HHS Public Access
Author manuscript
Ann Stat. Author manuscript; available in PMC 2018 September 24.

Published in final edited form as:
Ann Stat. 2018 August ; 46(4): 1742–1778. doi:10.1214/17-AOS1601.

Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model

David L. Donoho[*], **Matan Gavish**[†], and **Iain M. Johnstone**[*]

[*] Department of Statistics, Stanford University

[†] School of Computer Science and Engineering, Hebrew University of Jerusalem

Abstract

We show that in a common high-dimensional covariance model, the choice of loss function has a profound effect on optimal estimation.

In an asymptotic framework based on the Spiked Covariance model and use of orthogonally invariant estimators, we show that optimal estimation of the population covariance matrix boils down to design of an optimal shrinker η that acts elementwise on the sample eigenvalues. Indeed, to each loss function there corresponds a unique admissible eigenvalue shrinker η^* dominating all other shrinkers. The shape of the optimal shrinker is determined by the choice of loss function and, crucially, by inconsistency of both eigenvalues *and* eigenvectors of the sample covariance matrix.

Details of these phenomena and closed form formulas for the optimal eigenvalue shrinkers are worked out for a menagerie of 26 loss functions for covariance estimation found in the literature, including the Stein, Entropy, Divergence, Fréchet, Bhattacharya/Matusita, Frobenius Norm, Operator Norm, Nuclear Norm and Condition Number losses.

Keywords

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

1 Introduction

Suppose we observe p-dimensional Gaussian vectors $X_i \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_p)$, $i = 1, \ldots, n$, with $\Sigma = \Sigma_p$ the underlying p-by-p population covariance matrix. To estimate Σ, we form the empirical (sample) covariance matrix $S = S_{n,p} = n^{-1}\sum_{i=1}^{n} X_i X_i'$; this is the maximum likelihood

Proofs and Additional Results
In the supplementary material [40] we provide proofs omitted from the main text for space considerations and auxiliary lemmas used in various proofs. Notably, we prove Lemma 4, and provide detailed derivations of the 17 explicit formulas for optimal shrinkers, as summarized in Table 2. In addition, in the supplementary material we offer a detailed study of the large-λ asymptotics (asymptotic slope and asymptotic shift) of the optimal shrinkers discovered in this paper, and tabulate the asymptotic behavior of each optimal shrinker. We also study the asymptotic percent improvement of the optimal shrinkers over naive hard thresholding of the sample covariance eigenvalues.

estimator. Stein [1, 2] observed that the maximum likelihood estimator S ought to be improvable by eigenvalue shrinkage.

Write $S = V\Lambda V'$ for the eigendecomposition of S, where V is orthogonal and the diagonal matrix $\Lambda = \mathrm{diag}\left(\lambda_1, \dots, \lambda_p\right)$ contains the empirical eigenvalues. Stein [2] proposed to shrink the eigenvalues by applying a specific nonlinear mapping φ producing the estimate $\widehat{\Sigma}_\varphi = V\varphi(\Lambda)V'$, where φ maps the space of positive diagonal matrices onto itself. In the ensuing half century, research on eigenvalue shrinkers has flourished, producing an extensive literature. We can point here only to a fraction, with pointers organized into early decades [3, 4, 5, 6, 7, 8], the middle decades [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19], and the last decade [20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. Such papers 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 $\mathbb{E}L_p\left(\Sigma, \widehat{\Sigma}_\eta(S)\right)$.

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 set of assumptions about Σ known as the *Spiked Covariance Model* [30]. We study 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 [31, 32]. 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 two basic assumptions.

[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 and sample covariances in the n-th problem by $\Sigma = \Sigma_{p_n}$ and $S = S_{n,p_n}$ and assume that the eigenvalues ℓ_i of Σ_{p_n} satisfy:

[SPIKE($\ell_1, \dots, \ell_r$)] The r "spikes" $\ell_1 > \dots > \ell_r \geq 1$ are fixed independently of n and p_n, and $\ell_{r+1} = \dots = \ell_{p_n} = 1$.

The spiked model exhibits three important phenomena, not seen in classical fixed-p asymptotics, that play an essential role in the construction of optimal estimators. Drawing on results from [33, 34, 35, 36, 37, 38], we highlight:

a. *Eigenvalue spreading.* Consider model [Asy(γ)] in the null case $\ell_1 = \ldots = \ell_r = 1$. The empirical distribution of the sample eigenvalues $\lambda_{1n}, \ldots, \lambda_{pn}$ converges as $n \to \infty$ to a non-degenerate absolutely continuous distribution, the Marcenko-Pastur or 'quarter-circle' law [33]. The distribution, or 'bulk', is supported on a single interval, whose limiting 'bulk edges' are given by

$$\lambda_{\pm}(\gamma) = (1 \pm \sqrt{\gamma})^2 . \quad (1.1)$$

b. *Top eigenvalue bias.* Consider models [Asy(γ)] and [Spike($\ell_1, \ldots, \ell_r$)]. For $i = 1, \ldots, r$, the leading sample eigenvalues satisfy

$$\lambda_{in} \xrightarrow{a.s.} \lambda(\ell_i), \quad (1.2)$$

where the 'biasing' function

$$\lambda(\ell) = \ell + \gamma\ell/(\ell - 1), \ \ell \geq \ell_+(\gamma), \quad (1.3)$$

$\lambda(\ell) \equiv (1 + \sqrt{\gamma})^2 = \lambda_+(\gamma)$ for $\ell \leq \ell_+(\gamma)$, the *Baik-Ben Arous-Peché transition point*

$$\ell_+(\gamma) = 1 + \sqrt{\gamma} . \quad (1.4)$$

Thus the empirical eigenvalues λ_i are shifted upwards from their theoretical counterparts ℓ_i by an asymptotically predictable amount, of a size that exceeds γ even for very large signal strengths ℓ_i.

c. *Top eigenvector inconsistency.* Again consider models [Asy(γ)] and [Spike($\ell_1, \ldots, \ell_r$)], noting that $\ell_1 > \ldots > \ell_r$ are distinct. The angles between the sample eigenvectors $v_{1n}, \ldots, v_{pn}$, and the corresponding "true" population eigenvectors $u_{1n}, \ldots, u_{pn}$, have non-zero limits:

$$\left|\left\langle u_{in}, v_{jn}\right\rangle\right| \xrightarrow{a.s.} \delta_{i,j} \cdot c(\ell_i) \ \ 1 \leq i, j \leq r, \quad (1.5)$$

where the cosine function is given by

$$c(\ell) = \sqrt{\frac{1 - \gamma/(\ell - 1)^2}{1 + \gamma/(\ell - 1)}} \ \ \ell \geq \ell_+(\gamma), \quad (1.6)$$

and $c(\ell) = 0$ for $\ell \leq \ell_+(\gamma)$.

Loss functions and optimal estimation. Now consider a class of estimators for the population covariance , based on *individual* shrinkage of the sample eigenvalues. Specifically,

$$\widehat{\Sigma} = \widehat{\Sigma}_\eta = \eta(\lambda_1)v_1v_1' + \dots + \eta(\lambda_p)v_pv_p', \quad (1.7)$$

where v_i is the sample eigenvector with sample eigenvalue λ_i and $\eta(\lambda)$ is a *scalar nonlinearity*, $\eta : \mathbb{R}^+ \to [1, \infty)$, so that the *same* function acts on each sample eigenvalue. While this appears to be a significant restriction from Stein's use of vector functions φ [2], the discussion in Section 8 shows that nothing is lost in our setting by the restriction to scalar shrinkers.

Consider a family of loss functions $L = \{L_p\}_{p=1}^{\infty}$ and a fixed nonlinearity $\eta : [0, \infty) \to \mathbb{R}$. Define the asymptotic loss relative to L of the shrinkage estimator $\widehat{\Sigma}_\eta$ in model [Spike($\ell_1, \dots, \ell_r$)] by

$$L_\infty(\ell_1, \dots, \ell_r | \eta) = \lim_{n \to \infty} L_{p_n}\left(\Sigma_{p_n}, \widehat{\Sigma}_\eta\left(S_{n, p_n}\right)\right), \quad (1.8)$$

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

$$L_\infty(\ell_1, \dots, \ell_r | \eta^*) \le L_\infty(\ell_1, \dots, \ell_r | \eta) \quad (1.9)$$

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

In constructing estimators, it is natural to expect that the effect of the biasing function $\lambda(\ell)$ in (1.3) might be undone simply by applying its inverse function $\ell(\lambda)$ given by

$$\ell(\lambda) = \frac{(\lambda + 1 - \gamma) + \sqrt{(\lambda + 1 - \gamma)^2 - 4\lambda}}{2} \quad \lambda > \lambda_+(\gamma). \quad (1.10)$$

However, eigenvector inconsistency makes the situation more complicated (and interesting!), as we illustrate using Figure 1. Focus on the plane spanned by u_1, the top population eigenvector, and by v_1, its sample counterpart. We represent $\ell_1 u_1 u_1'$, the top rank one component of Σ, by the vector $\ell_1 u_1$. The corresponding top rank one component of S is $\lambda_1 v_1 v_1'$, represented by $\lambda_1 v_1$. If we apply the inverse function (1.10) to λ_1, we obtain $\ell(\lambda_1)v_1v_1'$. Since v_1 is not collinear with u_1, there is a non-vanishing error $\ell(\lambda_1)v_1v_1' - \ell_1 u_1 u_1'$ that remains, even though $\ell(\lambda_1) - \ell_1 = O_p\left(n^{-1/2}\right)$. As the picture suggests, it is quite possible that a different amount of shrinkage, $\eta(\lambda_1)v_1v_1'$ will lead to smaller error. However, we will see that *the optimal choice of* η *depends greatly on the particular error measure* $L_p\left(\Sigma, \widehat{\Sigma}\right)$ *that is chosen*.

To give the flavor of results to be developed systematically later, we now look at four error measures in common use. The first three, based on the operator, Frobenius and nuclear norms, use the singular values σ_j of $\widehat{\Sigma} - \Sigma$:

$$\begin{aligned} L^O\left(\Sigma, \widehat{\Sigma}\right) &= \left\| \widehat{\Sigma} - \Sigma \right\|_\infty = \max_i \sigma_i, \\ L^F\left(\Sigma, \widehat{\Sigma}\right) &= \left\| \widehat{\Sigma} - \Sigma \right\|_2 = \left(\sum_i \sigma_i^2 \right)^{1/2}, \\ L^N\left(\Sigma, \widehat{\Sigma}\right) &= \left\| \widehat{\Sigma} - \Sigma \right\|_1 = \sum_i \sigma_i, \\ L^{\mathrm{St}}\left(\Sigma, \widehat{\Sigma}\right) &= \mathrm{tr}\left(\Sigma^{-1} \widehat{\Sigma} - I\right) - \log \det\left(\Sigma^{-1} \widehat{\Sigma}\right). \end{aligned} \tag{1.11}$$

The fourth is Stein's loss, widely studied in covariance estimation [1, 9, 39].

For convenience, we begin with the single spike model **Spike(ℓ)**, so that $\Sigma = \Sigma_\ell = I + (\ell - 1)u_1 u_1'$. When η is continuous, the losses have a deterministic asymptotic limit $L_\infty(\ell | \eta)$ defined in−(1.8).

For many losses, including (1.11), this deterministic limiting loss has a simple form, and we can evaluate, often analytically, the optimal shrinkage function, namely the shrinkage function satisfying (1.9). For example, writing $\eta^*(\lambda) = \eta_*(\ell(\lambda))$, for the four popular loss functions (1.11) we find that on $\ell > 1 + \sqrt{\gamma}$ the corresponding four optimal shrinkers are

$$\begin{aligned} \eta_*^O(\ell) &= \ell & \eta_*^F(\ell) &= \ell c^2 + s^2 \\ \eta_*^N(\ell) &= \max\left(1 + (\ell - 1)\left(1 - 2s^2\right), 1\right) & \eta_*^{\mathrm{St}}(\ell) &= \ell / \left(c^2 + \ell s^2\right), \end{aligned} \tag{1.12}$$

where $s^2 = 1 - c^2$. Figure 2 shows these four optimal shrinkers as a function of the sample eigen value λ. These are just four examples; The full list of optimal shrinkers we discover in this paper appears in Table 2 below. In all cases, $\eta_*(\ell) \equiv 1$ for $\ell \leq 1 + \sqrt{\gamma}$. Figure 3 in Section 6 below shows all the full list of optimal shrinkers when $\gamma = 1$.

The main conclusion is that the *optimal shrinkage function depends strongly on the loss function* chosen. The operator norm shrinker η_*^O simply inverts the biasing function $\lambda(\ell)$, while the other functions shrink by much larger, and very different, amounts, with η_*^{St} typically shrinking most. There are also important qualitative differences in the optimal shrinkers: η_*^O is discontinuous at the bulk edge $\lambda = \lambda_+(\gamma)$. The others are continuous, but η_*^N has the additional feature that it shrinks a *neighborhood* of the bulk to 1.

