University of Pennsylvania
STAT 991: Random Matrix Theory Class Presentation

Asymptotic Risk of

High-Dimensional Regression

Behrad Moniri
bemoniri@seas.upenn.edu

Linear Regression

High-Dimensional Linear Regression

- **Data Generation**:

High-Dimensional Linear Regression

- **Data Generation**:

$$\begin{cases} \beta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \\ \\ \end{cases}$$

- **Data Generation**:

$$\begin{cases} \beta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \\ \{x_i\}_{i=1}^n \sim \mathcal{N}(0, \Sigma) \\ \end{cases}$$

High-Dimensional Linear Regression

- **Data Generation**:

$$\begin{cases} \beta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \\ \{x_i\}_{i=1}^n \sim \mathcal{N}(0, \Sigma) \\ y_i = \beta^\top x_i + \varepsilon_i \ \text{ where } \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{1}$$

High-Dimensional Linear Regression

- **Data Generation**:

$$\begin{cases} \beta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d} \\ \{x_i\}_{i=1}^n \sim \mathcal{N}(0, \Sigma) \\ y_i = \beta^\top x_i + \varepsilon_i \ \text{ where } \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{1}$$

- **Fit with ridge regression**:

High-Dimensional Linear Regression

- **Data Generation**:

$$\begin{cases} \beta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \\ \{x_i\}_{i=1}^n \sim \mathcal{N}(0, \Sigma) \\ y_i = \beta^\top x_i + \varepsilon_i \ \text{ where } \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{1}$$

- **Fit with ridge regression**:

$$\hat{\beta}_\lambda = \arg\min_{b \in \mathbb{R}^d} \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - b^\top x_i \right)^2 + \lambda ||b||_2^2 \right] \tag{2}$$

High-Dimensional Linear Regression

- **Data Generation**:

$$\begin{cases} \beta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d} \\ \{x_i\}_{i=1}^n \sim \mathcal{N}(0, \Sigma) \\ y_i = \beta^\top x_i + \varepsilon_i \ \text{ where } \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{1}$$

- **Fit with ridge regression**:

$$\hat{\beta}_\lambda = \arg\min_{b\in\mathbb{R}^d} \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - b^\top x_i \right)^2 + \lambda ||b||_2^2 \right] \tag{2}$$

- **Random design assumption**: β is random with

$$\mathbb{E}[\beta] = 0 \quad \text{and} \quad \mathrm{Cov}(\beta) = \frac{\alpha^2}{d} I_{d\times d} \tag{3}$$

Solution

- Let $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^n$ be the training data.

Solution

- Let $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^n$ be the training data.
- Ridge regression has a closed-form solution

$$\hat{\beta}_\lambda = (X^\top X + n\lambda I)^{-1} X^\top Y$$

Risk of the estimator

- **Question:** What is the risk of $\hat{\beta}_\lambda$?

Risk of the estimator

- **Question:** What is the risk of $\hat{\beta}_\lambda$?

$$
\begin{aligned}
r_\lambda(X) &= \mathbb{E}\left[\left(x^\top \hat{\beta}_\lambda - x^\top \beta - \varepsilon\right)^2 \middle| X\right] \\
&= 1 + \mathbb{E}\left[\left\{x^\top\left(\hat{\beta}_\lambda - \beta\right)\right\}^2 \middle| X\right] \\
&= 1 + \mathbb{E}\left[\left(\hat{\beta}_\lambda - \beta\right)^\top \Sigma \left(\hat{\beta}_\lambda - \beta\right) \middle| X\right]
\end{aligned}
$$

Risk of the estimator

- **Question:** What is the risk of $\hat{\beta}_\lambda$?

$$
\begin{aligned}
r_\lambda(X) &= \mathbb{E}\left[\left(x^\top \hat{\beta}_\lambda - x^\top \beta - \varepsilon\right)^2 \middle| X\right] \\
&= 1 + \mathbb{E}\left[\left\{x^\top \left(\hat{\beta}_\lambda - \beta\right)\right\}^2 \middle| X\right] \\
&= 1 + \mathbb{E}\left[\left(\hat{\beta}_\lambda - \beta\right)^\top \Sigma \left(\hat{\beta}_\lambda - \beta\right) \middle| X\right]
\end{aligned}
$$

