James–Stein estimation of the first principal component

Alex Shkolnik

Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, CA, 93106, USA

Correspondence
Alex Shkolnik, Department of Statistics and Applied Probability, University of California, Santa Barbara, 5501 South Hall (Building 528), Santa Barbara, CA 93106, USA.
Email: shkolnik@ucsb.edu

1 | INTRODUCTION

The Stein paradox has played an influential role in the field of high-dimensional statistics. This result warns that the sample mean, classically regarded as the “usual estimator,” may be suboptimal in high dimensions. In particular, Stein (1956) showed that the usual estimator of a location parameter $\theta \in \mathbb{R}^p$ from uncorrelated Gaussian observations becomes inadmissible when $p > 2$ under a mean-squared error criterion. That is, an estimator with a uniformly lower risk must exist. That estimator was established by James and Stein (1961) and eponymously named.

Among the numerous perspectives that motivate the James–Stein estimator,1 the empirical Bayes perspective (see Efron & Morris, 1975) is particularly elegant. Letting $\eta$ denote the sample mean based on $n$ measurements of an unknown $\theta \in \mathbb{R}$ and assuming an additive, normally distributed error $w$ that has zero mean and variance $\nu^2$ (e.g., $\nu = \delta/\sqrt{n}$ where each measurement has standard error $\delta$), we write

$$\eta = \theta + w.$$  (1)

Taking a Gaussian prior on the unknown $\theta$, that is independent of $w$, implies that

$$E(\theta|\eta) = E(\eta) + \left(1 - \frac{\nu^2}{\text{Var}(\eta)}\right) (\eta - E(\eta)).$$  (2)

the bivariate-normal conditional expectation formula. While, by the definition of conditional expectation, $E(\theta|\eta)$ is the best estimator of $\theta$ in the sense of mean-squared error, it cannot be implemented directly as the first two moments of $\eta$ are unknown.2 Stein’s paradox now amounts to the fact that “good” substitutes for $E(\eta)$ and $\text{Var}(\eta)$ are available only in higher dimensions (precisely, if and only if $\theta \in \mathbb{R}^p$ and $p > 2$).

Abbreviations: PCA, principal component analysis; HDLS, high-dimension and low-sample.

1A few examples include the Galtonian regression perspective promoted by Stigler (1990), the purely frequentist development of the estimator in Gupta and Peña (1991) and the geometrical explanation in Brown and Zhao (2012) that builds on Stein’s original heuristic argument (Stein, 1956, section 1).

2Often, $\nu$ is assumed to be known, but estimates $\hat{\nu}$ can also be used as done in James and Stein (1961).
Formula (2) extends easily to the multivariate\(^3\) case and motivates the estimator

\[ \eta(c) = m + c(\eta - m), \]

(3)

where \( m \) is an estimate (or guess) of the expected value of \( \eta \in \mathbb{R}^p \) and \( c \in (0, 1) \) is a shrinkage parameter. In words, (3) attempts to centre the entries of \( \eta \), shrinks the resulting entries and recentres at \( m \). Assuming \( \nu \) is known and \( p > 2 \), setting

\[ c = 1 - \frac{\nu^2}{s^2(\eta)} \left( \frac{p - 2}{p} \right), \]

(4)

where \( s^2(\eta) = \sum_{i=1}^p (\eta_i - m_i)^2 / p \) yields the James–Stein estimator. Remarkably, any fixed \( m \in \mathbb{R}^p \) (e.g., Stein, 1956) considers the origin\(^4\) results in an estimator (3) with a strictly smaller mean-squared error than \( \eta \) (Efron & Morris, 1975, section 1). While three is provably the critical dimension, Stein (1956) heuristically argued that normality is not essential for higher performance in higher dimensions.

The development of the James–Stein estimator has inspired a large literature on the theme of “shrinkage” in statistics. Just a small sampling of examples includes ridge regression (Hoerl & Kennard, 1970), the LASSO (Tibshirani, 1996), the Ledoit–Wolf covariance estimator (Ledoit & Wolf, 2004) and the Elastic Net (Zou & Hastie, 2005). Excellent textbook treatments of the ideas behind the Stein paradox and James–Stein shrinkage include Gruber (2017) and Fourdrinier et al. (2018). In this paper, we leverage these ideas to develop and analyse a James–Stein estimator for the first principal component of a sample covariance matrix. The results once again prove the efficacy of James–Stein estimation and do so for one of the cornerstone methods within high-dimensional statistics, principal component analysis (PCA).

Consider a \( p \times p \) sample covariance matrix \( S \) that is based on \( n \) observations of some random vector \( y \in \mathbb{R}^p \). Without loss of generality, we write

\[ S = s_p^2 hh^\top + G \]

for \( G = S - s_p^2 hh^\top \) and \( h \), the sample eigenvector with the largest eigenvalue, that is,

\[ Sh = s_p^2 h \quad \text{and} \quad s_p^2 = \max_{i=1}^n (x_i S x). \]

(6)

By convention, \( h \) has unit length and corresponds to a direction along which the variance of \( S \) (i.e., \( s_p^2 \)) is maximum (i.e., the first principal component). A substantial and rapidly growing literature exists to study the (and/or \( n \)) asymptotic behaviour of the eigenpair \((s_p^2, h)\) and the remaining eigenstructure in order to quantify either the estimation or the empirical error. See Wang and Fan (2017) for recent results and a systematic discussion of this literature. The topic of shrinkage estimators arises naturally in this context and was raised by Stein (1986), who suggested improving the usual estimate \( S \) via eigenvalue shrinkage. Indeed, there is by now a large literature on estimators that adjust the eigenvalues of sample covariance matrices to improve their performance with respect to some loss function (Donoho et al. 2018).

