Beyond James-Stein estimation for PCA.

Alex Shkolnik (shkolnik@ucsb.edu)
(joint with Youhong Lee)

6th International Conference on Statistics & Econometrics,
Waseda University, Tokyo, Japan.
August 1, 2023.

Department of Statistics & Applied Probability
University of California, Santa Barbara
Let $\langle x, y \rangle = \sum_{\ell=1}^m x_\ell y_\ell$ for $x, y \in \mathbb{R}^m$ and $|x| = \sqrt{\langle x, x \rangle}$.

\[ Y = \theta X^T + \mathcal{E} \]

(only $Y$ is observed)

$\theta \in \mathbb{R}^p$ (signal)

$X \in \mathbb{R}^n$ (latent factor)

$\mathcal{E}$ (p x n “noise” matrix)
Largest singular value decomposition involves $(s, w, v)$ with

$$Yw = (\theta X + \mathcal{E})w = sv$$

such that $s = |Yw|$ is maximized over all $w \in \mathbb{R}^n$ with $|w| = 1$. 
Largest singular value decomposition involves \((s, w, v)\) with

\[
Yw = (\theta X + \mathcal{E})w = s v
\]

such that \(s = |Yw|\) is maximized over all \(w \in \mathbb{R}^n\) with \(|w| = 1\).

- \(v \in \mathbb{R}^p\) is the left singular vector \((|v| = 1)\).
- \(w \in \mathbb{R}^n\) is the right singular vector \((|w| = 1)\).
- \(s \in \mathbb{R}\) is the largest singular value.
Largest singular value decomposition involves \((s, w, u)\) with

\[ Yw = (\theta X + \mathcal{E})w = s v \]

such that \(s = |Yw|\) is maximized over all \(w \in \mathbb{R}^n\) with \(|w| = 1\).

**GOAL.** Estimate \(\theta\) and \(X\) (upto unidentifiability).

- *Note*, \(v = \theta/|\theta|\) and \(w = X/|X|\) when \(\mathcal{E} = 0\).
- *And* \(s = |\theta|\) provided \(|X| = 1\) (w.l.o.g.).

“Behaviour” of \(\mathcal{E}\) determines the accuracy otherwise.
Asymptotics’ Horseshoe
Let $Y = \theta X^\top + \varepsilon$ regarding the length $|\theta|$ as the signal strength. Limit behaviour of the following quantity is insightful.

$$\tau_p = |\theta| \sqrt{n/p}$$

E.g., $\tau_p \to \infty$ implies consistency of many important estimates.
Let $Y = \theta X^\top + \varepsilon$ regarding the length $|\theta|$ as the signal strength. Limit behaviour of the following quantity is insightful.

$$\tau_p = |\theta| \sqrt{n/p}$$

e.g., $\tau_p \to \infty$ implies consistency of many important estimates.

**CONDITION 1.** $\tau_p$ converges in $(0, \infty)$ as $p \to \infty$.

Borrowed from Wang & Fan (2017).

Note that $n \geq 2$ may remain finite or tend to $\infty$ as well.
Our results will fail at RMT \((\sup_p |\theta| < \infty)\).
Inconsistency of the left singular vector
Let $Y = \theta X + \mathcal{E}$ and for the noise matrix $\mathcal{E}$, define

$$\Gamma = \mathcal{E}^\top \mathcal{E} / p.$$ 

**ASSUMPTION A.** Condition 1 holds with $\lim_{p \to \infty} |\theta| = \infty$. Additionally,

1. There exists $\nu$ which converges in $(0, \infty)$ and satisfies

$$|\Gamma - \nu^2 I| \leq C_p \sqrt{n/p}$$

for random sequence $C_p$ with $\limsup_p C_p < \infty$ almost surely,

2. We have $\limsup_p |\mathcal{E}^\top \theta|^2 / p < \infty$ almost surely,

3. We have $\lim_p \langle \mathcal{E}^\top \theta, X \rangle / p = 0$ almost surely,

4. $\langle X, X \rangle / n \neq 0$ almost surely for every $n \leq \infty$. 
Let $s_j$ be the $j$th singular value of $Y (= \theta X^T + \varepsilon)$.

Let $\psi$ be a function of the singular values specified as,

$$\psi^2 = 1 - \left( \sum_{j<p} \frac{s_j^2}{s_p^2} \right) \left( \frac{1 + n/p}{n - n/p - 1} \right).$$

Left singular vector $v$ estimates $u = \theta/|\theta|$ inconsistently.