Remark. The optimal shrinker also depends on γ, so we might write $\eta^*(\lambda, \gamma)$. In model [Asy(γ)], one can use the same γ for each problem size n. Alternatively, in the n-th problem, one might use $\gamma_n = p_n/n$. The former choice is simpler, as η^* can be regarded as a univariate function of λ, and so we make it in Sections 1–6. The latter choice is preferable technically, and perhaps also in practice, when one has p and n, but not γ. It does, however, require us to treat $\eta(\lambda, c)$ as a bivariate function – see Section 7.

1.2 Some key observations

The sections to follow construct a framework for evaluating and optimizing the asymptotic loss(1.8). We highlight here some observations that will play an important role. Beforehand, let us introduce a useful modification of (1.7) to a *rank-aware* shrinkage rule:

$$\hat{\Sigma}_{\eta, r} = \sum_{i=1}^{r} \eta(\lambda_i) v_i v_i' + \sum_{i=r+1}^{p} v_i v_i', \quad (1.13)$$

where the dimension r of the spiked model is taken as known. While our main results concern estimators $\hat{\Sigma}_\eta$ that naturally do not require r to be known in advance, it will be easier conceptually and technically to analyze rank-aware shrinkage rules as a preliminary step.

[Obs. 1] *Simultaneous block diagonalization.* (Lemmas 1 and 5). There exists a (random) basis W such that

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

where A_i and B_i are square blocks of equal size d_i, and $\Sigma d_i = 2r$. (Here and below, A ⊕ B denotes a block-diagonal matrix with blocks A and B).

[Obs. 2] *Decomposable loss functions.* The loss functions (1.11) and many others studied below satisfy

$$L_p\left(\Sigma, \hat{\Sigma}_{\eta, r}\right) = \sum_i L_{di}\left(A_i, B_i\right)$$

or the corresponding equality with *sum* replaced by *max*.

[Obs. 3] *Asymptotic deterministic loss.* (Lemmas 3 and 7). For rank-aware estimators, when η and L are suitably continuous, almost surely

$$L_\infty\left(\ell_1, \ldots, \ell_r \middle| \eta\right) = \lim_{p \to \infty} L_p\left(\Sigma, \hat{\Sigma}_{\eta, r}\right).$$

[Obs. 4] *Asymptotic equivalence of losses.*(Proposition 2). Conclusions derived for rank-aware estimators (1.13) carry over to the original estimators (1.7) because, under suitable condition

$$L_p\left(\Sigma, \widehat{\Sigma}_\eta\right) - L_p\left(\Sigma, \widehat{\Sigma}_{\eta, r}\right) \rightarrow_P 0.$$

This relies on the fact that in the [Spike($\ell_1, \ldots, \ell_r$)] model, the sample noise eigenvalues λ_{in}, $i \geq r+1$ "stick to the bulk" in an appropriate sense.

1.3 Organization of the paper

For simplicity of exposition, we assume a single spike, $r = 1$, in the first half of the paper. [Obs. 1], [Obs. 2] and [Obs. 3] are developed respectively in Sections 2, 3 and 4, arriving at an explicit formula for the asymptotic loss of a shrinker. Section 5 illustrates the assumptions with our list of 26 decomposable matrix loss functions. In Section 6 we use the 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 many of the 26 losses. Section 7 extends the results to the general case where $r > 1$ spikes are present. We develop [Obs. 4] , remove the rank-aware assumption and explore some new phenomena that arise in cases where the optimal shrinker turns out to be discontinuous. In Section 8 we show, at least for Frobenius and Stein losses, that our optimal univariate shrinkage estimator, which applies the same scalar function to each sample eigenvalue, in fact asymptotically matches the performance of the best *orthogonally-equivariant* covariance estimator under assumption [Spike($\ell_1, \ldots, \ell_r$)]. Section 9 extends to the more general spiked model with $\Sigma_p = diag\left(\ell_1, \ldots, \ell_r, \sigma^2, \ldots, \sigma^2\right)$ for $\sigma > 0$ known or unknown. Section 10 discusses our results in light of the high-dimensional covariance estimation work of El Karoui [24] and Ledoit and Wolf [26]. Some proofs and calculations are deferred to the supplementary article [40], where we also evaluate and document the strong signal (large-ℓ) asymptotics of the optimal shrinkage estimators, and the asymptotic percent improvement over naive hard thresholding of the sample covariance eigenvalues. Additional technical details and software are provided in the Code Supplement available online as a permanent URL from the Stanford Digital Repository [41].

2 Simultaneous Block-Diagonalization

We first develop [Obs. 1] in the simplest case, $r = 1$, assumping a rank-aware shrinker. In general, the estimator $\widehat{\Sigma}_\eta$ and estimand Σ are not simultaneously diagonalizable. However, in the particular case that both are rank-one perturbations of the identity, we will see that simultaneous block diagonalization is possible.

Some notation is needed. We denote the eigenvalues and eigenvectors of the spectral decompostion $S_{n, p_n} = V \Lambda V'$ by

$$spec\left(S_{n,p_n}\right) = \left[\left(\lambda_{1n}, \ldots, \lambda_{pn}\right), \left(v_{1n}, \ldots, v_{pn}\right)\right].$$

Whenever possible, we supress the index n and write e.g. S, λ_i and v_i instead. Similarly, we often write Σ_p or even Σ for Σ_{p_n}.

Lemma 1. *Let* Σ *and* $\widehat{\Sigma}$ *be (fixed, nonrandom) p-by-p symmetric positive definite matrices with*

$$spec(\Sigma) = \left[(\ell, 1, \ldots, 1), \left(u_1, \ldots, u_p\right)\right] \quad (2.1)$$

$$spec\left(\widehat{\Sigma}\right) = \left[(\eta, 1, \ldots, 1), \left(v_1, \ldots, v_p\right)\right]. \quad (2.2)$$

Let $c = \langle u_1, v_1 \rangle$ *and* $s = \sqrt{1 - c^2}$. *Then there exists an orthogonal matrix W, which depends on* Σ *and* $\widehat{\Sigma}$, *such that*

$$W'\Sigma W = A(\ell) \oplus I_{p-2}, \quad (2.3)$$

$$W' \widehat{\Sigma} W = B(\eta, c) \oplus I_{p-2}, \quad (2.4)$$

where the fundamental 2×2 *matrices A and B are given by*

$$A(\ell) = \begin{bmatrix} \ell & 0 \\ 0 & 1 \end{bmatrix}, \; B(\eta, c) = I_2 + (\eta - 1)\begin{bmatrix} c \\ s \end{bmatrix}\begin{bmatrix} c & s \end{bmatrix}. \quad (2.5)$$

Proof. Let $\Delta = \mathrm{diag}(\eta, 1, \ldots, 1) = I + (\eta - 1)e_1 e_1'$, where e_1 denotes the unit vector in the first co-ordinate direction. It is evident that

$$\Sigma = I + (\ell - 1)u_1 u_1', \; \widehat{\Sigma} = I + (\eta - 1)v_1 v_1'. \quad (2.6)$$

It is natural, then, to work in the "common" basis of u_1 and v_1. We apply one step of Gram-Schmidt if we can, setting

$$z = \begin{cases} \left(v_1 - c u_1\right)/s & \text{if } s \neq 0 \\ u_p & \text{if } s = 0. \end{cases}$$

In the second–exceptional–case, $v_1 = \pm u_1$, so we pick a convenient vector orthogonal to u_1. In either case, the columns of the $p \times 2$ matrix $W_2 = [u_1\ z]$ are orthonormal and their span contains both u_1 and v_1. Now fill out W_2 to an orthogonal matrix $W = \left[W_2\ \ W_2^{\perp}\right]$. Observe now that if y lies in the column span of W_2 and α is a scalar, then necessarily

$$W'\left(I_p + \alpha y y'\right)W = \left(I_2 + \alpha \check{y}\check{y}\right)\Big| \oplus I_{p-2}, \quad \check{y} = W_2' y .$$

The expressions (2.3) – (2.5) now follow from the rank one perturbation forms (2.6) along with

$$W_2' u_1 = \begin{bmatrix} u_1' u_1 \\ z' u_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } W'_2 v_1 = \begin{bmatrix} u_1' v_1 \\ z' v_1 \end{bmatrix} = \begin{bmatrix} c \\ s \end{bmatrix}.$$

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 $A = B$. A *loss family* is a sequence $L = \left\{L_p\right\}_{p=1}^{\infty}$, one for each matrix size p. We often write loss function and refer to the entire family. [Obs. 2] calls out a large class of loss functions which naturally exploit the simultaneously block-diagonalizability property of Lemma 1; we now develop this observation.

Definition 1. Orthogonal Invariance. We say the loss function $L_p(A, B)$ is *orthogonally invariant* if for each orthogonal p-by-p matrix O,

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

For given p and a given sequence of block $\{d_i\}$ sizes $\sum_i d_i = p$, consider block-diagonal matrix decompositions of p by p matrices A and B into blocks A^i and B^i of size d_i:

$$A = \oplus_i A^i \qquad B = \oplus_i B^i . \qquad (3.1)$$

Definition 2. Sum-Decomposability and Max-Decomposability. We say the loss function $L_p(A, B)$ is *sum-decomposable* if for all decompositions (3.1),

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

We say that it is *max-decomposable* if if for all decompositions (3.1),

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

Clearly, such loss functions can exploit the simultaneous block diagonalization of Lemma 1. Indeed,

Lemma 2. Reduction to Two-Dimensional Problem. *Consider an orthogonally invariant loss function, L_p, which is sum- or max-decomposable. Suppose that* Σ *and* $\widehat{\Sigma}$ *satisfy (2.1) and (2.2) respectively. Then*

$$L_p\left(\Sigma, \widehat{\Sigma}\right) = L_2(A(\ell), B(\eta, c)).$$

Proof. Lemma 1 provides a change of basis W yielding decompositions (2.3) and (2.4). From the invariance and decomposability hypotheses,

$$\begin{aligned} L_p\left(\Sigma, \widehat{\Sigma}\right) &= L_p\left(W' \Sigma W, W' \widehat{\Sigma} W\right) \\ &= L_p\left(A(\ell) \oplus I_{p-2}, B(\eta), c\right) \oplus I_{p-2}\right) \\ &= L_2(A(\ell), B(\eta, c)). \end{aligned}$$

4 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)]. The principal 2× 2 block estimator occurring in Lemmas 1 and 2 is $B(\eta(\lambda_{1n}), c_{1n})$ where λ_{1n} is the largest eigenvalue of S_n and $c_{1n} = \langle u_{1n}, v_{1n} \rangle$. If η is continuous, then the convergence results (1.2) and (1.5) imply that the principal block converges as $n \to \infty$. Specifically,

$$B\left(\eta\left(\lambda_{1n}\right), c_{1n}\right) \xrightarrow{a.s.} B(\eta(\lambda(\ell)), c(\ell)) =: B(\ell, \eta), \quad (4.1)$$

say, with the convergence occurring in all norms on 2 × 2 matrices.

In accord with [Obs. 3], we now show that the asymptotic loss (1.8) is a deterministic, explicit function of the population spike ℓ For now, we will continue to assume that the shrinker η is rank-aware. Alternatively, we can make a different simplifying assumption on η, which will be useful in what follows:

Definition 3. We say that a scalar function $\eta : [0, \infty) \to [1, \infty)$ is a *bulk shrinker* if $\eta(\lambda) = 1$ when $\lambda \le \lambda_+(\gamma)$, and a *neighborhood bulk shrinker* if for some $\epsilon > 0$, $\eta(\lambda) = 1$ whenever $\lambda \le \lambda_+(\gamma) + \epsilon$.

The neighborhood bulk shrinker condition on η is rather strong, but does hold for η_*^N in (1.12), for example. (Note that our definitions ignore the lower bulk edge $\lambda_-(\gamma)$, which is of less interest in the spiked model.)

Lemma 3. A Formula for the Asymptotic Loss. *Adopt models* [Asy(γ)] *and* [Spike(ℓ)] *with* $\ell_1, \ldots, \ell_r > \ell_+(\gamma)$. *Suppose (a) that the family* $L = \{L_p\}$ *of loss functions is orthogonally* *invariant and sum- or max-decomposable, and that* $B \mapsto L_2(A, B)$ *is continuous. Let* $\hat{\Sigma}_\eta = \hat{\Sigma}_\eta\left(S_{n,p_n}\right)$ *be given by (1.7), and let* $\hat{\Sigma}_{\eta,1}$ *be the corresponding rank-aware shrinkage rule (1.13) for* $r = 1$. *Suppose the scalar nonlinearity* η *is continuous on* $(\lambda_+(\gamma), \infty)$. *Then*

$$L_{p_n}\left(\Sigma_{p_n}, \hat{\Sigma}_{\eta,1}\right) \xrightarrow{a.s.} L_2(A(\ell), B(\ell, \eta)), \quad (4.2)$$

Furthermore, if (b) η is a neighborhood bulk shrinker, then $L_{p_n}\left(\Sigma_{p_n} \ \hat{\Sigma}_\eta\right)$ also has this limit a.s.

Each of the 26 losses considered in this paper satisfies conditions (a).