In this paper, we develop and analyse a James–Stein estimator for the first principal component of a high-dimensional and low-sample (HDLS) data set.\(^6\) The recipe for the estimator begins with \( h \) and \( s_p^2 \) in (6) and the next \((q - 1)\) largest sample eigenvalues \( s_p^2, \ldots, s_p^2\) \((\min(n, p) > q)\) corresponding to a model with \( q \) spikes.\(^7\)

Step 1. Set \( \eta = s_p h \), compute the sample statistics \( m(q) = \sum_{i=1}^p \eta_i / p \) and \( s^2(\eta) = \sum_{i=1}^p (\eta_i - m(q))^2 / p \) and define

\[ c = 1 - \frac{\nu^2}{s^2(\eta)} \quad \text{where} \quad \nu^2 = \left( \frac{\text{tr}(S) - (s_p^2 + \ldots + s_{p-q+1}^2)}{\min(n, p) - q} \right) / p. \]

Step 2. Return the estimator (corrected principal component)\(^8\)

\(^3\)As in Stein (1956), \( \eta \) is distributed as multivariate normal with mean \( \theta \) and covariance \( \Sigma + 1 \).

\(^4\)A natural choice for \( m \) takes the sample mean of \( \eta \) in each entry, referred to as shrinkage towards the “grand mean” (but would require \( p > 3 \) Efron & Morris, 1975).

\(^5\)The theme of a critical dimension is encountered frequently in statistics and probability. Brown (1971), for example, derives a close mathematical relationship between the admissibility of the James–Stein estimator and the transience of the Brownian motion in \( \mathbb{R}^p \), which also requires \( p > 2 \).

\(^6\)The HDLS framework was introduced in Hall et al. (2005) and Ahn et al. (2007) and is becoming increasingly relevant for data science applications.

\(^7\)Roughly speaking, \( q \) is the number of factors (or spikes) in the data, after which a sufficiently large eigengap (between the \( q \)th and the next eigenvalue) is observed (see Fan et al. 2020).

\(^8\)With a slight abuse of the notation, \( u - x = (u_1 - x_1, \ldots, u_p - x_p) \) for \( u \in \mathbb{R}^p \) and \( x \in \mathbb{R} \).
The vector $h^{JS}$ is the James–Stein estimator of the first principal component of the data. The numerator contains the shrinkage formula (3) while the divisor normalizes the shrunk vector to a unit length (by convention). The relationship to the shrinkage parameter $c$ in (4) is evident by treating $p$ as large. The estimate $\hat{c}^2$ corresponds to the bulk of the eigenvalue spectrum and may be viewed the “noise” in the context of a signal-to-noise ratio (SNR) that plays a prominent role of the results in Sections 2 and 3.

It is reasonable to suspect that a James–Stein-type shrinkage of the principal component $h$, a high-dimensional vector, could improve either the convergence rate or accuracy of the limit in some appropriate asymptotic regime. However, the standard orthonormal transformation and the eigengap partition of the sample eigenvectors, which is typically leveraged by their asymptotic analyses (e.g., Paul, 2007; Shen et al. 2016; Wang & Fan, 2017), can obscure the systematic nature of the sample bias. As sensibly pointed out by Wang and Fan (2017) in reference to the partition of the sample eigenvector, the “two parts intertwine in such a way that correction for the biases of estimating eigenvectors is almost impossible.” However, in the original (untransformed) coordinate system and the HDLS asymptotic regime, the bias can in fact be identified, characterized and (partially) corrected. This program was carried out in Goldberg et al. (2021), who adopt a factor model in an HDLS regime and utilize a portfolio theory application to motivate their analysis.

The main results of this paper rederive the adjustment of Goldberg et al. (2021) but within a James–Stein-type framework. In particular, we establish identity (1) in which $\eta = s_p h$ and $\theta$ is related to the associated population eigenvector. From here, the James–Stein shrinkage acts on the perturbation $w$ so that the estimator $h^{JS}$ outperforms $h$ on the mean-squared error and angle metrics as $p \to \infty$. The theoretical guarantees provided here are new and their proofs rely on a different set of mathematical tools than Goldberg et al. (2021). In particular, the new approach leverages Weyl’s inequality and the Davis–Kahan theorem from matrix perturbation theory to give simpler proofs and potentially expand the scope of applicability of the resulting estimator. The HDLS regime, in which the number of variables $p$ grows to infinity, and the number of observations $n$ to be fixed, plays a crucial role in the analysis.

The paper is organized as follows. Section 2 defines the spiked covariance model underlying our results. Section 3 develops the James–Stein estimator $h^{JS}$, and Section 4 proves the theoretical guarantees for this estimator. Appendix A contains proofs of the auxiliary results. The following notation is used throughout. Let $\langle u, v \rangle$ denote the standard inner product of $u, v \in \mathbb{R}^d$ so that $|u| = \sqrt{\langle u, u \rangle}$ and $m(u) = \langle u, e \rangle/d$ where $e = (1, \ldots, 1)^\top$ are the length and mean. Set $s^2(u) = |u - m(u)|^2/d$ and $\text{cov}(u, v) = \langle u - m(u), v - m(v) \rangle/d$ (see the notation of footnote 8). We use a subscript $1 \leq p \leq \infty$ to highlight the dependence on $p$ of various quantities, for example, $m(\eta) = m_p(\eta)$ for $\eta \in \mathbb{R}^p$ and $m_{\infty}(\eta)$ is the limit $\lim_{p \to \infty} m_p(\eta)$ when it exists.

