The Geometry of Estimation in High Dimensions

Alex Shkolnik (shkolnik@ucsb.edu)

Blessings of Dimensionality Workshop.

University of Connecticut, Storrs, CT.

July 15, 2024.

Department of Statistics & Applied Probability
University of California, Santa Barbara

Table of contents.

Notation.

- p – dimension
- $\langle u, v \rangle = \sum_{i=1}^{m} u_i v_i$
- $|u| = \sqrt{\langle u, u \rangle}$

Motivating example

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a p-variable function and consider,

$$\max_{x \in \mathbb{R}^p} f(x).$$

In practice f is unknown and we have an estimated surrogate,

$$\hat{f} : \mathbb{R}^p \to \mathbb{R}\,.$$

e.g., $f(x) = \mathrm{E}(\hat{f}(x))$

Let $\hat{x}$ be the maximizer of $\hat{f}$ and consider the objective value,

$$f(\hat{x}) \qquad \text{(realized “optimum”)}.$$

Closely related to the statistics notion of out-of-sample.

- $\hat{f}(\hat{x})$ *is the estimated optimum.*
- $\max_{x \in \mathbb{R}^p} f(x)$ *is the true optimum.*

An important example of f (a quadratic function of p-variables).

$$Q(x) = 1 + \langle \mu, x \rangle - \frac{1}{2}\langle x, \Sigma x \rangle \qquad (x \in \mathbb{R}^p)$$

- $\mu \in \mathbb{R}^p$ and Σ is a symmetric and pos. def. $(p \times p)$-matrix.

Applications.

- *optimization, graph theory, statistics and probability.*
- *Mean-variance portfolio optimization, robust (Capon) beamforming in signal processing, optimal fingerprinting in climate science.*
- *A Lagrangian for (linearly constrained) quadratic programs.*

Suppose some fixed number q of eigenvalues of $\Sigma = \Sigma_{p \times p}$ diverge in p and the remaining eigenvalues are bounded in $(0, \infty)$.

- *The diverging eigenvalues are often called spikes.*

Also suppose μ is not (eventually) an eigenvector for the spikes.

Their estimates ζ and $\hat{\Sigma}$ are assumed to inherit the same properties.

The maximizer of Q is unique and occurs at $\Sigma^{-1}\mu$ so that,

$$\max_{x \in \mathbb{R}^p} Q(x) = 1 + \frac{g_\mu^2}{2} \tag{1}$$

where $g_\mu^2 = \langle \mu, \Sigma^{-1}\mu \rangle = |\mu|^2 \langle v, \Sigma^{-1} v \rangle$ for $v = \mu/|\mu|$

- *The true optimum* (1) *grows in* p *in proportion to* $|\mu|^2$.
- *For effect, the natural assumption that* $|\mu|^2 = \sum_{i=1}^p \mu_i^2$ *grows in* p *to* $+\infty$ *is assumed but not needed (same for the estimate* ζ*).*

The *estimated objective* function $\hat{Q}$ takes the form,

$$\hat{Q}(x) = 1 + \langle \zeta, x \rangle - \frac{1}{2}\langle x, \hat{\Sigma} x \rangle$$

where we think of ζ and $\hat{\Sigma}$ as estimates of μ and Σ.

- *The case $\zeta = \mu$ is of practical utility but $\hat{\Sigma} \neq \Sigma$.*

The estimated optimum follows the logic of the previous slide.

Let $\hat{x}$ denote the unique maximizer ($\hat{\Sigma}^{-1}\zeta$) of $\hat{Q}$.

The *realized optimum* – true objective at the estimated maximizer.

$$\begin{aligned} Q(\hat{x}) &= 1 + \langle \mu, \hat{\Sigma}^{-1}\zeta \rangle - \frac{1}{2}\langle \hat{x}, \Sigma \hat{x} \rangle \\ &= 1 + \frac{\hat{g}_{\zeta}^{2}}{2} D_p \end{aligned}$$

where $\hat{g}_{\zeta}^{2} = \langle \zeta, \hat{\Sigma}^{-1}\zeta \rangle$ and we call D_p the *discrepancy*, i.e.,

$$\max_{x \in \mathbb{R}^p} Q(x) = 1 + \frac{g_{\mu}^{2}}{2}.$$

Unless $\hat{\Sigma}$ chosen "wisely", the discrepancy D_p tends to $-\infty$.

As $p \uparrow \infty$ the true objective $\max_{x \in \mathbb{R}^p} Q(x)$ tends to $+\infty$ but the realized objective $Q(\hat{x})$ tends to $-\infty$.