**Theorem.** Suppose Assumption A holds. Then, almost surely, $\langle v, u \rangle^2 \sim \psi^2 \quad (p \uparrow \infty)$

and eventually $\psi$ is in the interval $(0, 1)$. Moreover, the singular value $s$ is also inconsistent, but the right singular vector is consistent.
Theorem 1 holds if $\tau_p = |\theta| \sqrt{n/p}$ converges, but fails at RMT (where $\sup_p |\theta| < \infty$, i.e., Johnstone’s spike model (Johnstone & Lu 2009)).
INCONSISTENCY OF THE LEFT SINGULAR VECTOR.
Some James-Stein results
The James-Stein estimator for the left singular vector is

\[ v^{JS} \propto m + c(v - m) \quad \left( c = 1 - \frac{1 - \psi^2}{1 - \langle v, q \rangle^2} \right) \]

normalized to \(|v^{JS}| = 1\) for \(m = \langle v, q \rangle q\) and any unit vector \(q\).
The James-Stein left singular vector
Let \( q \in \mathbb{R}^p \) be any (sequence) of unit length vectors such that assumption A still holds with \( q \) replacing \( u = \theta / |\theta| \).

- Let \( d^2 = \frac{1 - \psi^2}{\psi^2} \).

**Theorem 2.** Suppose Assumption A holds. Then,

\[
\langle v^{JS}, u \rangle^2 - \langle v, u \rangle^2 \sim \langle v, q \rangle^2 d^2 (1 - c)
\]

*and c converges in \((0, 1)\) almost surely as \( p \to \infty \).*

**Remarks.**

- \( v^{JS} \) improves upon the left singular vector if \( \langle v, q \rangle \neq 0 \) ev.
- The result fails as we transition into the RMT regime.
Stepping stone to a CLT theorem for James-Stein.

- Let $\kappa^2 = \lim_{p \to \infty} \langle v^{JS}, u \rangle^2$.
- Let $d^2 = \frac{1 - \psi^2}{\psi^2}$.

**Theorem 3.** Suppose Assumption A holds with $n$ fixed. Then,

$$\langle v^{JS}, u \rangle^2 - \kappa^2 \sim (cd)^2 \mathcal{R}_1 + 2(cd)(1 - c) \mathcal{R}_2 + (1 - c)^2 \mathcal{R}_3$$

where every rate $\mathcal{R}_j \to 0$ as $p \to \infty$ almost surely.

**REMARK.** Each rate $\mathcal{R}_j$ may be reasonably assumed to decay at the rate $\sqrt{p}$ (weakly) or, more finely as $\sqrt{p \log \log p}$ (LIL).
Recall, $Y = \theta X + \varepsilon$ with $\Gamma = \varepsilon^\top \varepsilon / p \sim \nu^2 I$ and $\tau_p = |\theta| \sqrt{n/p}$. The first fundamental rate with coefficient $(cd)^2$ is given by,

$$R_1 = \left( \frac{\tau_p^2 - \tau_\infty^2}{\tau_\infty^2} \right) + \left( \frac{\langle X, (\Gamma - \nu^2 I)X \rangle}{|X|^2 \nu^2} \right).$$
Recall $Y = \theta X + \mathcal{E}$ and $u = \theta / |\theta|$.

The second fundamental rate with coefficient $2(cd)(1 - c)$ is given by,

$$R_2 = \frac{\langle X, \mathcal{E}^T u_q \rangle}{|X| \nu \sqrt{p}} \quad (u_q = \frac{u - \langle u, q \rangle q}{\sqrt{1 - \langle u, q \rangle^2}}).$$
Recall $Y = \theta X + \varepsilon$ and $u = \theta/|\theta|$.

The third fundamental rate with coefficient $(1 - c)^2$ is given by,

$$R_3 = \langle u, q \rangle - \lim_{p \to \infty} \langle u, q \rangle.$$
Theorem 3. Suppose Assumption A holds with $n$ fixed. Then,

$$\langle v^{JS}, u \rangle^2 - \kappa^2 \sim (cd)^2 R_1 + 2(cd)(1 - c) R_2 + (1 - c)^2 R_3$$

where every rate $R_j \to 0$ as $p \to \infty$ almost surely.

Working out the covariances of the $R_j$ allows for a CLT.
References.