Proof. In the rank-aware case $\hat{\Sigma}_\eta = \hat{\Sigma}_{\eta,1}$ satisfies

$$\operatorname{spec}\left(\hat{\Sigma}_\eta\right) = \left[\left(\eta\left(\lambda_{1n}\right), 1 \ldots, 1\right), \left(v_{1n}, \ldots, v_{pn}\right)\right],$$

Lemma 2 implies that

$$L_p\left(\Sigma, \hat{\Sigma}_\eta\right) = L_2\left(A(\ell), B\left(\eta\left(\lambda_{1n}\right), c_{1n}\right)\right) \xrightarrow{a.s.} L_2(A(\ell), B(\ell, \eta)),$$

where the limit on the right hand side follows from convergence (4.1) and the assumed continuity of L_2.

Now assume that η is a neighborhood bulk shrinker. From (1.2) we know that $\lambda_{1n} \xrightarrow{a.s.} \lambda(\ell)$ From eigenvalue interlacing (see (7.11) below) we have $\lambda_{2n} \le \mu_{1n}$, where μ_{1n} is the largest eigenvalue of a white Wishart matrix $W_{pn-1}(n,I)$, and satisfies $\mu_{1n} \xrightarrow{a.s.} \lambda_+$, from [42]. Let $\epsilon >$ 0 be small enough that $\lambda_+ + \in < \lambda(\ell)$ and also lies in the neighborhood shrunk to 1 by η. Hence, there exists a random variable $\hat{n}$ such that almost surely, $\lambda_{2n} < \lambda_+ + \epsilon < \lambda_{1n}$ for all $n > \hat{n}$. For such n, the first display above of this proof applies and we then obtain the second display as before.

5 Examples of Decomposable Loss Functions

Many of the loss functions that appear in the literature are *Pivot-Losses*. They can be obtained via the following common recipe:

Definition 4. Pivots. A *matrix pivot* is a matrix-valued function $\Delta(A, B)$ of two real positive definitee matrices A, B such that: (i) $\Delta(A, B) = 0$ if and only if $A = B$, (ii) Δ is orthogonally equivariant and (iii) Δ respects block structure in the sense that

$$\Delta\,(OAO', OBO') = O\,\Delta\,(A,B)O', \quad (5.1)$$

$$\Delta\left(\oplus A^i, \oplus B^i\right) = \oplus\ \Delta\left(A^i, B^i\right) \quad (5.2)$$

for any orthogonal matrix O of the appropriate dimension.

Matrix pivots can be symmetric-matrix valued, for example $\Delta(A, B) = A - B$, but need not be, for example $\Delta(A, B) = A^{-1}B - I$.

Definition 5. Pivot-Losses. Let g be a non-negative function of a symmetric matrix variable that is definite: $g(A) = 0$ if and only if $A = 0$, and orthogonally invariant: $g(O\Delta O^{\grave{}}) = g(\Delta)$ for any orthogonal matrix O. A symmetric-matrix valued pivot Δ induces an orthgonally-invariant *pivot loss*

$$L(A,B) = g(\,\Delta\,(A,B))\,. \quad (5.3)$$

More generally, for any matrix pivot Δ, set $|\,\Delta\,| = (\,\Delta'\ \ \Delta)^{1/2}$ and define

$$L(A,B) = g(|\,\Delta\,|(A,B))\,. \quad (5.4)$$

An orthogonally invariant function g depends on its matrix argument Δ or |Δ| only through its eigenvalues or singular values $\delta_1, \ldots, \delta_p$. We abuse notation to write $g(\Delta) = g(\delta_1, \ldots, \delta_p)$. Observe that if g has either of the forms

$$g\left(\delta_1, \ldots, \delta_p\right) = \sum_j g_1\left(\delta_j\right) \qquad \text{or} \qquad g\left(\delta_1, \ldots, \delta_p\right) = \max_j g_1\left(\delta_j\right),$$

for some univariate g_1, then the pivot loss $L(A, B) = g(\Delta(A, B))$ (symmetric pivot) or $L(A,B)=g(|\Delta|(A,B))$ (general pivot) is respectively sum- or max-decomposable. In case Δ is symmetric, the two definitions agree so long as g_1 is an even function of δ.

5.1 Examples of Sum-Decomposable Losses

There are different strategies to derive sum-decomposable pivot-losses. First, we can use statistical discrepancies between the Normal distributions $\mathcal{N}(0, A)$ and $\mathcal{N}(0, B)$:

1. *Stein Loss* [1, 9, 39]: Stein's Loss is defined as

$$L^{st}(A,B) = \mathrm{tr}\left(A^{-1}, B - I\right) - \log(\det(B)/\det(A))\,.$$

This is just twice the Kullback distance $D_{KL}(\mathcal{N}(0,B)||\mathcal{N}(0,A))$. Stein's loss is a pivot-loss with respect to $\Delta(A, B) = A^{-1/2}BA^{-1/2}$ and $g(\Delta) = \mathrm{tr}(\Delta - I) - \log\det(\Delta) = \Sigma_i\, g_1(\delta_i)$, where $g_1(\delta) = \delta - 1 - \log\,\delta$.

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 $L^{ent}(A, B) = L^{st}(B, A)$ [11, 15] and summing the Stein and Entropy losses gives *divergence* loss:

$$L^{div}(A,B) = L^{st}(A,B) + L^{st}(B,A) = \mathrm{tr}\left(A^{-1}, B - I\right) + \mathrm{tr}\left(B^{-1}A - I\right),$$

see [43, 18]. Each can be shown to be sum-decomposable, following the same argument as above.

3. *Bhattarcharya/Matusita Affinity* [44, 45]: Let

$$L^{aff}(A,B) = \frac{1}{2}\log\frac{|A+B|/2}{|A|^{1/2}|B|^{1/2}}.$$

This measures the statistical distinguishability of $\mathcal{N}(0,A)$ and $\mathcal{N}(0,B)$ 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 $\mathcal{N}(0,A)$ and $\mathcal{N}(0,B)$. Hence convergence of affinity loss to zero is equivalent to convergence of the underlying densities in Hellinger or Variation distance. This is a pivot-loss w.r.t $\Delta(A, B) = A^{-1/2}BA^{-1/2}$ and

$$g(\Delta) = \frac{1}{4}\log\left(\det\left(2I + \Delta + \Delta^{-1}\right)/4\right) = \sum_i g_1\left(\delta_i\right),$$

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

4. *Fréchet Discrepancy* [46, 47]: Let $L^{fre}(A, B) = \mathrm{tr}(A + B - 2A^{1/2}B^{1/2})$. This measures the minimum possible mean-squared difference between zero-mean random vectors with covariances A and B respectively. This is a pivot-loss w.r.t $\Delta(A, B) = A^{1/2} - B^{1/2}$, and $g(\Delta) = \mathrm{tr}\left(\Delta^2\right) = \Sigma_i\, g_1\left(\delta_i\right)$ with $g_1(\delta) = \delta^2$.

Second, we may obtain sum-decomposable pivot-losses $L(A, B) = g(\Delta(A, B))$ by simply taking g to be one of the standard matrix norms:

1. *Squared Error Loss* [3, 28, 25, 26]: Let $L^{F,1}(A,B) = \|A - B\|_F^2$. This is a pivot-loss w.r.t $\Delta(A, B) = A - B$ and $g(\Delta) = \mathrm{tr}\,\Delta'\Delta = \Sigma_i\, g_1\left(\delta_i\right)$ with $g_1(\delta) = \delta^2$.

2. *Squared Error Loss on Precision* [8]: Let $L^{F,2}(A,B) = \|A^{-1} - B^{-1}\|_F^2$. This is a pivot-loss w.r.t $\Delta(A, B) = A^{-1} - B^{-1}$ and $g(\Delta)=\mathrm{tr}\,\Delta'\Delta$.

3. *Nuclear Norm Loss.* Let $L^{N,1}(A,B) = \|A - B\|_*$ where $\| \Delta \|_*$ denotes the nuclear norm of the matrix Δ, i.e. the sum of its singular values. This is a kpivot-loss k w.r.t $\Delta(A, B) = A - B$ and $g(\Delta) = \sum_i |\delta_i|$.

4. Let $L^{F,3}(A,B) = \|A^{-1}B - I\|_F^2$. This is a pivot-loss w.r.t $\Delta(A, B) = A^{-1}B - I$. It was studied in [48, 6, 10] and later work.

5. Let $L^{F,7}(A,B) = \|\log A^{-1/2}BA^{-1/2}\|_F^2$, where log() denotes the matrix logarithm[1] [51, 49]. This is a pivot-loss w.r.t

$$\Delta(A,B) = \log\left(A^{-1/2}BA^{-1/2}\right).$$

5.2 Examples of Max-Decomposable Losses

Max-decomposable losses arise by applying the operator norm (the maximal singular value or eigenvalue of a matrix) to a suitable pivot. Here are a few examples:

1. *Operator Norm Loss* [52]: Let $L^{O,1}(A,B) = \|A - B\|_{op}$.This is a pivot-loss w.r.t $\Delta(A, B) = A - B$ and $g(\Delta) = \| \Delta \|_{op} = \max_i \delta_i$.

2. *Operator Norm Loss on Precision:* Let $L^{O,2}(A,B) = \|A^{-1} - B^{-1}\|_{op}$. This is a pivot-loss w.r.t. $\Delta(A, B) = A^{-1} - B^{-1}$.

3. *Condition Number Loss*: Let $L^{O,7}(A,B) = \left\|\log\left(A^{-1/2}BA^{-1/2}\right)\right\|_{op}$. This is a pivot-loss w.r.t $\Delta(A, B) = \log(A^{-1/2}BA^{-1/2})$, related to [29]. In the spiked kmodel discussed below, $L^{O,7}$ effectively measures the condition number of $A^{-1/2}BA^{-1/2}$.

We adopt the systematic naming scheme $L^{\text{norm,pivot}}$ where norm ∈ {F, O, N}, and pivot ∈ {and 1, ..., 7}. This set of 21 combinations covers the previous matrix norm examples and adds some more. Together with Stein's loss and the others based on statistical discrepancy mentioned above, we arrive at a set of 26 loss functions, Table 1, to be studied in this paper.

6 Optimal Shrinkage for Decomposable Losses

6.1 Formally Optimal Shrinker

Formula (4.2) for the asymptotic loss has only been shown to hold in the single spike model and only for a certain class of nonlinearities η. In fact, the same is true in the r-spike model and for a much broader class of nonlinearities η. To preserve the narrative flow of the paper, we defer the proof, which is more technical, to Section 7. Instead, we proceed under the single spike model, and simply assume that $L_\infty(\ell|\eta)$ from (4.2) is the correct limiting loss, and draw conclusions on the optimal shape of the shrinker η.

[1] The matrix logarithm transfers the matrices from the Riemannian manifold of symmetric positive semidefinite matrices to its tangent space at A. It can be shown that $L^{F,7}$ is the squared geodesic distance in this manifold. This metric between covariances has attracted attention, for example, in diffusion tensor MRI [49, 50].

Definition 6. Optimal Shrinker. Let $L = \{L_p\}_{p=1}^{\infty}$ be a given loss family and let $L_\infty(\ell|\eta)$ be the asymptotic loss corresponding to a nonlinearity η, as defined in (4.2), under assumption [Asy(γ)] . If η^* satisfies

$$L_\infty(\ell|\eta^*) = \min_\eta L_\infty(\ell|\eta), \ \forall \ell \geq 1, \quad (6.1)$$

and for any $\eta \neq \eta^*$ there exists $\ell \geq 1$ with $L_\infty(\ell|\eta^*) < L_\infty(\ell|\eta)$, then we say that η^* is the *formally optimal* shrinker for the loss family L and shape factor γ, and denote the corresponding shrinkage rule by $\lambda \mapsto \eta^*(\lambda; \gamma, L)$.

Below, we call formally optimal shrinkers simply "optimal". By definition, the optimal shrinkage 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 the single spiked model (and as we show later, generally in the spiked model) one never regrets using the optimal shrinker over any other (reasonably regular) univariate shrinker. In light of our results so far, an obvious characterization of an optimal shrinker is as follows.

Theorem 1. Characterization of Optimal Shrinker. *Let* $L = \{L_p\}_{p=1}^{\infty}$ *be a loss family. Define*

$$F(\ell, \eta) = L_2\left(\begin{bmatrix} \ell & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 + (\eta - 1)c^2 & (\eta - 1)cs \\ (\eta - 1)cs & 1 + (\eta - 1)s^2 \end{bmatrix}\right). \quad (6.2)$$

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 that for any $\ell > \ell_+(\gamma)$, there exists a unique minimizer

$$\eta^*(\ell) := argmin_{\eta \geq 1} F(\ell, \eta). \quad (6.3)$$

Further suppose that for every $1 \leq \ell \leq \ell_+(\gamma)$ we have $\text{argmin}_{\eta \geq 1} G(\eta) = 1$, where

$$G(\ell, \eta) = L_2\left(\begin{bmatrix} \ell & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \eta \end{bmatrix}\right). \quad (6.4)$$

Then the shrinker

$$\eta^*(\lambda) = \begin{cases} \eta^*(\ell(\lambda)) & \ell > \lambda_+(\gamma) \\ 1 & 1 \leq \ell \leq \lambda_+(\gamma) \end{cases},$$

where $\ell(\lambda)$ is given by (1.10), is the optimal shrinker of the loss family L.

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 shrinker numerically, by implementing in software a solver for the simple scalar optimization problem (6.3). Figure 3 portrays the optimal shrinkers for our 26 loss functions. We refer readers interested in computing specific individual shrinkers to our reproducibility advisory at the bottom of this paper, and invite the reader to explore the code supplement [41], consisting of online resources and code we offer.