- Which (or which part) of the estimates causes this?

$$Q(x) = 1 + \langle \mu, x \rangle - \frac{1}{2} \langle x, \Sigma x \rangle$$

$$\hat{Q}(x) = 1 + \langle \zeta, x \rangle - \frac{1}{2} \langle x, \hat{\Sigma} x \rangle$$

Let r_p be the rate at which the spiked eigenvalues diverge.

- *For simplicity, all non-spiked eigenvalues of $\hat{\Sigma}$ are identical.*

Theorem. (Gurdogan & Shkolnik, 2024)

$$D_p = O(1)\frac{|\mu|}{|\zeta|} - O(r_p)|\mathscr{E}_p(\mathscr{H})|^2 + \text{ 2nd order terms.}$$

We call the quantity $\mathscr{E}_p(\mathscr{H})$ the *quadratic optimization bias*.

Let $\mathscr{B}$ and $\mathscr{H}$ be the eigenvectors of Σ and $\hat{\Sigma}$ corresponding to the q spiked eigenvalues ($p \times q$ matrices with orthonormal columns).

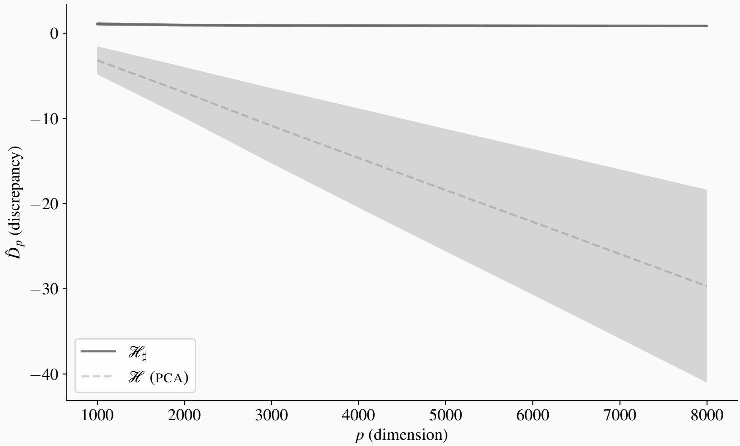
Quadratic optimization bias.

$$\mathscr{E}_p(\mathscr{H}) = \frac{\mathscr{B}^\top z - (\mathscr{B}^\top \mathscr{H})(\mathscr{H}^\top z)}{\sqrt{1 - |\mathscr{H}^\top z|^2}} \qquad \Big(z = \frac{\zeta}{|\zeta|}\Big).$$

- *Does not depend on any of the eigenvalues.*
- *Does not depend on the non-spiked eigenvectors.*
- *Tells us what has to be estimated accurately.*
- *i.e. find roots of* $\mathscr{E}_p : \mathbb{R}^{p\times q} \to \mathbb{R}^q$ *(perhaps asymptotically).*
- *The quantities* $\mathscr{B}^\top z$ *and* $\mathscr{B}^\top \mathscr{H}$ *are not observed.*
- *The most natural framework is* HDLSS.

Numerics on principal component analysis of a (simulated) financial data set with $q = 7$ spikes and a finite sample.

p	$\max_x Q(x)$	$\mathrm{E}\,Q(\hat{x})$	$\mathrm{E}\,D_p(\mathscr{H})$	$\mathrm{E}\,Q(\hat{x}_\sharp)$	$\mathrm{E}\,D_p(\mathscr{H}_\sharp)$
500	1.01	0.99	−1.16	1.00	1.22
2000	1.03	0.64	−7.11	1.01	0.93
8000	1.12	−4.98	−30.04	1.04	0.85
32000	1.47	−97.01	−121.81	1.18	0.86
128000	2.88	−1572.9	−486.92	1.70	0.87



Theoretical analysis and numerics in Gurdogan & Shkolnik (2024).
"*Quadratic Optimization Bias of Large Covariance Matrices*"

Mean estimation

There are so many methods for covariance estimation!

- *e.g., Yao, Zheng & Bai (2015)*

INADMISSIBILITY OF THE USUAL ESTIMATOR FOR THE MEAN OF A MULTIVARIATE NORMAL DISTRIBUTION

CHARLES STEIN
STANFORD UNIVERSITY

1. Introduction

If one observes the real random variables $X_1, \cdots, X_n$ independently normally distributed with unknown means $\xi_1, \cdots, \xi_n$ and variance 1, it is customary to estimate ξ_i