- We can also write

$$
\begin{aligned}
\hat{\beta}_\lambda - \beta &= \left(X^\top X + \lambda n I\right)^{-1} X^\top (X\beta + \varepsilon) - \beta \\
&= -\lambda n \left(X^\top X + \lambda n I\right)^{-1} \beta + \left(X^\top X + \lambda n I\right)^{-1} X^\top \varepsilon
\end{aligned}
$$

Risk of the estimator

- If we plug this back to the risk, we get

$$
\begin{aligned}
r_\lambda(X) = 1 &+ (\lambda n)^2 \, \mathbb{E}\left[\beta^\top \left(X^\top X + \lambda n I\right)^{-1} \Sigma \left(X^\top X + \lambda n I\right)^{-1} \beta \,\middle|\, X\right] \\
&+ \mathbb{E}\left[\varepsilon^\top X \left(X^\top X + \lambda n I\right)^{-1} \Sigma \left(X^\top X + \lambda n I\right)^{-1} X^\top \varepsilon \,\middle|\, X\right]
\end{aligned}
$$

Risk of the estimator

- If we plug this back to the risk, we get

$$
\begin{aligned}
r_\lambda(X) &= 1 + (\lambda n)^2\, \mathbb{E}\left[\beta^\top \left(X^\top X + \lambda n I\right)^{-1} \Sigma \left(X^\top X + \lambda n I\right)^{-1} \beta \,\middle|\, X\right] \\
&\quad + \mathbb{E}\left[\varepsilon^\top X \left(X^\top X + \lambda n I\right)^{-1} \Sigma \left(X^\top X + \lambda n I\right)^{-1} X^\top \varepsilon \,\middle|\, X\right] \\
&= 1 + \frac{(\lambda n)^2 \alpha^2}{d} \operatorname{Tr}\left(\Sigma \left(X^\top X + \lambda n I\right)^{-2}\right) \\
&\quad + \operatorname{Tr}\left(\Sigma \left(X^\top X + \lambda n I\right)^{-1} X^\top X \left(X^\top X + \lambda n I\right)^{-1}\right).
\end{aligned}
$$

Risk of the estimator

- If we plug this back to the risk, we get

$$\begin{aligned} r_\lambda(X) &= 1 + (\lambda n)^2 \,\mathbb{E}\left[\beta^\top \left(X^\top X + \lambda n I\right)^{-1} \Sigma \left(X^\top X + \lambda n I\right)^{-1} \beta \,\middle|\, X\right] \\ &\quad + \mathbb{E}\left[\varepsilon^\top X \left(X^\top X + \lambda n I\right)^{-1} \Sigma \left(X^\top X + \lambda n I\right)^{-1} X^\top \varepsilon \,\middle|\, X\right] \\ &= 1 + \frac{(\lambda n)^2 \alpha^2}{d} \operatorname{Tr}\left(\Sigma \left(X^\top X + \lambda n I\right)^{-2}\right) \\ &\quad + \operatorname{Tr}\left(\Sigma \left(X^\top X + \lambda n I\right)^{-1} X^\top X \left(X^\top X + \lambda n I\right)^{-1}\right). \end{aligned}$$

- Set $\hat{\Sigma} = n^{-1} X^\top X$ and $\gamma_d = d/n$:

$$r_\lambda(X) = 1 + \frac{\gamma_d}{d} \operatorname{Tr}\left(\Sigma \left(\widehat{\Sigma} + \lambda I\right)^{-1}\right) + \left(\lambda \alpha^2 - \gamma_d\right) \frac{\lambda}{d} \operatorname{Tr}\left(\Sigma \left(\widehat{\Sigma} + \lambda I\right)^{-2}\right)$$

Risk of the estimator

So, to analyze the asymptotic risk, we should look at the following two traces:

- **Trace 1**:
$$\frac{\gamma_d}{d} \operatorname{Tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I\right)^{-1}\right)$$
- **Trace 2**:
$$\left(\lambda\alpha^2-\gamma_d\right)\frac{\lambda}{d} \operatorname{Tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I\right)^{-2}\right)$$

Notation and Assumptions

- For a matrix A, define

$$F_A(x) = p^{-1} \sum_{i=1}^{p} \mathrm{I}\left(\lambda_i(A) \leq x\right).$$

Notation and Assumptions

- For a matrix A, define

$$F_A(x) = p^{-1} \sum_{i=1}^{p} \mathrm{I}\left(\lambda_i(A) \leq x\right).$$

- We will take the limit $\gamma_d = d/n \to \gamma > 0$.

