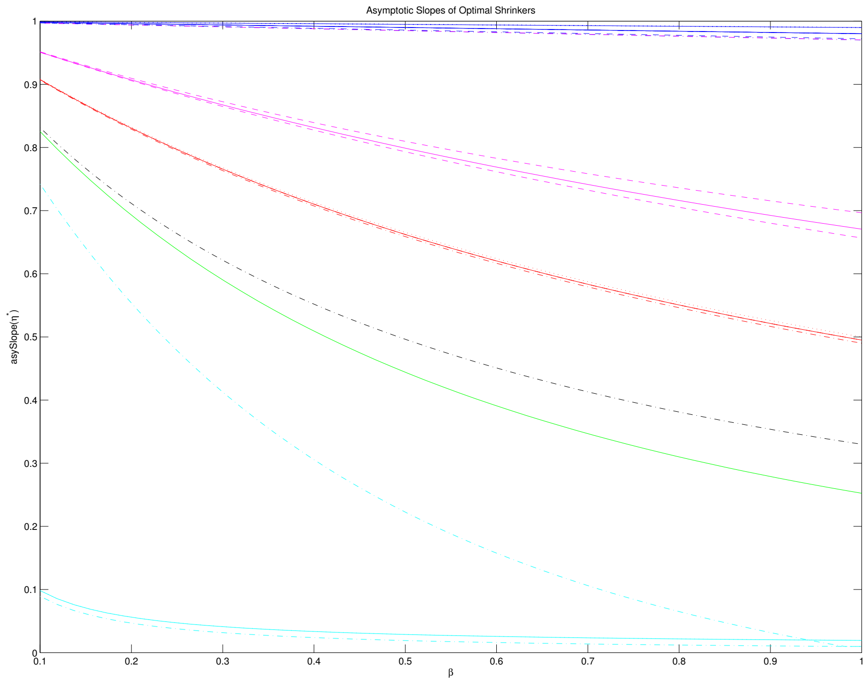

Figure 3: Approximate Asymptotic Slopes $\widehat{asySlope}(\eta, \gamma)$ for the 26 loss functions studied in this paper. In all cases shrinkage becomes more extreme as $\gamma \to 1$. The clustering of the 26 slopes into discrete clusters is evident. (Line and color convention is identical to Figure 2.) *Reproducibility advisory:* The data and code supplement [16] includes a script that generates any of these individual curves.

7.2 Asymptotic Shifts

Definition 7.2. *Conside a shrinker η with asymptotic slope 1. The* asymptotic shift *is*

$$\mathit{asyShift}(\eta) = \lim_{\lambda \to \infty} \eta(\lambda) - \lambda.$$

Observe that the asymptotic shift of a shrinker with asymptotic slope 1 is connected to its asymptotic bias by $asyShift(\eta) = \lim_{\ell \to \infty} bias(\eta, \ell) - \gamma$.

Lemma 7.3. *(**Asymptotic Shifts**) We have the asymptotic shifts in Table 5.*

The entries in this table are negative, meaning that asymptotically the optimal nonlinearity is pulling *down*. In cases where we could not determine asymptotic shift analytically, we resorted to numerical studies. One may compute numerical approximations to the asymptotic shifts by defining $\widehat{asyShift}(\eta) = (\eta(\lambda) - \lambda)|_{\lambda=100}$. Table 6 enumerates 4 additional cases with unit asymptotic slope, and the corresponding asymptotic shifts, determined numerically.

Loss	asyShift	asymptotic bias
$L^{F,1}$	-2γ	$-\gamma$
$L^{O,1}$	$-\gamma$	0
$L^{N,1}$	-3γ	-2γ
$L^{O,2}$	$-\gamma$	0
$L^{F,4}$	$-\gamma$	0
$L^{N,4}$	-2γ	$-\gamma$
L^{ent}	-2γ	$-\gamma$
L^{fre}	-3γ	-2γ

Table 5: Asymptotic Shifts; See Lemma 7.3

Loss	$\widehat{asyShift}(\eta; \gamma = 1)$	Inferred Formula
$L^{F,5}$	-2	-2γ
$L^{O,4}$	-1	$-\gamma$
$L^{O,5}$	-1	$-\gamma$
$L^{O,7}$	-1	$-\gamma$

Table 6: Approximate Asymptotic Shifts at $\gamma = 1$, rounded so eg. $-.990$ becomes -1.

7.3 Asymptotic Percent Improvement

One simple covariance estimation method is *Hard Thresholding* of the empirical eigenvalues, where the nonlinearity η applied to each λ_i is of the form $\eta : \lambda \mapsto 1 + (\lambda - 1)\mathbf{1}_{\lambda > \alpha}$ for some threshold α. A reasonable choice for the threshold α is the edge of the Marčenko-Pastur Bulk (4.1), leading to the bulk-edge hard threshold nonlinearity, $\eta^{\#}(\lambda) = 1 + (\lambda - 1)1_{\{\lambda > (1+\sqrt{\gamma})^2\}}$. This estimator appears implicitly in the multivariate analysis literature and can be traced back to the famous Scree Plot method [10] (see [31, ch. 6]).

To evaluate the improvement offered by the optimal shrinkers we propose, we computed percent improvement in asymptotic loss over the "naive" estimator $\eta^{\#}$.

Definition 7.3. *Let η^* denote the optimal shrinkage procedure and let $L_2(A(\ell_1), B(\ell_1, \eta^*))$ denote its component loss. The Possible Percent Improvement (PPI) measures the extent to which η^* delivers a reduction below the asymptotic loss suffered by bulk edge thresholding $\eta^{\#}$, with 100 meaning complete elimination of all loss:*

$$PPI = 100 \times \frac{L(A(\ell_1), B(\ell_1, \eta^{\#})) - L(A(\ell_1), B(\ell_1, \eta^*))}{L(A(\ell_1), B(\ell_1, \eta^{\#}))}.$$

Figure 4 shows that, for many Loss functions, $\eta^*(\cdot|L)$ succeeds in mitigating most of the asymptotic loss experienced by the naive covariance estimator for ℓ in the range $2 < \ell < 10$. In this range of moderate ℓ, the optimal estimator in some cases reduces the risk by 50% and in some cases much more.

Note that for some choices of L, the PPI appears to decay to zero with increasing ℓ, while in some others, it does not decay to zero. In the latter cases, the optimal rule provides a meaning-

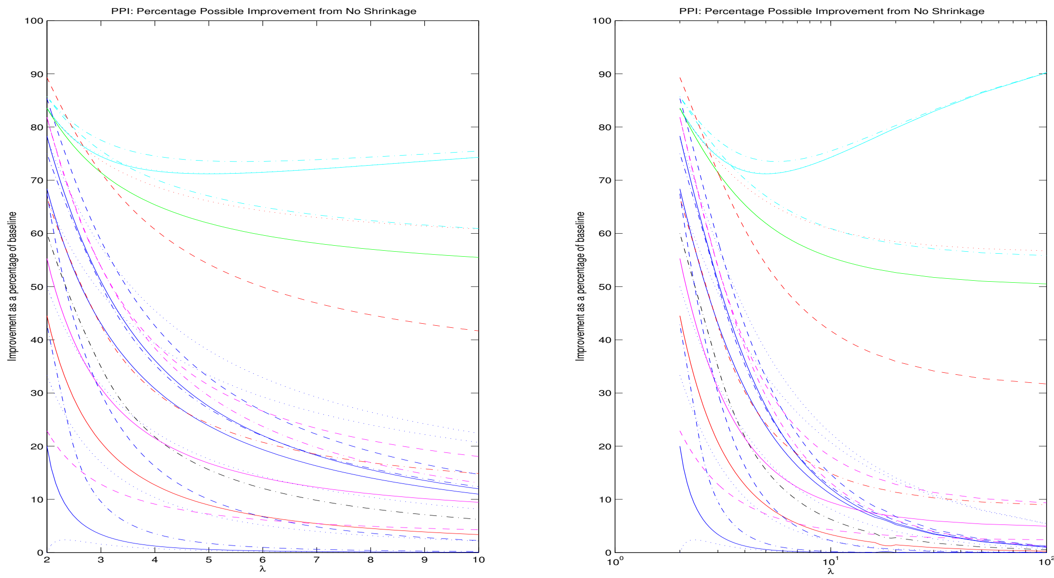

Figure 4: Loss Reduction as Percent of Possible Improvement (PPI), for each of the 26 loss functions, as a function of ℓ. Here $\gamma = 1$. In all cases, the asymptotic "weak signal" $\ell \to 2$ improvement over bulk-edge hard thresholding $\eta^{\#}$ is substantial. In most cases, the asymptotic "strong signal" $\lambda \to \infty$ improvement over $\eta^{\#}$ is negligible; but in several cases it is quite substantial. (Line and color convention is identical to Figure 2.) *Reproducibility advisory:* The data and code supplement [16] includes a script that generates any of these individual curves.

ful improvement on the naive rule *even when the signal is extremely strong*. Let the Asymptotic Percentage of Possible Improvement *asyPPI* denote the large-ℓ limit of $PPI(L, \ell)$. In many cases $asyPPI(L) = 0$; but Table 7 identifies 7 cases where $asyPPI > 0$; they are associated with Pivots $A^{-1}B - I$, $A^{-1/2}BA^{-1/2} - I$ and with Stein Loss.