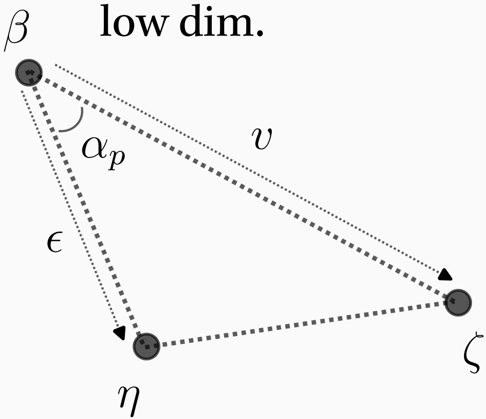
– Stein (1955)

Third Berkeley symposium on mathematical statistics and probability

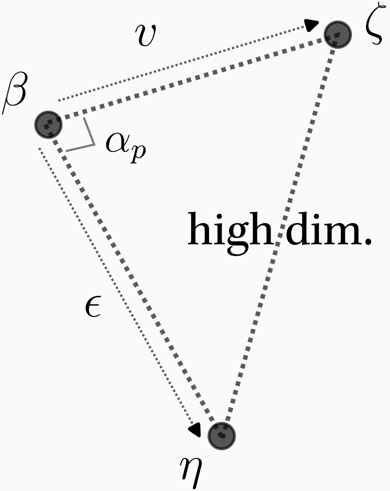
Let $\eta = \beta + \epsilon$ be a noisy observation of a vector $\beta \in \mathbb{R}^p$.

- *Critical dimension* $p = 3$.
- *This is a HDLSS framework where* $p \uparrow \infty$.

Let $\epsilon = \eta - \beta$ represent "noise".



The vector $\zeta \in \mathbb{R}^p$ is treated as nonrandom (or independent).


υ
ζ
β
α_p
high dim.
ϵ
η

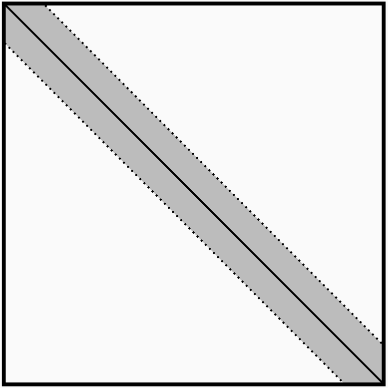
Geometric LLN.

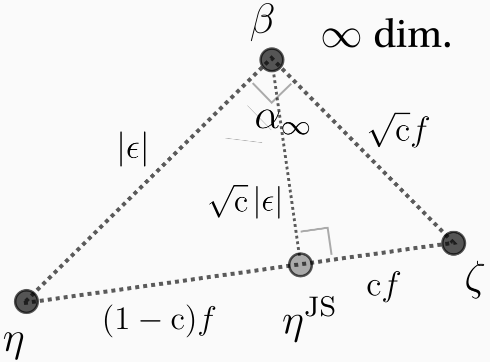
$$\langle \eta - \beta, \zeta - \beta \rangle / p = \langle \epsilon, v \rangle / p = \frac{1}{p} \sum_{i=1}^{p} \epsilon_i v_i \to 0$$

As $p \uparrow \infty$ the angle α_p between ϵ and $v = \zeta - \beta$ tends to 90°.

- *Noise (pure randomness) is orthogonal to (or independent of) any high dimensional vector not corrupted by it.*
- *Very similar to perhaps the first concentration of measure result due to Borel 1914 (mass of hypercube concentrates on equator).*

Borel 1914 – Concentration of mass of hypercube on its equator.





Once we reach dimension ∞ the geometry is computed in terms of the limits of $f = \frac{|\epsilon|}{\sqrt{1-c}}$ and $c = 1 - \frac{|\epsilon|^2}{|\eta-\zeta|^2}$.

The point η^{JS} is an estimator of β strictly better than η with respect to ℓ_2-error (originally due to James & Stein (1961) for $p > 2$).

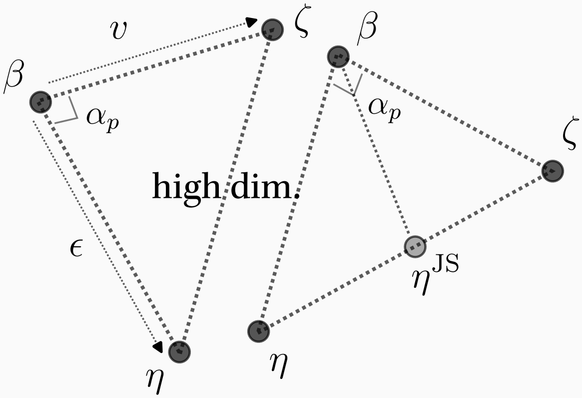
Letting ν^2 be a p-consistent estimate of $|\epsilon|^2 = |\eta - \beta|^2$ (may be obtained with just $n = 2$ weakly dependent observations), define