## 2 A SCARCELY SAMPLED SPIKED MODEL

We use a spiked covariance model borrowed from the HDLS literature. We also restrict ourselves to a single unbounded spike in the “boundary case” (see Jung et al. 2012) wherein the largest eigenvalue of the covariance matrix grows linearly in $p$. In particular, consider a mean-zero (w.l.o.g.) random vector $y \in \mathbb{R}^p$ with a $p \times p$ covariance matrix $\Sigma = \text{Var}(y)$ and let

\[
\Sigma = \Gamma + \beta \beta^\top, \tag{7}
\]

for a symmetric and positive-semidefinite $p \times p$ matrix $\Gamma$ and vector $\beta \in \mathbb{R}^p$. The following affirms that $\beta = \beta/|\beta|$ is an eigenvector of $\Sigma$ with eigenvalue $\langle \beta, \beta \rangle$.

**Assumption 1** w.l.o.g., $\Gamma b = 0$ and $m_p(b) \geq 0$ for any $p$.

To state our additional assumptions on the model, we project the data vector $y \in \mathbb{R}^p$ onto the eigenvectors of $\Sigma$. More precisely, define

\[
\psi_p = \langle \hat{\phi}; y \rangle / \langle \hat{\phi}, \hat{\phi} \rangle. \tag{8}
\]

It is immediate that $E(\psi_p) = 0$ and $\text{Var}(\psi_p) = 1$. For every eigenvalue $\alpha_i$ of $\Gamma$, let

\[
\phi_i = \langle \gamma; y \rangle, \quad \text{where} \quad \gamma \in \mathbb{R}^p : \Gamma \gamma = \alpha_i \gamma. \tag{9}
\]

Note that $E(\phi_i) = 0$ and $\text{Var}(\phi_i) = \alpha_i$. As the dimension $p$ grows, we obtain a sequence $\{\phi_i\}_{i=1}^\infty$. As a technical remark, $\phi_i = \phi_{ip}$ and $\alpha_i = \alpha_{ip}$ depend on $p$. We consider a sequence of models (7) constructed from sequences $\{\beta_i\}_{i=1}^\infty$ and $\{\Gamma_p\}$.
Assumption 2. For constants μ ∈ ℝ and σ, δ ∈ (0, ∞) as p → ∞, we have

(i) m_μ (β) = μ and s^2_δ (β) = δ^2.
(ii) ψ_∞ = lim_{p→∞} ψ_p exists as a ℝ-valued random variable almost surely.
(iii) m_ω (φ) = m_ω (ψ) = 0 almost surely for |ψ_p| ≤ 1 with φ_p = ϕ_p |ψ_p|
(iv) s^2_δ (φ) = m_ω (α) = δ^2 almost surely.

Condition (i) imposes regularity on the sequence {ψ_p} and implies that the largest eigenvalue of Σ (i.e., ⟨β, β⟩) grows linearly with p. The random variable ψ_∞ in (ii) is closely related to a principal component score (in the limit p → ∞), and it captures the randomness along the first (population) principal component. Conditions (iii) and (iv) are related to certain requirements on a measure of sphericity of the model (e.g., sphericity). In particular, (iii) may be viewed as laws of large numbers for {ψ_p} and summarizes the conclusions of Jung and Marron (2009). In particular, (iii) as a measure of sphericity where the spike eigenvalue (β, β) is bounded in p, and per (iii), the eigenvectors γ^j do not have entries biased towards a non-zero mean, unlike b (an exception is μ = 0, which is the case when the James–Stein estimator will turn out to be ineffective; see Remark 2).

The following assumption is a standard one in statistical data analysis but may be relaxed in the HDLS set-up (e.g., Jung et al. 2012).

Assumption 3. There are a fixed n ≥ 2 i.i.d. observations of y ∈ ℝ^p.

Our forthcoming results hold even when only two observations are available, hence, a scarcely sampled model. Let Y be the p × n data matrix with the kth column containing the kth observation of y, and define the sample covariance matrix S by

S = YY^T /n. \hfill (10)

We let h ∈ ℝ^p denote the eigenvector of S with the largest eigenvalue s^2_1 (see 6). It is unique only up to sign (and |h| = 1), motivating the following (c.f. Assumption 1).

Assumption 4. w.l.o.g., m_p (h) ≥ 0 for any p.

We write S = G + ηγ γ^T in analogy to (7) (taking G = S − ηγ γ^T) and set

η = s_p h. \hfill (11)

Next, define the following measure of finite-sample distortion,

\( \chi^2_n = \langle \chi, \chi \rangle /n \quad \text{and} \quad \chi = \lim_{p \to \infty} Y^T \beta \)

The latter limit exists by Assumption 2 while Assumption 3 implies that X ∈ ℝ^p has i.i.d. entries (distributed as ψ_∞). Consequently, we have \( \chi^2_n \to 1 \) as n → ∞.

We can measure the error in any estimator η of θ by the mean-squared error, as would be consistent with the James–Stein framework.