6.2 Optimal Shrinkers Collapse 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 the supplemental article [40]:

Lemma 4. 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 $[1, \ell_+(\gamma)]$

As part of the proof of Lemma 4, in Table 6 in the supplemental article [40], we explicitly calculate the fundamental loss function $G(\ell, \eta)$ of (6.4) for many of the loss families discussed in this paper.

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

6.3 Optimal Shrinkers by Computer

The scalar optimization problem (6.3) is easy to solve numerically, so that one can always compute the optimal shrinker at any desired value λ. In the code supplement [41] we provide Matlab code to compute the optimal nonlinearity for each of the 26 loss families discussed. In the sibling problem of singular value shrinkage for matrix denoising, [53] demonstrates numerical evaluation of optimal shrinkers for the Schatten-p norm, where analytical derivation of optimal shrinkers appears to be impossible.

6.4 Optimal Shrinkers in Closed Form

We were able to obtain simple analytic formulas for the optimal shrinker η^* in each of 18 loss families from Section 3. While the optimal shrinkers are of course functions of the empirical eigenvalue λ, in the interest of space, we state the lemmas and 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. If $\lambda \leq \lambda_+(\gamma)$ set $\eta^*(\lambda) = 1$. Otherwise:
2. Calculate $\ell(\lambda)$ using (1.10).
3. Calculate $c(\lambda) = c(\ell(\lambda))$ using (1.6) and (1.10).
4. Calculate $s(\lambda) = s(\ell(\lambda))$ using $s(\ell) = \sqrt{1 - c^2(\ell)}$.

5. Substitute $\ell(\lambda)$ $c(\lambda)$ and $s(\lambda)$ into the formula provided to get $\eta^*(\lambda)$.

The closed forms we provide are summarized in Table 2. Note that ℓ, c and s refer to the functions $\ell(\lambda)$, $c(\ell(\lambda))$ and $s(\ell(\lambda))$ These formulae are formally derived in a sequence of lemmas that are stated and proved in the supplemental article [40]. The proofs also show that these optimal shrinkers are unique, as in each case the optimal shrinker is shown to be the unique minimizer, as in (6.3), of (6.2). We make some remarks on these optimal shrinkers by focusing first on operator norm loss for covariance and precision matrices:

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

This asymptotic relationship reflects the classical fact that in finite samples, the top empirical eigen value is always biased upwards of the underlying population eigenvalue [54, 55]. 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 words, within the top branch of (6.5), the effect of *operator-norm optimal shrinkage is to debias the top eigenvalue*:

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

On the other hand, within the bottom branch, the effect is to *shrink the bulk to 1*. In terms of Definition 3 we see that η^* is a bulk shrinker, but not a neighborhood bulk shrinker.

One might expect asymptotic debiasing from *every* loss function, but, perhaps surprisingly, precise asymptotic debiasing is exceptional. In fact, none of the other optimal nonlinearities in Table 2 is precisely debiasing.

In the supplemental article [40] we also provide a detailed investigation of the large-λ asymptotics of the optimal shrinkers, including their asymptotic slopes, asymptotic shifts and asymptotic percent improvement.

7 Beyond Formal Optimality

The shrinkers we have derived and analyzed above are formally optimal, as in Definition 6, in the sense that they minimize the 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)], and only over neighborhood bulk shrinkers.

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(ℓ_1, …,ℓ_r)], and over a large class of bulk shrinkers, which are possibly not neighborhood bulk shrinkers.

We start by establishing the rank r analog of Lemma 1. For a vector $\ell \in \mathbb{R}^r$, let $\Delta_r(\ell)$ = diag(ℓ_1,…,ℓ_r)

Lemma 5. *Assume that* Σ *and* $\widehat{\Sigma}$ *are fixed matrices with*

$$spec(\Sigma) = \left[\left(\ell_1 \dots \ell_r, 1 \dots, 1\right), \left(u_1, \dots, u_p\right)\right]$$
$$spec(\widehat{\Sigma}) = \left[\left(\eta_1 \dots \eta_r, 1 \dots, 1\right), \left(v_1, \dots, v_p\right)\right].$$

Let U_r and V_r denote the p-by-r matrices consisting of the top r eigenvectors of Σ *and* $\widehat{\Sigma}$ *respectively. Suppose that* $[U_r\ V_r]$ *has full rank* $2r$*, and consider the QR decomposition*

$$\left[U_r V_r\right] = QR,$$

where Q has $2r$ *orthonormal columns and the* $2r \times 2r$ *matrix R is upper triangular. Let R_2 denote the* $2r \times r$ *submatrix formed by the last r columns of R. Fill out Q to an orthogonal matrix* $W = [Q\ Q^{\perp}]$*. Then in the transformed basis we have the simultaneous block decompositions*

$$W'\ \Sigma\ W = \Sigma^{\circ}_{2r} \oplus I_{p-2r}, \quad \Sigma^{\circ}_{2r} = \Delta_r(\ell) \oplus I_r \qquad (7.1)$$

$$W'\ \widehat{\Sigma}\ W = \widehat{\Sigma}^{\circ}_{2r} \oplus I_{p-2r}, \quad \widehat{\Sigma}^{\circ}_{2r} = I_{2r} + R_2\ \Delta_r(\eta - 1)R_2'. \qquad (7.2)$$

Proof. We start with observations about the structure of Q and R. Since the first r columns of Q are identically those of U_r, we let Z_r be the n-by-r matrix such that $Q = [U_r\ Z_r]$. For the same reason, R 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 $V_r = U_r R_{12} + Z_r R_{22}$, so that

$$R_{12} = U'_r V_r \quad R_{22} = Z'_r V_r. \qquad (7.3)$$

Since V_r has orthogonal columns, we have

$$I_r = V'_r V_r = R'_{12} R_{12} + R'_{22} R_{22} \qquad (7.4)$$
$$R'_{22} R_{22} = I - R'_{12} R_{12}.$$

Let H be a $p \times r$ matrix whose columns lie in the column span of Q and let Δ be an $r \times r$ diagonal matrix. Observe that

$$W'(I + H \Delta H')W = I + W'H \Delta H'W$$
$$= \left(I_{2r} + Q'H \Delta H'Q\right) \oplus I_{p-2r} = C_{2r} \oplus I_{p-2r},$$

say, since the columns of $Q^{\perp}$ are orthogonal to those of H. By analogy to (2.6), we may write

$$\Sigma = I + U_r(\Delta_r(\ell) - I_r)U'_r, \qquad \hat{\Sigma} = I + V_r(\Delta_r(\eta) - I_r)V'_r \qquad (7.5)$$

and so both of the form $I + H\Delta H'$, with $H = U_r$ and V_r respectively. We find that

$$Q'U_r = \begin{bmatrix} I_r \\ 0 \end{bmatrix}, \quad Q'V_r = \begin{bmatrix} R_{12} \\ R_{22} \end{bmatrix} = R_2,$$

We can then compute the value of C_{2r} in the two cases to be given by Σ°_{2r} and $\hat{\Sigma}^{\circ}_{2r}$ respectively, which establishes (7.1) and (7.2), and hence the lemma.

We intend to apply Lemma 5 to Σ and $\hat{\Sigma} = \hat{\Sigma}_{\eta,r}$, the “rank-aware” modification (1.13) of the estimator $\hat{\Sigma}_\eta$ in (1.7). Assume now that $\hat{\Sigma}$ and the $p \times r$ matrix $V_{r,n}$ formed by the top eigenvectors of V are random.

Lemma 6. *The rank of* $[U_r\ V_{r,n}]$ *equals* 2r *almost surely.*

Proof. Let $\prod_r(V)$ be the projection that picks out the first r columns of an orthogonal matrix V. For a fixed r-frame U_r, we consider the event

$$A = \left\{V \in O_p : \operatorname{rank}\left(\left[U_r \prod_r(V)\right]\right) < 2r\right\},$$

where O_p is the group of orthogonal p-by-p matrices. Let $P_\Sigma(d\Lambda, dV)$ denote the joint distribution of eigenvalues $\Lambda = \operatorname{diag}(\lambda_1,...,\lambda_p)$ and eigenvectors V when $S \sim W_p(n, \Sigma)$. As shown by [56], P_Σ is absolutely continuous with respect to $v_p \times \mu_p$, the product of Lebesgue measure on $\mathbb{R}^p$ and Haar measure on $O(p)$. Since $\mu_p(A) = 0$, it follows that $P_\Sigma(A) = 0$.

Lemma 7. Adopt models [Asy(γ)] [Spike($\ell_1,\ldots,\ell_r$)]and with $\ell_1,\ldots,\ell_r > \ell_r > \ell_+(\gamma)$. Suppose the scalar nonlinearity η is continuous on $(\lambda_+(\gamma), \infty)$. For each p there exists w.p. 1 an orthogonal change of basis W such that

$$W'\Sigma W = \Sigma_{2r} \oplus I_{p-2r}, \quad W'\hat{\Sigma}_{\eta,r} W = \hat{\Sigma}_{2r} \oplus I_{p-2r}, \quad (7.6)$$

where the $2r \times 2r$ *matrices* Σ_{2r}, $\hat{\Sigma}_{2r}$ *satisfy*

$$\Sigma_{2r} = \oplus_{i=1}^{r} A(\ell_i), \ \hat{\Sigma}_{2r} \xrightarrow{a.s.} \oplus_{i=1}^{p} B(\ell_i, \eta), \quad (7.7)$$

and the 2 × 2 matrices $A(\ell)$, $B(\ell, \eta)$ are defined at (2.5).

Suppose also that the family $L = \{L_p\}$ *of loss functions is orthogonally invariant and sum- or maxdecomposable, and that* $B \to L_{2r}(A, B)$ *is continuous. Then*

$$L_p\left(\Sigma, \hat{\Sigma}_{\eta,r}\right) \xrightarrow{a.s.} \left(\sum / \max\right)_{i=1,\ldots r} L_2\left(A(\ell_i), B(\ell_i, \eta)\right). \quad (7.8)$$

If η *is a neighborhood bulk shrinker, then* $L_p\left(\Sigma, \hat{\Sigma}_{\eta}\right)$ *also has this limit a.s.*

This is the rank r analog of Lemma 3. The optimal nonlinearity η^* is continuous on $[0,\infty)$ for all losses except the operator norm ones, for which η^* is continuous except at $\lambda = \lambda_+(\gamma)$. Our result (7.7) requires only continuity on $(\lambda_+(\gamma), \infty)$ and so is valid for all 26 loss functions, as is the deterministic limit (7.8) for the rank-aware $\hat{\Sigma}_{\eta,r}$. However, as we saw earlier, only the nuclear norm based loss functions yield optimal functions that are neighborhood bulk shrinkers. To show that (7.8) holds for $L_p\left(\Sigma, \hat{\Sigma}_{\eta}\right)$ for most other important shrinkage functions, some further work is needed – see Section 7.1 below.

Proof. We apply Lemma 5 to Σ and $\hat{\Sigma}_{\eta,r}$ on the set of probability 1 provided by Lemma 6. First, we rewrite (7.2) to show the subblocks of R:

$$\hat{\Sigma}^{\circ}_{2r} = I_{2r} + \begin{bmatrix} R_{12} \\ R_{22} \end{bmatrix} \Delta_r \left(\eta^{(n)} - 1\right)\left[R'_{12} \, R'_{22}\right],$$

where we write $\eta^{(n)} = (\eta(\lambda_{1,n}),\ldots, \eta(\lambda_{r,n}))$ to show explicitly the dependence on n. The limiting behavior of R may be derived from (7.3) and (7.4) along with spiked model properties (1.2) and (1.5), so we have[2], as $n \to \infty$,

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

$$R_{12} = U'_r V_{r,n} \rightarrow_{a.s.} \Delta_r(c)$$
$$R_{22} R'_{22} = I - R_{12} R'_{12} \rightarrow_{a.s.} \Delta_r\left(s^2\right) \quad (7.9)$$
$$R_{22} \rightarrow_{a.s.} \Delta_r(s).$$

Here $c = (c(\ell_1),\ldots, c(\ell_r))$ and $s = (s(\ell_1),\ldots, s(\ell_r))$.