	MatrixNorm		
Pivot	Frobenius	Operator	Nuclear
$A^{-1}B - I$	100	100	50
$A^{-1/2}BA^{-1/2} - I$	50	56	56
	Statistical Measures		
Stein	30		

Table 7: Asymptotic PPI; $\gamma = 1$. All Figures in percentage points, max 100. Figures not listed are 0.

8 Beyond Formal Optimality

The shrinkers we have derived and analyzed above are formally optimal, as in Defintion 6.1, in the sense that they minimize a formal expression $L_\infty(\ell|\eta)$. So far we have only shown that formally optimal shrinkers actually minimize the asymptotic loss (namely, are asymptotically unique admissible) in the single-spike case, under assumptions **[Asy(γ)]** and **[Spike(ℓ_1)]**, and only over nonlinearities η that collapse the vicinity of the bulk to 1.

In this section, we show that formally optimal shrinkers in fact minimize the asymptotic loss in the general Spiked Covariance model, namely under assumptions **[Asy(γ)]** and **[Spike($\ell_1, \ldots, \ell_r$)]**, and over a large class of nonlinearities that collapse the bulk, but possibly not the vicinity of the bulk, to 1.

8.1 The Multiple Spike Case

Lemma 8.1. A deterministic formula for the asymptotic loss: multiple non-degenerate spikes. *Let $L = \{L_p\}$ be a family of loss functions satisfying the assumptions of Lemma 5.2. Consider a problem sequence with $n, p \to \infty$, with $p = p_n$ obeying $p_n/n \to \gamma$, with theoretical covariance matrices Σ_{p_n} obeying the multiple-spike model with fixed top r eigenvalues $\ell = (\ell_1, \ldots, \ell_r)$* all distinct, *such that $\ell_i > \ell_+(\gamma)$, $i = 1, \ldots, r$. Suppose the scalar nonlinearity η collapses the vicinity of the bulk to 1 and is continuous. Then*

$$L_{p_n}\left(\Sigma_{p_n}\,,\,\hat{\Sigma}_\eta(S_{n,p_n})\right) \to_P L_\infty(\ell_1, \ldots, \ell_r|\eta) \equiv \sum_{i=1}^{r} L_2\left(A(\ell_i)\,,\,B\left(\eta(\lambda(\ell_i), c(\ell_i), s(\ell_i))\right)\right)\,, \tag{8.1}$$

where $\lambda(\ell)$, $c(\ell)$ and $s(\ell)$ are as in Lemma 5.2 above.

Corollary 8.1. *The optimal shrinker η^* in the single-spike model is also optimal in the multiple-spike model:*

$$L_\infty(\ell_1, \ldots, \ell_r|\eta^*) = \min_\eta L_\infty(\ell_1, \ldots, \ell_r|\eta), \qquad \forall \ell \geq 1\,,$$

where η^* is the optimal shrinker of Definition 6.2 corresponding to the loss L.

It follows that the optimal shrinkers we derived correspond to the unique admissible rule, in the asymptotic sense, among rules of the form $\hat{\Sigma}_\eta(S_{n,p}) = V\eta(\Lambda)V'$ in the general Spiked Covariance Model (with $r \geq 1$ spikes), and over nonlinearities that collapse the vicinity of the bulk to 1.

Lemma 8.1 is a technical generalization of Lemma 5.2. To prove it, one proceeds as in the proof of Lemma 5.2 above, except that the block-diagonalizing basis W is constructed more generally by applying the Gram-Schmidt process to

$$u_{1,n}, u_{2,n}, \ldots, u_{r,n}, v_{1,n}, v_{2,n}, \ldots, v_{r,n}, v_{r+1,n}, v_{r+2,n}, \ldots, v_{p-r,n}$$

and then permuting outputs, interleaving the output vectors as $1, r+1, 2, r+2, \ldots, r, 2r, 2r+1, \ldots, p-2r$. Denote the resulting basis vectors by $w_1, \ldots w_p$. The interleaving associates the sequence of index pairs $(1,2), (3,4), \ldots, (2r-1, 2r)$ to a block structure. Lemma 8.1 now follows from the following result, proved in Appendix A.2:

Lemma 8.2. *Assume* **[Asy(γ)]** *and* **[Spike($\ell_1, \ldots, \ell_r$)]**, *with $\ell_1, \ldots, \ell_r > \ell_+(\gamma)$ all distinct. As before, write $\Sigma_p = diag(\ell_1, \ldots, \ell_r) \oplus I_{p-2r}$ for the population covariance matrix at a given value of p. Let W denote* *the basis described above and denote $c_i = c(\ell_i)$, $s_i = s(\ell_i)$ and $\eta_i = \eta(\lambda(\ell_i))$. Also let $A(\ell) = diag(\ell, 1)$ and let $B(\eta, c, s)$ be as in* (2.5). *Then*

$$W' \, \Sigma_p \, W = (\oplus_{i=1}^r A(\ell_i)) \oplus I_{p-2r} \,.$$

Let η collapse the vicinity of the bulk to zero. Then

$$\| W' \, \hat{\Sigma}_\eta(S_{n,p_n}) \, W - \widetilde{\Sigma}_{p_n} \|_F \to_P 0 \,,$$

where

$$\widetilde{\Sigma}_p = (\oplus_{i=1}^r B(\eta_i, c_i, s_i)) \oplus I_{p-2r} \,. \tag{8.2}$$

Empirically and heuristically, Lemma 8.1 remains valid in the degenerate case, where the nontrivial population eigenvalues are not necessarily distinct. The construction of W would have to be different in the degenerate case; see discussion in [17].

8.2 Nonlinearities that Only Collapse the Bulk

So far we have shown that the optimal shrinkers we derived are asymptotically unique admissible over nonlinearities that collapse the vicinity of the bulk to 1. The following result, proved in Appendix A.2, shows that they remain asymptotically unique admissible over a larger class of nonlinearities, some of which collapse the bulk, but not the vicinity of the bulk, to 1.

Lemma 8.3. *Assume that the scalar nonlinearity η is continuous, and assume one of the following conditions:*

1. *η collapses the vicinity of the bulk to 1.*
2. *η collapses the bulk to 1, is monotone non-decreasing in a neighborhood of the bulk edge, and is α-Hölder at the bulk edge for some $0 < \alpha \le 1$.*

Then Lemma 8.1 holds.

It follows that the optimal shrinkers we derived are asymptotically unique admissible over nonlinearities that either collapse the vicinity of the bulk to 1, or collapse the bulk to 1 and are Hölder and non-decreasing at the bulk edge.

9 Optimality Among Equivariant Procedures

The notion of optimality in asymptotic loss, with which we have been concerned, is fairly weak. Also, the class of covariance estimators we have considered, namely procedures that apply the *same* univariate shrinker to all empirical eigenvalues, is fairly restricted.

Naturally the reader will at this point suppose that the procedures we proposed are optimal *only* with respect to their respective limiting loss, and *only* within the class of estimators of the form $\hat{\Sigma}(S) = V\eta(\Lambda)V'$, where η is a single nonlinearity applied in turn to each of the empirical eigenvalues λ_i, and where, as before, the columns of V are the empirical eigenvectors.

Consider the much broader class of orthogonally-equivariant procedures for covariance estimation [58, 45, 49], in which estimates take the form $\hat{\Sigma} = V \, \Delta \, V'$. Here, $\Delta = \Delta(\Lambda)$ is any diagonal matrix that depends on the empirical eigenvalues Λ in possibly a more complex way than simple scalar element-wise shrinkage $\eta(\Lambda)$. One might imagine that the extra freedom available with more general shrinkage rules would lead to improvements in loss, relative to our optimal scalar nonlinearity; certainly the proposals of [58, 45, 43] are of this more general type.

The smallest achievable loss by any orthogonally equivariant procedure is obtained with the "oracle" procedure $\hat{\Sigma}^{oracle} = V\,\Delta^{oracle}\,V'$, where Δ^{oracle} is chosen to be the minimizer

$$\Delta^{oracle} = \text{argmin}_{\Delta} L(\Sigma, V\,\Delta\,V'),$$

the minimum being taken over diagonal matrices with diagonal entries ≥ 1. Clearly, this optimal performance is not attainable, since the minimization problem explicitly demands perfect knowledge of Σ. This knowledge is never available to us in practice – hence the label *oracle* [5]. Nevertheless, this optimal performance is a legitimate benchmark.

Interestingly, at least in the popular Frobenius case $L^{F,1}$ and in the Stein Loss case L^{st}, the optimal nonlinearities η^*, which we have derived, deliver oracle-level performance – asymptotically. The following theorem is proved in Appendix A.3.

Theorem 9.1. *(***Asymptotic-loss optimality among all equivariant procedures.***) Let L denote either the direct Frobenius Loss $L^{F,1}$ or the Stein Loss L^{st}. Consider a problem sequence satisfying assumptions* **[Asy(γ)]** *and* **[Spike($\ell_1, \ldots, \ell_r$)]** *where the fixed top r eigenvalues $\ell = (\ell_1, \ldots, \ell_r)$ are* all *distinct. For Σ_{p_n} and $\hat{\Sigma}_{p_n}^{oracle}$, the p_n-by-p_n matrices in our sequence of statistical estimation problems, we have*

$$\lim_{n\to\infty} L_{p_n}(\Sigma_{p_n}, \hat{\Sigma}_{p_n}^{oracle}) =_P L_\infty(\ell_1 \ldots, \ell_r | \eta^*)\,, \tag{9.1}$$

where η^* is the optimal shrinker (Definition 6.2) corresponding to the loss L.