Notation and Assumptions

- For a matrix A, define

$$F_A(x) = p^{-1} \sum_{i=1}^{p} \mathrm{I}\left(\lambda_i(A) \leq x\right).$$

- We will take the limit $\gamma_d = d/n \to \gamma > 0$.
- Assume that the spectral distribution F_Σ of Σ converges to a limit H supported on $[0, \infty)$.

Notation and Assumptions

- We defined the Stieltjes transform of a measure G as

$$m_G(z) = \int_{l=0}^{\infty} \frac{dG(l)}{l - z}, \quad z \in \mathbb{C} \backslash \mathbb{R}^+.$$

Notation and Assumptions

- We defined the Stieltjes transform of a measure G as

$$m_G(z) = \int_{l=0}^{\infty} \frac{dG(l)}{l-z}, \quad z \in \mathbb{C} \backslash \mathbb{R}^+.$$

- Stieltjes transform of the spectral measure of $\hat{\Sigma}$

$$m_{\widehat{\Sigma}}(z) = \frac{1}{p} \operatorname{tr}\left(\left(\widehat{\Sigma} - zI_{p\times p}\right)^{-1}\right) \text{ converges to } m(z)$$

Notation and Assumptions

- We defined the Stieltjes transform of a measure G as

$$m_G(z) = \int_{l=0}^{\infty} \frac{dG(l)}{l-z}, \quad z \in \mathbb{C}\backslash\mathbb{R}^+.$$

- Stieltjes transform of the spectral measure of $\hat{\Sigma}$

$$m_{\widehat{\Sigma}}(z) = \frac{1}{p} \operatorname{tr}\left(\left(\widehat{\Sigma} - zI_{p\times p}\right)^{-1}\right) \text{ converges to } m(z)$$

- The companion Stieltjes transform is defined as

$$\gamma(m(z) + 1/z) = v(z) + 1/z \text{ for all } z \in \mathbb{C}\backslash\mathbb{R}^+$$

A result of Ledoit and Péché (2011)

The main step to derive the limiting risk, is proving the following theorem:

Theorem (Ledoit and Péché (2011))

Assume that $m_{\widehat{\Sigma}}(z) \to m(z)$ *and let* $v(z)$ *be the companion transform for* $m(z)$*. We have*

$$\frac{1}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right) \to_{a.s.} \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right) \quad \textit{as } n, d \to \infty$$

A result of Ledoit and Péché (2011)

The main step to derive the limiting risk, is proving the following theorem:

Theorem (Ledoit and Péché (2011))

Assume that $m_{\widehat{\Sigma}}(z) \to m(z)$ and let $v(z)$ be the companion transform for $m(z)$. We have

$$\frac{1}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right) \to_{a.s.} \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right) \quad \text{as } n, d \to \infty$$

Statements like this are nontrivial. It is clear that these quantities converge, but there is no general theory to tell us what the limit is.

A result of Ledoit and Péché (2011)

The main step to derive the limiting risk, is proving the following theorem:

Theorem (Ledoit and Péché (2011))

Assume that $m_{\widehat{\Sigma}}(z) \to m(z)$ and let $v(z)$ be the companion transform for $m(z)$. We have

$$\frac{1}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right) \to_{a.s.} \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right) \quad \textit{as } n, d \to \infty$$

Statements like this are nontrivial. It is clear that these quantities converge, but there is no general theory to tell us what the limit is.

Before we prove it, lets use it to derive the asymptotic risk.

Limiting Risk of Ridge Regression

- **First Trace**: Ledoit and Péché (2011) proves that

$$\frac{1}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right)\to_{a.s.}\frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right):=\kappa(\lambda)$$

Limiting Risk of Ridge Regression

- **First Trace**: Ledoit and Péché (2011) proves that

$$\frac{1}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right)\to_{a.s.}\frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right):=\kappa(\lambda)$$

- **Second Functional**:

$$\left(\lambda\alpha^2-\gamma_d\right)\frac{\lambda}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-2}\right)$$

Limiting Risk of Ridge Regression

- **First Trace**: Ledoit and Péché (2011) proves that

$$\frac{1}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right) \to_{a.s.} \frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right) := \kappa(\lambda)$$

- **Second Functional**:

$$\left(\lambda\alpha^2-\gamma_d\right)\frac{\lambda}{d}\operatorname{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-2}\right) \to_{a.s.} -\left(\lambda\alpha^2-\gamma_d\right)\lambda\kappa'(\lambda)$$

