Random Matrix Theory for Modern Machine Learning: New Intuitions, Improved Methods, and Beyond: Part 3

Short Course @ Institut de Mathématiques de Toulouse, France

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

July 3rd, 2024

Outline

Definition (Affine-transformed model)

For $\mathbf{Z} \in \mathbb{R}^{p \times n}$ having i.i.d. sub-gaussian entries of zero mean and unit variance, and let $\mathbf{A} \in \mathbb{R}^{q \times n}$ and $\mathbf{C} \in \mathbb{R}^{q \times p}$ be two deterministic matrices, we say $\mathbf{X}$ is a affine transformed random matrix model

$$\mathbf{X} = \mathbf{A} + \mathbf{C}\mathbf{Z} \in \mathbb{R}^{q \times n}. \tag{1}$$

- this extends SCM, and can be used to derive results for a wide range of **linear ML** methods
- exhibit different behaviors and intuitions, on **classical** or **proportional** regime, analogous to SCMs

Table: Roadmap of linear ML models considered.

ML Problem	Classical Regime	Proportional Regime
Low rank approximation $\hat{\mathbf{X}}$ of info-plus-noise matrix $\mathbf{X}$	smooth decay of $\lVert \mathbf{X} - \hat{\mathbf{X}} \rVert_2 / \lVert \mathbf{X} \rVert_2 \simeq (1+\ell)^{-1}$ Proposition 1 Item (i)	sharp transition of $\lVert \mathbf{X} - \hat{\mathbf{X}} \rVert_2 / \lVert \mathbf{X} \rVert_2$ at $\ell = c + \sqrt{c}$ Proposition 1 Item (ii)
Classification of binary Gaussian mixtures of distance in means $\Delta\boldsymbol{\mu}$	pairwise $\simeq$ spectral approach Proposition 2 Item (i)	pairwise $\ll$ spectral approach Proposition 2 Item (ii)
Linear least squares regression risk as $n \uparrow$	bias $= 0$ and variance $\propto n^{-1}$ Proposition 3 Item (i)	monotonic bias and non-monotonic variance Proposition 3 Item (ii)

Affine-transformed model

Definition (Affine-transformed model)

For $\mathbf{Z} \in \mathbb{R}^{p \times n}$ having i.i.d. sub-gaussian entries of zero mean and unit variance, and let $\mathbf{A} \in \mathbb{R}^{q \times n}$ and $\mathbf{C} \in \mathbb{R}^{q \times p}$ be two deterministic matrices, we say $\mathbf{X}$ is an *affine transformed random matrix model*

$$\mathbf{X} = \mathbf{A} + \mathbf{C}\mathbf{Z} \in \mathbb{R}^{q \times n}. \tag{2}$$

- matrix version of an affine transformation of a vector: for $\mathbf{z} \in \mathbb{R}^p$ having independent entries of zero mean and unit variance, deterministic $\mathbf{a} \in \mathbb{R}^q$ and matrix $\mathbf{C} \in \mathbb{R}^{q \times p}$,

$$\mathbf{x} = \mathbf{a} + \mathbf{C}\mathbf{z} \in \mathbb{R}^q, \tag{3}$$

  is an affine transformation of $\mathbf{z}$ with mean $\mathbb{E}[\mathbf{x}] = \mathbf{a}$ and covariance $\mathrm{Cov}[\mathbf{x}] = \mathbf{C}\mathbf{C}^\mathsf{T} \succeq \mathbf{0}$
- due to the "structure" in $\mathbf{X}$, we shall see:

(i) the limiting eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ can significantly diverge from the Marčenko-Pastur law

(ii) depending on the dimension ratio $c = p/n$, a few eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ may isolate from the rest of eigenvalue **bulk**, for which a **phase transition** behavior can be observed

- can be assessed via the proposed **Deterministic Equivalent for resolvent** approach in a unified fashion

Deterministic Equivalents for resolvent of affine SCM

Theorem (Asymptotic Deterministic Equivalent for resolvent of affine-transformed model)

For random matrix $\mathbf{Z} \in \mathbb{R}^{p\times n}$ having i.i.d. sub-gaussian entries of zero mean and unit variance, let $\mathbf{X} = \mathbf{A} + \mathbf{C}\mathbf{Z}$ be an affine-transformed model, for deterministic $\mathbf{A} \in \mathbb{R}^{q\times n}, \mathbf{C} \in \mathbb{R}^{q\times p}$ such that $\|\mathbf{C}\|_2 \leq C$, $\|\mathbf{A}\|_2 \leq C\sqrt{n}$, and $\|\mathbf{a}_i\| \leq C$ for some universal constant $C > 0$, with $\mathbf{a}_i \in \mathbb{R}^q$ the i^{th} column of $\mathbf{A}$. Then, one has, for $z \in \mathbb{C}$ not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ and as $p, q, n \to \infty$ at the same pace, the following asymptotic Deterministic Equivalent,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = \left(\frac{\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T}}{1+\delta(z)} - z\mathbf{I}_q \right)^{-1} \tag{4}$$

for the resolvent $\mathbf{Q}(z) \equiv (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_q)^{-1}$, with $\delta(z)$ the unique Stieltjes transform solution to the fixed point equation

$$\delta(z) = \frac{1}{n} \operatorname{tr} \mathbf{C}^\mathsf{T} \bar{\mathbf{Q}}(z) \mathbf{C}. \tag{5}$$

- *For the co-resolvent $\tilde{\mathbf{Q}}(z) \equiv (\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - z\mathbf{I}_n)^{-1}$, one has instead*

$$\tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z), \quad \bar{\tilde{\mathbf{Q}}}(z) = -\frac{\mathbf{I}_n}{z(1+\delta(z))}. \tag{6}$$

Useful lemmas: recap

Lemma (Resolvent identity)

For invertible matrices $\mathbf{A}$ *and* $\mathbf{B}$, *we have* $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$.

Lemma (Woodbury)

For $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p \times n}$, *such that both* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{U}\mathbf{V}^\mathsf{T}$ *are invertible, we have*

$$(\mathbf{A} + \mathbf{U}\mathbf{V}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_n + \mathbf{V}^\mathsf{T}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^\mathsf{T}\mathbf{A}^{-1}.$$

In particular, for $n = 1$, *i.e.,* $\mathbf{U}\mathbf{V}^\mathsf{T} = \mathbf{u}\mathbf{v}^\mathsf{T}$ *for* $\mathbf{U} = \mathbf{u} \in \mathbb{R}^p$ *and* $\mathbf{V} = \mathbf{v} \in \mathbb{R}^p$, *the above identity specializes to the following Sherman–Morrison formula,*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\mathsf{T}\mathbf{A}^{-1}}{1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}, \quad \text{and } (\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T})^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}.$$

And the matrix $\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T} \in \mathbb{R}^{p \times p}$ *is invertible if and only if* $1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u} \neq 0$.

Heuristic derivation via "leave-one-out"

- propose $\bar{\mathbf{Q}} = (\mathbf{F} - z\mathbf{I}_q)^{-1}$ for some deterministic $\mathbf{F} \in \mathbb{R}^{q \times q}$ to be determined, and try to "guess" $\mathbf{F}$
- by resolvent identity

$$\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}] = \mathbb{E}\left[\mathbf{Q}\left(\mathbf{F} - \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right)\right]\bar{\mathbf{Q}} = \mathbb{E}[\mathbf{Q}]\mathbf{F}\bar{\mathbf{Q}} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\mathbf{Q}}$$
$$= \mathbb{E}[\mathbf{Q}]\mathbf{F}\bar{\mathbf{Q}} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1 + \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right]\bar{\mathbf{Q}}$$

  with $\mathbf{x}_i = \mathbf{a}_i + \mathbf{C}\mathbf{z}_i \in \mathbb{R}^q$ the i^{th} column of $\mathbf{X} \in \mathbb{R}^{q \times n}$ for $\mathbf{a}_i \in \mathbb{R}^q$ the i^{th} column of $\mathbf{A} \in \mathbb{R}^{q \times n}$ and $\mathbf{z}_i \in \mathbb{R}^p$ the i^{th} column of $\mathbf{Z}$, $\mathbf{Q}_{-i} = (\frac{1}{n}\sum_{j \neq i}\mathbf{x}_j\mathbf{x}_j^\mathsf{T} - z\mathbf{I}_p)^{-1}$ independent of $\mathbf{x}_i$,
- in the denominator

$$\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i = \frac{1}{n}(\mathbf{a}_i + \mathbf{C}\mathbf{z}_i)^\mathsf{T}\mathbf{Q}_{-i}(\mathbf{a}_i + \mathbf{C}\mathbf{z}_i) \simeq \frac{1}{n}\mathbf{a}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{a}_i + \frac{1}{n}\mathbf{z}_i^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{C}\mathbf{z}_i$$
$$\simeq \frac{1}{n}\operatorname{tr}(\mathbf{C}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{C}) \simeq \frac{1}{n}\operatorname{tr}(\mathbf{C}^\mathsf{T}\bar{\mathbf{Q}}\mathbf{C}) \equiv \delta(z),$$

- ignore the cross terms (of the form $2\mathbf{a}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{C}\mathbf{z}_i/n$, which, when conditioned on $\mathbf{Q}_{-i}$, is sub-gaussian with zero mean and variance $4\mathbf{a}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{C}\mathbf{C}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{a}_i/n^2 \leq 4n^{-2}\|\mathbf{a}_i\|^2 \cdot \|\mathbf{Q}_{-i}\|_2^2 \cdot \|\mathbf{C}\|_2^2 = O(n^{-2})$)
- approximate the term $\frac{1}{n}\mathbf{z}_i^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{C}\mathbf{z}_i$ by its expectation (e.g., Hanson-Wright) and use Deterministic Equivalent relations $\mathbf{Q}_{-i} \leftrightarrow \mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}$

Heuristic derivation via “leave-one-out”

- the Deterministic Equivalent relations $\mathbf{Q}_{-i} \leftrightarrow \mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}$ holds since

$$0 \preceq \mathbb{E}[\mathbf{Q}_{-i} - \mathbf{Q}] = \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right] \preceq \frac{1}{n}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}] = \frac{1}{n}\mathbb{E}\left[\mathbf{Q}_{-i}(\mathbf{a}_i\mathbf{a}_i^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T})\mathbf{Q}_{-i}\right], \tag{7}$$

for $\|\mathbf{a}_i\| = O(1)$ and $\|\mathbf{C}\|_2 = O(1)$.