MSE_p (η|θ) = ⟨η − θ, η − θ⟩ /n. \hfill (13)

Proposition 1. Let θ = ρ = 1 and suppose Assumptions 1–4 hold. Then,

MSE_p (η|θ) = \frac{δ^2}{n}. \hfill (14)

Remark 1. If θ = 0, then the right side would be multiplied by the factor, 1 + SNR \( \frac{2}{12}\left(\frac{2-1}{2}\right)^2 \), where we define SNR and signal-coherence r as
For SNR, we regard $\sigma \chi \sqrt{n}$ as a distorted signal and $\delta / \sqrt{n}$ as noise that vanishes as $n$ grows (also, $\sigma \chi \rightarrow 0$). The signal-incoherence $r_\infty$ is the limit of $r_p = r_p(\beta) = s_p(\beta) / |\beta|$ (per Assumption 2) determined by the SNR $\frac{\delta}{\sigma}$ of the vector $\beta$. A large value of $r_\infty$ corresponds to more variation, or “incoherence” in the entries $|\beta|_{i=1}$.

A more standard way to evaluate the goodness of a sample eigenvector $h$ is via its angle away from its population counterpart $b = \beta / |\beta|$. To this end, let

$$SPH_p(h;b) = SPH_p(\eta;\theta) = \sin^2 \left( \arccos \frac{\langle \eta, \theta \rangle}{|\eta| |\theta|} \right).$$

**Proposition 2.** Let $\theta = \chi \beta$ and suppose Assumptions 1–4 hold. Then,

$$SPH_\infty(\eta;\theta) = \frac{r_\infty^2}{r_\infty^2 + \text{SNR}^2}.$$  \hspace{1cm} (17)

**Remark 2.** Clearly, (17) also holds with $\theta = \beta$.

### 3 | James–Stein Estimation of Sample Eigenvectors

Having established that the sample eigenvector corresponding to the spike (i.e., the largest eigenvalue) carries finite-sample error, it is natural to ask whether James–Stein shrinkage can improve this “usual” estimator. The key to this question is the (to be established) identity

$$\eta = \theta + w \quad \text{and} \quad \theta = \chi \beta$$

for a random vector $w \in \mathbb{R}^p$ specified in (27) of Section 4.1. As in definition (11), we have $\eta = s_p h$ where $h$ is the sample eigenvector with the largest eigenvalue, $s_p^2$. The perturbation of $\theta$ turns out to be such that the shrinkage of $\eta$ is effective. We remark that the recipe of Section 1 extends our derivation of the James–Stein estimator below to the case of multiple (there, $q$) spikes in a natural way. This extension is effective because the eigenvectors corresponding to the spikes are mutually orthogonal, but it is suboptimal. An optimal estimator in the multipiked set-up is left for future work.

#### 3.1 | The JS estimator

Equation (18) establishes a relationship between the sample and population eigenvectors that suggests a James–Stein estimator may be derived. An informal derivation proceeds as follows. Consider the shrinkage parameter

$$c = 1 - \frac{\hat{\nu}^2}{\nu^2(\eta)}$$

based on (4) with $\hat{\nu}$, an estimate of the “noise.” It is reasonable to assign the latter to be the average of the non-spiked, non-zero eigenvalues of $S$. That is,

$$\hat{\nu}^2 = \left( \frac{\text{tr}(S) - s_p^2}{n - 1} \right) / p \quad (p \leq n),$$

where the scaling by $p$ turns out to be necessary due to the counterintuitive behaviour of the HDLS asymptotics. When $p < n$, the divisor $n - 1$ must be replaced by $p - 1$. 

$\text{SNR} = \left( \frac{\sigma}{\delta} \right) \chi \sqrt{n}$ and $r_\infty = \frac{1}{\sqrt{1 + (\mu/\sigma)^2}}$.  \hspace{1cm} (15)
This paves the way for the James–Stein sample eigenvector estimate, 

\[ \eta^{JS} = m(\eta) + c(\eta - m(\eta)) \]

of the unnormalized eigenvector and by convention, we take unit length version, 

\[ h^{JS} = \frac{\eta^{JS}}{|\eta^{JS}|} = \frac{1}{\sqrt{\beta}} \left( \frac{m(h) + c(h - m(h))}{\sqrt{m^2(h) + c^2s^2(h)}} \right), \]  

as the James–Stein estimator of the population eigenvector \( b = \beta/|\beta| \).

The following James–Stein-type theorems characterize the improvement due to shrinkage in the original mean-squared sense and in the angle metrics.

**Theorem 1.** Suppose Assumptions 1–4 hold. Then, almost surely, 

\[ \text{MSE}_w(\eta^{JS}|\theta) = c_w \text{MSE}_w(\eta|\theta), \]

where \( c_w \in (0, 1) \) is the limit of \( c_p = c \) in (19) with SNR defined in (15) and 

\[ c_w = \frac{\text{SNR}^2}{1 + \text{SNR}^2} \]  

**Theorem 2.** Suppose Assumptions 1–4 hold. Then, almost surely, 

\[ \text{SPH}_w(\eta^{JS}|\theta) = \text{SPH}_w(h^{JS}|b) = d_w \text{SPH}_w(h|b), \]

where \( d_w \in [c_w, 1] \) where \( c_w \) is in (22) and with SNR and \( r_m \) in (15), we have 

\[ d_w = c_w + \frac{r_m^2}{1 + \text{SNR}^2}. \]  

A related result is available in Goldberg et al. (2021), but the metrics there are motivated by the solutions of certain quadratic programs used in portfolio theory.

**Remark 3.** Note that \( d_w = 1 \) (i.e., no improvement in angle) if and only if \( \mu = 0 \).

### 3.2 | The geometry of Stein’s paradox

We shed insight into the James–Stein estimator in (21) by deriving general conditions under which Theorems 1 and 2 hold. Our analysis adopts the pure frequentist perspective in Gupta and Peña (1991) and supplements it by illustrating the Euclidean and the spherical geometry of the estimator. The two geometries reflect the definitions of the error (MSE and SPH) in the two theorems.