In short, the shrinker $\eta^*()$, which has been designed to minimize the *limiting* loss, asymptotically delivers the same performance as the oracle procedure, which has the lowest possible loss over the entire class of covariance estimators by arbitrary high-dimensional shrinkage rules. On the other hand, by definition, the oracle procedure outperforms every orthogonally-equivariant statistical estimator. We conclude that η^* – as one such orthogonally-invariant estimator – is indeed optimal (in the sense of having the lowest limiting loss) among all orthogonally invariant procedures. While we only discuss the cases $L^{F,1}$ and L^{st}, we suspect that this theorem holds true for many, or all, of the 26 loss functions considered.

10 Discussion

In the present paper, we consider the problem of covariance estimation in high dimensions, where the dimension p is comparable to the number of observations n. We choose a fixed-rank principal subspace, and let the dimension of the problem grow large. A different asymptotic framework for covariance estimation would choose a principal subspace that is a fixed fraction of the problem dimension; i.e. the rank of the principal subspace is growing rather than fixed. (In the sibling problem of matrix denoising, compare the "spiked" setup [15, 54] with the "fixed fraction" setup of [14].)

In the fixed fraction framework, some of underlying phenomena remain qualitatively similar to those governing the spiked model, while new effects appear. Importantly, the relationships used

[5]The oracle procedure does not attain zero loss since it is "doomed" to use the eigenbasis of the empirical covariance, which is a random basis corrupted by noise, to estimate the population covariance.

in this paper, predicting the location of the top empirical eigenvalues, as well as the displacement of empirical eigenvectors, in terms of the top theoretical eigenvalues, no longer hold. Instead, one has a complex nonlinear relationship between the limiting distribution of the empirical eigenvalues and the limiting distribution of the theoretical eigenvalues, as expressed by the Marčenko-Pastur (MP) relation between their Stieltjes transforms [47, 2].

Covariance shrinkage in the proportional rank model should then, naturally, make use of the so-called *MP Equation*. Noureddine El Karoui [34] proposed a method for debiasing the empirical eigenvalues, namely, for estimating (in a certain specific sense) their corresponding population eigenvalues; Olivier Ledoit and Sandrine Peché [41] developed analytic tools to also account for the inaccuracy of empirical eigenvectors, and Ledoit and Michael Wolf [43] have implemented such tools and applied them in this setting.

We admire the depth of insight involved in the study of the proportional rank case, which is really quite subtle and beautiful. Yet, the fixed-rank case deserves to be worked out carefully. In particular, the shrinkers we have obtained here in the fixed-rank case are extremely simple to implement, requiring just a few code lines in any scientific computing language. In comparison, those covariance estimation ideas, based on powerful and deep insights from MP theory, require a delicate, nontrivial effort to implement in software, and call for expertise in numerical analysis and optimization. As a result, the simple shrinkage rules we propose here are more likely to be applied correctly in practice, and to work as expected, even in relatively small sample sizes.

An analogy of this situation can be made to shrinkage in the normal means problem, for example [18]. In applications of that problem, often a full Bayesian model applies, and in principle a Bayesian shrinkage would provide an optimal result [8]. Yet, in applications one often wants a simple method which is easy to implement correctly, and which is able to deliver much of the benefit of the full Bayesian approach. In literally thousands of cases, simple methods of shrinkage - such as thresholding - have been chosen over the full Bayesian method for precisely that reason.

Reproducible Research

In the data and code supplement [16] we offer a Matlab software library that includes:

1. A function to compute the value of each of the 26 optimal shrinkers discussed to high precision.
2. A function to compute the optimal risk corresponding to these optimal shrinkers.
3. A function to test the correctness of each of the 17 analytic shrinker fomulas provided.
4. Tabulated values of the optimal shrinkers, and a function for fast evaluation of each of the 26 optimal shrinkers based on interpolation of tabulated values.
5. Scripts that generate each of the figures in this paper, or subsets of them for specified loss functions.

Acknowledgements

We thank Amit Singer, Andrea Montanari, Sourav Chatterjee and Boaz Nadler for helpful discussions and advice. This work was partially supported by NSF DMS-0906812 (ARRA). MG was partially supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship.

A Proofs

A.1 Closed Forms of Optimal Shrinkers

Here we derive the closed forms for the optimal shrinkers, summarized in Table 2. In the interest of space, only key steps are discussed in detail. Additional details may be found in the supplemental article [17].

Proof of Lemma 6.1

Let L be any of the loss functions considered. The following condition can be easily directly verified for each of our 26 losses in turn: For any $\ell \in [1, 1+\sqrt{\gamma}]$ and any $\eta \geq 1$, the function $G(\eta, \ell)$ of (6.4) satisfies $G(\eta, \ell) = g_1(\eta) + g_2(\ell)$ or $G(\eta, \ell) = \max\{g_1(\eta)\,,\, g_2(\ell)\}$, where g_1,g_2 are continuous and strictly increasing on $[1, \sqrt{\gamma}]$ and $[1, \infty)$, respectively, and where $g_1(1) = g_2(1) = 0$.

Now, assume such a decomposition of $G(\ell, \eta)$. Let η be any random variable with values in $[1, 1+\sqrt{\gamma}]$. It is enough to show that

$$\mathbb{E}G(\ell, \eta) \geq G(\ell, 1)\,,$$

with strict inequality holding for some sub-interval $[1, \ell'] \subset [1, \sqrt{\gamma}]$. Indeed, if $G = g_1 + g_2$ then $\mathbb{E}g_2(\eta) > 0$ unless $\eta \equiv 1$ almost surely. And if $G = \max\{g_1, g_2\}$, then we have $G(\ell, \eta) - G(\ell, 1) = \max\{g_1(\ell), g_2(\eta)\} - g_1(\ell)$ and it follows that

$$D(\ell) = \mathbb{E}G(\ell, \eta) - G(\ell, 1) \geq \mathbb{E}(g_2(\eta) - g_1(\ell))_+ \geq 0$$

is a continuous function on $[1, 1+\sqrt{\gamma}]$ with $D(1) = \mathbb{E}g(\eta) > 0$ unless $\eta \equiv 1$ almost surely. □

Turning now to derive the explicit formulas of the optimal shrinkers for $\ell > \ell_+(\gamma)$, we repeatedly will use two simple observations regarding 2-by-2 matrices:

Lemma A.1. *Let S be a symmetric 2-by-2 matrix. Then $\|S\|_F^2 = \mathrm{trace}(S)^2 - 2\det(S)$.*

Lemma A.2. *The eigenvalues of any 2-by-2 matrix M with trace $\mathrm{trace}(M)$ and determinant $det(M)$ are given by*

$$\lambda_{\pm}(M) = \frac{1}{2}\left(\mathrm{trace}(M) \pm \sqrt{\mathrm{trace}(M)^2 - 4\det(M)}\right) \tag{A.1}$$

These have useful corollaries:

Lemma A.3. *Assume that $\Delta(\eta)$ is a 2-by-2 symmetric matrix and $\eta^* \geq 1$ is such that*

$$\mathrm{trace}(\Delta(\eta^*)) \cdot \frac{d}{d\eta}\mathrm{trace}(\Delta(\eta))\Big|_{\eta=\eta^*} = \frac{d}{d\eta}\det(\Delta(\eta))\Big|_{\eta=\eta^*}\,. \tag{A.2}$$

Then η^* solves the Frobenius norm minimization problem: $\eta^* = argmin_{\eta\geq 1}\|\Delta(\eta)\|_F$.

Proof. Differentiate with respect to η both sides of the equality in Lemma A.1 and equate to zero. □

Lemma A.4. *Assume that $\Delta(\ell, \eta)$ is a 2-by-2 matrix and such that $\det(\Delta(\eta)) \leq 0$ for all $\eta \geq 1$, and assume that $\eta^* \geq 1$ is such that*

$$\mathrm{trace}(\Delta(\eta^*)) \cdot \frac{d}{d\eta}\mathrm{trace}(\Delta(\eta))\Big|_{\eta=\eta^*} = 2 \cdot \frac{d}{d\eta}\det(\Delta(\eta))\Big|_{\eta=\eta^*}\,. \tag{A.3}$$

Then η^* solves the nuclear norm minimization problem: $\eta^* = argmin_{\eta\geq 1}\|\Delta(\eta)\|_*$.

Proof. The condition $\det(\Delta) \leq 0$ is equivalent to the condition $|\mathrm{trace}(\Delta)| \leq \sqrt{\mathrm{trace}(\Delta)^2 - 4\det(\Delta)}$, which implies that $\lambda_+(\Delta) \geq 0$ and $\lambda_-(\Delta) \leq 0$. Hence,

$$\|\Delta\|_*^2 = (|\lambda_+| + |\lambda_-|)^2 = (\lambda_+(\Delta) - \lambda_-(\Delta))^2 = \mathrm{trace}(\Delta)^2 - 4\det(\Delta)\,,$$

using Lemma A.2. Now differentiate both sides and equate to zero. □

Finally, we will use the following fact regarding singular values of 2-by-2 matrices:

Lemma A.5. *Let Δ be a 2-by-2 matrix with singular values $\sigma_+ \geq \sigma_- > 0$. Define $t = \mathrm{trace}(\Delta'\Delta) = \|\Delta\|_F^2$, $d = \det(\Delta)$ and $r^2 = t^2 - 4e^2$. Assume that Δ depends on a parameter η and let $\dot\sigma_\pm, \dot t, \dot e$ denote the derivative of these quantities w.r.t the parameter η. Then*

$$r\left(\dot\sigma_+ + \dot\sigma_-\right)\left(\dot\sigma_+ - \dot\sigma_-\right) = 2\left(\dot t + 2\dot e\right)\left(\dot t - 2\dot e\right)\,.$$

Proof. By Lemma A.2 we have $2\sigma_\pm^2 = t \pm r$ and therefore

$$\sqrt{2}\dot\sigma_\pm = \frac{\dot t \pm \dot r}{2\sqrt{t \pm r}}\,.$$

Differentiating and expanding $\dot\sigma_+ \pm \dot\sigma_-$ we obtain the relation

$$(\dot\sigma_+ + \dot\sigma_-) = \frac{(8d/r)(\dot t^2 - 4\dot e^2)}{(t^2 - r^2)(\dot\sigma_+ - \dot\sigma_-)}$$

and the result follows.