$$\begin{aligned}
\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}] &= \mathbb{E}[\mathbf{Q}]\mathbf{F}\bar{\mathbf{Q}} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\mathbf{Q}} \simeq \mathbb{E}[\mathbf{Q}]\mathbf{F}\bar{\mathbf{Q}} - \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]}{1+\delta(z)}\bar{\mathbf{Q}} \\
&= \mathbb{E}[\mathbf{Q}]\mathbf{F}\bar{\mathbf{Q}} - \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}[\mathbf{Q}_{-i}](\mathbf{a}_i\mathbf{a}_i^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T})}{1+\delta(z)}\bar{\mathbf{Q}} \\
&\simeq \mathbb{E}[\mathbf{Q}]\left(\mathbf{F} - \frac{\frac{1}{n}\sum_{i=1}^n(\mathbf{a}_i\mathbf{a}_i^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T})}{1+\delta(z)}\right)\bar{\mathbf{Q}} \\
&= \mathbb{E}[\mathbf{Q}]\left(\mathbf{F} - \frac{\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T}}{1+\delta(z)}\right)\bar{\mathbf{Q}}
\end{aligned}$$

- independence between $\mathbf{Q}_{-i}$ and $\mathbf{x}_i$ in the third line
- to have $\mathbb{E}[\mathbf{Q}] \simeq \bar{\mathbf{Q}}$, just take $\mathbf{F} = \frac{\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T}}{1+\delta(z)}$

Remark: on the low-rankness of $\mathbf{A}$

- we consider $\mathbb{E}[\mathbf{X}] = \mathbf{A} \in \mathbb{R}^{q \times n}$ satisfies (i) $\|\mathbf{A}\|_2 \leq C\sqrt{n}$ and (ii) $\|\mathbf{a}_i\| \leq C$ for all $i \in \{1, \ldots, n\}$, $\mathbf{a}_i \in \mathbb{R}^q$ the i-th column of $\mathbf{A} \in \mathbb{R}^{q \times n}$, and some constant $C > 0$

(i) the first is just proper scaling, so that $\|\mathbf{A}\|_2$ and $\|\mathbf{CZ}\|_2$ are of the same order

(ii) the second bound on the Euclidean norm of *all* columns of $\mathbf{A}$ is more subtle: taking $\|\mathbf{A}\|_2 = C_1\sqrt{n}$ and $\|\mathbf{a}_i\| = C_{2,i}$ for $C_1, C_{2,i} > 0$,

$$\sum_{i=1}^{n} \|\mathbf{a}_i\|^2 = \sum_{i=1}^{n} C_{2,i}^2 = \|\mathbf{A}\|_F^2 = \sum_{i=1}^{\operatorname{rank}(\mathbf{A})} \sigma_i^2(\mathbf{A}) = \Theta(n) \tag{8}$$

with $\sigma_1(\mathbf{A}) \geq \ldots \geq \sigma_{\operatorname{rank}(\mathbf{A})}(\mathbf{A})$ the (nonzero) singular values of $\mathbf{A}$ arranged in a non-increasing order. Since $\sigma_1^2(\mathbf{A}) = \|\mathbf{A}\|_2^2 = \Theta(n)$, the following two typical scenarios:

(1) $\operatorname{rank}(\mathbf{A}) = \Theta(n)$, a majority (of size $\Theta(n)$) of singular values $\sigma_i(\mathbf{A}) = O(1)$, so that the matrix $\mathbf{A}$ has a fast decay in its singular values; or

(2) $\operatorname{rank}(\mathbf{A}) = \Theta(1)$, a few singular values $\sigma_i(\mathbf{A}) = \Theta(n)$, and $\mathbf{A}$ is exactly of low rank.

- This is in consistent with common ML assumptions, e.g., that the data are drawn from one or a mixture (when in a classification context) of distributions, and the mean $\mathbf{A}$ is of low rank.
- existing RMT results, e.g., on spiked model [BS06; BGN11], mostly focuses on exactly low rank $\mathbf{A}$.
- However, if one further relaxes the assumption $\|\mathbf{a}_i\| = O(1)$ and let $\mathbf{A}$ have a slow singular decay, the result collapses.

Remark: Stieltjes transform can not capture few important eigenvalues

Lemma ([SB95, Lemma 2.6])

For $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p \times p}$ symmetric and nonnegative definite, $\mathbf{u} \in \mathbb{R}^p$, $\tau > 0$ and $z < 0$,

$$\left| \operatorname{tr} \mathbf{A}(\mathbf{M} + \tau \mathbf{u}\mathbf{u}^\mathsf{T} - z\mathbf{I}_p)^{-1} - \operatorname{tr} \mathbf{A}(\mathbf{M} - z\mathbf{I}_p)^{-1} \right| \leq \frac{\|\mathbf{A}\|_2}{|z|}.$$

- for low-rank $\mathbf{A}$, $\delta(z)$ is asymptotically independent on $\mathbf{A}$.

$$\delta(z) = \frac{1}{n} \operatorname{tr} \mathbf{C}\mathbf{C}^\mathsf{T} \left(\frac{\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{C}\mathbf{C}^\mathsf{T}}{1+\delta(z)} - z\mathbf{I}_q \right)^{-1} = \frac{1}{n} \operatorname{tr} \mathbf{C}\mathbf{C}^\mathsf{T} \left(\frac{\mathbf{C}\mathbf{C}^\mathsf{T}}{1+\delta(z)} - z\mathbf{I}_q \right)^{-1} + O(n^{-1}). \tag{9}$$

- same holds for $\frac{1}{q} \operatorname{tr} \bar{\mathbf{Q}}(z) = \frac{1}{q} \operatorname{tr} \left(\frac{\mathbf{C}\mathbf{C}^\mathsf{T}}{1+\delta(z)} - z\mathbf{I}_q \right)^{-1} + O(n^{-1})$ for n, p, q large
- while the Deterministic Equivalent $\bar{\mathbf{Q}}(z)$ is itself dependent on $\mathbf{A}$, its normalized trace is NOT
- this independence of $\delta(z)$ and $\frac{1}{q} \operatorname{tr} \bar{\mathbf{Q}}(z)$ on $\mathbf{A}$ is also a limitation of the Stieltjes transform approach, does **not** allow for a characterization of a negligible proportion (of order $o(n)$) of eigenvalues (e.g., due to $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T}$).
- **contrasts with** Deterministic Equivalents approach: $\mathbf{Q}(z)$ and $\tilde{\mathbf{Q}}(z)$ remain dependent on $\mathbf{A}$, and thus can capture the influence of the low rank $\mathbf{A}$

Remarks

Remark (DE-SCM as a corollary of the Linear Master Theorem)

The Deterministic Equivalents for resolvents of SCM, can be derived from our Linear Master Theorem above: Taking $q = p$, $c = p/n$, $\mathbf{A} = \mathbf{0}$ *and* $\mathbf{C} = \mathbf{I}_p$,

$$\bar{\mathbf{Q}}(z) = \frac{1}{-z + \frac{1}{1+cm(z)}} \mathbf{I}_p \equiv m(z)\mathbf{I}_p, \tag{10}$$

where we denote $m(z) \equiv \frac{1}{p} \operatorname{tr} \bar{\mathbf{Q}}(z)$ *that satisfies the following quadratic equation*

$$czm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{11}$$

Table: Overview of upcoming results, illustrating the connection between the Linear Master Theorem different random matrix models, and applications.

$\mathbf{A}$	$\mathbf{C}$	z	RMT results	Related ML applications
$\mathbf{0}$	$\mathbf{I}_p$	complex	Distribution of eigenvalues (Marčenko-Pastur law)	Previous results on SCM
low rank	$\mathbf{I}_p$	complex	Extreme eigenvalues (Additive spiked eigenvalues in Theorem 12)	Low rank approximation
low rank	$\mathbf{I}_p$	complex	Extreme eigenvectors (Info-plus-noise spiked eigenvectors in Theorem 10)	Classification
$\mathbf{0}$	$\mathbf{I}_p$	real	Resolvent matrix (Deterministic Equivalent in Theorem 3)	Linear least squares

Information-plus-noise spiked model

- $\mathbf{C} = \mathbf{I}_p$, random matrix $\mathbf{Z}$ for homogeneous "**noise**", and $\mathbf{A} \in \mathbb{R}^{p \times n}$ informative "**signal**" matrix, low rank

Definition (Information-plus-noise spiked model)

We say a symmetric random matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$ follows an *information-plus-noise spiked model* if

$$\mathbf{X} = \frac{1}{n}(\mathbf{A} + \mathbf{Z})(\mathbf{A} + \mathbf{Z})^\mathsf{T}, \tag{12}$$

for some *deterministic* matrix $\mathbf{A} \in \mathbb{R}^{p \times n}$ and random matrix $\mathbf{Z} \in \mathbb{R}^{p \times n}$ with $\mathbb{E}[\mathbf{Z}] = \mathbf{0}$.

- determine when the "information in $\mathbf{A}$ can be "found," and when it is "lost" due to the noise in $\mathbf{Z}$
- for $\mathbf{A} \neq \mathbf{0}$, expect a few eigenvalues "jumping" out the Marčenko-Pastur support (due to $\mathbf{A}$, refer to as the spikes) and isolate from the main eigenvalue bulk $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$

$$\frac{1}{n}\mathbb{E}[(\mathbf{A} + \mathbf{Z})(\mathbf{A} + \mathbf{Z})^\mathsf{T}] = \frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \frac{1}{n}\mathbb{E}[\mathbf{Z}\mathbf{Z}^\mathsf{T}] = \frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{I}_p \tag{13}$$

- so for $n \gg p$, the information-plus-noise spiked model $\frac{1}{n}(\mathbf{A} + \mathbf{Z})(\mathbf{A} + \mathbf{Z})^\mathsf{T}$ is close to $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{I}_p$, the largest r eigenvalues are $1 + \lambda_i(\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T})$
- in the case of $n \sim p \gg 1$ both large, expects the top eigenvalues/eigenvectors of $\frac{1}{n}(\mathbf{A} + \mathbf{Z})(\mathbf{A} + \mathbf{Z})^\mathsf{T}$ still somewhat relates to those of $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T}$

Eigenvalue characterization for the information-plus-noise spiked model

- already know that if $\mathbf{Z} \in \mathbb{R}^{p \times n}$ is a random matrix having i.i.d. entries of zero mean and unit variance, then as $n, p \to \infty$, the limiting eigenvalue distribution of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ is the Marčenko-Pastur law
- it does not guarantee that **no eigenvalue** lies outside of the support of the Marčenko-Pastur law (i.e., outside the interval $[(1-\sqrt{c})^2, (1+\sqrt{c})^2])$
- e.g., only states that the averaged number of eigenvalues of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ lying within $[a,b] \subset [(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ converges to $\mu([a,b])$—more precisely, is of the order $p \times \mu([a,b]) + o(p)$
- remains unclear, e.g., **whether there could be a number of order** $o(p)$ **"leaking"** from the limiting Marčenko-Pastur support $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$, even for n, p sufficiently large

Theorem ("No eigenvalue outside the support" in the absence of information, [BS98])