**Lemma 1.** Let \( \eta = \theta + w \) for \( \theta, w \in \mathbb{R}^p \). Then, the solutions of the optimizations \( \min_{c \in \mathbb{R}} \text{MSE}(c|\theta) \) and \( \min_{c \in \mathbb{R}} \text{SPH}(c|\theta) \) (see 13 and 16) are given by 

\[ c_{\text{MSE}} = \frac{\text{cov}(\theta, \eta)}{S^2(\eta)} \quad \text{and} \quad c_{\text{SPH}} = \frac{m(\eta)}{m(\theta)} c_{\text{MSE}}. \]  

The next assumptions may be viewed as laws of large numbers in the random setting or regularity conditions in a deterministic one. They concern the sequences \( \{\theta_i\}_{i=1}^m \) and \( \{w_i\}_{i=1}^m \) and allow for dependence on \( p \) i.e., \( \theta_i = \theta_i^{(p)} \) and \( w_i = w_i^{(p)} \).
Assumption 5. For constants $m \in \mathbb{R}$ and $\nu, \xi \in (0, \infty)$ as $p \uparrow \infty$, we have:

- (i) $m(\theta) = m$ and $s^2_{m}(\theta) = \xi^2$,
- (ii) $m(w) = 0$ and $s^2_{m}(w) = \nu^2$,
- (iii) $\text{cov}_m(\theta, w) = 0$, and
- (iv) there exists an estimator $\hat{\nu} = \hat{\nu}_p$ for each $p$ with $\hat{\nu}_\infty = \nu$.

The following identities follow by direct calculation.

**Lemma 2.** Suppose $\{\theta_i\}$ and $\{w_i\}$ satisfy Assumption 5 and $\eta_i = \theta_i + w_i$. Then (almost surely), $m(\eta) = m$, $\text{cov}(\eta, \theta) = \xi^2$ and $s^2_{m}(\eta) = \xi^2 + \nu^2$.

We define the SNR and the signal-incoherence $r_\infty$ as

$$
\text{SNR} = \frac{\xi}{\nu} \quad \text{and} \quad r_\infty = \frac{1}{\sqrt{1 + (m/\xi)^2}},
$$

which are compatible with (15) upon taking $\theta = \chi_n \beta$ and $\nu = \delta/\sqrt{n}$. The following result establishes the conclusions of Theorems 1 and 2 in our abstract setting.

**Proposition 3.** Let $\eta = \theta + w$ where $\theta, w \in \mathbb{R}^p$ and an estimator $\hat{\eta}$ satisfy Assumption 5. Then, for $c_\infty$ and $d_\infty$ defined in (22) and (23) but with SNR and $r_\infty$ in (25) the estimate $\eta(c) = \eta + c(\eta - m(\eta))$ with parameter $c = 1 - \xi^2/\nu^2$ satisfies

$$
\text{MSE}_\infty(\eta(c) | \theta) = c_\infty \text{MSE}_\infty(\eta | \theta) \quad \text{and} \quad \text{SPH}_\infty(\eta(c) | \theta) = d_\infty \text{SPH}_\infty(\eta | \theta).
$$

Moreover, the optimal parameters $c_{\text{MSE}}$ and $c_{\text{SPH}}$ in (24) converge as $p \uparrow \infty$ to $c_\infty$.

Figure 1 illustrates the geometry of the estimator $\eta(c)$ of the vector $\theta$.

**4 | PROOFS THE MAIN RESULTS**

We proceed along the following three main steps.

![Figure 1](image-url)
1. We establish the key identity \( \eta = \theta + w \) with \( \theta = \chi \beta \) per (18) in Section 4.1.
2. We derive the convergence properties in the HDLS regime of the eigenvalues and eigenvectors of \( S \) under our spiked covariance model setting in Section 4.2.
3. We verify that \( \theta, w \) in (1) and \( \bar{\nu} \) in (20) satisfy the conditions of Assumption 5, which leads to the guarantees for James–Stein shrinkage in Proposition 3.

Theorems 1 and 2 and then corollaries of Proposition 3; the latter proved in Appendix A. We will make use of two classic results in matrix perturbation theory.

**Theorem 3** Weyl. Let \( A \) and \( (A + \Delta) \) be (real) symmetric \( n \times n \) matrices with eigenvalues \( \alpha_1 \geq \ldots \geq \alpha_n \) and \( \zeta_1 \geq \ldots \geq \zeta_n \), respectively. Then,

\[
\max_{1 \leq i \leq n} |\alpha_i - \zeta_i| \leq |\Delta|.
\]

For a proof, see Horn and Johnson (2013) (also Weyl, 1912).

**Theorem 4** Davis–Kahan. Let \( A \) and \( (A + \Delta) \) be (real) symmetric \( n \times n \) matrices with \( A \alpha = \alpha \alpha \) and \( (A + \Delta) \beta = \beta \beta \) for eigenvectors \( \alpha, \beta \in \mathbb{R}^n \) and eigenvalues \( \alpha, \beta \in \mathbb{R} \). Suppose \( \alpha_1 \geq \ldots \geq \alpha_n \) and \( \beta_1 \geq \ldots \geq \beta_n \) with the convention \( \alpha_0 = \infty = -\alpha_{n+1} \) and assume \( \gamma_j = \min(\alpha_{j-1} - \alpha_j, \alpha_j - \alpha_{j+1}) > 0 \). Then,

\[
|\alpha - \beta| \leq \frac{2}{\gamma_j} |\Delta| \quad \text{provided (w.l.o.g.)} \ (\alpha', \beta') \geq 0.
\]

This result is proved in Yu et al. (2015, corollary 1).