□

Keeping in mind that $c^2 + s^2 = 1$, the reader can verify the following auxilliary facts by direct calculation:

Lemma A.6. *Define the following auxiliary quantities:*

$$\begin{aligned}\tilde\ell &= \ell - 1\\ \bar\ell &= \ell^{-1} - 1\\ \tilde\eta &= \eta - 1\\ \bar\eta &= \eta^{-1} - 1\,.\end{aligned}$$

Then for the matrix $A(\ell) = diag(\ell, 1)$ and the matrix $B(\ell)$ (which depends on η) of (5.1) *we have:*

Proof of Lemma 6.3

We use Lemma A.6 as needed. For $L^{F,1}$ and $L^{F,6}$, observe that $\Delta_1(\tilde\eta)$ and $\Delta_6(\tilde\eta)$ are symmetric and solve equation (A.2) for the matrices $\Delta_1(\tilde\eta)$ and $\Delta_6(\tilde\eta)$, respectively. Note that minimizing w.r.t $\tilde\eta$ is equivalent to minimizing w.r.t η. For $L^{F,2}$, observe that $\Delta_2(\bar\eta)$ is symmetric and solve equation (A.2) for the matrix $\Delta_2(\bar\eta)$. Note that minimizing w.r.t $\bar\eta$ is equivalent to minimizing w.r.t η.

Next, define the 2-by-2 matrix-valued function

$$\delta(\eta, \ell, \alpha) = \begin{bmatrix}\ell + \alpha\eta c^2 & \alpha\eta cs\\ \eta cs & \eta s^2\end{bmatrix}\,. \tag{A.4}$$

Direct calculation shows that

$$\mathrm{argmin}_{\eta\geq 1}\|\delta(\eta, \ell, \alpha)\|_F^2 = \frac{-\ell\alpha c^2}{s^2 + \alpha^2 c^2}\,.$$

Since $\Delta_3 = \delta(\tilde\eta, \bar\ell, 1/\ell)$ and $\Delta_4 = \delta(\bar\eta, \tilde\ell, \ell)$, substitution yields the minimizers of $L^{F,3}$ and $L^{F,4}$. □

$$
\begin{array}{llll}
B = & \begin{bmatrix} 1+\tilde{\eta}c^2 & \tilde{\eta}cs \\ \tilde{\eta}cs & 1+\tilde{\eta}s^2 \end{bmatrix} & \mathrm{trace}(B) = \eta + 1 & \det(B) = \eta \\
B^{-1} = & \begin{bmatrix} 1+\bar{\eta}c^2 & \bar{\eta}cs \\ \bar{\eta}cs & 1+\bar{\eta}s^2 \end{bmatrix} & \mathrm{trace}(B^{-1}) = \eta^{-1} + 1 & \det(B^{-1}) = \eta^{-1} \\
\Delta_1 \equiv A - B = & \begin{bmatrix} \tilde{\ell}-\tilde{\eta}c^2 & \tilde{\eta}cs \\ -\tilde{\eta}cs & -\tilde{\eta}s^2 \end{bmatrix} & \mathrm{trace}(\Delta_1) = \tilde{\ell} - \tilde{\eta} & \det(\Delta_1) = -\tilde{\ell}\tilde{\eta}s^2 \\
\Delta_2 \equiv A^{-1} - B^{-1} = & \begin{bmatrix} \bar{\ell}-\bar{\eta}c^2 & \bar{\eta}cs \\ -\bar{\eta}cs & -\bar{\eta}s^2 \end{bmatrix} & \mathrm{trace}(\Delta_2) = \bar{\ell} - \bar{\eta} & \det(\Delta_2) = -\bar{\ell}\bar{\eta}s^2 \\
\Delta_3 \equiv A^{-1}B - I = & \begin{bmatrix} \bar{\ell}-\tilde{\eta}c^2/\ell & \tilde{\eta}cs/\ell \\ -\tilde{\eta}cs & -\tilde{\eta}s^2 \end{bmatrix} & \mathrm{trace}(\Delta_3) = \bar{\ell} + \tilde{\eta}(c^2/\ell + s^2) & \det(\Delta_3) = \bar{\ell}\tilde{\eta}s^2 \\
\Delta_4 \equiv B^{-1}A - I = & \begin{bmatrix} \tilde{\ell}-\bar{\eta}\ell c^2 & \bar{\eta}\ell cs \\ \bar{\eta}cs & \bar{\eta}s^2 \end{bmatrix} & \mathrm{trace}(\Delta_4) = \tilde{\ell} + \bar{\eta}(\ell c^2 + s^2) & \det(\Delta_4) = \tilde{\ell}\bar{\eta}s^2 \\
\Delta_6 \equiv A^{-1/2}BA^{-1/2} - I = & \begin{bmatrix} \bar{\ell}+\ell^{-1}\tilde{\eta}c^2 & \ell^{-1/2}\tilde{\eta}cs \\ \ell^{-1/2}\tilde{\eta}cs & \tilde{\eta}s^2 \end{bmatrix} & \mathrm{trace}(\Delta_6) = \bar{\ell} + \tilde{\eta}(c^2/\ell + s^2) & \det(\Delta_6) = \bar{\ell}\tilde{\eta}s^2
\end{array}
$$

Proof of Lemma 6.4

We use Lemma A.6 as needed. For $L^{N,1}$, $L^{N,2}$ and $L^{N,6}$, observe that Δ_i is symmetric and $\det(\Delta_i) \leq 0$, and it is therefore enough to solve (A.3). Now, in the cases $L^{N,1}$ and $L^{N,6}$, solve (A.3) for $\tilde{\eta}$ and note that minimizing w.r.t $\tilde{\eta}$ is equivalent to minimizing w.r.t η. In the case $L^{N,2}$, solve equation (A.3) for $\bar{\eta}$ and note that minimizing w.r.t $\bar{\eta}$ is equivalent to minimizing w.r.t η.

For $L^{N,3}$, in the notation of Lemma A.5 we have $\dot{t} \pm 2\dot{e} = 2\left(c^2 + \ell^2 s^2\right)(\tilde{\eta} - \tilde{\eta}_\pm)$, where

$$\tilde{\eta}_\pm = \frac{c^2 + \ell s^2}{c^2 + \ell^2 s^2}\tilde{\ell}\,.$$

One can check that here $r^2 > 0$ for all $\eta \geq 1$, and moreover that $\dot{\sigma}_+(\tilde{\eta}_+) + \dot{\sigma}_-(\tilde{\eta}_+) \neq 0$ and $\dot{\sigma}_+(\tilde{\eta}_+) - \dot{\sigma}_-(\tilde{\eta}_-) \neq 0$. It follows that the only zero of $\dot{\sigma}_+(\tilde{\eta}_+) + \dot{\sigma}_-(\tilde{\eta}_-)$ is at $\tilde{\eta} = \tilde{\eta}_-$, which we solve for η.

For $L^{N,4}$, again in the notation of Lemma A.5, we have $\dot{t} \pm 2\dot{e} = -2\left(\ell^2 c^2 + s^2\right)(\bar{\eta} - \bar{\eta}_\pm)$, where

$$\bar{\eta}_\pm = -\frac{\ell c^2 + s^2}{\ell^2 c^2 + s^2}\bar{\ell}\,.$$

One can check that here $r^2 > 0$ for all $\eta \geq 1$, and moreover that $\dot{\sigma}_+(\bar{\eta}_+) + \dot{\sigma}_-(\bar{\eta}_+) \neq 0$ and $\dot{\sigma}_+(\bar{\eta}_+) - \dot{\sigma}_-(\bar{\eta}_-) \neq 0$. It follows that the only zero of $\dot{\sigma}_+(\bar{\eta}_+) + \dot{\sigma}_-(\bar{\eta}_-)$ is at $\bar{\eta} = \bar{\eta}_-$, which we solve for η. □

Proof of Lemma 6.2

For $L^{O,1}$, by Lemma A.2 the eigenvalues of $\Delta_1(\tilde{\eta})$ are given by Eq. (A.1). Fix $\ell \geq \ell_+(\gamma)$ and consider the functions $\eta \mapsto \lambda_+(\eta)$ and $\eta \mapsto \lambda_-(\eta)$. The following facts are verified by simple algebra:

- For all $\eta \geq 0$ we have $\lambda_+(\eta) \geq 0$, $\lambda_+(\eta) \geq \lambda_-(\eta)$ and $\lambda_-(\eta) < 0$.
- For $\eta > \ell$ we have $\lambda_+(\eta) \leq -\lambda_-(\eta)$ and for $\eta < \ell$ we have $\lambda_+(\eta) \geq -\lambda_-(\eta)$.
- For all $\eta > \ell$ we have $d\lambda_-(\eta)/d\eta < 0$ and for all $\eta < \ell$ we have $d\lambda_+(\eta)/d\eta < 0$.