Let $\mathbf{X}_{\mathbf{A}=\mathbf{0}}$ be the information-plus-noise spiked model with $\mathbf{A} = \mathbf{0}$, and random noise matrix $\mathbf{Z} \in \mathbb{R}^{p \times n}$ having independent entries of zero mean, unit variance, and κ-kurtosis, then as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, with probability one, the empirical spectral measure $\mu_{\mathbf{X}_{\mathbf{A}=\mathbf{0}}}$ of $\mathbf{X}_{\mathbf{A}=\mathbf{0}}$, converges weakly to the Marčenko-Pastur law and

(i) *if $\kappa < \infty$, then*

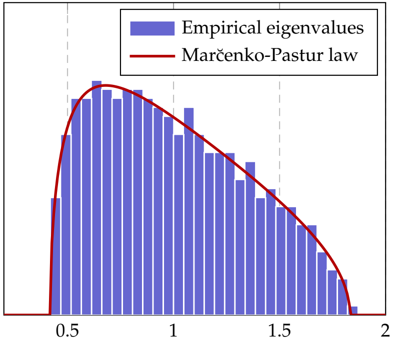
$$\lambda_{\min}\left(\mathbf{X}_{\mathbf{A}=\mathbf{0}}\right) \to (1-\sqrt{c})^2, \quad \lambda_{\max}\left(\mathbf{X}_{\mathbf{A}=\mathbf{0}}\right) \to (1+\sqrt{c})^2 \tag{14}$$

that is, no eigenvalue of $\mathbf{X}_{\mathbf{A}=\mathbf{0}} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ appears outside the limiting Marčenko-Pastur support; and

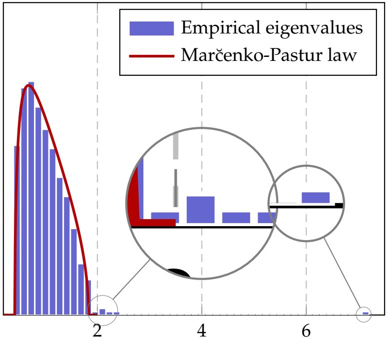
(ii) *if $\kappa = \infty$, then*

$$\lambda_{\max}\left(\mathbf{X}_{\mathbf{A}=\mathbf{0}}\right) \to \infty. \tag{15}$$

Eigenvalue characterization for the information-plus-noise spiked model



(a) Gaussian $\mathbf{Z}$



(b) Student-t $\mathbf{Z}$ with degree of freedom three

Figure: Eigenvalue distribution of sample covariance matrix $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ for Gaussian (**left**) and Student-t (**right**) $\mathbf{Z}$, versus the *same* limiting Marčenko-Pastur law, with $p = 512$ and $n = 8p$.

(i) in the Gaussian case (**left**), no eigenvalue outside the Marčenko-Pastur support; and

(ii) in the Student-t case (**right**), a few eigenvalues are observed to "leak" from the Marčenko-Pastur support, even in the noise -only model with $\mathbf{A} = \mathbf{0}$, in line with the "no eigenvalue outside the support" result

Eigenvalue characterization for the information-plus-noise spiked model

Theorem (Information-plus-noise spiked eigenvalues, [BS06])

Let $\mathbf{Z} \in \mathbb{R}^{p\times n}$ be a random matrix having i.i.d. sub-gaussian entries of zero mean and unit variance, and let $\mathbf{A} \in \mathbb{R}^{p\times n}$ be a deterministic matrix of rank r with $\|\mathbf{A}\| \le C\sqrt{n}$ for some constants $r, C > 0$. Then, for $\mathbf{X} = \mathbf{A} + \mathbf{Z} \in \mathbb{R}^{p\times n}$ and $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T} = \sum_{i=1}^r \ell_i \mathbf{u}_i \mathbf{u}_i^\mathsf{T}$ the spectral decomposition of $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T}$, one has, as $n, p \to \infty$ with $p/n \to c \in (0,\infty)$, that

$$\lambda_i\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right) \to \bar{\lambda}_i = \begin{cases} 1 + c + \ell_i + \frac{c}{\ell_i}, & \ell_i > \sqrt{c} \\ (1+\sqrt{c})^2 \equiv E_+, & \ell_i \le \sqrt{c}. \end{cases} \tag{16}$$

almost surely, for $\lambda_i(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})$ and ℓ_i the i^{th} largest eigenvalue of the information-plus-noise spiked model $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ in Theorem 7 and of $\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T}$, respectively.

---

[1] Jinho Baik and Jack W. Silverstein. "Eigenvalues of large sample covariance matrices of spiked population models". In: *Journal of Multivariate Analysis* 97.6 (2006), pp. 1382–1408

Proof using the Linear Master Theorem

- it follows from Woodbury identity the following Deterministic Equivalent holds

$$
\begin{aligned}
\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) &= \left(\frac{\frac{1}{n}\mathbf{A}\mathbf{A}^\top + \mathbf{I}_p}{1+\delta(z)} - z\mathbf{I}_p \right)^{-1} \\
&= \frac{1+\delta(z)}{1-z-z\delta(z)} \left(\mathbf{I}_p - \mathbf{U} \left((1-z-z\delta(z))\mathbf{L}^{-1} + \mathbf{I}_r \right)^{-1} \mathbf{U}^\top \right). \qquad (17)
\end{aligned}
$$

- here, $\frac{1}{n}\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{L}\mathbf{U}^\top = \sum_{i=1}^r \ell_i \mathbf{u}_i \mathbf{u}_i^\top$ is the spectral decomposition of $\frac{1}{n}\mathbf{A}\mathbf{A}^\top$, for $\{\ell_i\}_{i=1}^r$ the (non-zero) eigenvalue, $\mathbf{u}_i \in \mathbb{R}^p$ the corresponding eigenvectors, and $\delta(z)$ the unique valid Stieltjes transform solution to the quadratic equation

$$
z\delta^2(z) - (1-c-z)\delta(z) + c = 0. \qquad (18)
$$

- To locate a possibly **isolated** eigenvalue of the information-plus-noise random matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$ outside the Marčenko-Pastur support, we are looking for $z \in \mathbb{R}$ such that $\delta(z)$ in Equation (18) is well defined (so that it is "**outside**" the limiting bulk) **but** the Deterministic Equivalent $\bar{\mathbf{Q}}(z)$ in Equation (17) is **undefined** (so that z is an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$).
- check that $\delta(z) = z^{-1} - 1$ is not a solution to Equation (18), so that the denominator of $\bar{\mathbf{Q}}(z)$ is not zero, and the real z that we are looking for must satisfy

$$
z(1+\delta(z)) = 1 + \ell_i. \qquad (19)
$$

Proof using the Linear Master Theorem

Location of spiked eigenvalues: real z such that

$$\boxed{z(1+\delta(z)) = 1+\ell_i.} \tag{20}$$

- determine the condition under which this equation has a solution: for $z \in \mathbb{R}$ the function $z\delta(z) = \int \frac{z}{t-z}\mu(dt)$ is increasing on its domain of definition and

$$\lim_{z \downarrow (1+\sqrt{c})^2} z(1+\delta(z)) = 1+\sqrt{c}. \tag{21}$$

- admits a solution (that corresponds to an isolated eigenvalue) if and only if

$$\ell_i \geq \sqrt{c}. \tag{22}$$

- Plugging back, this leads to the following explicit solution

$$\boxed{z = 1+\ell_i + c + \frac{c}{\ell_i} \geq (1+\sqrt{c})^2}. \tag{23}$$

Phase transition in spiked eigenvalues

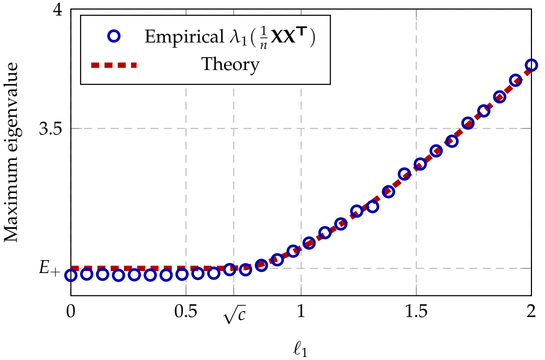


Figure: Phase transition behavior of the largest eigenvalue $\lambda_1(\mathbf{X}\mathbf{X}^\mathsf{T}/n)$ of the information-plus-noise model $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, as a function of ℓ_1, with $\mathbf{X} = \mathbf{A} + \mathbf{Z}$, $\mathbf{A} = \sqrt{\ell_1} \cdot \mathbf{u}_1 \mathbf{1}_n^\mathsf{T}$ for $\|\mathbf{u}_1\| = 1$, so that $\lambda_1(\mathbf{A}\mathbf{A}^\mathsf{T}/n) = \ell_1$, for $p = 512$ and $n = 1\,024$.

Phase transition: depending on "signal strength" $\ell_1 = \|\frac{1}{n}\mathbf{A}\mathbf{A}^\mathsf{T}\|_2$,

(i) if $\ell_1 \leq \sqrt{c}$: largest eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ asymptotically the same as $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ and independent of ℓ_1

(ii) if $\ell_1 > \sqrt{c}$: larger than that of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$, and increases as ℓ_1 becomes large

Eigenvector characterization for the information-plus-noise spiked model

Theorem (Information-plus-noise spiked eigenvectors, [Pau07])

In the setting of Theorem 9, assume that the eigenvalues ℓ_i of $\frac1n \mathbf{A}\mathbf{A}^\mathsf{T}$ are all distinct and satisfy $\ell_1 > \ldots > \ell_r > 0$, and let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r$ be the eigenvectors associated with the r largest eigenvalues $\lambda_1(\frac1n \mathbf{X}\mathbf{X}^\mathsf{T}) > \ldots > \lambda_r(\frac1n \mathbf{X}\mathbf{X}^\mathsf{T})$ of the information-plus-noise model $\frac1n \mathbf{X}\mathbf{X}^\mathsf{T}$. Then, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ deterministic vectors of unit norm,

$$\mathbf{a}^\mathsf{T} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\mathsf{T} \mathbf{b} \to \eta_i = \begin{cases} \frac{1 - c\ell_i^{-2}}{1 + c\ell_i^{-1}} \cdot \mathbf{a}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{b}, & \ell_i > \sqrt{c}; \\ 0, & \ell_i \leq \sqrt{c}. \end{cases} \tag{24}$$

almost surely as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, for $\mathbf{u}_i$ the eigenvector associated with ℓ_i of $\frac1n \mathbf{A}\mathbf{A}^\mathsf{T}$. In particular, taking $\mathbf{a} = \mathbf{b} = \mathbf{u}_i$ leads to

$$(\hat{\mathbf{u}}_i^\mathsf{T} \mathbf{u}_i)^2 \to \eta_i = \begin{cases} \frac{1 - c\ell_i^{-2}}{1 + c\ell_i^{-1}}, & \ell_i > \sqrt{c}; \\ 0, & \ell_i \leq \sqrt{c}. \end{cases} \tag{25}$$