$$\eta^{\mathrm{JS}} = \zeta + \mathrm{c}\,(\eta - \zeta)\,, \qquad \left(\mathrm{c} = 1 - \frac{\nu^2}{|\eta - \zeta|^2}\right)$$

$$= \mathrm{c}\,\eta + (1 - \mathrm{c})\,\zeta$$

Theorem. If $|\beta|^2/p$ and $|\epsilon|^2/p$ are bounded in $(0, \infty)$ as $p \to \infty$ and the Geometric LLN holds, $\sqrt{\mathrm{c}}$ is eventually in the interval $(0, 1)$ and,

$$|\eta^{\mathrm{JS}} - \beta| \sim \sqrt{\mathrm{c}}\,|\eta - \beta| = \sqrt{\mathrm{c}}\,|\epsilon|\,.$$



The vector ζ is called the "shrinkage target".

Remarks.

A similar (but not asymptotic) geometry is illustrated in Efron (1978).

The first p-asymptotic analysis of JS estimator appears in is Casella & Hwang (1982) but without any mention of geometry.

Fourdrinier, Strawderman & Wells (2018) – excellent treatment of the theoretical aspects of the James-Stein (JS) estimator.

Efron & Morris (1975) – "*Difficulties in adapting the James-Stein estimator to the many special cases that invariably arise in practice.*"

Covariance estimation

Eigenvector estimation.

As sensibly pointed out by Wang & 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."*

The original problem in Stein (1955) concerns mean estimation from finitely many Gaussian observations $y = \mu + \epsilon$

- *The vector $\mu \in \mathbb{R}^p$ is a population mean.*
- *The usual estimate is the sample mean.*
- *The metric is quadratic loss (expected mean-squared error).*
- *The estimator y^{JS} is not optimal (just better than y).*
- *The shrinkage target is arbitrary (Stein's paradox).*

Its arguable that the more natural framework is that of spiked covariance estimation (eigenvector estimation in particular).

- *The vector $\beta \in \mathbb{R}^p$ is a population eigenvector.*
- *The usual estimate is a sample eigenvector.*
- *The metric is the quadratic optimization bias.*
- *The* JS *estimator will be optimal wrt this metric.*
- *There is a natural choice of shrinkage target.*

To move over to covariance estimation, correlate our variables via

$$y = \mu + \beta x + \epsilon$$

where x is mean-zero of unit variance and uncorrelated from ϵ.

Letting $V = (y - \mu)(y - \mu)^\top$ we return to our quadratic function,

$$\begin{aligned} Q(x) &= \mathrm{E}(\hat{Q}(x)) \\ &= \mathrm{E}\big(1 + \langle y, x\rangle - \frac{1}{2}\langle x, Vx\rangle\big) \\ &= 1 + \langle \mu, x\rangle - \frac{1}{2}\langle x, \Sigma x\rangle \end{aligned}$$

where $\Sigma = \mathrm{E}(V) = \textsc{var}(y) = \beta\beta^\top + \Gamma$ for $\Gamma = \textsc{var}(\epsilon)$.

- *Assume all eigenvalues of Γ are bounded in* $(0, \infty)$.

We take the simple estimate $\hat{\Sigma} = \eta\eta^\top + \nu^2 I$ and some ζ.

$$\hat{Q}(x) = 1 + \langle \zeta, x \rangle - \frac{1}{2}\langle x, \hat{\Sigma} x \rangle$$

Our discrepancy theorem is restated for this example as,

$$D_{\boldsymbol{p}} = -O(|\beta|^2)|\mathscr{E}_{\boldsymbol{p}}(\eta)| + O(1)\frac{|\mu|}{|\zeta|}$$

where $\mathscr{E}_{\boldsymbol{p}}(\eta)$ is the optimization bias (for $q = 1$).

Lets embed everything into a unit sphere in $\mathbb{R}^{\boldsymbol{p}}$.

$$h = \frac{\eta}{|\eta|} \qquad b = \frac{\beta}{|\beta|} \qquad z = \frac{\zeta}{|\zeta|}$$

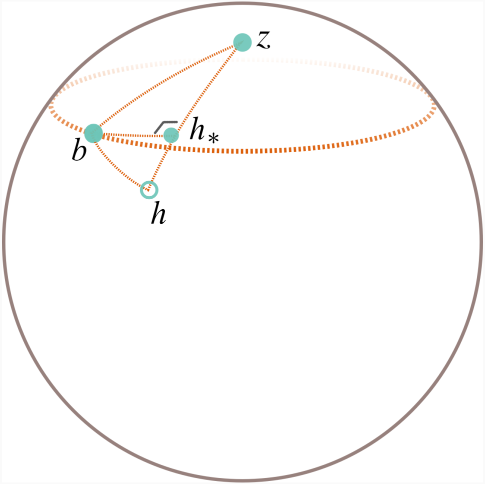
In this single eigenvector special case the optimization bias is,

$$\mathscr{E}(h) = \frac{\langle b, z\rangle - \langle h, b\rangle\langle h, z\rangle}{\sqrt{1 - \langle h, z\rangle^2}}$$

First identified in Goldberg, Papanicolaou & Shkolnik (2022).