The first two facts imply that

$$\|\Delta_1\|_{op} = \max\{\lambda_+(\eta)\,,\,|\lambda_-(\eta)|\} = \begin{cases} \lambda_+(\eta) & \eta < \ell \\ -\lambda_-(\eta) & \eta > \ell \end{cases} .$$

The third fact implies that indeed $\min_\eta \|\Delta_1\|_{op} = \ell$.

An identical proof for $\Delta_2(\bar{\eta})$ instead of $\Delta_1(\tilde{\eta})$ shows that the same nonlinearity minimizes $L^{O,2}$ as well. □

Proof of Lemma 6.5

We use Lemma A.6 as needed. For L^{st}, we have $2L^{st} = \bar{\ell} + \tilde{\eta}(c^2/\ell + s^2) - \log(\eta/\ell)$. Now solve $dL^{st}/d\eta = 0$ for η. For L^{ent}, we have $2L^{ent} = \tilde{\ell} + \bar{\eta}(\ell c^2 + s^2) - \log(\ell(\bar{\eta}+1))$. Now solve $dL^{ent}/d\bar{\eta} = 0$ for $\bar{\eta}$. For L^{div}, we have $2L^{div} = \tilde{\ell} + \bar{\ell} + \bar{\eta}(\ell c^2 + s^2) + \tilde{\eta}(c^2/\ell + s^2)$. Now solve $dL^{ent}/d\eta = 0$ for η.

Proof of Lemma 6.6

We use Lemma A.6 as needed. Define $\hat{\eta} = \sqrt{\eta} - 1$ and

$$M = \begin{bmatrix} 1 + \hat{\eta}c^2 & \hat{\eta}cs \\ \hat{\eta}cs & 1 + \hat{\eta}s^2 \end{bmatrix} .$$

Direct calculation shows that $M^2 = B$, namely, $M = \sqrt{B}$. Therefore $\mathrm{trace}(\sqrt{A}\sqrt{B}) = \sqrt{\ell}(1 + \hat{\eta}c^2) + 1 + \hat{\eta}s^2$, and $L^{fre} = \eta + \ell + 2 - 2\left(\mathrm{trace}(\sqrt{A}\sqrt{B})\right)$. Now solve $dL^{fre}/d\hat{\eta} = 0$ for $\hat{\eta}$. □

A.2 Beyond Formal Optimality

Proof of Lemma 8.2

We may assume that all the eigenvalues of $\hat{\Sigma}_\eta(S_{n,p_n})$ lie in $[1,\infty)$, since the probability of this event tends to 1 as $n \to \infty$. Let U_r and $V_{r,n}$ denote the n-by-r matrices consisting of the top r eigenvectors of Σ_p and $\hat{\Sigma}_\eta(S_{n,p_n})$, respectively. Let W_{2r} denote the n-by-$2r$ matrix, whose columns comprise the first $2r$ vectors of the W basis. Observe that the matrix W_{2r} can be obtained as the output of the following three-stage process.

Stage 1. Form the matrix $W_{2r}^{(0)} = [U_r\ V_{r,n}]$ by stacking the matrices U_r and $V_{r,n}$.

Stage 2. Let the matrix $W_{2r}^{(1)}$ be the factor Q in the QR decomposition of $W_{2r}^{(0)}$, namely

$$W_{2r}^{(0)} = W_{2r}^{(1)}\ R\,, \tag{A.5}$$

where $W_{2r}^{(1)}$ has orthogonal columns and the matrix R is upper triangular.

Stage 3. Form $W_{2r}^{(2)}$ by permuting the columns of $W_{2r}^{(1)}$ as follows. Let π_{2r} be the permutation defined by

$$\pi_{2r} : (1,\ldots,2r) \mapsto (1, r+1, 2, r+2, 3, \ldots, 2r)\ ,$$

and let Π_{2r} be the permutation matrix corresponding to π_{2r}. Now define $W_{2r}^{(2)} = W_{2r}^{(1)}\,\Pi_{2r}$.

Since the first r columns of $W_{2r}^{(1)}$ are identically those of U_r, we let Z_r be the n-by-r matrix such that $W_{2r}^{(1)} = [U_r\ Z_r]$. Also note that the upper-diagonal matrix R in (A.5) has the block structure

$$R = \begin{bmatrix} I_{r\times r} & R_{12} \\ 0_{r\times r} & R_{22} \end{bmatrix} ,$$

where the matrices R_{12} and R_{22} satisfy

$$V_{r,n} = U_r R_{12} + Z_r R_{22} ,$$

so that

$$R_{12} \quad = \quad U_r' \, V_{r,n} \tag{A.6}$$
$$R_{22} \quad = \quad Z_r' \, V_{r,n} \, . \tag{A.7}$$

Since $V_{r,n}$ has orthogonal columns, we have

$$\begin{aligned} V_{r,n}' \, V_{r,n} &= I = R_{12}' \, R_{12} + R_{22}' \, R_{22} \\ R_{22}' \, R_{22} &= I - R_{12}' \, R_{12} \, . \end{aligned}$$

Observe that the subspace spanned by the columns of $W_{2r}^{(i)}$ is stable under action of the population covariance Σ_p, for $i = 1, 2$. Let $\Sigma_{2r}^{(i)}$ denote the representation of the population covariance Σ_p, restricted to the image of $W_{2r}^{(i)}$, in the basis provided by the columns of $W_{2r}^{(i)}$. Similarly, let $\hat{\Sigma}_{2r}^{(i)}$ denote the representation of $\hat{\Sigma}_\eta(S_{n,p_n})$, restricted to the image of $W_{2r}^{(i)}$, in the basis provided by the columns of $W_{2r}^{(i)}$. Formally,

$$\begin{aligned} \Sigma_{2r}^{(1)} &= (W_{2r}^{(1)})' \, \Sigma_p \, W_{2r}^{(1)} = diag\,(\ell_1, \ldots, \ell_r) \oplus I_r \\ \Sigma_{2r}^{(2)} &= (W_{2r}^{(2)})' \, \Sigma_p \, W_{2r}^{(2)} = \Pi_{2r}' \, \Sigma_{2r}^{(1)} \, \Pi_{2r} = \oplus_{k=1}^{r} A(\ell_k) \, . \end{aligned}$$

To simplify notation, we write $\eta_{i,n}$ for the shrunken empirical eigenvalue $\eta(\lambda_{i,n})$, and let $\hat{\Sigma}_\eta(S_{n,p_n}) = V\, diag(\eta_{1,n}, \ldots, \eta_{p,n}) \, V'$ be the full eigen-decomposition of $\hat{\Sigma}(S_{n,p_n})$. From (A.6) and (A.7), since $\hat{\Sigma}_\eta(S_{n,p_n}) = V \cdot (diag(\eta_{1,n}, \ldots, \eta_{p,n}) - I_p) \cdot V' + I_p$, we have

$$\hat{\Sigma}_{2r}^{(1)} = (W_{2r}^{(1)})' \, \hat{\Sigma}_\eta(S_{n,p_n}) \, W_{2r}^{(1)} = \begin{bmatrix} R_{12} \\ R_{22} \end{bmatrix} (diag(\eta_{1,n}, \ldots, \eta_{p,n}) - I_r) \begin{bmatrix} R_{12}' & R_{22}' \end{bmatrix} + I_{2r} \, .$$

Now, by Theorem 4.1, as $n \to \infty$, we have [6]

$$\begin{aligned} R_{12} = U_r' \, V_{r,n} \quad &\to_P \quad diag(c(\ell_1), \ldots, c(\ell_r)) \\ R_{22} \, R_{22}' = I - R_{12} \, R_{12}' \quad &\to_P \quad diag\left(s^2(\ell_1), \ldots, s^2(\ell_r)\right) \\ R_{22} \quad &\to_P \quad diag\,(s(\ell_1), \ldots, s(\ell_r)) \, . \end{aligned}$$

Therefore, as $n \to \infty$ we have

$$\hat{\Sigma}_{2r}^{(1)} \to_P B_{2r} ,$$

[6]For simplicity, we chose the QR decomposition to make the sign of $s(\ell_i)$ positive.

where

$$B_{2r} = \begin{bmatrix} I_r + diag\left((\eta_1-1)c_1^2,\ldots,(\eta_r-1)c_r^2\right) & diag\left((\eta_1-1)c_1s_1,\ldots,(\eta_r-1)c_rs_r\right) \\ diag\left((\eta_1-1)c_1s_1,\ldots,(\eta_r-1)c_rs_r\right) & I_r + diag\left((\eta_1-1)s_1^2,\ldots,(\eta_r-1)s_r^2\right) \end{bmatrix}$$

is a fixed (non-random) $2r$-by-$2r$ matrix. Since the convergence occurs in each coordinate, it also occurs in Frobenius norm:

$$\|\hat{\Sigma}_{2r}^{(1)} - B_{2r}\|_F \to_P 0\,.$$

Moreover, since the Frobenius norm is invariant under left and right orthogonal rotations, and in particular under permutations, we have

$$\|\Pi_{2r}'\left(\hat{\Sigma}_{2r}^{(1)} - B_{2r}\right)\Pi_{2r}\|_F \to_P 0\,.$$