[2] Debashis Paul. "Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model". In: *Statistica Sinica* 17.4 (2007), pp. 1617–1642

Proof using the Linear Master Theorem

- consider the i^{th} eigenvalue ℓ_i of $\frac{1}{n}\mathbf{A}\mathbf{A}^\top$ that satisfies $\ell_i > \sqrt{c}$ above the phase transition threshold
- by Cauchy's integral formula

$$\mathbf{a}^\top \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top \mathbf{b} = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i}} \mathbf{a}^\top \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top - z\mathbf{I}_p \right)^{-1} \mathbf{b}\, dz \tag{26}$$

for Γ_{λ_i} a positively oriented contour enclosing only the i^{th} eigenvalue of $\lambda_i(\frac{1}{n}\mathbf{X}\mathbf{X}^\top)$

- according to Theorem 9, this converges almost surely to $\bar{\lambda}_i = 1 + c + \ell_i + \frac{c}{\ell_i}$ as $n, p \to \infty$
- by our Linear Master Theorem

$$\begin{aligned}
\mathbf{a}^\top \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top - z\mathbf{I}_p \right)^{-1} \mathbf{b} &\simeq \frac{1+\delta(z)}{1-z-z\delta(z)} \mathbf{a}^\top \left(\mathbf{I}_p - \mathbf{U} \left((1-z-z\delta(z))\mathbf{L}^{-1} + \mathbf{I}_r \right)^{-1} \mathbf{U}^\top \right) \mathbf{b} \\
&= \frac{1+\delta(z)}{1-z-z\delta(z)} \mathbf{a}^\top \mathbf{b} - \frac{1+\delta(z)}{1-z-z\delta(z)} \sum_{j=1}^{r} \frac{\mathbf{a}^\top \mathbf{u}_j \mathbf{u}_j^\top \mathbf{b}}{1 + (1-z-z\delta(z))\ell_j^{-1}}
\end{aligned}$$

with $\frac{1}{n}\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{L}\mathbf{U}^\top = \sum_{i=1}^{r} \ell_i \mathbf{u}_i \mathbf{u}_i^\top$ the spectral decomposition of $\frac{1}{n}\mathbf{A}\mathbf{A}^\top$, and $\delta(z)$ unique solution to

$$z\delta^2(z) - (1-c-z)\delta(z) + c = 0. \tag{27}$$

- $\frac{1+\delta(z)}{1-z-z\delta(z)}\mathbf{a}^\top\mathbf{b}$ has **no pole** outside the Marčenko-Pastur support (i.e., the denominator $1 - z - z\delta(z) \neq 0$).

Proof using the Linear Master Theorem

- we further deduce that

$$\mathbf{a}^\mathsf{T} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\mathsf{T} \mathbf{b} \simeq \frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i}} \frac{1+\delta(z)}{1-z-z\delta(z)} \frac{\mathbf{a}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{b}}{1+(1-z-z\delta(z))\ell_i^{-1}} dz, \tag{28}$$

  which has a **pole** satisfying $1+(1-z-z\delta(z))\ell_i^{-1}=0$ and corresponds to spike location $z=\bar{\lambda}_i$ above
- one can evaluate the above expression by residue calculus at $z=\bar{\lambda}_i$ as

$$\begin{aligned} \mathbf{a}^\mathsf{T} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\mathsf{T} \mathbf{b} &\simeq \mathbf{a}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{b} \cdot \lim_{z\to\bar{\lambda}_i} \frac{(z-\bar{\lambda}_i)(1+\delta(z))}{(1-z-z\delta(z))+(1-z-z\delta(z))^2\ell_i^{-1}} \\ &= \frac{1+\delta(\bar{\lambda}_i)}{1+\delta(\bar{\lambda}_i)+\bar{\lambda}_i\delta'(\bar{\lambda}_i)} \cdot \mathbf{a}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{b}, \end{aligned}$$

  by L'Hôpital's rule, where we denote $\delta'(z)$ the derivative of $\delta(z)$ with respect to z, given by

$$\delta'(z) = \frac{\delta(z)(1+\delta(z))}{1-c-z-2z\delta(z)}. \tag{29}$$

- This is $\boxed{\mathbf{a}^\mathsf{T} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\mathsf{T} \mathbf{b} \to \frac{1-c\ell_i^{-2}}{1+c\ell_i^{-1}} \cdot \mathbf{a}^\mathsf{T} \mathbf{u}_i \mathbf{u}_i^\mathsf{T} \mathbf{b}.}$

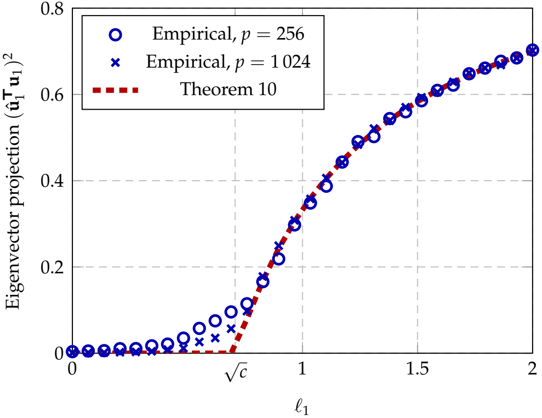


Figure: Phase transition behavior of the eigenvector projection $(\hat{\mathbf{u}}_1^\mathsf{T}\mathbf{u}_1)^2$ of the top eigenvector $\hat{\mathbf{u}}_i$ associated with the largest eigenvalue of the information-plus-noise model $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, as a function of ℓ_1, with $\mathbf{X} = \mathbf{A} + \mathbf{Z}$, $\mathbf{A} = \sqrt{\ell_1}\mathbf{u}_1\mathbf{1}_n^\mathsf{T}$ for $\|\mathbf{u}_1\| = 1$, so that $\lambda_1(\mathbf{A}\mathbf{A}^\mathsf{T}/n) = \ell_1$, for different values of p, n with $n = 2p$.

(i) empirical transitions for $p = 256, 1\,024$ not sharp, $\mathbf{u}_1^\mathsf{T}\hat{\mathbf{u}}_1 > 0$ even below threshold $\ell_1 \leq \sqrt{c}$;
(ii) become **closer** to the limiting theoretical one as the dimensions n, p grow large

The additive spiked model

Definition (Additive spiked model)

We say a symmetric random matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$ follows an *additive spiked model* if

$$\mathbf{X} = \mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}, \tag{30}$$

for some *deterministic* symmetric matrix $\mathbf{B} \in \mathbb{R}^{p \times p}$ and random matrix $\mathbf{Z} \in \mathbb{R}^{p \times n}$ with $\mathbb{E}[\mathbf{Z}] = \mathbf{0}$.

- useful (and low rank) information $\mathbf{B}$ buried by random symmetric noise matrix $\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}$
- of interest in low-rank approximation of noise matrices for data science applications of, e.g., recommendation system or LoRA technique in Large Language Models (LLMs) [Hu+21]

[3] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations.* Oct. 2021

Eigenvalue characterization for the information-plus-noise spiked model

- recall from "no eigenvalue outside the support" that in the absence of the additive term $\mathbf{B} = \mathbf{0}$ and sub-gaussian $\mathbf{Z}$, no eigenvalue of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}$ is outside the Marčenko-Pastur support

Theorem (Additive spiked eigenvalues, [BGN11])

Let $\mathbf{Z} \in \mathbb{R}^{p \times n}$ be a random matrix having i.i.d. sub-gaussian entries of zero mean and unit variance, and let $\mathbf{B} \in \mathbb{R}^{p \times p}$ be a symmetric deterministic matrix of rank r with $\|\mathbf{B}\|_2 \leq C$ for some constants $r, C > 0$. Then, for additive spiked model $\mathbf{X} = \mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^{\mathsf{T}} \in \mathbb{R}^{p \times p}$ in Theorem 11 with symmetric $\mathbf{B} = \sum_{i=1}^{r} \ell_i \mathbf{u}_i \mathbf{u}_i^{\mathsf{T}}$ the spectral decomposition of $\mathbf{B}$, one has, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, that

$$\lambda_i(\mathbf{X}) \to \bar{\lambda}_i = \begin{cases} 1 + \ell_i + \frac{c}{\ell_i - c}, & \ell_i > c + \sqrt{c} \\ (1 + \sqrt{c})^2, & \ell_i \leq c + \sqrt{c}. \end{cases} \tag{31}$$

almost surely, for $\lambda_i(\mathbf{X})$ and ℓ_i the i^{th} largest eigenvalue of the additive spiked model $\mathbf{X}$ and of $\mathbf{B}$, respectively.

[4] Florent Benaych-Georges and Raj Rao Nadakuditi. "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1 (2011), pp. 494–521

Proof using the Linear Master Theorem

- to locate a possibly isolated eigenvalue of $\mathbf{X}$ outside the (limiting) Marčenko-Pastur support (of the eigenvalues of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$), look for $z \in \mathbb{R}$ solution to the following determinant equation

$$0 = \det\left(\mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p\right) = \det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p\right) \cdot \det\left(\mathbf{I}_p + \mathbf{Q}(z)\mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T}\right). \tag{32}$$

- Here, $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ is the resolvent of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$, and $\mathbf{B} = \mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T}$ is the spectral decomposition of $\mathbf{B}$, with $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{p \times r}$ and $\mathbf{L} = \operatorname{diag}\{\ell_i\}_{i=1}^r$
- looking for $z \in \mathbb{R}$ outside the main bulk, so that $\mathbf{Q}(z)$ is well defined and $\det \mathbf{Q}^{-1}(z) \neq 0$,

$$0 = \det\left(\mathbf{I}_p + \mathbf{Q}(z)\mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T}\right) \Leftrightarrow 0 = \det\left(\mathbf{I}_r + \mathbf{L}\mathbf{U}^\mathsf{T}\mathbf{Q}(z)\mathbf{U}\right), \tag{33}$$

- apply the Linear Master Theorem to approximate

$$\mathbf{U}^\mathsf{T}\mathbf{Q}(z)\mathbf{U} \simeq \mathbf{U}^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{U} = m(z)\mathbf{I}_r, \tag{34}$$

with $m(z)$ the unique Stieltjes transform solution to the Marčenko-Pastur equation,

$$0 = \det\left(\mathbf{I}_p + \mathbf{Q}(z)\mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T}\right) \Leftrightarrow 0 = \det(\mathbf{I}_r + m(z)\mathbf{L}) \leftrightarrow \boxed{m(z) = -\ell_i^{-1}}. \tag{35}$$

Proof using the Linear Master Theorem

Spiked eigenvalues $z \in \mathbb{R}$ such that $\boxed{m(z) = -\ell_i^{-1}}$.