- *Now frequently called the* GPS *paper.*
- *The map $\mathscr{E}_{\boldsymbol{p}}(\cdot)$ has at least 2 distinct roots.*
- **Problem 1** – *characterize all the roots.*
- **Problem 2** – *which may estimated from data p-asymptotically.*

By the spherical law of cosines $\mathscr{E}_p(h_*) = 0$.



Why is this applicable for eigenvectors?

- *Suppose our data (columns) are generated* n *observations of*

$$y = \mu + \beta x + \epsilon$$

  which are weakly dependent (say i.i.d.).

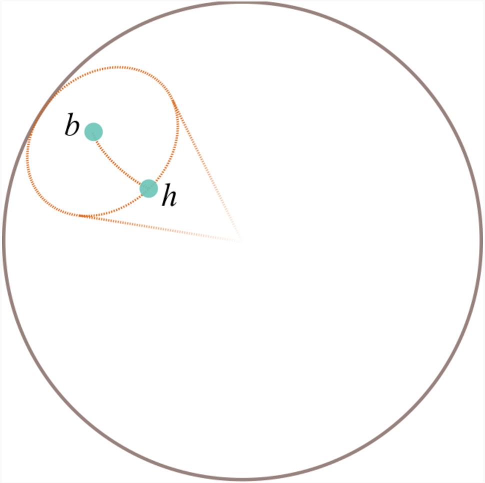
- *We will need* $n \geq 3$ *observations (the very first one, one to center the data and one more to estimate the optimal shrinkage amount).*

Making a (centered) $p \times p$ sample covariance matrix, we can extract an eigenvector h with the largest eigenvalue $\mathscr{s}^2$.

$$\eta = \mathscr{s}h = \chi_n \beta + \epsilon'$$

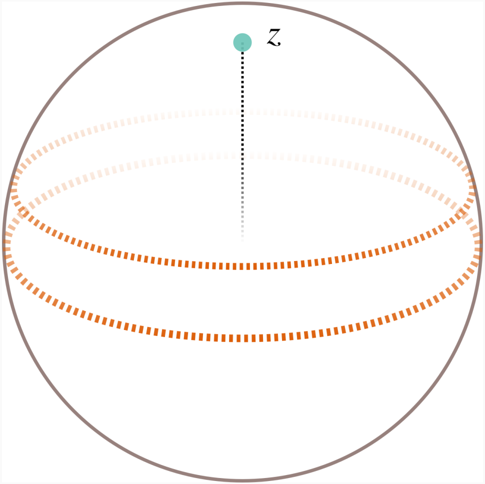
for a random variable χ_n related to $\langle h, b \rangle$ and where $\epsilon' \neq \epsilon$ that has sufficient weak dependency in its entries to satisfy the Geometric LLN.

Bias of high-dimensional vectors ($h = \psi b + \epsilon''$).



A question about geometry – from the “right” perspective.

Levy's concentration of measure on the sphere.



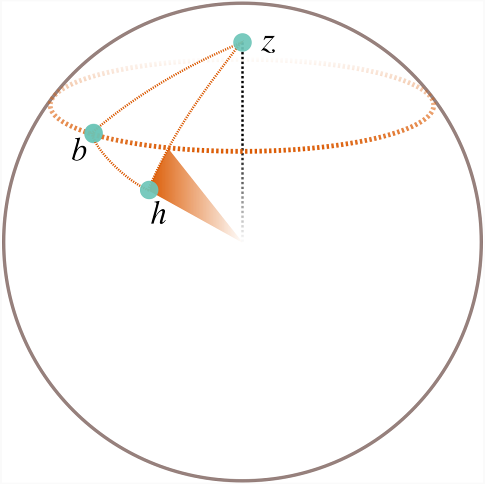
Let $\mathscr{A}$ be an area of size at least $1/2$ on $\mathbb{S}^{p-1}$ (Lévy 1919). Then,

$$\mu(x \in \mathbb{S}^{p-1} : \mathrm{d}(x, \mathscr{A}) \geq r) \leq 2e^{-pr^2/64}$$

where μ is the uniform surface area measure and $\mathrm{d}(x, \mathscr{A})$ denotes the Euclidean geodesic from x to $\mathscr{A}$.