From $\hat{\Sigma}_{2r}^{(2)} = \Pi_{2r}'\,\hat{\Sigma}_{2r}^{(1)}\,\Pi_{2r}$ and $\Pi_{2r}'\,B_{2r}\,\Pi_{2r} = \Sigma_{2r}^{(2)}$ we obtain

$$\|\hat{\Sigma}_{2r}^{(2)} - \Sigma_{2r}^{(2)}\|_F \to_P 0\,.$$

As the Frobenius norm is sum-decomposable, we conclude that as $n \to \infty$,

$$\|\left(\hat{\Sigma}_{2r}^{(2)} \oplus I_{p-2r}\right) - \left(\Sigma_{2r}^{(2)} \oplus I_{p-2r}\right)\|_F \to_P 0\,.$$

Finally, note that by definition

$$\begin{aligned} \hat{\Sigma}_{2r}^{(2)} \oplus I_{p-2r} &= W'\,\hat{\Sigma}_\eta(S_{n,p_n})\,W \\ \Sigma_{2r}^{(2)} &= \Pi_{2r}'\,B_{2r}\,\Pi_{2r} = \oplus_{i=1}^r B(\eta_i, c_i, s_i)\,, \end{aligned}$$

and the lemma follows. □

Proof of Lemma 8.3

Consider the single spike case, namely assume **[Asy(γ)]** and **[Spike(ℓ_1)]**. Observe that the proof of Lemma 5.2 remains valid if the nonlinearity η collapses the bulk to 1, and if, in addition,

$$\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=2}^{p_n} (\eta(\lambda_{i,n}) - 1)^2\right) = 0\,. \tag{A.8}$$

We now show that (A.8) holds. Fix n. Let $\Pi : \mathbb{R}^{p_n} \to \mathbb{R}^{p_n-1}$ denote the projection on the last $p_n - 1$ coordinates, and let X be a p_n-by-n matrix whose rows are i.i.d draws from $\mathcal{N}(0, \Sigma_{p_n})$, so that the eigenvalues of XX'/n are just $\lambda_{1,n} \geq \ldots \geq \lambda_{p_n,n}$. Now, let $\tilde{\lambda}_{1,n} \geq \ldots \tilde{\lambda}_{p_n-1,n}$ denote the eigenvalues of $\Pi X(\Pi X)'/n$. By the Cauchy interlacing Theorem, we have $\lambda_{i,n} \leq \tilde{\lambda}_{i-1,n}$ for $2 \leq i \leq p_n$. Let $\lambda_+ = (1+\sqrt{\beta})^2$ denote the bulk edge. Since η is α-Hölder at the bulk edge, there exists a constant $C > 0$ such that $|\eta(\lambda) - \eta(\lambda_+)| \leq C|\lambda - \lambda_+|^\alpha$ for all λ. It follows that $(\eta(\lambda)-1)^2 = (\eta(\lambda) - \eta(\lambda_+))^2 \leq C^2(\lambda - \lambda_+)^{2\alpha}$ for any $\lambda > \lambda_+$. Since η is monotone non-decreasing,

$$\sum_{i=2}^{p_n} (\eta(\lambda_{i,n}) - 1)^2 \leq \sum_{i=1}^{p_n-1} (\eta(\tilde{\lambda}_{i,n}) - 1)^2 \leq C^2 \sum_{i=1}^{p_n-1} (\tilde{\lambda}_{i,n} - \lambda_+)^{2\alpha}\,.$$

However, as the columns of ΠX are just i.i.d samples from the standard multivariate normal distribution in $\mathbb{R}^{p_n-1}$, the empirical distribution of $\{\tilde{\lambda}_{i,n}\}$ converges to the Marčenko-Pastur bulk and in fact

$$\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{p_n-1}(\tilde{\lambda}_{i,n}-\lambda_+)^{2\alpha}\right)=0\,,$$

which implies (A.8). The multiple spike case is proved similarly. □

A.3 Optimality Among Equivariant Procedures

Proof of Theorem 9.1: Direct Frobenius Loss case $L = L^{F,1}$

The Theorem follows from Lemma A.7 and Lemma A.8 below, which explicilty evaluate both sides of (9.1) to show that they are equal. □

Lemma A.7.

$$L_\infty^{F,1}(\ell_1,\ldots,\ell_r|\eta^*) \quad = \quad \sum_{k=1}^{r}(\ell_k-1)^2(1-c^4(\ell_k))\,.$$

Proof. By Lemma 8.1, it is enough to prove

$$L_2^{F,1}(A(\ell),B(\eta^*,c,s))=(\ell-1)^2(1-c^4(\ell))\,. \tag{A.9}$$

Recall that $(\eta^*-1)=(\ell-1)c^2(\ell)$. The pivot matrix is

$$\begin{aligned}
\Delta\left(A(\ell)\,,\,B(\eta^*,c,s)\right) &= \begin{bmatrix}(\ell-1)-(\ell-1)c^4 & -(\ell-1)c^3s\\ -(\ell-1)c^3s & (\ell-1)c^2s^2\end{bmatrix}\\
&= (\ell-1)\begin{bmatrix}1-c^4 & -c^3s\\ -c^3s & c^2s^2\end{bmatrix}.
\end{aligned} \tag{A.10}$$

Taking the squared Frobenius norm of (A.10), a straightforward calculation yields (A.18). □

Lemma A.8. *As $n\to\infty$,*

$$L_{p_n}^{F,1}(\Sigma_{p_n},\hat{\Sigma}_{p_n}^{oracle}) \to_P \sum_{k=1}^{r}(\ell_k-1)^2(1-c^4(\ell_k))\,. \tag{A.11}$$

Proof. To simplify some of the formal steps below, we use the familiar asymptotic notation o_P and O_P, with respect to convergence in probability as $n\to\infty$.

By definition,

$$L_{p_n}^{F,1}(\Sigma_{p_n},\hat{\Sigma}_{p_n}^{oracle})=\min_{\Delta}\|diag(\ell_1,\ldots,\ell_r,1,\ldots,1)-V\,\Delta\,V'\|_F^2\,,$$

where the columns of V are the eigenvectors of S_{n,p_n} and the minimum is taken over all p_n-by-p_n diagonal matrices. Fix $(\ell_1,\ldots,\ell_r)$ and define $L=diag(\ell_1,\ldots,\ell_r,1,\ldots,1)$. Since the Frobenius norm is orthogonally invariant, for any diagonal matrix Δ we have

$$\begin{aligned}
\|L-V\,\Delta\,V'\|_F^2 &= \|(\Delta-I)-V'(L-I)V\|_F^2\\
&= \sum_i\left((\Delta-I)-V'(L-I)V\right)_{i,i}^2+\sum_{i\neq j}(V'(L-I)V)_{i,j}^2\,,
\end{aligned} \tag{A.12}$$

where we have separated summation over the diagonal and non-diagonal entries.

Clearly, the minimum of (A.12) over Δ is attained when the left term vanishes, namely

$$\begin{aligned}
\min_{\Delta} \|diag(\ell_1,\ldots,\ell_r,1,\ldots,1) - V\,\Delta\,V'\|_F^2 &= \sum_{i\neq j}(V'(L-I)V)_{i,j}^2 \\
&= \sum_{i\neq j}\left(\sum_{k=1}^{n}(\ell_k-1)v_k(i)v_k(j)\right)^2 \\
&= \sum_{i\neq j}\sum_{k,k'=1}^{n}(\ell_k-1)(\ell_{k'}-1)v_k(i)v_k(j)v_{k'}(i)v_{k'}(j) \\
&= \sum_{k,k'=1}^{n}(\ell_k-1)(\ell_{k'}-1)\sum_{i\neq j}v_k(i)v_k(j)v_{k'}(i)v_{k'}(j)\,,
\end{aligned}$$

where $v_{k'}$ denotes the k'-th column of V' and v_k denotes its transpose. To evaluate the last term, first note that

$$\sum_{i,j}v_k(i)v_k(j)v_{k'}(i)v_{k'}(j) = \left(\sum_i v_k(i)v_{k'}(i)\right)^2 = \delta_{k,k'}\,.$$

Now, by Lemma A.9, stated and proved below, we have

$$\sum_i v_k^2(i)v_{k'}^2(i) = c^4(\ell_k)(1+o_P(1))\delta_{k,k'} + o_P(1)\,.$$

Since

$$\sum_{i\neq j}v_k(i)v_k(j)v_{k'}(i)v_{k'}(j) = \sum_{i,j}v_k(i)v_k(j)v_{k'}(i)v_{k'}(j) - \sum_i v_k^2(i)v_{k'}^2(i)$$

we have

$$\sum_{i\neq j}v_k(i)v_k(j)v_{k'}(i)v_{k'}(j) = \big(1-c^4(\ell_k)\big)\cdot(1+o_P(1))\cdot\delta_{k,k'} + o_P(1)\,.$$

It thus follows that

$$\min_{\Delta}\|diag(\ell_1,\ldots,\ell_r,1,\ldots,1) - V\,\Delta\,V'\|_F^2 = \sum_k(\ell_k-1)^2\left(1-c^4(\ell_k)\right) + o_P(1)$$

as desired.