- Since $m(z) = \int \frac{\mu(dt)}{t-z}$ is an increasing function of z on its domain of definition and

$$\lim_{z \downarrow (1+\sqrt{c})^2} m(z) = -\frac{1}{c+\sqrt{c}}, \tag{36}$$

  the equation $m(z) = -\ell_i^{-1}$ admits a solution *if and only if*

$$\ell_i > c + \sqrt{c}, \tag{37}$$

  with explicit solution (and therefore the spike location)

$$\boxed{z = 1 + \ell_i + \frac{c}{\ell_i - c} \geq (1+\sqrt{c})^2.} \tag{38}$$

Comparison of spiked eigenvalues for information-plus-noise versus additive model

- for **information-plus-noise spiked model** $\mathbf{X} = \frac{1}{n}(\mathbf{A}+\mathbf{Z})(\mathbf{A}+\mathbf{Z})^{\mathsf{T}}$:

$$\lambda_i(\mathbf{X}) \to \bar{\lambda}_i = 1 + c + \ell_i + \frac{c}{\ell_i}, \quad \ell_i > \sqrt{c}, \quad \ell_i = \lambda_i\left(\frac{1}{n}\mathbf{A}\mathbf{A}^{\mathsf{T}}\right); \tag{39}$$

- for **additive spiked model** $\mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}$:

$$\lambda_i(\mathbf{X}) \to \bar{\lambda}_i = 1 + \ell_i + \frac{c}{\ell_i - c}, \quad \ell_i > c + \sqrt{c}, \quad \ell_i = \lambda_i(\mathbf{B}); \tag{40}$$

- connected via the "change-of-variable" $\lambda_i(\mathbf{A}\mathbf{A}^{\mathsf{T}}/n) + c \sim \lambda_i(\mathbf{B})$ with $c = p/n$, in the sense that:

(i) the phase transition condition is $\lambda_i(\mathbf{A}\mathbf{A}^{\mathsf{T}}/n) \geq \sqrt{c}$ for the information-plus-noise model and $\lambda_i(\mathbf{B}) \geq c + \sqrt{c}$ for the additive model; and

(ii) above phase transition, the isolated eigenvalues of the information-plus-noise model are given by $1 + c + \lambda_i(\mathbf{A}\mathbf{A}^{\mathsf{T}}/n) + c/\lambda_i(\mathbf{A}\mathbf{A}^{\mathsf{T}}/n)$, while those of the additive model are given by $1 + \lambda_i(\mathbf{B}) + c/\left(\lambda_i(\mathbf{B}) - c\right)$.

Theorem (Additive spiked eigenvectors, [BGN11])

In the setting of Theorem 12, assume that the eigenvalues ℓ_i of $\mathbf{B}$ are all distinct and satisfy $\ell_1 > \ldots > \ell_r > 0$, and let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r$ be the eigenvectors associated with the r largest eigenvalues $\lambda_1(\mathbf{X}) > \ldots > \lambda_r(\mathbf{X})$ of the additive model $\mathbf{X} = \mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$,

$$(\hat{\mathbf{u}}_i^\mathsf{T} \mathbf{u}_i)^2 \to \eta = \begin{cases} 1 - \frac{c}{(\ell_i - c)^2}, & \ell_i > c + \sqrt{c} \\ 0, & \ell_i \leq c + \sqrt{c}. \end{cases} \tag{41}$$

almost surely, for $\mathbf{u}_i$ the eigenvector associated with the eigenvalue ℓ_i of $\mathbf{B}$.

---

[5] Florent Benaych-Georges and Raj Rao Nadakuditi. “The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices”. In: *Advances in Mathematics* 227.1 (2011), pp. 494–521

Proof using the Linear Master Theorem

- follow the same line of arguments as in the proof of information-plus-noise spiked model
- write, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of unit norm,

$$\mathbf{a}^\mathsf{T} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\mathsf{T} \mathbf{b} = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i}} \mathbf{a}^\mathsf{T} \left(\mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p \right)^{-1} \mathbf{b} \, dz, \tag{42}$$

for Γ_{λ_i} a positively oriented contour enclosing only the i^{th} eigenvalue of $\mathbf{X} = \mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ (that admits the almost sure limit $\bar{\lambda}_i = 1 + \ell_i + \frac{c}{\ell_i - c}$)
- let $\mathbf{B} = \mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T} = \sum_{i=1}^r \ell_i \mathbf{u}_i \mathbf{u}_i^\mathsf{T}$ be the spectral decomposition of $\mathbf{B}$, then

$$\mathbf{a}^\mathsf{T} \left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p + \mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{b} = \mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{b} - \mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{U}(\mathbf{L}^{-1} + \mathbf{U}^\mathsf{T}\mathbf{Q}(z)\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}\mathbf{Q}(z)\mathbf{b},$$

with $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p)^{-1}$
- applying the Deterministic Equivalent result $\mathbf{Q}(z) \leftrightarrow m(z)\mathbf{I}_p$

$$\mathbf{a}^\mathsf{T} \left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p + \mathbf{U}\mathbf{L}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{b} \simeq m(z)\mathbf{a}^\mathsf{T}\mathbf{b} - m^2(z)\mathbf{a}^\mathsf{T}\mathbf{U} \left(m(z)\mathbf{I}_r + \mathbf{L}^{-1} \right)^{-1} \mathbf{U}^\mathsf{T}\mathbf{b},$$

with $m(z)$ unique solution to

$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{43}$$

- the first term $m(z)\mathbf{a}^\mathsf{T}\mathbf{b}$ has no pole outside the Marčenko-Pastur support

- So

$$\mathbf{a}^\mathsf{T}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}\mathbf{b} \simeq \frac{1}{2\pi\imath}\oint_{\Gamma_{\lambda_i}} \frac{m^2(z)\cdot\mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b}}{m(z)+\ell_i^{-1}}dz. \tag{44}$$

- This has a pole satisfying $m(z) = -\ell_i^{-1}$ and corresponds to spike location at $z = \bar{\lambda}_i = 1 + \ell_i + \frac{c}{\ell_i - c}$ characterized in Theorem 12.
- evaluate this expression by the residue calculus at $z = \bar{\lambda}_i$ as

$$\mathbf{a}^\mathsf{T}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}\mathbf{b} \simeq \mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b}\cdot\frac{m^2(\bar{\lambda}_i)}{m'(\bar{\lambda}_i)} = \mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b}\left(1 - \frac{c}{(\ell_i - c)^2}\right), \tag{45}$$

  with $m'(z)$ the derivative of $m(z)$ with respect to z satisfying

$$m'(z) = \frac{m^2(z)}{1 - \frac{cm^2(z)}{(1+cm(z))^2}}. \tag{46}$$

- Plugging in we conclude the proof.

Take-away of this section

- a Master Theorem: Deterministic Equivalent for resolvent for affine-transformed SCM model $\mathbf{X} = \mathbf{A} + \mathbf{CZ}$
- **information-plus-noise spiked model** $\mathbf{X} = \frac{1}{n}(\mathbf{A} + \mathbf{Z})(\mathbf{A} + \mathbf{Z})^\mathsf{T}$: phase transition in spiked eigenvalues and eigenvectors
- **additive spiked model** $\mathbf{B} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$: phase transition in spiked eigenvalues and eigenvectors

Table: Roadmap of linear ML models considered.

ML Problem	Classical Regime	Proportional Regime
Low rank approximation $\hat{\mathbf{X}}$ of info-plus-noise matrix $\mathbf{X}$	smooth decay of $\Vert\mathbf{X} - \hat{\mathbf{X}}\Vert_2 / \Vert\mathbf{X}\Vert_2 \simeq (1+\ell)^{-1}$ Proposition 1 Item (i)	sharp transition of $\Vert\mathbf{X} - \hat{\mathbf{X}}\Vert_2 / \Vert\mathbf{X}\Vert_2$ at $\ell = c + \sqrt{c}$ Proposition 1 Item (ii)
Classification of binary Gaussian mixtures of distance in means $\Delta\boldsymbol{\mu}$	pairwise $\simeq$ spectral approach Proposition 2 Item (i)	pairwise $\ll$ spectral approach Proposition 2 Item (ii)
Linear least squares regression risk as $n \uparrow$	bias $= 0$ and variance $\propto n^{-1}$ Proposition 3 Item (i)	monotonic bias and non-monotonic variance Proposition 3 Item (ii)

Definition (Rank-one matrix recovery)

Taking $\mathbf{B} = \ell \mathbf{u}\mathbf{u}^\mathsf{T}$ in Theorem 11 of the additive spiked model, we have

$$\mathbf{X} = \ell \mathbf{u}\mathbf{u}^\mathsf{T} + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} \in \mathbb{R}^{p\times p}, \tag{47}$$

for $\mathbf{u} \in \mathbb{R}^p$ some deterministic signal of unit norm, i.e., $\|\mathbf{u}\| = 1$, $\ell \geq 0$ the informative "signal strength," and $\mathbf{Z} \in \mathbb{R}^{p\times n}$ a random "noise" matrix having i.i.d. entries of zero mean and unit variance.

- known from **Eckart-Young-Mirsky theorem** that the "best" low-rank approximation of a given matrix $\mathbf{X}$, measured by any unitarily invariant matrix norm (including the Frobenius and the spectral/operator norm) is given by retaining the **top singular/eigenvalue decomposition**
- let $\mathbf{X} = \sum_{i=1}^p \lambda_i(\mathbf{X})\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}$, be the eigenvalue-eigenvector decomposition of a symmetric and nonnegative definite matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, with $\lambda_1(\mathbf{X}) \geq \ldots \geq \lambda_p(\mathbf{X}) \geq 0$ listed in a non-increasing order. Then, for $k \leq \operatorname{rank}(\mathbf{X})$, the solution to

$$\hat{\mathbf{X}}_* = \operatorname*{arg\,min}_{\operatorname{rank}(\hat{\mathbf{X}})=k} \|\mathbf{X} - \hat{\mathbf{X}}\| = \sum_{i=1}^k \lambda_i(\mathbf{X})\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}, \tag{48}$$

for any unitarily invariant norm $\|\cdot\|$.

- evaluate the relative spectral norm error $\|\mathbf{X}-\hat{\mathbf{X}}\|_2/\|\mathbf{X}\|_2$ of rank-one approximation under rank-one matrix recovery model, for input $\mathbf{X} \in \mathbb{R}^p$ drawn from additive spiked model, and $\hat{\mathbf{X}} = \lambda_1(\mathbf{X})\hat{\mathbf{u}}_1\hat{\mathbf{u}}_1^\mathsf{T}$ the optimal rank-one approximation of $\mathbf{X}$ given by its top eigenvalue-eigenvector pair $(\lambda_1(\mathbf{X}), \hat{\mathbf{u}}_1)$.