- *Leads to a concentration around the equator.*

The following is true in high dimensions (whp).



Scale ζ to ensure $\zeta = \langle \eta, z \rangle z$.

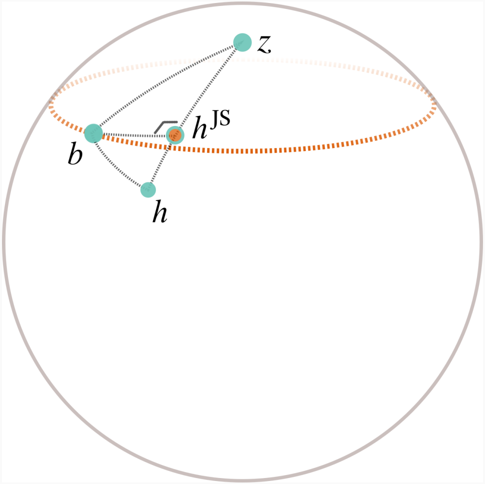
Theorem. The JS estimator of the leading eigenvector is a p-asymptotic root of the quadratic optimization bias $\mathscr{E}(\cdot)$ for fixed $n \geq 3$.

- *Gurdogan & Shkolnik (2024) extend this to any number of spiked eigenvectors (Quadratic Optimization Bias of Large Covariance Matrices).*

Replacing the eigenvector h with h^{JS} ensures the optimization bias is zero and bounds the discrepancy in L_2.

- *The latter is closely related to Conjecture 1 in the GPS paper.*
- *In this covariance estimation + optimization bias framework the JS estimator is optimal and has a natural shrinkage target.*

$$\hat{Q}(x) = 1 + \langle \zeta, x \rangle - \frac{1}{2} \langle x, \hat{\Sigma} x \rangle$$


z
h^{JS}
b
h

Beyond the simple

We now consider a general quadratic program (QP)

$$\min_{x \in \mathbb{R}^p} \frac{1}{2}\langle x, \hat{\Sigma} x \rangle$$
$$\mathscr{C}^\top x \leq c$$

The solution is the maximizer of the Lagrangian,

$$\hat{Q}(x) = c_0 + \langle x, \zeta \rangle - \frac{1}{2}\langle x, \hat{\Sigma} x \rangle$$

where ζ is a linear combination of the k columns $\zeta_j \in \mathbb{R}^p$ of $\mathscr{C}$,

$$\zeta = \ell_1 \zeta_1 + \cdots + \ell_k \zeta_k$$

in terms of the (right) multipliers ℓ_j.

We now consider a general quadratic program (QP)

$$\min_{x \in \mathbb{R}^p} \frac{1}{2}\langle x, \hat{\Sigma} x\rangle$$
$$\mathscr{C}^\top x \le c$$

The solution $\hat{x}$ is the maximizer of the Lagrangian,

$$\hat{Q}(x) = c_0 + \langle x, \zeta\rangle - \frac{1}{2}\langle x, \hat{\Sigma} x\rangle \, .$$

As before, of interest is the behaviour of the realized optimum $Q(\hat{x})$.

- *Again the key quantity is the optimization bias.*

Let $\mathscr{B}$ and $\mathscr{H}$ be the eigenvectors of Σ and $\hat{\Sigma}$ corresponding to the q spiked eigenvalues ($p \times q$ matrices with orthonormal columns).

Quadratic optimization bias.

$$\mathscr{E}_p(\mathscr{H}) = \frac{\mathscr{B}^\top z - (\mathscr{B}^\top \mathscr{H})(\mathscr{H}^\top z)}{\sqrt{1 - |\mathscr{H}^\top z|^2}} \qquad \Big(z = \frac{\zeta}{|\zeta|}\Big).$$

- **Now, it is a sum over the ζ_j (columns of $\mathscr{C}$).**
- *Does not depend on any of the eigenvalues.*
- *Does not depend on the non-spiked eigenvectors.*
- *Tells us what has to be estimated accurately.*
- *i.e. find roots of $\mathscr{E}_p : \mathbb{R}^{p \times q} \to \mathbb{R}^q$ (perhaps asymptotically).*
- *The quantities $\mathscr{B}^\top z$ and $\mathscr{B}^\top \mathscr{H}$ are not observed.*
- *The most natural framework is HDLSS.*

Letting ν^2 be a p-consistent estimate of $|\epsilon|^2 = |\eta - \beta|^2$ (may be obtained with just $n = 2$ weakly dependent observations), define

$$\eta^{\mathrm{JS}} = \eta \mathrm{c} + \zeta(1 - \mathrm{c}) \qquad \left(\mathrm{c} = 1 - \nu^2 J^{-1}\right)$$

where $J = (\eta - \zeta)^\top (\eta - \zeta)$.