### 4.1 Establishing the key identity

A key tool for random matrix theory in the HDLS regime is the dual sample covariance matrix. This is \( n \times n \) matrix \((n \geq 2)\) is fixed,

\[
L = Y^\top Y / p.
\]

The next result is well known and relates the spectra of \( S = Y Y^\top / n \) and \( L \).

**Lemma 3.** Let \( Lu = \ell^2 u \) where \( \ell^2 \in (0, \infty) \) and \( u \in \mathbb{R}^n \). Then, \( Sv = s^2 v \) where \( v = Y u / (\sqrt{\ell p}) \) and \( s^2 = \ell^2 p / n \). Conversely, let \( Sv = s^2 v \) where \( s^2 \in (0, \infty) \) and \( v \in \mathbb{R}^p \). Then, \( Lu = \ell^2 u \) where \( u = Y^\top v / (\sqrt{s n}) \) and \( \ell^2 = s^2 n / p \).

**Proof.** Multiplying the identity \( Lu = \ell^2 u \) by \( Y \) from both sides, we obtain

\[
Y Lu = \ell^2 Y u \quad \Rightarrow \quad SY u = \left( \frac{\ell^2 p}{n} \right) Y u.
\]

Note that \((Yu)^\top (Yu) = (u^\top Lu) p = \ell^2 p \), so \( v = (Yu) / (\sqrt{\ell p}) \) has unit length. Dividing by \( \sqrt{\ell p} \) yields \( Sv = s^2 v \) as required. The converse has an identical argument.

The spike model \( \Sigma = \Gamma + \beta \beta^\top \) has a full basis of eigenvectors given by \( b \beta / |\beta| \) and \( \{\psi_i\}_{i=1}^{p-1} \), the latter corresponding to the non-zero eigenvalues \( \{\alpha_i\}_{i=1}^{p-1} \) of \( \Gamma \). Thus,

\[
y = (b, y)b + \sum_{i=1}^{p} \langle \psi_i, y \rangle \psi_i = \beta \psi + \epsilon,
\]

where \( \epsilon = \sum_{i=1}^{p} \psi_i b \psi / B \) and \( \psi = \psi_p = (b, y) / (\beta, \beta) \) as in (8). Consequently, letting \( Y \) denote the \( p \times n \) matrix of i.i.d. observations of \( y \in \mathbb{R}^p \), we have
where $X = Y^T / \beta$, $\beta \in \mathbb{R}^p$ consists of i.i.d. observations of $\psi$ and $\varepsilon$ is a $p \times n$ matrix with i.i.d. columns consisting of the observations of $\epsilon$ as defined above.

By orthogonality, we obtain that $L = Y^T Y / p = (\langle \beta, \beta \rangle / p) XX^T + \varepsilon \varepsilon^T / p$. Let $x_p$ be the eigenvector of $L$ with the largest eigenvalue, $\beta \varepsilon^2$, and $x_\infty = X / X_n \in \mathbb{R}^p$ (the unit length normalization of $X$), where $\lambda$ and $x_n$ are defined in (12). By Lemma 3,

$$h = \frac{Y x_p}{\sqrt{p} \sqrt{p}} = \frac{(\langle \beta, \beta \rangle / p) x_p \varepsilon^2 (x_p x_n)}{\sqrt{p} / \sqrt{p}} + \frac{\varepsilon x_p}{\sqrt{p} / \sqrt{p}}$$

$$= \beta \left( \frac{x_p}{\sqrt{p}} \right) + \beta \left( \frac{x_p}{\sqrt{p}} \right) (x_p x_n - X_n) + \frac{\varepsilon x_p}{\sqrt{p}}$$

We deduce that $h = s_p h = x_p \beta + w$ as required by (18) where

$$w = x_p \beta (x_p x_n - X_n) + \frac{\varepsilon x_p}{\sqrt{p}}$$

(27)

### 4.2 Convergence of the eigenvalues/vectors

It is not difficult to establish that the limit of $L = L^{(p)}$ as $p \to \infty$ (in any norm on $\mathbb{R}^{p \times n}$) takes the following form:

$$L^{(\infty)} = (\sigma^2 + \mu^2) (x^T_n x_n) + \delta^2 I.$$  

The first term is the limit of $\langle \beta, \beta \rangle / p$ under Assumption 2 (note that $\langle \beta, \beta \rangle = \sigma^2 + \mu^2$) and the definitions of $X$ and $x_n$ above. By Assumption 3, the columns of $\varepsilon$ are i.i.d. copies of $\epsilon$ with $E(\epsilon) = 0$, by definition, and the strong law of large numbers confirms that the off-diagonal entries of the second term are zero. That $\delta^2$ determines all the diagonal entries is again a consequence of Assumption 2. This entails proving that $\langle \epsilon, \epsilon \rangle / p$ converges almost surely to $\sigma^2$ as is done in Section 4.3 item (ii).

The matrix $L^{(\infty)}$ has an easily described spectrum. Its largest eigenvalue is given by $\epsilon^{(\infty)} = (\sigma^2 + \mu^2) (x^T_n x_n) + \delta^2$ and has the eigenvector $x_\infty$. All remaining eigenvalues equal $\delta^2$. Since $L^{(p)}$ converges (in any norm) to $L^{(\infty)}$, the largest eigenvalue $\epsilon^{(p)}$ converges to $\epsilon^{(\infty)}$ almost surely. All the remaining eigenvalues converge to $\delta^2$.