Proposition (Relative spectral error of low-rank approximation)

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be an additive spiked random matrix, for $\mathbf{Z}$ having i.i.d. sub-gaussian entries of zero mean and unit variance, and let $\hat{\mathbf{X}} = \lambda_1(\mathbf{X})\hat{\mathbf{u}}_1\hat{\mathbf{u}}_1^\mathsf{T}$ the optimal rank-one approximation of $\mathbf{X}$ given by its top eigenvalue-eigenvector pair $(\lambda_1(\mathbf{X}), \hat{\mathbf{u}}_1)$. Then, one has,

(i) *in the* ***classical*** *regime, for p fixed and $n \to \infty$ that*

$$\frac{\|\mathbf{X}-\hat{\mathbf{X}}\|_2}{\|\mathbf{X}\|_2} \to f_{n\gg p}(\ell) \equiv \frac{1}{1+\ell}, \tag{49}$$

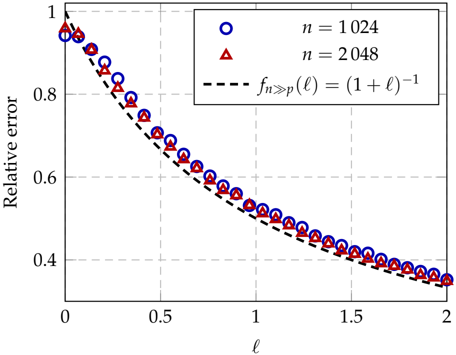
almost surely; and

(ii) *in the* ***proportional*** *regime, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$ that*

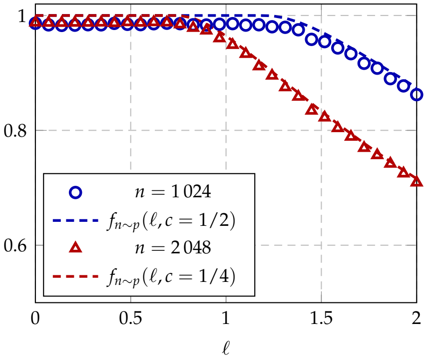
$$\frac{\|\mathbf{X}-\hat{\mathbf{X}}\|_2}{\|\mathbf{X}\|_2} \to f_{n\sim p}(\ell, c) \equiv \begin{cases} \frac{(1+\sqrt{c})^2}{1+\ell+\frac{c}{\ell-c}}, & \ell > c+\sqrt{c} \\ 1, & \ell \leq c+\sqrt{c} \end{cases} \tag{50}$$

almost surely.

Numerical results



(a) $p = 4$



(b) $p = 512$

- sharp phase transition of the relative error as the signal strength ℓ increases
- for p large and fixed, transition thresholds in ℓ are different for different values of n, and they become smaller as the dimension n increases from 1 024 to 2 048

Proof in the classical regime

- evoking the LLN, one has
$$\mathbf{X} \to \mathbb{E}[\mathbf{X}] = \mathbf{I}_p + \ell \mathbf{u}\mathbf{u}^\mathsf{T}, \tag{51}$$
almost surely as $n \to \infty$ for p fixed
- in the classical $n \gg p$ regime, $\mathbf{X}$ is close, in both a max and a spectral norm sense, to its expectation $\mathbb{E}[\mathbf{X}] = \mathbf{I}_p + \ell \mathbf{u}\mathbf{u}^\mathsf{T}$, and the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$, when arranged in a non-increasing order, are (asymptotically and approximately) given by
$$\|\mathbf{X}\|_2 \approx \lambda_1(\mathbf{X}) = 1 + \ell \geq \lambda_2(\mathbf{X}) = \ldots = \lambda_p(\mathbf{X}) \approx 1. \tag{52}$$
- for $n \gg p$ that
$$\frac{\|\mathbf{X} - \hat{\mathbf{X}}\|_2}{\|\mathbf{X}\|_2} \approx \frac{\lambda_2(\mathbb{E}[\mathbf{X}])}{\lambda_1(\mathbb{E}[\mathbf{X}])} = \frac{1}{1+\ell} \equiv f_{n \gg p}(\ell). \tag{53}$$
The approximation "$\approx$" can be replaced by an almost sure convergence in the limit of $n \to \infty$ for p fixed

Proof in the proportional regime

In the proportional $n \sim p$ regime:

(i) by Marčenko-Pastur law, in the absence of information signal $\ell \mathbf{u}\mathbf{u}^\mathsf{T}$ (i.e., $\ell = 0$), the eigenvalues of $\mathbf{X}$ have a Marčenko-Pastur shape;

(ii) by Theorem 12, in the presence of the rank-one informative signal $\ell \mathbf{u}\mathbf{u}^\mathsf{T}$ in Equation (47), that depending the "signal strength" $\|\ell \mathbf{u}\mathbf{u}^\mathsf{T}\|_2 = \ell > 0$, the largest eigenvalue of $\mathbf{X}$ establishes a phase transition behavior and is no longer a smooth function of ℓ (as opposed to its classical counterpart in Item (i) of Proposition 1)

For additive spiked model, one has

$$\|\mathbf{X}\|_2 \to \bar{\lambda}_1 = \begin{cases} 1 + \ell + \frac{c}{\ell - c}, & \ell > c + \sqrt{c} \\ (1+\sqrt{c})^2, & \ell \leq c + \sqrt{c}. \end{cases} \tag{54}$$

almost surely as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$. Since $\|\mathbf{X} - \hat{\mathbf{X}}\|_2 = \lambda_2(\mathbf{X})$ and $\lambda_2(\mathbf{Z}\mathbf{Z}^\mathsf{T}/n) \leq \lambda_2(\mathbf{X}) \leq \lambda_1(\mathbf{Z}\mathbf{Z}^\mathsf{T}/n)$ (Weyl's inequality), one has also

$$\|\mathbf{X} - \hat{\mathbf{X}}\|_2 \to (1+\sqrt{c})^2, \tag{55}$$

almost surely, so that by Slutsky's Theorem, one has $\frac{\|\mathbf{X} - \hat{\mathbf{X}}\|_2}{\|\mathbf{X}\|_2} \to f_{n \sim p}(\ell, c)$.

Definition (Gaussian Mixture Model, GMM)

We say $\mathbf{x} \in \mathbb{R}^p$ follows a two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) Gaussian Mixture Model if it is drawn from one of the two multivariate Gaussian distribution, that is

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_p), \quad \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I}_p); \quad \Delta\boldsymbol{\mu} \equiv \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \quad \|\Delta\boldsymbol{\mu}\| = \Theta(1). \tag{56}$$

Proposition (Fundamental limits of GMM classification: pairwise versus spectral approach)

For Gaussian mixture classification between $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_p)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I}_p)$, with $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, one has, for some constant $C > 0$ independent of p,

(i) *based on a pairwise (Euclidean) distance comparison approach, one is able to separate binary Gaussian mixtures satisfying $\|\Delta\boldsymbol{\mu}\| \geq C p^{1/4}$; and*

(ii) *based on an eigenspectral approach, one is able to separate a closer distance of $\|\Delta\boldsymbol{\mu}\| \geq C$, which is, up to a constant factor, the minimum distance possible.*

Illustration

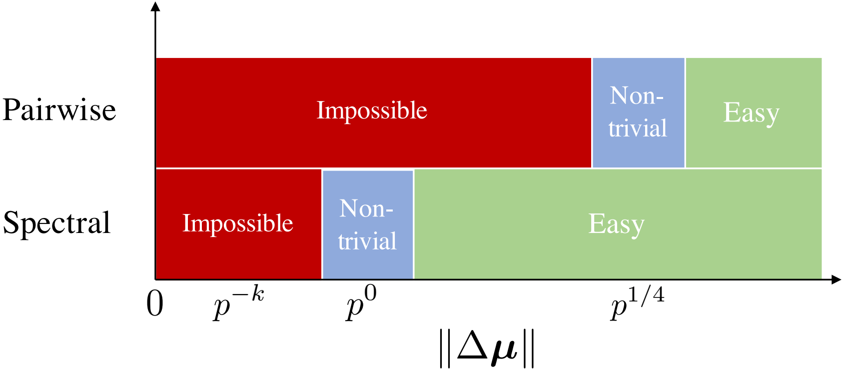


Figure: Illustration of different regimes in separating a binary GMM based on the distance in means $\|\Delta\boldsymbol{\mu}\|$, with $k > 0$, for both pairwise and spectral approaches.

Proof in the classical regime

- classification of the binary Gaussian mixture

$$\mathcal{C}_1 : \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_p) \quad \text{versus} \quad \mathcal{C}_2 : \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I}_p), \quad \boxed{\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2.} \tag{57}$$

- for two distinct data vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$, belonging to class $\mathcal{C}_a$ and $\mathcal{C}_b$, $a, b \in \{1, 2\}$, we have $\mathbf{x}_i = \boldsymbol{\mu}_a + \mathbf{z}_i \in \mathcal{C}_a$ and $\mathbf{x}_j = \boldsymbol{\mu}_b + \mathbf{z}_j \in \mathcal{C}_b$, for standard Gaussian $\mathbf{z}_i, \mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Then, their (normalized) Euclidean distance is given by

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b + \mathbf{z}_i - \mathbf{z}_j\|^2, \tag{58}$$

which is also the (i, j) entry of the Euclidean distance matrix $\mathbf{E} \equiv \{\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\}_{i,j=1}^n$.

- so

$$\begin{aligned} \frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 &= \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 + \frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j) \\ &= \frac{1}{p}\|\mathbf{z}_i\|^2 + \frac{1}{p}\|\mathbf{z}_j\|^2 - \frac{2}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j + \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 + \frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j). \end{aligned} \tag{59}$$

Proof in the classical regime

- in expectation, we have $\frac{1}{p}\mathbb{E}\left[\|\mathbf{x}_i - \mathbf{x}_j\|^2\right] = 2 + \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2$, for $i \neq j$, where we used the fact that $\mathbb{E}[\mathbf{z}_i^\mathsf{T}\mathbf{z}_i]/p = \operatorname{tr}(\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^\mathsf{T}])/p = 1$;

$$\begin{aligned}\operatorname{Var}\left[\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right] &= \operatorname{Var}\left[\frac{1}{p}(\Delta\mathbf{z} + 2(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b))^\mathsf{T}\Delta\mathbf{z}\right] \\ &= \frac{4}{p^2}\mathbb{E}\left[(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}\Delta\mathbf{z}\Delta\mathbf{z}^\mathsf{T}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b) + (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}\Delta\mathbf{z}\Delta\mathbf{z}^\mathsf{T}\Delta\mathbf{z}\right] + \frac{1}{p^2}\operatorname{Var}[\|\Delta\mathbf{z}\|^2] \\ &= \frac{8}{p^2}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 + \frac{8}{p} \leq \frac{16}{p}\end{aligned}$$

  for $\Delta\mathbf{z} \equiv \mathbf{z}_i - \mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, 2\mathbf{I}_p)$ and $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| \leq \sqrt{p}$.
- to ensure that the pairwise approach works, one must have that the distances between data points $\mathbf{x}_i, \mathbf{x}_j$ from the *same* Gaussian (with $a = b$) are, with non-trivial probability, smaller than those from *different* Gaussian (with $a \neq b$). This requires that

$$2 \pm \sqrt{Cp^{-1}} \leq 2 + \|\Delta\boldsymbol{\mu}\|^2/p \pm \sqrt{Cp^{-1}} \tag{60}$$

  and therefore

$$\boxed{\|\Delta\boldsymbol{\mu}\| \geq C'p^{1/4},} \tag{61}$$

  for some $C, C' > 0$ independent of p.