Theorem. If $|\beta|^2/p$ and $|\epsilon|^2/p$ are bounded in $(0, \infty)$ as $p \to \infty$ and the Geometric LLN holds, $\sqrt{\mathrm{c}}$ is eventually in the interval $(0, 1)$ and,

$$|\eta^{\mathrm{JS}} - \beta| \sim \sqrt{\mathrm{c}}\,|\eta - \beta| = \sqrt{\mathrm{c}}\,|\epsilon|.$$

- *For eigenvectors we took $\eta = \mathscr{s} \times h$ and $\zeta = \langle \eta, z \rangle z$*
- *We can generalize to a $p \times q$ matrix of eigenvectors $\mathscr{H}$. Write $\boldsymbol{H}$ for $\mathscr{H}$ with columns scaled by $\sqrt{\text{eigenvalue}}$.*

Computing the pseudo-inverse $\mathscr{C}^{+} = (\mathscr{C}^{\top}\mathscr{C})^{-1}\mathscr{C}^{\top}$,

$$\begin{aligned} H^{\mathrm{JSQP}} &= HC + Z(I - C) \\ Z &= \mathscr{C}\mathscr{C}^{+}H\,, \\ C &= I - \nu^{2}J^{-1}\,, \\ J &= (H - M)^{\top}(H - M)\,. \end{aligned}$$

The theory for why $\mathscr{E}_{p}(H^{\mathrm{JSQP}}) \to 0$ is in Gurdogan & Shkolnik (2024).

How is ν computed?

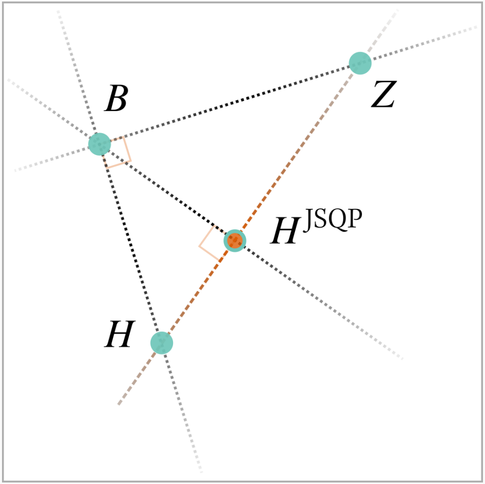
Given a centered sample covariance matrix $S = HH^\top + G$,

$$\nu^2 = \frac{\mathrm{tr}(G)}{n_+ - q} \qquad \text{(Noise)}$$

where n_+ is the number of nonzero eigenvalues of S.

- *q is the number of spikes.*

Thanks D. Hilbert.



Appendix

We are going to expand $\textsc{col}(H)$ by e and define

$$\mathscr{H}_z = \left(\mathscr{H} \qquad \frac{z - z_{\mathscr{H}}}{|z - z_{\mathscr{H}}|}\right) \qquad\qquad (z = \mathrm{e}/\sqrt{p}).$$

We try to see if a linear transformation can hit the root of $\mathscr{E}_{\boldsymbol{p}}(\cdot)$.

$$T \mapsto \mathscr{H}_z T\,, \qquad (T^\top T \in \mathbb{R}^{q\times q} \text{ invertible}).$$

This leads to the following transformation of the optimization bias.

$$\mathscr{E}_p(\mathscr{H}_z T) = \frac{\mathscr{B}^\top z - \mathscr{B}^\top \mathscr{H}_z T T^\dagger \mathscr{H}_z^\top z}{1 - |z_{\mathscr{H}_z T}|^2}$$

Slick observation: $T^\top T T^\dagger = T^\top$ (i.e., $T^\dagger = (T^\top T)^{-1} T^\top$).

This suggests we set $T = T_* = \mathscr{H}_z^\top \mathscr{B}$ above.

This leads to the following transformation of the optimization bias.

$$\mathscr{E}_p(\mathscr{H}_z T) = \frac{\mathscr{B}^\top z - \mathscr{B}^\top \mathscr{H}_z T T^\dagger \mathscr{H}_z^\top z}{1 - |z_{\mathscr{H}_z T}|^2}$$

Slick observation: $T^\top T T^\dagger = T^\top$ (i.e., $T^\dagger = (T^\top T)^{-1} T^\top$).

This suggests we set $T = T_* = \mathscr{H}_z^\top \mathscr{B}$ above.