By Weyl's inequality, setting $A = \Omega$ and $B = (L - \Omega)$ so that $A + B = L$, we have

$$|\epsilon^{(p)} - (\sigma^2 + \mu^2) | (x^T_n x_n) | 1_{(i-1)}^2 + \delta^2 | \leq |L^{(p)} - L^{(\infty)}| \quad 1 \leq i \leq n.$$  

(28)

We immediately deduce (by Lemma 3) the following result, variants of which appear in Jung et al. (2012), Shen et al. (2016) and Goldberg et al. (2021). Let $s_i^{(p)}$ denote the i-th largest eigenvalue of $S$ (for $i > \min(p, n)$ all eigenvalues are zero).

**Proposition 4.** Fix $n \geq 2$ and suppose Assumptions 1,2 and 3 hold. Then, $\lim_{p \to \infty} s_i^{(p)} / p = 1_{(i-1)} x^T_n (\sigma^2 + \mu^2) + \frac{\delta^2}{n}$ almost surely for fixed $1 \leq i \leq n$.

Next, by the Davis–Kahan theorem with $L^{(p)} = L^{(\infty)} + \Delta$ and $\Delta = L^{(p)} - L^{(\infty)}$, for $x_p$ and $x_\infty$ the eigenvectors of $L^{(p)}$ and $L^{(\infty)}$, respectively,

$$|x_p - x_\infty| \leq \frac{3}{\delta^2} |L^{(p)} - L^{(\infty)}|.$$  

(29)

Note the condition $\langle \epsilon, \epsilon \rangle \geq 0$ is without loss of generality as the orientation of the eigenvectors is always arbitrary. The following result follows immediately.

**Proposition 5.** Fix $n \geq 2$ and suppose that Assumptions 1,2 and 3 hold. Then, we have $|x_p - x_\infty| \to 0$ as $p \to \infty$ almost surely.
4.3 Verifying Assumption 5

We make use of Assumptions 1 and 2. There are four items to verify for \( \theta = \chi \beta \) and \( w \) in (27). Take \( m = \chi m, \xi = \chi \sigma \) and \( \nu = \delta / \sqrt{n} \).

(i) We have \( m_m(\theta) = \chi m, \) and \( s_w^2(\theta) = \chi^2 \sigma^2 \) by Assumption 2 part (i).

(ii) To see that \( m_m(w) = 0 \), we compute \( m(w) \) using (27) which gives

\[
m(w) = \chi m(\beta) \langle x_p, x_p - x_m \rangle + m(\xi \phi) / \sqrt{n}.
\]

The first term vanishes by Proposition 5 since \( \langle x_p, x_p - x_m \rangle \leq |x_m - x_p| \) and \( m_m(\beta) \) and \( \chi m \) are finite almost surely by Assumption 2 part (ii).

The second term vanishes because \( m(\xi \phi) \) may be written as a linear combination of a fixed \( n \) realizations of \( m(\epsilon) \) with coefficients being the entries of \( x_p \) and \( |x_m| \) is bounded. To this end, \( m(\epsilon) = m_p(\epsilon) = \sum_{i=1}^{p} \phi_i m(\gamma_i) = m_p(\phi) \) which tends to zero as \( p \to \infty \) (almost surely) by Assumption 2 part (iii).

Similarly, to verify that \( s_w^2(w) = \nu \), we use (27) again to calculate that

\[
s^2(w) = \chi^2 \langle \langle \beta, \beta \rangle / p \rangle \langle x_p, x_p - x_m \rangle^2 + \langle \xi \phi, \xi \phi \rangle / (pn) - m^2(w),
\]

where we have used the fact that \( \langle \beta, \beta \rangle = 0_n \) by Assumption 1. Since \( \chi m \) and the limit \( \chi^2(n^2 + \mu^2) \) of \( \langle \beta, \beta \rangle / p \), as above, by Proposition 5, the first term tends to zero as \( p \to \infty \). For the second term, we note that \( \langle \xi \phi, \xi \phi \rangle \) may be written as a convex combination of a fixed \( n \) realizations of \( \langle \epsilon, \epsilon \rangle \) with coefficients being the entries of \( x_p \) squared, and \( |x_m| \) is bounded. We have \( \langle \epsilon, \epsilon \rangle / p = \sum_{i=1}^{p} \phi_i^2 / p = s^2(\phi) + m^2(\phi) \) as the \( \langle \beta_i \rangle \) are orthonormal. By Assumption 2 parts (iii) and (iv), we have \( s_w^2(\phi) = \delta^2 \) and \( m_m(\phi) = 0 \). It follows that the second term converges to \( \delta^2 |x_m|^2 / n = \delta^2 / n = \nu^2 \). The last term tends to zero since \( m_m(w) = 0 \) as above, and the claim now follows.

(iii) We have \( \operatorname{cov}(\theta, w) = \langle \theta, w \rangle / p - m(\theta) m(w) \), and since \( m_m(\theta) \) is finite, the second term vanishes as \( m_m(w) \) as above. Again by (27) and since \( \langle \beta, \beta \rangle = 0_n \),

\[
\langle \theta, w \rangle / p = \chi^2 \langle \langle \beta, \beta \rangle / p \rangle \langle x_p, x_p - x_m \rangle,
\]

which vanishes in the limit by the same arguments as in (ii) above.

(iv) To see that \( \nu = \hat{\nu} \) in (20) is an asymptotically exact estimate of \( \nu = \delta / \sqrt{n} \), we use Proposition 4. Since \( (\operatorname{tr}(S) - s_p^2) / p = \sum_{i=1}^{p} s_p^2 / p \), under the hypotheses of Proposition 4 converges almost surely to \( \delta^2 \hat{\nu}^{1/2} \).