Again by (1.2) $\lambda_{i,n} \rightarrow_{a.s.} \lambda(\ell_i) > \lambda_+(\gamma)$ and so continuity of η above $\lambda_+(\gamma)$ assures that $\Delta_r(\eta^{(\eta)}-1) \rightarrow \Delta_r(\eta-1)$, where $\eta = (\eta_i)$ and $\eta_i = \eta(\lambda(\ell_i))$. Together with (1.5), we obtain simplified structure in the limit,

$$\hat{\Sigma}^{\circ}_{2r} \rightarrow_{a.s.} I_{2r} + \begin{bmatrix} \Delta_r\left((\eta-1)c^2\right) & \Delta_r((\eta-1)cs) \\ \Delta_r((\eta-1)cs) & \Delta_r\left((\eta-1)s^2\right) \end{bmatrix}. \quad (7.10)$$

To rewrite the limit in block diagonal form, let Π_{2r} be the permutation matrix corresponding to the permutation defined by

$$(1,\ldots,2r) \longmapsto (1, r+1, 2, r+2, 3,\ldots, 2r).$$

Permuting rows and columns in (7.1) and (7.10) using Π_{2r} to obtain

$$\Sigma_{2r} := \prod\nolimits'_{2r} \Sigma^{\circ}_{2r} \prod\nolimits_{2r} = \oplus_{i=1}^{r} A(\ell_i),$$
$$\hat{\Sigma}_{2r} := \prod\nolimits'_{2r} \hat{\Sigma}^{\circ}_{2r} \prod\nolimits_{2r} \rightarrow_{a.s.} \oplus_{i=1}^{p} B(\ell_i, \eta),$$

we obtain (7.7). Using (7.6), the orthogonal invariance and sum/max decomposability, along with the continuity of $L_{2r}(A,\cdot)$, we have

$$\begin{aligned} L_p\left(\Sigma_p, \hat{\Sigma}_{\eta,r}\right) &= L_p\left(\Sigma_{2r} \oplus I_{p-2r}, \hat{\Sigma}_{2r} \oplus I_{p-2r}\right) \\ &= L_{2r}\left(\Sigma_{2r}, \hat{\Sigma}_{2r}\right) \\ &= L_{2r}\left(\prod\nolimits'_{2r} \Sigma_{2r} \prod\nolimits_{2r}, \prod\nolimits'_{2r} \hat{\Sigma}_{2r} \prod\nolimits_{2r}\right) \rightarrow_{a.s.} L_{2r}\left(\oplus_{i=1}^{r} A(\ell_i), \oplus_{i=1}^{p} B(\ell_i, \eta)\right) \\ &= \left(\sum/\max\right)_{i=1,\ldots r} L_2\left(A(\ell_i), B(\ell_i, \eta)\right), \end{aligned}$$

which completes the proof of Lemma 7.

7.1 Removing the rank-aware condition

In this section we prove Proposition 2 below, whereby the asymtotic losses coincide for a given estimator sequence $\hat{\Sigma}_\eta$ and the rank-aware versions $\hat{\Sigma}_{\eta,r}$. This result is plausible because of two observations:

1. Null eigenvalues *stick to the bulk*, i.e. for $i \geq r + 1$ exceptions , most eigenvalues λ_{in} $\lambda_+(\gamma)$ and the few exceptions are not much larger. Hence, if η is a continuous bulk shrinker, we expect $\widehat{\Sigma}_\eta$ to be close to $\widehat{\Sigma}_{\eta, r}$,

2. under a suitable continuity assumption on the loss functions L_p, $L\left(\Sigma, \widehat{\Sigma}_\eta\right)$should then be close to $L\left(\Sigma, \widehat{\Sigma}_{\eta, r}\right)$.

Observation 1 is fleshed out in two steps. The first step is eigenvalue comparison: The sample eigenvalue λ_{in} arise as eigenvalues of XX'/n when X is a p_n-by-n matrix whose rows are i.i.d draws from $\mathcal{N}\left(0, \Sigma_{p_n}\right)$. (0, $\sum_{pn}$). Let $\prod : \mathbb{R}^{p_n} \to \mathbb{R}^{p_n - r}$ denote the projection on the last $p_n - r$ coordinates in $\mathbb{R}^{p_n}$ and let $\mu_{1n} \geq \ldots \geq \mu_{pn-r,n}$ denote the eigenvalues of $\prod X(\prod X)'/n$. By the Cauchy interlacing Theorem (e.g. [57, p. 59]), we have

$$\lambda_{in} \leq \mu_{i - r, n} \quad \text{for } r + 1 \leq i \leq p_n, \quad (7.11)$$

where the (μ_{in}) are the eigenvalues of a white Wishart matrix $W_{pn-r}(n, I)$.

The second step is a bound on eigenvalues of a white Wishart that exit the bulk. Before stating it, we return to an important detail introduced in the Remark concluding Section 1.1.

Definition 3 of a bulk shrinker depends on the parameter $\gamma = \lim p/n$ through $\lambda_+(\gamma)$. Making that dependence explicit, we obtain a bivariate function $\eta(\lambda, c)$. In model [Asy(γ)]and in the n-th problem, we might use $\eta(\lambda, c_n)$ either with $c_n = \gamma$ or $c_n = p/n$. For Proposition 1 below, it will be more natural to use the latter choice. We also modify Definition 3 as follows.

Definition 7. We call $\eta : [0, \infty) \times (0, 1] \to [1, \infty)$ a *jointly continuous bulk shrinker* if $\eta(\lambda, c)$ is jointly continuous in λ and c, satisfies $\eta(\lambda, c) = 1$ for $\lambda \leq \lambda_+(c)$ and is dominated: $\eta(\lambda, c) \leq M\lambda$ for some M and all λ.

The following result is proved in [58, Theorem 2(a)].

Proposition 1. Let $\left(\mu_{in}\right)_{i=1}^N$ denote the sample eigenvalues of a matrix distributed as $W_N(n, I)$, with $N/n \to \gamma > 0$. Suppose that $\eta(\lambda, c)$ is a jointly continuous bulk shrinker and that $c_n - N/n = O(n{-}2/3)$. Then for $q > 0$,

$$\left\|\eta\left(\mu_{in}, c_n\right) - 1\right\|_{\ell_q\left(\mathbb{R}^N\right)} \to P\, 0. \quad (7.12)$$

The continuity assumption on the loss functions may be formulated as follows. Suppose that A, B_1, B_2 are p-by-p positive definite matrices, with A satisfying assumption [Spike($\ell_1, \ldots, \ell_r$)] and spec(B_k) = [(η_{ki}), (v_i)], thus B_1 and B_2 have the same eigenvectors. Set $\eta_1 = \max\{\eta_{11}, \eta_{21}\}$. We assume that for some $q \in [1, \infty]$ and some continuous function $C(\ell_1, \eta_1)$ not depending on p, we have

$$\left|L_p(A, B_1) - L_p(A, B_2)\right| \le C(\ell_1, \eta_1)\left\|\eta_1 - \eta_2\right\|_{\ell_q(\mathbb{R}^p)} \quad (7.13)$$

whenever $\|\eta_1 - \eta_2\|_{\ell_q(\mathbb{R}^p)} \le 1$. Condition (7.13) is satisfied for all 26 of the loss functions of Section 3, as is verified in Proposition 1 in SI.

In the next proposition we adopt the convention that estimators $\widehat{\Sigma}_\eta$ of (1.7) and $\widehat{\Sigma}_{\eta,r}$ of (1.13) are constructed with a jointly continuous bulk shrinker, which we denote $\eta(\lambda, c_n)$.

Proposition 2. Adopt models [Asy(γ)] *and* [Spike($\ell_1,\ldots,\ell_r$)]. Suppose also that the family $L = \{L_p\}$ of loss functions is orthogonally invariant and sum- or max- decomposable, and satisfies continuity condition (7.13). If $\eta(\lambda, c_n)$ is a jointly continuous bulk shrinker with $c_n = p_n/n$, then

$$L_p\left(\Sigma, \widehat{\Sigma}_\eta\right) - L_p\left(\Sigma, \widehat{\Sigma}_{\eta,r}\right) \to_P 0,$$

and so $L_p\left(\Sigma, \widehat{\Sigma}_\eta\right)$ converges in probability to the deterministic asymptotic loss (7.8).

proof in the left side of (7.13), substitute $A = \Sigma$, $B_1 = \widehat{\Sigma}_\eta$ *and* $B_2 = \widehat{\Sigma}_{\eta,r}$. By definition ,$\widehat{\Sigma}_\eta$ and $\widehat{\Sigma}_{\eta,r}$ share the same eigenvectors. The components of $\eta_1 - \eta_2$ then satisfy

$$\eta_{1i} - \eta_{2i} = \begin{cases} \eta(\lambda_{in}, c_n) - 1 & i \ge r+1 \\ 0 & 1 \le i \le r. \end{cases}$$

We now use (7.11)to compare the eigenvectors λ_{in} of the spiked model to those of a suitable white Wishart matrix to which Proposition 1 applies. The fuction $\eta^{\uparrow}(\mu, c) = \max\{\eta(\lambda, c), 1 \le \lambda \le \mu\}$ and is non-decreasing and jointly continuos. Hence $\eta(\lambda_{in}, c_n) \le \eta^{\uparrow}(\lambda_{in}, c_n) \le \eta^{\uparrow}(\mu_{i-r,n in}, c_n)$, and so

$$\sum_{i=r+1}^{p}\left[\eta(\lambda_{in}, c_n) - 1\right]^q \le \sum_{j=1}^{p-r}\left[\eta^{\uparrow}(\mu_{jn}, c_n) - 1\right]^q,$$

with a corresponding bound for $q = \infty$. From continuity condition (7.13),

$$\left|L_p\left(\Sigma, \widehat{\Sigma}_\eta\right) - L_p\left(\Sigma, \widehat{\Sigma}_{\eta,r}\right)\right| \le C(\ell_1, \eta(\lambda_{1n}, c_n))\left\|\eta^{\uparrow}(\mu_{jn}, c_n) - 1\right\|_{\ell_q(\mathbb{R}^{p-r})}.$$

The constant $C(\ell_1, \eta(\lambda_{1n}, c_n))$ remains bounded by (1.2). The ℓ_q norm converges to 0 in probability, applying Proposition 1 to the eigenvalues of $W_{pn-r}(n, I)$, with $N = p_n - r$, noting that $c_n - N/n = r/n = O(n^{-2/3})$.

7.2 Asymptotic loss for discontinuous optimal shrinkers

Formula (6.5) showed that the optimal shrinker $\eta^*(\lambda, \gamma)$ for operator norm losses $L^{O,1}$,$L^{O,2}$ is *discontinuous* at $\ell = \ell_+(\gamma) = 1 + \sqrt{\gamma}$. In this section, we show that when η^* is used, a deterministic asymptotic loss exists for $L^{O,1}$, but *not* for $L^{O,2}$. The reason will be seen to lie in the behavior of the optimal component loss $F_*(\ell) = L_2[A(\ell),B(\ell,\eta^*)]$. Indeed, calculation based on (6.2), (6.5) shows that for $\ell \geq \ell_+$,

$$F_*(\ell) = \left[\frac{\ell^a \gamma(\ell - 1)}{\ell - 1 + \gamma}\right]^{1/2} \to F_*(\ell_+) = \begin{cases} \sqrt{\gamma} & a = 1 \\ \dfrac{\sqrt{\gamma}}{1 + \sqrt{\gamma}} & a = -1 \end{cases}$$

as $\ell \downarrow \ell_+$, where indices $a = 1$ and -1 correspond to $F_*^{O,1}$ and $F_*^{O,2}$ respectively. Importantly, $F_*^{O,1}$ is strictly increasing on $[\ell_+,\infty)$ while $F_*^{O,2}$ is strictly decreasing there.

Proposition 3. *Adopt models* $[\text{Asy}(\gamma)]$ *and* $[\text{Spike}(\ell_1,\ldots,\ell_r)]$ *with* $\ell_r > \ell_+(\gamma)$. *Consider the optimal shrinker* $\eta^*(\lambda, \gamma)$ *with* $\gamma_n = p_n/n$ *given by (6.5) for both* $L^{O,1}$ *and* $L^{O,2}$. *For* $L^{O,1}$, *the asymptotic loss is well defined:*

$$\left\| \hat{\Sigma}_\eta - \Sigma \right\|_\infty - \left\| \hat{\Sigma}_{\eta,r} - \Sigma \right\|_\infty \to_P 0. \quad (7.14)$$

However, for $L^{O,2}$,

$$\left\| \hat{\Sigma}_\eta^{-1} - \Sigma^{-1} \right\|_\infty - \left\| \hat{\Sigma}_{\eta,r}^{-1} - \Sigma^{-1} \right\|_\infty \xrightarrow{\mathcal{D}} W. \quad (7.15)$$

where W has a two point distribution in which

$$W = \begin{cases} F_*^{O,2}(\ell_+) - F_*^{O,2}(\ell_r) & \textit{with prob } 1 - F_1(0) \\ 0 & \textit{otherwise}, \end{cases}$$

and $F_1(0) = \mathbb{P}\{TW_1 \leq 0\}$ *for a real Tracy-Widom variate* TW_1 *[59].*

Roughly speaking, there is positive limiting probability that the largest noise eigenvalue will exit the bulk distribution, and in such cases the corresponding component loss $F_*(\ell_+)$ – which is due to noise alone – exceeds the largest component loss due to any of the r spikes, namely $F_*(\ell_r)$. Essentially, this occurs because precision losses $L^{\{O,F,N\},2}(a\Sigma, a\hat{\Sigma})$ *decrease* as signal strength a increases. The effect is not seen for $L^{\{F,N\},2}$ because the optimal shrinkers in those cases are continuous at ℓ_+!