Provided $T_*^\top T_*$ is invertible and $|z_{\mathscr{H}_z T_*}| < 1$, for $T_* = \mathscr{H}_z^\top \mathscr{B}$,

$$\mathscr{E}_p(\mathscr{H}_z T_*) = \frac{\mathscr{B}^\top z - T_*^\top \mathscr{H}_z^\top z}{1 - |z_{\mathscr{H}_z T_*}|^2} = \frac{\mathscr{B}^\top z - \mathscr{B}^\top \mathscr{H}_z \mathscr{H}_z^\top z}{1 - |z_{\mathscr{H}_z T_*}|^2} = 0\,.$$

Major obstacle – there is no way to estimate T_* from Y.

THEOREM (Gurdogan & Shkolnik 2024).

There does not exist a function $f : \mathbb{R}^{p \times n} \to \mathbb{R}^{q \times q}$ *with* $q \geq 2$ *for which,*

$$\mathscr{B}^\top \mathscr{H} \sim f(\mathrm{Y})$$

without "very strong assumptions" (e.g., $\mathrm{X}^\top \mathrm{X} = \mathrm{I}$*).*

- This does not mean we do not have limit theorems for the angles between the sample and population principal component angles.

Recall $\Psi^2 = \mathrm{I} - \kappa_p^2 \, \mathcal{S}_p^{-2}$ (constructed from sample eigenvalues).

THEOREM (Gurdogan & Shkolnik 2024).

Suppose Assumption A holds. Then, almost surely,

$$\mathscr{H}^\top \mathscr{B}\mathscr{B}^\top \mathscr{H} \sim \Psi^2 \quad \textit{and} \quad \mathscr{H}^\top z \sim (\mathscr{H}^\top \mathscr{B})\mathscr{B}^\top z\,.$$

Moreover, every diagonal entry of Ψ *is eventually in* $(0, 1)$ *wp1.*

LEMMA. *For any invertible matrix* K *we have* $\mathrm{e}_H = \mathrm{e}_{HK}$.

As a corollary, for any invertible matrix K, we also have

$$\mathscr{E}(\mathscr{H}_z T_*) = \mathscr{E}(\mathscr{H}_z \mathscr{H}_z^\top \mathscr{B}) = \mathscr{E}(\mathscr{H}_z \mathscr{H}_z^\top \mathscr{B} K)$$

A good choice turns out to be $K = \mathscr{B}^\top \mathscr{H}$.

Recap.

- We found that $\mathscr{E}_p(\mathscr{H}_z T_*) = 0$ for $T_* = \mathscr{H}_z^\top \mathscr{B}$.
- Key lemma: $\mathscr{E}_p(H) = \mathscr{E}_p(HK)$ for invertible K.
- Choosing $K = \mathscr{B}^\top \mathscr{H}$ leads to $T_{**} = T_* \mathscr{B}^\top \mathscr{H}$ with

$$T_{**} = \mathscr{H}_z \mathscr{B}\mathscr{B}^\top \mathscr{H}$$

  which may be estimated due to the PCA angles theorem

$$\mathscr{H}^\top \mathscr{B}\mathscr{B}^\top \mathscr{H} \sim \Psi^2 \quad \text{and} \quad \mathscr{H}^\top z \sim (\mathscr{H}^\top \mathscr{B})\mathscr{B}^\top z\,.$$

- These estimates lead to our $H_\sharp$ with the guarantee,

$$\lim_p \mathscr{E}_p(H_\sharp) = 0_q\,.$$

References.

Casella, G. & Hwang, J. T. (1982), 'Limit expressions for the risk of james-stein estimators', *Canadian Journal of Statistics* **10**(4), 305–309.

Efron, B. (1978), 'Controversies in the foundations of statistics', *The American Mathematical Monthly* **85**(4), 231–246.

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

Fourdrinier, D., Strawderman, W. E. & Wells, M. T. (2018), *Shrinkage estimation*, Springer.

Goldberg, L. R., Papanicolaou, A. & Shkolnik, A. (2022), 'The dispersion bias', *SIAM Journal on Financial Mathematics* **13**(2), 521–550.

Gurdogan, H. & Shkolnik, A. (2024), 'The quadratic optimization bias of large covariance matrices'.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', The Regents of the University of California.

Stein, C. (1955), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', The Regents of the University of California.

Wang, W. & Fan, J. (2017), 'Asymptotics of empirical eigenstructure for high dimensional spiked covariance', *The Annals of Statistics* **45**(3), 1342–1374.

Yao, J., Zheng, S. & Bai, Z. (2015), 'Sample covariance matrices and high-dimensional data analysis', *Cambridge UP, New York* .