Proof in the proportional regime

- consider the more challenging setting of $\|\Delta\boldsymbol{\mu}\| = \Theta(1)$ in the proportional regime, that classification remains doable via an eigenspectral approach on Euclidean distance matrix $\mathbf{E} = \{\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\}_{i,j=1}^n$
- for $\|\Delta\boldsymbol{\mu}\| = \Theta(1)$ and n, p both large, it follows from the expansion in Equation (59) that

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2 + \underbrace{\psi_i + \psi_j - \frac{2}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})} + \underbrace{\frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 + \frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j)}_{O(p^{-1})} \tag{62}$$

where we denote $\psi_i \equiv \|\mathbf{z}_i\|^2/p - 1$ with $\mathbb{E}[\psi_i] = 0$ and $\mathrm{Var}[\psi_i] = 2/p$.
- in matrix form,

$$\mathbf{E} = 2 \cdot \mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\boldsymbol{\psi}^\mathsf{T} - \frac{2}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z} + \frac{1}{p}\mathbf{J}\begin{bmatrix} 0 & \|\Delta\boldsymbol{\mu}\|^2 \\ \|\Delta\boldsymbol{\mu}\|^2 & 0 \end{bmatrix}\mathbf{J}^\mathsf{T} + \boldsymbol{\Theta} - \mathrm{diag}(\cdot) \tag{63}$$

where we denote $\mathbf{J} = \begin{bmatrix}\mathbf{j}_1 & \mathbf{j}_2\end{bmatrix} \in \mathbb{R}^{n\times 2}$ for $\mathbf{j}_a \in \mathbb{R}^n$ the label vector of class $\mathcal{C}_a$ such that $[\mathbf{j}_a]_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$, $\boldsymbol{\psi} \in \mathbb{R}^n$ a random vector containing ψ_i as its i-th entry, $\boldsymbol{\Theta} \equiv \{2(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j)/p\}_{i,j=1}^n$, and we use the notation $\mathbf{X} - \mathrm{diag}(\cdot)$ to remove the diagonal of a given matrix $\mathbf{X}$.

$$\mathbf{E} = 2 \cdot \mathbf{1}_n \mathbf{1}_n^\mathsf{T} + \boldsymbol{\psi} \mathbf{1}_n^\mathsf{T} + \mathbf{1}_n \boldsymbol{\psi}^\mathsf{T} - \frac{2}{p} \mathbf{Z}^\mathsf{T} \mathbf{Z} + \frac{1}{p} \mathbf{J} \begin{bmatrix} 0 & \|\Delta \boldsymbol{\mu}\|^2 \\ \|\Delta \boldsymbol{\mu}\|^2 & 0 \end{bmatrix} \mathbf{J}^\mathsf{T} + \boldsymbol{\Theta} - \operatorname{diag}(\cdot) \tag{64}$$

- a low-rank non-informative matrix $2 \cdot \mathbf{1}_n \mathbf{1}_n^\mathsf{T} + \boldsymbol{\psi} \mathbf{1}_n^\mathsf{T} + \mathbf{1}_n \boldsymbol{\psi}^\mathsf{T}$ of spectral norm of order $O(n)$
- a sample covariance-type random matrix $2\mathbf{Z}^\mathsf{T}\mathbf{Z}/p$ for $\mathbf{Z} \in \mathbb{R}^{p \times n}$ having i.i.d. standard Gaussian entries, the spectrum of which follows a Marčenko-Pastur shape (and is of order $O(1)$)
- a low-rank informative matrix $\frac{1}{p} \mathbf{J} \begin{bmatrix} 0 & \|\Delta \boldsymbol{\mu}\|^2 \\ \|\Delta \boldsymbol{\mu}\|^2 & 0 \end{bmatrix} \mathbf{J}^\mathsf{T} + \boldsymbol{\Theta}$ that depends on the label vector $\mathbf{j}_1, \mathbf{j}_2 \in \mathbb{R}^n$ (so of interest for classification) and the statistical difference (in means) $\Delta \boldsymbol{\mu}$, also of spectral norm order $O(1)$

(i) while in the critical regime $\|\Delta \boldsymbol{\mu}\| = \Theta(1)$, data vectors $\mathbf{x}_i, \mathbf{x}_j$ are pairwise indistinguishable based on their Euclidean distance, due to the dominant order of the random $\mathbf{z}_i^\mathsf{T} \mathbf{z}_j / p = O(p^{-1/2})$ over the informative term $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2/p = \Theta(p^{-1})$ in Equation (62);

(ii) they can still be "clustered" into two classes with a spectral approach based on the global observation of the large Euclidean distance matrix $\mathbf{E}$, since the sample covariance-type random matrix and the low-rank informative matrix are both of spectral norm order $O(1)$, and thus comparable for n, p large.

Numerical results

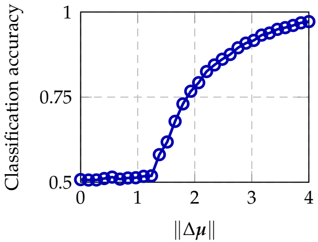


Figure: Phase transition behavior of the classification accuracy using the sign of the second top eigenvector $\mathbf{v}_2$ of the Euclidean distance matrix $\mathbf{E}$, as a function of the statistical difference $\|\Delta\boldsymbol{\mu}\|$ in the non-trivial $\|\Delta\boldsymbol{\mu}\| = \Theta(1)$ regime, for $p = 512$, $n = 4p$, and $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$. Results averaged over 10 independent runs.

"More refined" sharp phase transition, the second dominant eigenvector $\mathbf{v}_2$ of $\mathbf{E}$:

(i) for n, p fixed and large, when $\|\Delta\boldsymbol{\mu}\|$ below threshold, $\mathbf{v}_2$ does not contain data class information, the clustering/classification based on $\operatorname{sign}(\mathbf{v}_2)$ **random guess**

(ii) above the phase transition threshold, the eigenvector $\mathbf{v}_2$ contains data class information $\mathbf{j}_a$, and the classification accuracy increases as $\|\Delta\boldsymbol{\mu}\|$ and/or n/p becomes large.

Noisy linear model

Consider a given set of data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of size n, composed of the (random) input data $\mathbf{x}_i \in \mathbb{R}^p$ and its corresponding output target $y_i \in \mathbb{R}$, drawn from the following noisy linear model.

Definition (Noisy linear model)

We say a data-target pair $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ follows a noisy linear model if it satisfies

$$y = \boldsymbol{\beta}_*^\mathsf{T} \mathbf{x} + \epsilon \tag{65}$$

for some deterministic (ground-truth) vector $\boldsymbol{\beta}_* \in \mathbb{R}^p$, and random variable $\epsilon \in \mathbb{R}$ independent of $\mathbf{x} \in \mathbb{R}^p$, with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$.

- aim to find a regressor $\boldsymbol{\beta} \in \mathbb{R}^p$ that best describes the linear relation $y_i \approx \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i$, by minimizing the ridge-regularized mean squared error (MSE)

$$L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i)^2 + \gamma \|\boldsymbol{\beta}\|^2 = \frac{1}{n} \|\mathbf{X}^\mathsf{T} \boldsymbol{\beta} - \mathbf{y}\|^2 + \gamma \|\boldsymbol{\beta}\|^2 \tag{66}$$

for $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^n$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, and some regularization penalty $\gamma \geq 0$

Out-of-sample prediction risk

- unique solution given by

$$\boldsymbol{\beta}_\gamma = \left(\mathbf{X}\mathbf{X}^\mathsf{T} + n\gamma\mathbf{I}_p\right)^{-1}\mathbf{X}\mathbf{y} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X} + n\gamma\mathbf{I}_n\right)^{-1}\mathbf{y}, \quad \gamma > 0 \tag{67}$$

- in the $\gamma = 0$ setting, the minimum ℓ_2 norm least squares solution

$$\boldsymbol{\beta}_0 = \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{+}\mathbf{X}\mathbf{y} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{+}\mathbf{y}, \tag{68}$$

  where $(\mathbf{A})^+$ denotes the Moore–Penrose pseudoinverse, also **"ridgeless" least squares solution**.
- "statistical quality" of $\boldsymbol{\beta}$, as a function of dimensions n, p, noise level σ^2, and the regularization γ
- evaluating the **out-of-sample prediction risk** (or simply, **risk**)

$$R_\mathbf{X}(\boldsymbol{\beta}) = \mathbb{E}[(\boldsymbol{\beta}^\mathsf{T}\hat{\mathbf{x}} - \boldsymbol{\beta}_*^\mathsf{T}\hat{\mathbf{x}})^2 \mid \mathbf{X}] = \underbrace{(\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}] - \boldsymbol{\beta}_*)^\mathsf{T}\mathbf{C}(\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}] - \boldsymbol{\beta}_*)}_{\equiv B_\mathbf{X}(\boldsymbol{\beta})} + \underbrace{\operatorname{tr}\left(\operatorname{Cov}[\boldsymbol{\beta} \mid \mathbf{X}]\mathbf{C}\right)}_{\equiv V_\mathbf{X}(\boldsymbol{\beta})} \tag{69}$$

  for an independent test data point. We denote $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbf{C}$, and $B_\mathbf{X}(\boldsymbol{\beta}), V_X(\boldsymbol{\beta})$ the **bias** as well as **variance** of the solution $\boldsymbol{\beta}$.