□

Lemma A.9. *Let $1\le k,k'\le r$. As $n\to\infty$ we have*

$$\sum_{i=1}^{n} v_k^2(i)v_{k'}^2(i) \to_P \begin{cases} c^4(\ell_k) & k=k' \\ 0 & k\neq k' \end{cases}\,.$$

Proof. Let $1\le k,k'\le r$. By Theorem 4.1,

$$\sum_{i=1}^{r} v_k^2(i)v_{k'}^2(i) \to_P c^4(\ell_k)\delta_{k,k'}\,.$$

It therefore remains to show that, as $n \to \infty$,

$$\sum_{i=r+1}^{p} v_k^2(i) v_{k'}^2(i) \to_P 0\,. \tag{A.13}$$

Indeed, by Lemma A.10, stated and proved below,

$$\max_{r+1 \le i \le p} |v_k(i)| = O_P\left(\frac{\sqrt{\log p}}{p}\right)\,,$$

so that

$$v_k^2(i) v_{k'}^2(i) = O_P\left(\frac{\log^2(p)}{p^4}\right)\,,$$

and (A.13) follows. □

Lemma A.10. *Let v_k denote the empirical eigenvector of $S_{n,p}$ corresponding to the k-largest empirical eigenvalue. Define the statistic*

$$M_{n,p,r} = \max_{j>r} |\langle v_k, e_j \rangle|\,.$$

Then

$$M_{n,p,r} = O_P\left(\frac{\sqrt{\log p}}{p}\right)\,.$$

Proof. Define, for a vector v and a matrix U,

$$M_r(v, U) = \max_{j>r} |\langle v, U e_j \rangle|$$

so that our quantity of interest satisfies

$$M_{n,p,r} \equiv M_r(v_k, I)\,.$$

Consider the $\mathbb{R}^p$-valued function of a matrix variable defined by $S_{n,p} \mapsto v_k$. Observe that it is invariant under orthogonal rotations of $\mathbb{R}^p$, in the sense that

$$v_k(U\, S_{n,p}\, U') = U'\, v_k(S_{n,p})\,. \tag{A.14}$$

Here and below, the symbol $=_D$ denotes equality in distribution. Observe that, if U_0 is any rotation that leaves Σ invariant (in the sense that $U_0\, \Sigma\, U_0' = \Sigma$), then

$$v_k\left(U_0\, S_{n,p}\, U_0'\right) =_D v_k(S_{n,p})\,, \tag{A.15}$$

and hence

$$U_0'\, v_k(S_{n,p}) =_D v_k(S_{n,p})\,. \tag{A.16}$$

Combining (A.14), (A.15) and (A.16) we obtain

$$M_r(v_k, U_0) =_D M_r(v_k, I) \equiv M_{n,p,r}\,. \tag{A.17}$$

Consider the case $U_0 = I_r \oplus Q_0$, where Q_0 is a Haar-distributed $(p-r)$-by-$(p-r)$ orthogonal matrix. Note that almost surely, U_0 leaves Σ invariant. Now let w be any fixed vector in $\mathbb{R}^p$ and let $P_{>r}$ be any orthoprojector of $\mathbb{R}^p$ onto $span\,\{e_{r+1},\ldots,e_p\}$. Then $P_{>r}w$ sends the first r coordinates to zero and we have

$$P_{>r}w = 0_r \oplus w_0\,,$$

where $0_r \in \mathbb{R}^r$ is the zero vector and $w_0 \in \mathbb{R}^{p-r}$. Clearly,

$$M_1(w,U_0) = M_r(P_{>r}w,U_0) = M_r(w_0,Q_0)\,.$$

We now apply Lemma A.11, stated and proved below, to $M_1(w_0,Q_0)$ with $q = p-r$ and obtain

$$M_1(w_0,Q_0) = O_P\left(\frac{\sqrt{\log(p-r)}}{(p-r)}\right)$$

where the implied constants on the right hand side are uniform across all vectors $w_0 \in \mathbb{R}^{p-r}$ for which $\|w_0\|_2 \le 1$.

In the above construction, let $S_{n,p}$ be some fixed realization, let $v_k(S_{n,p})$ be the corresponding fixed eigenvector in $\mathbb{R}^p$ and set $w = v_k(S_{n,p})$. Then the above implies

$$M_r(v_k,U_0) = O_P\left(\frac{\sqrt{\log p}}{p}\right)\,,$$

where the implied constants on the right hand side are independent of the specific realization of $S_{n,p}$. Now apply (A.17) and conclude that

$$M_r(v_k) = O_P\left(\frac{\sqrt{\log p}}{p}\right)\,,$$

as required. ☐

Lemma A.11. *Let Q be a Haar-distributed q-by-q orthogonal matrix and let $w \in \mathbb{R}^q$ be a fixed vector with $\|w\|_2 \le 1$. Then*

$$\max_{1\le j\le q} |\langle w\,,\,Q\,e_j\rangle| = O_P\left(\frac{\sqrt{\log q}}{q}\right)\,,$$

where the implied constants on the right hand side are independent of the choice of w.

Proof. Let U be a rotation that satisfies $Uw = e_1$. Then $U'\,Q =_D Q$. Hence,

$$\max_{1\le j\le q} |\langle w\,,\,Q\,e_j\rangle| =_D \max_{1\le j\le q} |\langle e_1\,,\,Q\,e_j\rangle|\,.$$

Now, the map

$$Q \mapsto |\langle e_1\,,\,Q\,e_j\rangle|$$

is 1-Lipschitz w.r.t Frobenius norm, and the Haar matrix Q obeys concentration of measure [40], hence

$$P\left\{|\langle e_1\,,\,Q\,e_j\rangle| > t\right\} \le e^{-qt^2/2}\,.$$

Therefore,

$$P\left\{\max_{1\le j\le q} |\langle e_1\,,\,Q\,e_j\rangle| > t\right\} \le q\,e^{-qt^2/2}\,.$$

Picking $t = c\sqrt{\log q}/q$ we get that the right hand side is bounded by $e^{-c^2/2}$. ☐

Proof of Theorem 9.1: Stein Loss case $L = L^{st}$

The Theorem follows from Lemma A.12 and Lemma A.13 below, which explicilty evaluate the both sides of (9.1) to show that they are equal. □

Lemma A.12.

$$L^{st}_{\infty}(\ell_1, \ldots, \ell_r | \eta^*) \quad = \quad \sum_{k=1}^{r} \left[(\ell_k^{-1} - 1) s^2(\ell_k) + \log\left(c(\ell_k)^2 + \ell_k s^2(\ell_k) \right) \right]$$

Proof. By Lemma 8.1, it is enough to prove

$$L^{st}_2(A(\ell), B(\eta^*, c, s)) = (\ell^{-1} - 1) s^2(\ell) + \log\left(c(\ell)^2 + \ell s^2(\ell) \right) . \tag{A.18}$$

Indeed,

$$L^{st}_2(A(\ell), B(\eta, c, s)) = \mathrm{trace}(A^{-1}(\ell)\, B(\eta, c, s) - I) - \log\left(\frac{|B(\eta, c, s)|}{|A(\ell)|} \right) .$$

Direct calculation shows that

$$\begin{aligned} \mathrm{trace}\left(A^{-1}(\ell)\, B(\eta, c, s) - I \right) &= \left(\ell^{-1} - 1 \right) + (\eta - 1)\left(c^2/\ell + s^2 \right) \\ \log\left(|B(\eta, c, s)| / |A(\ell)| \right) &= \log(\eta/\ell) \end{aligned}$$

Above we have seen that the optimal shrinker η^* for the Stein Loss satisfies $\eta^*(\ell) = \left(c^2(\ell) + s^2 \right)^{-1}$. Substituting, we obtain

$$\begin{aligned} \mathrm{trace}\left(A^{-1}(\ell)\, B(\eta^*(\ell), c(\ell), s(\ell)) - I \right) &= \left(\ell^{-1} - 1 \right) s^2(\ell) \\ \log\left(|B(\eta^*(\ell), c(\ell), s(\ell))| / |A(\ell)| \right) &= -\log\left(c^2(\ell) + \ell s^2(\ell) \right) , \end{aligned}$$

as required. □

Lemma A.13. *As $n \to \infty$,*

$$L^{st}_{p_n}(\Sigma_{p_n}, \hat{\Sigma}^{oracle}_{p_n}) \to_P \sum_{k=1}^{r} \left[(\ell_k^{-1} - 1) s^2(\ell_k) + \log\left(c(\ell_k)^2 + \ell_k s^2(\ell_k) \right) \right] . \tag{A.19}$$

Proof. Define $L = diag(\ell_1, \ldots, \ell_r, 1, \ldots, 1)$. We first show that the oracle estimator for Stein Loss satisfies $\Delta^{oracle} = diag(d_1^{oracle}, \ldots, d_n^{oracle})$ with

$$d_i^{oracle} = \frac{1}{1 - \sum_{k=1}^{n} \left(1 - \ell_k^{-1} \right) v_k^2(i)} . \tag{A.20}$$

Indeed, for any diagonal matrix Δ we have

$$\begin{aligned} L^{st}_p\left(L\, ,\, V\, \Delta\, V' \right) &= \mathrm{trace}\left(L^{-1}\, V\, \Delta\, V' - I \right) - \log\left(|\Delta| / |L| \right) \\ &= \mathrm{trace}(V'(L^{-1} - I)\, V\, \Delta + \Delta - I) - \log(|\Delta|/|\Lambda|) \\ &= \sum_{i=1}^{p} \left[(d_i - 1) - d_i \sum_{k=1}^{r} \left(1 - \ell_k^{-1} \right) v_k^2(i) - \log(d_i / L_{i,i}) \right] \end{aligned} \tag{A.21}$$

where we have used the cyclic property of the trace.