Proof. For the proof, write $\|\cdot\|$ for $\|\cdot\|_\infty$. Let $W = [W_1 \; W_2]$ be the orthogonal change of basis matrix constructed in Lemma 7, with W_1 containing the first $2r$ columns. We treat the two losses $L^{O,1}$ and $L^{O,2}$ at once using an exponent $a = \pm 1$, and write $\eta^a(\lambda)$ for $\eta^a(\lambda, \gamma_n)$. Thus, let

$$\Delta = \Delta_n = \hat{\Sigma}^a_\eta - \hat{\Sigma}^a_{\eta, r} = \sum_{i = r + 1}^{p} \left[\eta^a(\lambda_i) - 1\right] v_i v_i',$$

and observe that the loss of the rank-aware estimator

$$\Psi = \Psi_n = \hat{\Sigma}^a_{\eta, r} - \hat{\Sigma}^a = \sum_{i = 1}^{r} \left[\eta^a(\lambda_i) - 1\right] v_i v_i' - \sum_{i = 1}^{r} \left[\ell_i^a - 1\right] u_i u_i'$$

lies in the column span of W_1. We have $\hat{\Sigma}^a_\eta - \Sigma^a = \Psi_n + \Delta_n$, and the main task will be to show that for $a = \pm 1$,

$$\| \Psi_n + \Delta_n \| = \max\left(\| \Psi_n \|, \| \Delta_n \|\right) + o_P(1). \quad (7.16)$$

Assuming the truth of this for now, let us derive the proposition. The quantities of interest in(7.14), (7.15) become First, note from Lemma 7 that

$$\begin{aligned} \left\| \hat{\Sigma}^a_\eta - \hat{\Sigma}^a \right\| - \left\| \hat{\Sigma}^a_{\eta, r} - \hat{\Sigma}^a \right\| &= \| \Psi_n + \Delta_n \| - \| \Psi_n \| \\ &= \max\left(\| \Delta_n \| - \| \Psi_n \|, 0\right) + o_P(1). \end{aligned}$$

First, note from Lemma 7 that

$$\| \Psi_n \| \to_{a.s.} \max_{1 \le i \le r} F_*(\ell_i). \quad (7.17)$$

Observe that for *both* $a = 1$ and -1,

$$\| \Delta_n \| = \max_{i \ge r + 1} \left| \eta^{*a}(\lambda_{in}) - 1 \right| = \left| \eta^a(\lambda_{r + 1, n}) - 1 \right|.$$

The rescaled noise eigenvalue $p^{2/3}\left(\lambda_{r + 1, n} - \lambda_+(\gamma_n)\right) \xrightarrow{\mathcal{D}} \sigma(\gamma) W$ has a limiting real Tracy-Widom distribution with scale factor $\sigma(\gamma) > 0$ [60, Prop. 5.8]. Hence, using the discontinuity of the optimal shrinker η^*, and the square root singularity from above

$$\eta^*(\lambda_{r + 1, n}, \gamma_n) = \begin{cases} \ell_+(\gamma_n) + O_P\left(p^{-1/3}\right) & \lambda_{r + 1, n} > \lambda_+(\gamma_n) \\ 1 & \lambda_{r + 1, n} \le \lambda_+(\gamma_n). \end{cases}$$

Consequently, recalling that $F_*(\ell_+) = \left|(1 + \sqrt{\gamma})^a - 1\right|$, we have

$$\| \Delta_n \| \to_P F_*(\ell_+) I(TW > 0). \quad (7.18)$$

For $L^{O,1}$, with $a = 1$, $F_*(\ell)$ is strictly increasing and so from (7.17) and (7.18), we obtain $\| \Psi_n \| \geq \| \Delta_n \| + o_P(1)$ and hence (7.14).For $L^{O,2}$, with $a = -1$, $F_*(\ell)$ is strictly so on the event $TW > 0$,

$$\| \Delta_n \| - \| \Psi_n \| \xrightarrow{\mathscr{D}} F_*(\ell_+) - F_*(\ell_r) > 0,$$

which leads to (7.15) and hence the main result.

It remains to prove (7.16). For a symmetric block matrix,

$$\max(\|A\|, \|C\|) \leq \left\| \begin{pmatrix} A & B \\ B' & C \end{pmatrix} \right\| \leq \max(\|A\|, \|C\|) + \|B\|. \quad (7.19)$$

Apply this to $W'(\Psi + \Delta)W$ with

$$\begin{aligned} A_n &= W_1'(\Psi + \Delta)W_1, \\ B_n &= W_1'(\Psi + \Delta)W_2 = W_1' \Delta W_2, \\ C_n &= W_2'(\Psi + \Delta)W_2 = W_2' \Delta W_2, \end{aligned}$$

since $\Psi W_2 = 0$. Hence

$$\| \Psi_n + \Delta_n \| = \max(\|A_n\|, \|C_n\|) + O_P(\|B_n\|). \quad (7.20)$$

We now show that $\|\Delta W_1\| \to_P 0$. Using notation from Lemma 5,

$$W_1 = [U_r V_r] R^{-1} = [U_r (V_r - U_r R_{12}) R_{22}^{-1}].$$

Since $\Delta v_k = 0$ for $k = 1, \ldots, r$,

$$\| \Delta W_1 \| \leq \| \Delta U_r \| (1 + \|R_{12} R_{22}^{-1}\|).$$

From (7.9), we have $\|R_{12} R_{22}^{-1}\| \to \| \Delta_r (c/s) \| = c(\ell_1)/s(\ell_1)$, and hence is bounded. Observe that $\Delta_{u_k} = \sum_{i = r+1}^{p} \delta_{in}^a (v_i' u_k) v_i$, where we have set $\delta_{in} = \eta(\lambda_i, \gamma_n) - 1$. Note from (6.5) that $\delta_{in} = 0$ unless $\lambda_i > \lambda_+(\gamma_n)$. With $N_n = \#\{i \geq r + 1 : \lambda_{in} > \lambda_+(\gamma_n)\}$, we then have

$$\| \Delta U_r\| \le \sqrt{r} \max_{k = 1, \ldots, r} \| \Delta_{u_k} \|2 \le \sqrt{r} \| \Delta \| N_n \max_{k \le r; i > r} |v_i' u_k|. \quad (7.21)$$

From (7.18) we have $\|\Delta_n\| = O_P(1)$. Since each v_i, $i > r$ is uniformly distributed on S^{p-1}, a simple union bound based on (7.23) below yields

$$\max_{i > r, k \le r} (v_i' u_k)^2 = O_P\left(\frac{\log p}{p}\right). \quad (7.22)$$

It remains to bound N_n. From the interlacing inequality (7.11),

$$N_n \le \tilde{N}_n = \#\{j \ge 1 : \mu_{jn} > \lambda_+(\gamma_n)\},$$

where $\{\mu_{jn}\}$ are the eigenvalues of a white Wishart matrix $W_{pn-r}(n, I)$. This quantity is bounded in [58, Theorem 2(c)], which says that $\tilde{N}_n = O_p(1)$. In more detail, we make the correspondences $N \leftarrow p_n - r$, $\gamma_N \leftarrow (p_n - r)/n$ and $c_N \leftarrow p_n/n$ so that $c_N - \gamma_N = r/n = o(n^{-2/3})$ and obtain $E\tilde{N}_n \rightarrow c_0 = 0.17$.

From (7.21) and the preceding two paragraphs, we conclude that $\| \Delta U_r\| = O_P\left(p^{-1/2}\sqrt{\log p}\right)$ and so $\| \Delta W_1\| \rightarrow_P 0$.

Returning to (7.20), we deduce now that $\|Bn\| \le \|\Delta W_1\| \rightarrow_P 0$. From the definition of W_1 we have $\|W_1' \Psi W_1\| = \| \Psi \|$ and hence the inequalities

$$\left| \|A_n\| - \| \Psi_n \| \right| \le \|W_1' \Delta W_1\| \rightarrow P\, 0.$$

Now observe that $\|C_n\| \le \|\Delta_n\|$. Apply (7.19) to $W' \Delta W$ to get

$$\| \Delta_n \| \le \|C_n\| + \|W_1' \Delta W_1\| + \|W_2' \Delta W_1\|,$$

and hence that $\|C_n\| \ge \|\Delta_n\| - o_P(1)$. Thus $\|C_n\| = \|\Delta_n\| + o_P(1)$. Inserting these results into (7.20), we obtain

$$\| \Psi_n + \Delta_n\| = \max\left(\|A_n\|, \|C_n\|\right) + o_P(1) = \max\left(\| \Psi_n \|, \| \Delta_n \|\right) + o_P(1),$$

which completes the proof of (7.16), and hence of Proposition 3.

Finally, we record a concentration bound for the uniform distribution on spheres. While more sophisticated results are known [61], an elementary bound suffices for us.

Lemma 8. *If U is uniformly distributed on S^{n-1} and $u \in S^{n-1}$ if fixed, then for M < 0 and n ≥ 4,*

$$P\left(|\langle U, u\rangle| \geq 2\sqrt{Mn^{-1}\log n}\right) \leq \sqrt{\pi/2} \cdot n^{1/2 - M}. \quad (7.23)$$

Proof. Since $U_1^2 := \langle U, u\rangle^2$ has the Beta $\left(\frac{1}{2}, \frac{n-1}{2}\right)$ distribution,

$$P\left(U_1^2 \geq a\right) \leq B\left(\frac{1}{2}, \frac{n-1}{2}\right)^{-1} \int_a^1 t^{-\frac{1}{2}}(1-t)^{\frac{n-3}{2}} dt \leq \gamma_n (1-a)^{\frac{n}{2}-1},$$

where by Gautschi's inequality [62, 63, (5.6.4)]

$$\gamma_n = B\left(\frac{1}{2}, \frac{1}{2}\right)/B\left(\frac{1}{2}, \frac{n-1}{2}\right) = \sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)/\,\Gamma\left(\frac{n-1}{2}\right) < \sqrt{\pi n/2}$$

Since $(1 - x/m)^m < e^{-x}$ for x, $m > 0$, and $4/n \geq 2/(n-2)$ for $n \geq 4$,

$$P\left(U_1^2 \geq 4Mn^{-1}\log n\right) < \sqrt{\pi n/2}\left(1 - \frac{M\log n}{n/2 - 1}\right)^{n/2 - 1} < \sqrt{\pi/2} \cdot n^{1/2 - M}.$$

8 Optimality Among Equivariant Procedures

The notion of optimality in asymptotic loss, with which we have been concerned so far, is relatively 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.

Consider the much broader class of orthogonally-equivariant procedures for covariance estimation [2, 19, 64], in which estimates take the form $\hat{\Sigma} = V \Delta V'$ Here, Δ = Δ (Λ) is *any* diagonal matrix that depends on the empirical eigenvalues ∧ in possibly a more complex way than the simple scalar element-wise shrinkage η(∧) we have considered so far. 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 [2, 19, 26] 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

$$\Delta^{oracle} = \mathrm{argmin}_{\Delta} L(\Sigma, V \Delta V'), \quad (8.1)$$

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, precisely the object that we aim to recover. This knowledge is never available to us in practice – hence the label *oracle*[3]. Nevertheless, this optimal performance is a legitimate benchmark.

Interestingly, at least for the popular Frobenius and Stein losses, our optimal nonlinearities η^* deliver oracle-level performance – asymptotically. To state the result, recall expression (6.2) for these losses: $F(\ell, \Delta) = L_2(A(\ell), B(\ell, \Delta))$.

Theorem 2. *(***Asymptotic 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($\ell_1,\dots,\ell_r$)]*. We have*

$$\lim_{n\to\infty} L_{p_n}\left(\Sigma, \hat{\Sigma}^{oracle}\right) =_P L_\infty\left(\ell_1\dots,\ell_r \middle| \eta^*\right) = \sum_{i=1}^{r} F\left(\ell_i, \eta^*\right),$$

where η^* *is the optimal shrinker for the losses* $L^{F,1}$ *or* L^{st} *in Table 2.*

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, in finite-n, 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 of the 26 loss functions considered.

Proof. We first outline the approach. We can write Σ and Σ^{-1} in the form $I + F$, and $\hat{\Sigma}_\Delta = I + \tilde{\Delta}$ with

$$F = \sum_{k=1}^{r} \beta_k u_k u_k', \quad \tilde{\Delta} = \sum_{i=1}^{p} \tilde{\Delta}_i v_i v_i'$$

where $\beta_k = \ell_k - 1$ for $L^{F,1}$ and $\ell_k^{-1} - 1$ for L^{St} and $\tilde{\Delta}_i = \Delta_i - 1$. Write

$$\operatorname{tr} F\tilde{\Delta} = \sum_{i=1}^{p} \tilde{\Delta}_i b_i, \quad b_i := \sum_{k=1}^{r} \beta_k \left(u_k' v_i\right)^2. \quad (8.2)$$

For both $L = L^{F,1}$ and L^{St}, we establish a decomposition

[3]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.

$$L_p\left(\Sigma,\ \widehat{\Sigma}_\Delta\right) = \sum_{i=1}^{r} F\left(\ell_i,\ \Delta_i\right) + a\left(\Delta_i - 1\right)\epsilon_i + \sum_{i=r+1}^{p} H\left(b_i,\ \Delta_i\right). \quad (8.3)$$

Here, a is a constant depending only on the loss function,

$$\epsilon_i = b_i - \beta_i c\left(\ell_i\right)^2, \quad (8.4)$$

and

$$H(b,\ \Delta) = \begin{cases} (\Delta - 1)^2 - 2(\Delta - 1)b & \text{for } L^{F,\,1} \\ (\Delta - 1)(1 + b) - \log \Delta & \text{for } L^{St}. \end{cases} \quad (8.5)$$

Decomposition (8.3) shows that the oracle estimator (8.1) may be found term by term, using just univariate minimization over each Δ_i. Consider the first sum in (8.3), and let $\widetilde{F}\left(\ell_i,\ \Delta_i\right)$ denote the summand. We will show that

$$\min_{\Delta_i} \widetilde{F}\left(\ell_i,\ \Delta_i\right) \xrightarrow{P} \min_{\Delta_i} F\left(\ell_i,\ \Delta_i\right), \quad (8.6)$$

and that

$$\sum_{i=r+1}^{p} \min_{\Delta_i} H\left(b_i,\ \Delta_i\right) = O_P\left(\frac{\log^2 p}{p}\right). \quad (8.7)$$

Together (8.6) and (8.7) establish the Theorem.