Risk of linear ridge regression

Proposition (Risk of linear ridge regression)

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random data matrix having i.i.d. sub-gaussian entries of zero mean and unit variance (so that $\mathbf{C} = \mathbf{I}_p$). Then, under the linear model and for the out-of-sample prediction risk $R_{\mathbf{X}}$ of the linear ridge regressor $\boldsymbol{\beta}_\gamma$ given in Equation (67), one has $R_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) + V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma)$ and

(i) *in the **classical** regime, for p fixed and $n \to \infty$ that*

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) - \left(\frac{\gamma}{1+\gamma} \right)^2 \|\boldsymbol{\beta}_*\|^2 \to 0, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) - \frac{p}{n} \frac{\sigma^2}{(1+\gamma)^2} \to 0, \tag{70}$$

almost surely, so that $R_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) - R_{n \gg p}(\gamma) \to 0$, with $R_{n \gg p}(\gamma) \equiv \frac{\gamma^2 \|\boldsymbol{\beta}_*\|^2 + \frac{p}{n}\sigma^2}{(1+\gamma)^2}$;

(ii) *in the **proportional** regime, as $n, p \to \infty$ with $p/n \to c \in (0,1) \cup (1, \infty)$ that*

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) - \gamma^2 \|\boldsymbol{\beta}_*\|^2 m'(-\gamma) \to 0, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) - \sigma^2 c \left(m(-\gamma) - \gamma m'(-\gamma) \right) \to 0, \tag{71}$$

almost surely, with $m'(-\gamma) = \frac{m(-\gamma)(cm(-\gamma)+1)}{2c\gamma m(-\gamma)+1-c+\gamma}$ by differentiating the Marčenko-Pastur equation

$$R_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) - R_{n \sim p}(\gamma) \to 0, \quad \text{with} \quad R_{n \sim p}(\gamma) \equiv \sigma^2 c m(-\gamma) + \gamma m'(-\gamma) \left(\sigma^2 c - \gamma \|\boldsymbol{\beta}_*\|^2 \right). \tag{72}$$

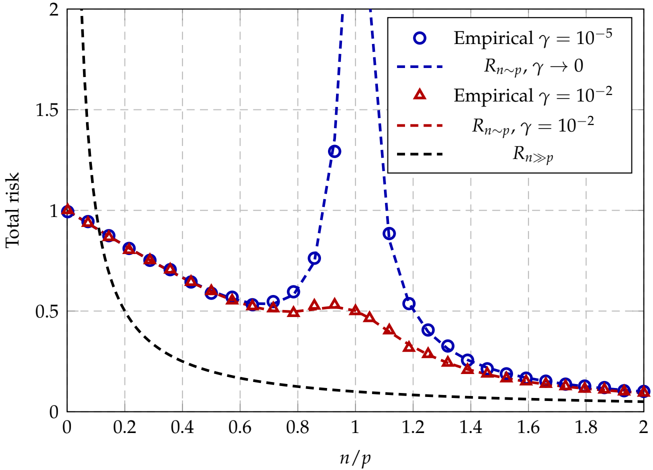


Figure: Out-of-sample risk $R_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) + V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma)$ of the ridge regression solution $\boldsymbol{\beta}_\gamma$ defined in Equation (67) as a function of the dimension ratio n/p, for fixed $p = 512$, $\|\boldsymbol{\beta}_*\| = 1$, and different regularization penalty $\gamma = 10^{-2}$ and $\gamma = 10^{-5}$, Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 = 0.1)$.

Remark

for relatively small regularization $\gamma = 10^{-5}$ and as the sample size n increases, that the total risk $R_{\mathbf{X}}(\boldsymbol{\beta}_\gamma)$:

1. first decreases and then increases as n approaches the input dimension p in the under-determined $n < p$ regime; and
2. reaches a singular "**peak point**" at $n = p$ with a large risk; and
3. decreases again monotonically as n continues to increase, in the over-determined $n > p$ regime.
4. This phenomenon is largely alleviated, yet still visible, for larger regularization of $\gamma = 10^{-2}$, and is referred to as the "**double descent**" test curve.

Proof in the classical regime

- denote $\mathbf{Q}(-\gamma) \equiv (\hat{\mathbf{C}} + \gamma \mathbf{I}_p)^{-1}$ the **resolvent** of the (un-centered) sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{Q}(\gamma = 0) = \lim_{\gamma \downarrow 0} \mathbf{Q}(-\gamma) = \hat{\mathbf{C}}^+$.
- we can write

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \boldsymbol{\beta}_*^\mathsf{T} \left(\mathbf{I}_p - \mathbf{Q}(-\gamma)\hat{\mathbf{C}}\right) \mathbf{C} \left(\mathbf{I}_p - \mathbf{Q}(-\gamma)\hat{\mathbf{C}}\right) \boldsymbol{\beta}_*, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \frac{\sigma^2}{n} \operatorname{tr}\left(\mathbf{Q}(-\gamma)\hat{\mathbf{C}}\mathbf{Q}(-\gamma)\mathbf{C}\right). \tag{73}$$

- for $\gamma > 0$, one has $\mathbf{I}_p - \mathbf{Q}(-\gamma)\hat{\mathbf{C}} = \mathbf{I}_p - \mathbf{Q}(-\gamma)(\hat{\mathbf{C}} + \gamma\mathbf{I}_p - \gamma\mathbf{I}_p) = \gamma\mathbf{Q}(-\gamma)$, so that

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \gamma^2 \boldsymbol{\beta}_*^\mathsf{T} \mathbf{Q}^2(-\gamma)\boldsymbol{\beta}_* = -\gamma^2 \frac{\partial \boldsymbol{\beta}_*^\mathsf{T} \mathbf{Q}(-\gamma)\boldsymbol{\beta}_*}{\partial \gamma}, \tag{74}$$

$$V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \sigma^2 \left(\frac{1}{n} \operatorname{tr} \mathbf{Q}(-\gamma) - \frac{\gamma}{n} \operatorname{tr} \mathbf{Q}^2(-\gamma) \right) = \sigma^2 \left(\frac{1}{n} \operatorname{tr} \mathbf{Q}(-\gamma) + \frac{\gamma}{n} \frac{\partial \operatorname{tr} \mathbf{Q}(-\gamma)}{\partial \gamma} \right), \tag{75}$$

where we used the fact that $\mathbf{C} = \mathbf{I}_p$ and $\partial \mathbf{Q}(-\gamma)/\partial\gamma = -\mathbf{Q}^2(-\gamma)$.

Proof in the classical regime

- by LLN, we have, in the classical regime for fixed p and as $n \to \infty$ that $\hat{\mathbf{C}} \to \mathbf{C} = \mathbf{I}_p$, and therefore

$$\mathbf{Q}(-\gamma) \to (\mathbf{C} + \gamma \mathbf{I}_p)^{-1} = \frac{\mathbf{I}_p}{1+\gamma}. \tag{76}$$

- we have that

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) \to -\gamma^2 \frac{\partial}{\partial \gamma} \frac{\|\boldsymbol{\beta}_*\|^2}{1+\gamma} = \left(\frac{\gamma}{1+\gamma} \right)^2 \|\boldsymbol{\beta}_*\|^2,$$
$$V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) \to \sigma^2 \left(\frac{p}{n} \frac{1}{1+\gamma} + \gamma \cdot \frac{p}{n} \frac{\partial}{\partial \gamma} \frac{1}{1+\gamma} \right) = \frac{p}{n} \frac{\sigma^2}{(1+\gamma)^2}.$$

- in the ridgeless setting with $\gamma = 0$

$$B_{\mathbf{X}}(\boldsymbol{\beta}_0) = 0, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_0) = \frac{\sigma^2}{n} \operatorname{tr} \left(\mathbf{Q}(\gamma = 0) \mathbf{C} \right) \to \sigma^2 \frac{p}{n}, \tag{77}$$

Proof in the proportional regime

- it follows from our Linear Master Theorem that

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) \to -\gamma^2 \|\boldsymbol{\beta}_*\|^2 \frac{\partial m(-\gamma)}{\partial \gamma} = \gamma^2 \|\boldsymbol{\beta}_*\|^2 m'(-\gamma),$$

$$V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) \to \sigma^2 \cdot \frac{p}{n} \left(m(-\gamma) - \gamma m'(-\gamma) \right),$$

  with $m'(z) = -\frac{m(z)(cm(z)+1)}{2czm(z)-1+c+z}$ the derivative of the Stieltjes transform $m(z)$
- in the ridgeless setting as $\gamma \to 0$, one has $m(\gamma) = \frac{1}{1-c} > 0$ only if $c < 1$ and $\lim_{\gamma \to 0} m(\gamma)$ undefined otherwise, but satisfying $\lim_{\gamma \to 0} \gamma m(\gamma) = \frac{c-1}{c} > 0$, in the under-determined regime with $n < p$.

$$B_{\mathbf{X}}(\boldsymbol{\beta}_0) \to 0, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_0) \to \sigma^2 \frac{c}{1-c}, \text{ for } c < 1 \tag{78}$$

$$B_{\mathbf{X}}(\boldsymbol{\beta}_0) - \|\boldsymbol{\beta}_*\|^2 \left(1 - \frac{1}{c}\right) \to 0, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_0) \to \sigma^2 \frac{1}{c-1}, \text{ for } c > 1 \tag{79}$$

- **Note**: for $c > 1$, $V_{\mathbf{X}}(\boldsymbol{\beta}_0)$ more involved, as one cannot take the limit $\gamma \to 0$. Instead,

$$V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \frac{\sigma^2}{n^2} \operatorname{tr}\left(\tilde{\mathbf{Q}}(-\gamma) \mathbf{X}^\mathsf{T} \mathbf{C} \mathbf{X} \tilde{\mathbf{Q}}(-\gamma) \right), \quad \tilde{\mathbf{Q}}(-\gamma) \equiv \left(\frac{1}{n} \mathbf{X}^\mathsf{T} \mathbf{X} + \gamma \mathbf{I}_n \right)^{-1}. \tag{80}$$

  which is more convenient to work with in the $c > 1$ regime.

Take-away messages of this section

Table: Roadmap of linear ML models considered.

ML Problem	Classical Regime	Proportional Regime
$\hat{\mathbf{X}}$ of info-plus-noise matrix $\mathbf{X}$	smooth decay of $\|\mathbf{X}-\hat{\mathbf{X}}\|_2/\|\mathbf{X}\|_2 \simeq (1+\ell)^{-1}$ Proposition 1 Item (i)	sharp transition of $\|\mathbf{X}-\hat{\mathbf{X}}\|_2/\|\mathbf{X}\|_2$ at $\ell = c+\sqrt{c}$ Proposition 1 Item (ii)
Classification of binary Gaussian mixtures of distance in means $\Delta\boldsymbol{\mu}$	pairwise $\simeq$ spectral approach Proposition 2 Item (i)	pairwise $\ll$ spectral approach Proposition 2 Item (ii)
Linear least squares regression risk as $n \uparrow$	bias $= 0$ and variance $\propto n^{-1}$ Proposition 3 Item (i)	monotonic bias and non-monotonic variance Proposition 3 Item (ii)

- Linear Master Theorem provides a **unified** analysis framework to
- low rank approximation: phase transition in spiked eigenvalue
- classification: phase transition in spiked eigenvector
- linear least squares: double descent as phase transition in resolvent

Thank you! Q & A?