Minimizing (A.21) w.r.t Δ is thus achieved by minimizing w.r.t each d_i. Setting the derivative to zero and solving for d_i, we arrive at (A.21). A simple calculation now shows that

$$\begin{aligned} L^{st}_{p_n}(\Sigma_{p_n}, \hat{\Sigma}^{oracle}_{p_n}) &= L^{st}_{p_n}(L\,,\, V\,\Delta^{oracle}\,V') \\ &= \sum_{i=1}^{p_n} \log\left(\frac{\left(1-\sum_{k=1}^{r}\left(1-L_{k,k}^{-1}\right)v_k^2(i)\right)^{-1}}{L_{k,k}}\right) \\ &= \sum_{i=1}^{r} \log\left(\frac{\left(1-\sum_{k=1}^{r}\left(1-\ell_k^{-1}\right)v_k^2(i)\right)^{-1}}{\ell_k}\right) \qquad \text{(A.22)} \\ &- \sum_{i=r+1}^{p_n} \log\left(1-\sum_{k=1}^{r}\left(1-\ell_k^{-1}\right)v_k^2(i)\right)\,. \qquad \text{(A.23)} \end{aligned}$$

We first evaluate the limiting value of (A.22). By Theorem 4.1, as $n \to \infty$ we have

$$v_k^2(i) \to_P c^2(\ell_k)\,\delta_{k,i}\,.$$

It follows that, as $n \to \infty$,

$$\sum_{i=1}^{r} \log\left(\frac{\left(1-\sum_{k=1}^{r}\left(1-\ell_k^{-1}\right)v_k^2(i)\right)^{-1}}{\ell_k}\right) \to_P \sum_{i=1}^{r} \log\left(c^2(\ell_k)+\ell_k s^2(\ell_k)\right)\,.$$

To evaluate (A.23), first observe that, by Lemma A.10,

$$\max_{r<i<p} \sum_{k=1}^{r} v_k^2(i)(1-\ell_k^{-1}) = O_P\left(\frac{\log p}{p}\right)\,,$$

and therefore, as $n \to \infty$,

$$\sum_{i=r+1}^{p_n} \log\left(1-\sum_{k=1}^{r}\left(1-\ell_k^{-1}\right)v_k^2(i)\right) \to_P \sum_{i=r+1}^{p_n}\sum_{k=1}^{r}\left(1-\ell_k^{-1}\right)v_k^2(i)\,.$$

However, $\|v_k\|^2 = 1$, hence by Theorem 4.1, as $n \to \infty$ we have as $n \to \infty$

$$\sum_{i=r+1}^{p_n} v_k^2(i) \to_P 1 - c^2(\ell_k) = s^2(\ell_k)\,,$$

so that, as $n \to \infty$,

$$\sum_{i=r+1}^{p_n} \log\left(1-\sum_{k=1}^{r}\left(1-\ell_k^{-1}\right)v_k^2(i)\right) \to_P \sum_{k=1}^{r} s^2(\ell_k)(1-\ell_k^{-1})\,,$$

and the lemma follows. □

References

[1] Zhidong Bai and Jack W Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.

[2] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Springer, 2010.

[3] Zhidong Bai and Jian-Feng Yao. Central limit theorems for eigenvalues in a spiked population model. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 44(3):447–474, June 2008.

[4] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.

[5] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

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

[7] James Berger. Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. *Statistical Decision Theory and Related Topics III*, 1:109–141, 1982.

[8] Lawrence D. Brown and Eitan Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 37(4):pp. 1685–1704, 2009.

[9] Theophilos Cacoullos and Ingram Olkin. On the bias of functions of characteristic roots of a random matrix. *Biometrika*, 52(1/2):pp. 87–94, 1965.

[10] RB Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1966.

[11] Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for MMSE covariance estimation. *Signal Processing, IEEE Transactions on*, 58(10):5016–5029, 2010.

[12] Michael J Daniels and Robert E Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184, 2001.

[13] Dipak K Dey and C Srinivasan. Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, pages 1581–1591, 1985.

[14] D. L. Donoho and M. Gavish. Minimax Risk of Matrix Denoising by Singular Value Thresholding. *ArXiv e-prints*, April 2013.

[15] D. L. Donoho and M. Gavish. The Optimal Hard Threshold for Singular Values is 4/sqrt(3). *ArXiv e-prints*, May 2013.

[16] David L. Donoho, Matan Gavish, and Iain M. Johnstone. Code and Data supplement to "Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model". `http://purl.stanford.edu/xy031gt1574`, 2013. accessed 10 October 2013.

[17] David L. Donoho, Matan Gavish, and Iain M. Johnstone. Supporting Information for "Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model". `http://purl.stanford.edu/xy031gt1574`, 2013.

[18] David L Donoho and Iain M Johnstone. Minimax risk over ℓ_p-balls for ℓ_p-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

[19] DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[20] Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):pp. 1102–1123, 2009.

[21] Bradley Efron and Carl Morris. Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics*, 4(1):pp. 22–32, 1976.

[22] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[23] Wolfgang Förstner and Boudewijn Moonen. A metric for covariance matrices. *Quo vadis geodesia*, pages 113–128, 1999.

[24] AK Gupta and Samuel Ofori-Nyarko. Improved minimax estimators of normal covariance and precision matrices. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(1):19–25, 1995.

[25] LR Haff. Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *The Annals of Statistics*, pages 1264–1276, 1979.

[26] LR Haff. An identity for the wishart distribution with applications. *Journal of Multivariate Analysis*, 9(4):531–544, 1979.

[27] LR Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 8(3):586–597, 1980.

[28] Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[29] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

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

[31] I Jolliffe. *Principal component analysis*. Springer, New York, 2nd edition, 2005.

[32] Thomas Kailath. The divergence and Bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52–60, 1967.

[33] Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008.

[34] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, pages 2757–2790, 2008.

[35] Yoshihiko Konno. On estimation of a matrix of normal means with unknown covariance matrix. *Journal of Multivariate Analysis*, 36(1):44–55, 1991.

[36] K. Krishnamoorthy and A. K. Gupta. Improved minimax estimation of a normal precision matrix. *Canadian Journal of Statistics*, 17(1):91–102, 1989.

[37] K Krishnamoorthy and AK Gupta. Improved minimax estimation of a normal precision matrix. *Canadian Journal of Statistics*, 17(1):91–102, 1989.

[38] Tatsuya Kubokawa. Improved estimation of a covariance matrix under quadratic loss. *Statistics & Probability Letters*, 8(1):69 – 71, 1989.

[39] Tatsuya Kubokawa and Yoshihiko Konno. Estimating the covariance matrix and the generalized variance under a symmetric loss. *Annals of the Institute of Statistical Mathematics*, 42(2):331–343, 1990.

[40] Michel Ledoux. The concentration of measure phenomenon. *American Mathematical Society, Surveys and Monographs* Vol. 89 , 2001

[41] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

[42] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[43] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[44] Christophe Lenglet, Mikaël Rousson, Rachid Deriche, and Olivier Faugeras. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor mri processing. *Journal of Mathematical Imaging and Vision*, 25(3):423–444, 2006.

[45] S.F. Lin and M.D Perlman. A monte carlo comparison of four estimators for a covariance matrix. In *Multivariate Analysis VI (P.R. Krishnaiah, ed.)*, pages 411–429. North Holland, Amsterdam, 1985.

[46] Wei-Liem Loh. Estimating covariance matrices. *The Annals of Statistics*, pages 283–296, 1991.

[47] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483, 1967.

[48] Kameo Matusita. On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics*, 19:181–192, 1967.

[49] Robb J Muirhead. Developments in eigenvalue estimation. In *Advances in Multivariate Statistical Analysis*, pages 277–288. Springer, 1987.

[50] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[51] N Pal. Estimating the normal dispersion matrix and the precision matrix from a decision-theoretic point of view: a review. *Statistical Papers*, 34(1):1–26, 1993.

[52] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.

[53] Jegadevan Balendran Selliah. *Estimation and testing problems in a Wishart distribution*. Department of Statistics, Stanford University., 1964.

[54] Andrey A. Shabalin and Andrew B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, July 2013.

[55] Divakar Sharma and K. Krishnamoorthy. Empirical Bayes estimators of normal covariance matrix. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 47(2):pp. 247–254, 1985.

[56] BK Sinha and M. Ghosh. Inadmissibility of the best equivariant estimators of the variance-covariance matrix, the precision matrix, and the generalized variance under entropy loss. *Statist. Decisions*, 5:201–227, 1987.

[57] Charles Stein. Some problems in multivariate analysis. Technical report, Department of Statistics, Stanford University, 1956.

[58] Charles Stein. Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences*, 34(1):1373–1403, 1986.

[59] D. Sun and X. Sun. Estimation of the multivariate normal precision and covariance matrices in a star-shape model. *Annals of the Institute of Statistical Mathematics*, 57(3):455–484, 2005.

[60] H. R. van der Vaart. On certain characteristics of the distribution of the latent roots of a symmetric random matrix under general conditions. *The Annals of Mathematical Statistics*, 32(3):pp. 864–873, 1961.

[61] Joong-Ho Won, Johan Lim, Seung-Jean Kim, and Bala Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2012.

[62] Ruoyong Yang and James O Berger. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, pages 1195–1211, 1994.