ACKNOWLEDGEMENTS

I am indebted to Lisa Goldberg for conjecturing the result derived in this work. I thank the participants of the 2021 Symposium on Data Science and Statistics and the 2021 Joint Statistical Meetings where this work was presented for their comments.

CONFLICT OF INTEREST

The author declares no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Data sharing not applicable - no new data generated, or the article describes entirely theoretical research

ORCID

Alex Shkolnik https://orcid.org/0000-0003-1106-5689

REFERENCES

Ahn, J., Marron, J. S., Muller, K. M., & Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika, 94(3), 760–766.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. The Annals of Mathematical Statistics, 42(3), 855–903.

Brown, L. D., & Zhao, L. H. (2012). A geometrical explanation of Stein shrinkage. Statistical Science, 27(1), 24–30.

Donoho, D. L., Gavish, M., & Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. Annals of Statistics, 46(4), 1742.

Efron, B., & Morris, C. (1975). Data analysis using Stein’s estimator and its generalizations. Journal of the American Statistical Association, 70(350), 311–319.

Fan, J., Guo, J., & Zheng, S. (2020). Estimating number of factors by adjusted eigenvalues thresholding. Journal of the American Statistical Association, 1–10.

Fouad, D., Strawderman, W. E., & Wells, M. T. (2018). Shrinkage estimation: Springer.
APPENDIX A: AUXILIARY PROOFS

As shown in Section 4.3, the hypotheses of Propositions 1 and 2 (i.e., Assumptions 1–4) guarantee that the conditions on \(\{\theta_t\}_{t=1}^{\infty}\) and \(\{w_t\}_{t=1}^{\infty}\) in Assumption 5 are satisfied. Consequently, the proofs of these two results, as well as that of Proposition 3 that requires Assumption 5 directly, reduce to the calculations below. The proof of Lemma 2 is omitted as it is elementary and that of Lemma 1 is a direct consequence of some of the expressions below.

For any \(c \in \mathbb{R}\) and \(\eta(c) = m(\eta) + c(\eta - m(\eta))\), by direct calculation,

\[
\text{MSE}(\eta(c), \theta) = (m(\eta) - m(\theta))^2 + s^2(\theta) + c^2s^2(\eta) - 2ccov(\eta, \theta).
\]

When \(c = 1\) for which \(\eta(1) = \eta\) and applying Lemma 2 yields

\[
\text{MSE}_\text{w}(\eta|\theta) = s^2 + (\xi^2 + \eta^2) - 2\xi^2 = \xi^2 / n,
\]

which proves Proposition 1. The sine of the angle squared metric is computed as

\[
\text{SPH}(\eta(c), \theta) = 1 - \left(\frac{\eta(c), \theta}{\eta(c)}\right)^2 = 1 - \frac{(m(\eta) + cccov(\eta, \theta))^2}{(m^2(\eta) + c^2s^2(\eta))(m^2(\theta) + s^2(\theta))}.
\]

Using the raw estimate \(\eta = \eta(1)\) for which \(c = 1\), we deduce by Lemma 2 that

\[
\text{SPH}_\text{w}(\eta|\theta) = 1 - \frac{m^2 + \xi^2}{m^2 + \xi^2 + \eta^2} = \frac{\eta^2}{\xi^2 + m^2 + \eta^2} = \frac{r^2}{\text{SNR}^2 + r^2_w}.
\]
which establishes Proposition 2 with SNR and \( r_\infty \) in (15) (c.f. 25).

Note that minimizing the expressions for \( \text{MSE}(\eta(c)|\theta) \) and \( \text{SPH}(\eta(c)|\theta) \) above over \( c \in \mathbb{R} \) yields the \( c_{MSE} \) and \( c_{SPH} \) in (24) proving Lemma 1. The limits as \( p \to \infty \) of these quantities are easily verified as \( c_\infty = \frac{\text{SNR}}{1+\text{SNR}} \) in (22) using Lemma 2. This establishes the last part of Proposition 3. To prove the first part, we again apply Lemma 2 to deduce that \( c = c_p = 1 - \frac{\xi_2}{\nu_2} \approx 1 - \frac{\xi}{\nu} = c_\infty \) as above. Also,

\[
\text{MSE}_\infty(\eta(c)|\theta) = \xi^2 + c_\infty^2 (\xi^2 + \nu^2) - 2c_\infty \xi^2 = \xi^2 + c_\infty \xi^2 - 2c_\infty \xi^2 = \xi^2 (1 - c_\infty) = \xi^2 \frac{1}{1 + \text{SNR}^2} = c_\infty \xi^2 = \text{cMSE}_\infty(\eta|\theta),
\]

as required. Similarly, by Lemma 2, we derive that

\[
\text{SPH}_\infty(\eta(c)|\theta) = 1 - \frac{(m^2 + c_\infty^2 \xi^2)^2}{(m^2 + c_\infty^2 (\xi^2 + \nu^2)) (\mu^2 + \sigma^2)} = (1 - c_\infty) r_\infty^2 = (1 - c_\infty) (\text{SNR}^2 + r_\infty^2) \text{SPH}_\infty(\eta|\theta) = \left( \frac{\text{SNR}^2 + r_\infty^2}{1 + \text{SNR}^2} \right) \text{SPH}_\infty(\eta|\theta) = d_\infty \text{SPH}_\infty(\eta|\theta),
\]

for \( d_\infty \) as in (23). This concludes the proof of Proposition 3.