Turning to the details, we begin by showing (8.3). For Frobenius loss, we have from our definitions and (8.2) that

$$\left\| \widehat{\Sigma}_\Delta - \Sigma \right\|_F^2 = \mathrm{tr}\left(\widetilde{\Delta} - F\right)\left(\widetilde{\Delta} - F\right)' = \sum_{i=1}^{p} \left(\Delta_i - 1\right)^2 - 2\left(\Delta_i - 1\right)b_i + \sum_{i=1}^{r} \left(\ell_i - 1\right)^2.$$

For $i \geq r + 1$, each summand in the first sum equals $H(b_i, \Delta_i)$ and for $i \leq r$, we use the decomposition $b_i = (\ell_i - 1)c(\ell_i)^2 + \epsilon_i$. We obtain decomposition (8.3) with $a = -2$ and

$$F(\ell,\ \Delta) = (\ell - 1)^2 - 2(\ell - 1)(\Delta - 1)c^2 + (\Delta - 1)^2.$$

For Stein's loss, our definitions yield

$$L^{St}\left(\Sigma,\ \hat{\Sigma}_{\Delta}\right) = \mathrm{tr}\tilde{\Delta} + \mathrm{tr}F + \mathrm{tr}F\tilde{\Delta} - \log\left(\left|\hat{\Sigma}_{\Delta}\right|/|\Sigma|\right)$$
$$= \sum_{i=1}^{p} \tilde{\Delta}_i\left(1+b_i\right) - \log \Delta_i + \sum_{k=1}^{r} \beta_k + \log \ell_k .$$

Again, for each $i \geq r+1$, each summand in the first sum equals $H(b_i, \Delta_i)$ and with $b_i = (\ell_i - 1)c(\ell_i)^2 + \epsilon_i$ we obtain (8.3) with $a = 1$ and

$$F(\ell,\ \Delta) = \left(\ell^{-1} - 1\right) + (\ \Delta\ - 1)\left(c^2/\ell + s^2\right) - \log(\ \Delta\ /\ell) .$$

It remains to verify (8.6) and (8.7). Theorem 1 says that for $1 \leq i \leq r$,

$$\epsilon_i = \sum_{k=1}^{r} \beta_k\left[\left(u_k' v_i\right)^2 - \delta_{k,\ i}c\left(\ell_i\right)^2\right] \xrightarrow{P} 0,$$

which yields (8.6). From (8.5), we observe that in our two cases

$$h(b) := \min_{\Delta} H(b,\ \ \Delta) = \begin{cases} -b^2 \\ -b + \log(1+b) \end{cases} = O\left(b^2\right), \quad (8.8)$$

Now, using (8.2) and (7.22), we get

$$\max_{r+1 \leq i \leq p} \left|b_i\right| \leq r \max_{1 \leq k \leq r} |\beta_k| \cdot \max_{i > r, k \leq r} \left(u_k' v_i\right)^2 = O_P\left(\frac{\log p}{p}\right).$$

From the previous two displays, we conclude

$$\sum_{i=r+1}^{p} \min_{\Delta_i} H\left(b_i,\ \ \Delta_i\right) = \sum_{i=r+1}^{p} h\left(b_i\right) = O_P\left(\frac{\log^2 p}{p}\right).$$

which is (8.7), and so completes the full proof.

9 Optimal Shrinkage with common variance $\sigma^2 \neq 1$

Simply put, the Spiked Covariance Model is a proportional growth independent-variable Gaussian model, where all variables, except the first r, have common variance σ. Literature on the spiked model often simplifies the situation by assuming $\sigma^2 = 1$, as we have done in our assumption [Spike($\ell_1,\ldots,\ \ell_r$)] above. To consider optimal shrinkage in the case of general common variance $\sigma^2 > 0$, assumption [Spike($\ell_1,\ldots,\ \ell_r$)] has to be replaced by

[Spike($\ell_1,\ldots,\ \ell_r|\sigma^2$)] The population eigenvalues in the n-th problem, namely the eigenvalues of Σ_{p_n}, are given by ($\ell_1,\ldots,\ \ell_r,\ \sigma^2,\ldots,\ \sigma^2$), where the number of “spikes” r and their amplitudes $\ell_1 > \ldots > \ell_r \geq 1$ are fixed independently of n and p_n.

In this section we show how to use an optimal shrinker, designed for the spiked model with common variance $\sigma^2 = 1$, in order to construct an optimal shrinker for a general common variance σ^2, namely, under assumptions [Asy(γ)] and [Spike($\ell_1,\ldots,\ell_r|\sigma^2$)].

9.1 σ^2 known

Let Σ_p and $S_{n,p}$ be population and sample covariance matrices, respectively, under assumption [Spike($\ell_1,\ldots,\ell_r|\sigma^2$)]. When the value of σ is known, the matrices $\tilde{\Sigma}_p = \Sigma_p/\sigma^2$ and the sample covariance matrix $\tilde{S}_{n,p} = S_{n,p}/\sigma^2$ constitute population and sample covariance matrices, respectively, under assumption [Spike($\ell_1,\ldots,\ell_r$)]. Let L be any of the loss families considered above and let η be a shrinker. Define the shrinker $\tilde{\eta}$ corresponding to η by

$$\tilde{\eta} : \lambda \mapsto \sigma^2 \cdot \eta\left(\lambda/\sigma^2\right). \quad (9.1)$$

Observe that for each of the loss families we consider, $L_p(\sigma^2 A, \sigma^2 B) = \sigma^{2\kappa} L_p(A, B)$, where $\kappa \in \{-2, -1, 0, 1, 2\}$ depends on the family $\{L_p\}$ alone. Hence

$$L_p\left(\Sigma_p, \hat{\Sigma}_{\tilde{\eta}}\left(S_{n,p}\right)\right) = \sigma^{2\kappa} L_p\left(\tilde{\Sigma}_p, \hat{\Sigma}_{\tilde{\eta}}\left(\tilde{S}_{n,p}\right)\right)$$

It follows that if η^* is the optimal shrinker for the loss family L, in the sense of Definition 6, under Assumption [Spike($\ell_1,\ldots,\ell_r$)] , then $\tilde{\eta}^*$ is the optimal shrinker for L under Assumption [Spike($\ell_1,\ldots,\ell_r|\sigma^2$)]. Formula (9.1) therefore allows us to translate each of the optimal shrinkers given above to a corresponding optimal shrinker in the case of a general common variance $\sigma^2 > 0$.

9.2 σ^2 unknown

In practice, even if one is willing to assume a common variance σ^2 and subscribe to the spiked model, the value of σ^2 is usually unknown. Assume however that we have a sequence of estimators $\left\{\hat{\sigma}_n\right\}_{n=1,2\ldots,}$ where for each n, $\hat{\sigma}_n$ is a real function of a p_n-by-p_n positive definite symmetric matrix argument. Assume further that under the spiked model with general common variance σ^2, namely under assumptions [Asy(γ)] and [Spike($\ell_1,\ldots,\ell_r|\sigma^2$)], the sequence of estimators is consistent in the sense that $\hat{\sigma}_n\left(S_{n,p_n}\right) \to \sigma$, almost surely. For a continuous shrinker η, define a sequence of shrinkers $\left\{\tilde{\eta}_n\right\}_{n=1,2,\ldots}$ by

$$\tilde{\eta}_n : \lambda \mapsto \hat{\sigma}_n^2 \cdot \eta\left(\lambda/\hat{\sigma}_n^2\right). \quad (9.2)$$

Again for each of the loss families we consider, almost surely,

$$\lim_{n\to\infty} L_{p_n}\left(\Sigma_{p_n}, \hat{\Sigma}_{\tilde{\eta}_n}\left(S_{n,p_n}\right)\right) = \sigma^{2\kappa} \lim_{n\to\infty} L_{p_n}\left(\tilde{\Sigma}_{p_n}, \hat{\Sigma}_{\eta}\left(\tilde{S}_{n,p_n}\right)\right).$$

We conclude that, using (9.2), any consistent sequence of estimators $\hat{\sigma}_n$ yields a sequence of shrinkers with the same asymptotic loss as the optimal shrinker for known σ^2. In other words, at least inasmuch as the asymptotic loss is concerned, under the spiked model, there is no penalty for not knowing σ^2.

Estimation of σ^2 under Assumption [Spike($\ell_1,\dots,\ell_r|\sigma^2$)] has been considered in [65, 66, 31] where several approaches have been proposed. As an simple example of a consistent sequence of estimators $\hat{\sigma}_n$, we consider the following simple and robust approach based on matching of medians [32]. The underlying idea is that for a given value of n the sample eignevalues $\lambda_{r+1},\dots,\lambda_{pn}$ form an approximate Marčenko-Paster bulk inflated by σ^2, and that a median sample eigenvalue is well suited to detect this inflation as it is unaffected by the sample spikes $\lambda_1,\dots,\lambda_r$.

Define, for a symmetric p-by-p positive definite matrix S with eigenvalues $\lambda_1,\dots,\lambda_p$ the quantity

$$\mu(S) = \frac{\lambda_{med}}{\mu_\gamma}, \quad (9.3)$$

where λ_{med} is a median of $\lambda_1,\dots,\lambda_p$ and μ_γ is the median of the Marčenko-Pastur distribution, namely, the unique solution in $\lambda_-(\gamma) \le x \le \lambda_+(\gamma)$ to the equation

$$\int_{\lambda_-(\gamma)}^{x} \frac{\sqrt{\left(\lambda_+(\gamma) - t\right)\left(t - \lambda_-(\gamma)\right)}}{2\pi\gamma t}\, dt = \frac{1}{2},$$

where as before $\lambda_\pm(\gamma) = (1 \pm \sqrt{\gamma})^2$. Note that the median μ_γ is not available analytically but can easily be obtained numerically, for example using remarks on the Marčenko-Pastur cumulative distribution function included in SI. Now for a sequence $\{S_{n,pn}\}$ of sample covariance matrices,define the sequence of estimators

$$\hat{\sigma}_n : \left(S_{n,\,p_n}\right) \mapsto \sqrt{\mu\left(S_{n,\,p_n}\right)}. \quad (9.4)$$

Lemma 9. *Let* $\sigma^2 > 0$*, and assume* [Asy(γ)] *and* [Spike($\ell_1, \dots, \ell_r|\sigma^2$)]. *Then almost surely*

$$\lim_{n \to \infty} \hat{\sigma}_n\left(S_{n,\,p_n}\right) = \sigma.$$

In summary, using (9.1) (for σ^2 known) or (9.2) with (9.4) (for σ^2 unknown) one can use the optimal shrinkers for each of the loss families discussed above, designed for the case $\sigma = 1$, to construct a shrinker that is optimal, for the same loss family, under the spiked model with common variance $\sigma^2 \neq 1$.

10 Discussion

In this paper, we considered covariance estimation in high dimensions, where the dimension p is comparable to the number of observations n. We chose 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 whose rank 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 [32, 31, 53] with the “fixed fraction” setup of [67].)

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 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, a complex nonlinear relation exists 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 [33, 68].

Covariance shrinkage in the proportional rank model should then, naturally, make use of the so-called *MP Equation*. Noureddine El Karoui [24] 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 [25] developed analytic tools to also account for the inaccuracy of empirical eigenvectors, and Ledoit and Michael Wolf [26] have implemented such tools and applied them in this setting.

The proportional rank case is indeed 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, the covariance estimation ideas of [24, 26], 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 may be more likely to be applied correctly in practice, and to work as expected, even in relatively small sample sizes.

An analogy can be made to shrinkage in the normal means problem, for example [69]. In that problem, often a full Bayesian model applies, and in principle a Bayesian shrinkage would provide an optimal result [70]. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

We thank Amit Singer, Andrea Montanari, Sourav Chatterjee and Boaz Nadler for helpful discussions. We also thank the anonymous referees for significantly improving the manuscript through their helpful comments. 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.

References

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

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

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

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

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

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

[7]. Berger James. 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]. Haff LR. Estimation of the inverse covariance matrix: Random mixtures of the inverse wishart matrix and the identity. The Annals of Statistics, pages 1264–1276, 1979.

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

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

[11]. Sinha BK and Ghosh M. 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.

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

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

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

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

[16]. Pal N. 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.

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

[18]. Gupta AK 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.

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

[20]. Daniels Michael J and Kass Robert E. Shrinkage estimators for covariance matrices. Biometrics, 57(4):1173–1184, 2001. [PubMed: 11764258]

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

[22]. Sun D and Sun X. 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 cited By (since 1996)7.

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

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

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

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

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

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

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

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

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

[32]. Gavish M and Donoho DL. The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$. IEEE Transactions on Information Theory, 60(8):5040–5053, 2014.

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

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

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

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

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

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

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

[40]. Donoho David L., Gavish Matan, and Johnstone Iain M.. Supplementary Material for "Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model". http://purl.stanford.edu/xy031gt1574, 2016.

[41]. Donoho David L., Gavish Matan, and Johnstone Iain M.. Code supplement to "Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model". http://purl.stanford.edu/xy031gt1574, 2016.

[42]. Geman Stuart. A limit theorem for the norm of random matrices. Annals of Probability, 8:252–261, 1980.

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

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

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

[46]. Olkin Ingram and Pukelsheim Friedrich. The distance between two randomvectorswith given dispersion matrices Linear Algebra and its Applications, 48:257–263, 1982.

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

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

[49]. Lenglet Christophe, Rousson Mikaël, Deriche Rachid, and Faugeras Olivier. 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.

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

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

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

[53]. Gavish M and Donoho DL. Optimal Shrinkage of Singular Values. ArXiv 1405.7511, 2016.

[54]. van der Vaart HR. 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.

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

[56]. James AT. Normal multivariate analysis and the orthogonal group. Annals of Mathematical Statistics, 25(1):40–75, 1954.

[57]. Bhatia Rajendra. Matrix analysis, volume 169 of Graduate Texts in Mathematics. Springer-Verlag, New York, 1997.

[58]. Johnstone Iain M.. Tail sums of wishart and gue eigenvalues beyond the bulk edge. Australian and New Zealand Journal of Statistics, 2017 to appear.

[59]. Tracy CA and Widom H. On orthogonal and symplectic matrix ensembles. Comm Math Phys, 177:727–754, 1996.

[60]. Benaych-Georges F, Guionnet A, and Maida M. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. Electron. J. Probab, 16:no. 60, 1621–1662, 2011.

[61]. Ledoux Michel. The concentration of measure phenomenon. Number 89. American Mathematical Society, 2001.

[62]. NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.9 of 2014–08-29. Online companion to [63].

[63]. Olver FWJ, Lozier DW, Boisvert RF, and Clark CW, editors. NIST Handbook of Mathematical Functions. Cambridge University Press, New York, NY, 2010 Print companion to [62].

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

[65]. Kritchman Shira and Nadler Boaz. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. Signal Processing, IEEE Transactions, 57(10): 3930– 3941, 2009.

[66]. Passemier Damien and Yao Jian-feng. Variance estimation and goodness-of-fit test in a high-dimensional strict factor model. arXiv:1308.3890, 2013.

[67]. Donoho DL and Gavish M. Minimax Risk of Matrix Denoising by Singular Value Thresholding. ArXiv e-prints, 2013.

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

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

Author Manuscript

Author Manuscript

Author Manuscript

[70]. Brown Lawrence D. and Greenshtein Eitan. 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.

Author Manuscript

Author Manuscript

Author Manuscript

Reproducible Research

In the code supplement [41] 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 test the correctness of each of the 18 analytic shrinker fomulas provided.
3. Scripts that generate each of the figures in this paper, or subsets of them for specified loss functions.

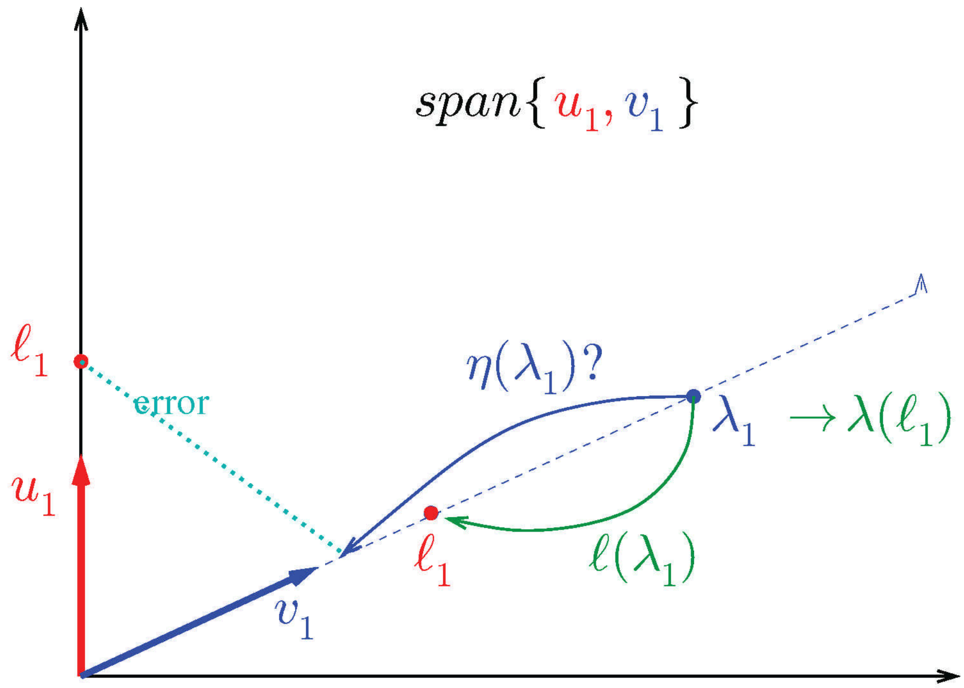

Figure 1:
Shrinking empirical eigenvalue λ_1 to a value $\eta(\lambda_1)$ that is smaller than the inverse function $\ell(\lambda_1)$ may reduce the error of estimation.

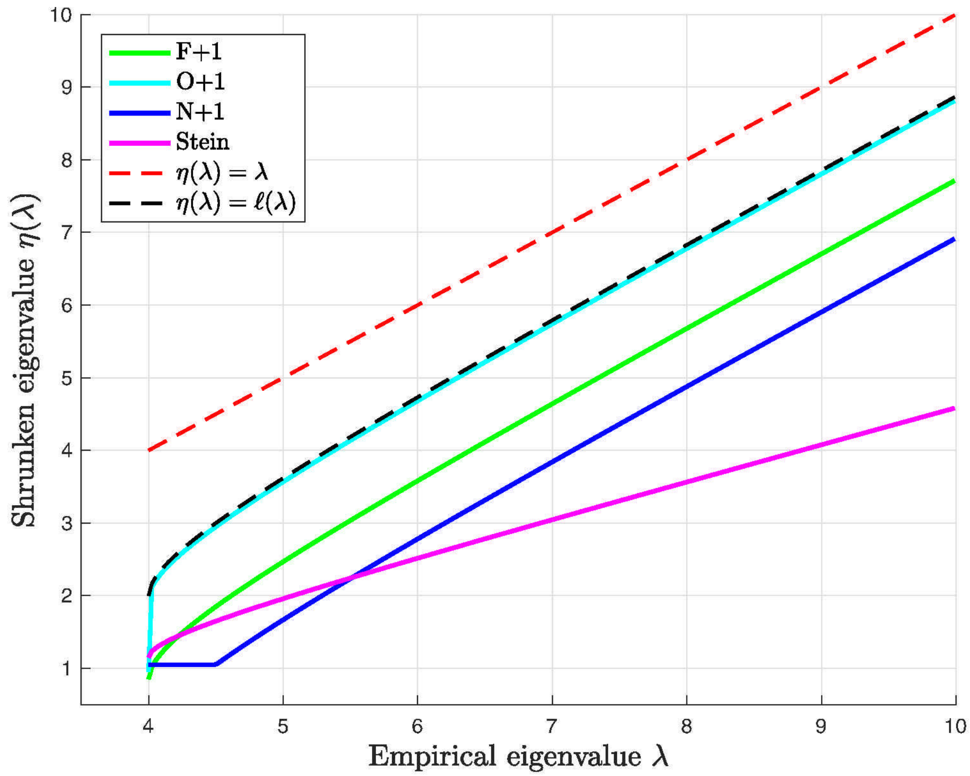

Figure 2:
Vertical axis: optimal shrinkers η_* from (1.12), shown as functions $\eta_*(\ell(\lambda))$ of the empirical eigenvalue λ, horizontal axis. Here $\gamma = \lim p_n/n = 1$, so $\lambda_+(\gamma) = 4$. (Color online.)

Author Manuscript
Author Manuscript
Author Manuscript

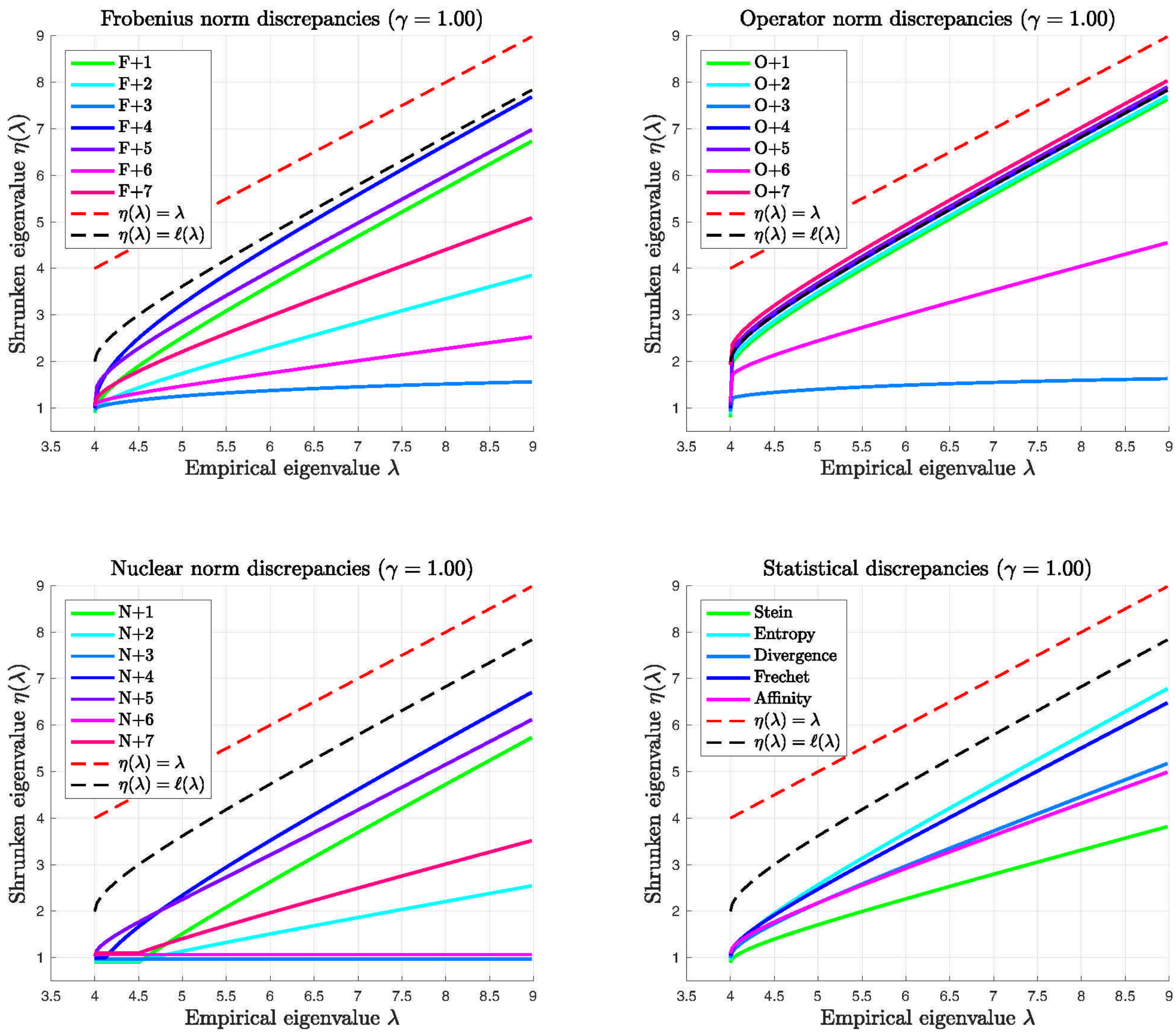

Figure 3:

Optimal Shrinkers for 26 Component Loss Functions for $\gamma = 1$ and $4 \leq \lambda \leq 10$. Upper Left:Frobenius-norm-based losses; Lower Left: Nuclear-Norm based losses; Upper Right: Operator-norm-based losses; Lower Right: Statistical Discrepancies. (Color online; curves jittered in vertical axis to avoid overlap.) The supplemental article [40] contains an larger version of these plots. *Reproducibility advisory:* The code supplement [41] includes a script that reproduces any one of these individual curves.

Table 1:

Systematic notation for the 26 loss functions considered in this paper.

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

Table 2:

Optimal shrinkers $\eta^*(\lambda)$ for 18 of the loss families L discussed. Values shown are shrinkers for $\lambda > \lambda_+(\gamma)$. All shrinkers obey $\eta^*(\lambda) = 1$ for $\lambda \le \lambda_+(\gamma)$. Here, ℓ c and s depend on λ (and implicitly on γ) according to (1.10), (1.6) and $s = \sqrt{1 - c^2}$. In cases marked "N/A" the optimal shrinker does not seem to admit a simple closed form, but can be easily calculated numerically.

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