Limiting Risk of Ridge Regression

- **First Trace**: Ledoit and Péché (2011) proves that

$$\frac{1}{d}\,\mathrm{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-1}\right)\to_{a.s.}\frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right):=\kappa(\lambda)$$

- **Second Functional**:

$$\left(\lambda\alpha^2-\gamma_d\right)\frac{\lambda}{d}\,\mathrm{tr}\left(\Sigma\left(\widehat{\Sigma}+\lambda I_{d\times d}\right)^{-2}\right)\to_{a.s.}-\left(\lambda\alpha^2-\gamma_d\right)\lambda\kappa'(\lambda)$$

- After simplification, the limiting risk converges almost surely

$$r_\lambda(X)\to_{a.s.}\frac{1}{\lambda v(-\lambda)}\left\{1+\left(\frac{\lambda\alpha^2}{\gamma}-1\right)\left(1-\frac{\lambda v^\top(-\lambda)}{v(-\lambda)}\right)\right\}.$$

Properties of the Solution

- From the formula, the optimal ridge parameter λ is

$$\lambda^\star = \gamma\alpha^{-2},$$

and we have

$$r_{\lambda^\star}(X) \to_{a.s.} \frac{1}{\lambda^\star v(-\lambda^\star)}.$$

Properties of the Solution

- From the formula, the optimal ridge parameter λ is

$$\lambda^\star = \gamma\alpha^{-2},$$

and we have

$$r_{\lambda^\star}(X) \to_{a.s.} \frac{1}{\lambda^\star v(-\lambda^\star)}.$$

- With $\Sigma = I_{d\times d}$, by the Marchenko-Pastur theorem we have

$$r_\lambda(X) \to 1 + \gamma m_I(-\lambda;\gamma) + \lambda\left(\lambda\alpha^2 - \gamma\right) m_I^\top(-\lambda;\gamma)$$

where $m_I(\cdot;\gamma)$ is the Stieltjes transform of the Marchenko Pastur law.

Properties of the Solution

- From the formula, the optimal ridge parameter λ is

$$\lambda^\star = \gamma\alpha^{-2},$$

  and we have

$$r_{\lambda^\star}(X) \to_{a.s.} \frac{1}{\lambda^\star v(-\lambda^\star)}.$$

- With $\Sigma = I_{d\times d}$, by the Marchenko-Pastur theorem we have

$$r_\lambda(X) \to 1 + \gamma m_I(-\lambda;\gamma) + \lambda\left(\lambda\alpha^2 - \gamma\right) m_I^\top(-\lambda;\gamma)$$

  where $m_I(\cdot;\gamma)$ is the Stieltjes transform of the Marchenko Pastur law.
- In this case, the optimal risk can be written in a closed form.

Computing in the General Case

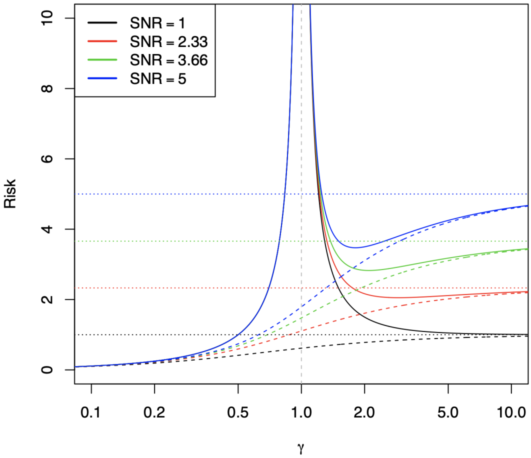
Silverstein and Choi (1995) show that $v(z)$ is the unique solution with positive imaginary part of the Silverstein equation:

$$-\frac{1}{v(z)} = z - \gamma \int \frac{t dH(t)}{1 + t v(z)}, z \in \mathbb{C}^+.$$

This equation can be solved by a fixed-point algorithm to compute $v(z)$ for all $z \in \mathbb{C}^+$.

Plots!

Assume $\Sigma = I$. This is the risk plot as a function of γ for ridgeless $\lambda \to 0$ (dashed) and optimal ridge (solid), for different SNRs.



Proof of Ledoit and Péché (2011)

[We will use bold symbols for matrices from now on]

- Remember that $\mathbf{X} \in \mathbb{R}^{n \times d}$ and define $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. Let $\mathbf{G} = (\lambda\mathbf{I} + \hat{\mathbf{\Sigma}})^{-1}$ be the resolvent.

Proof of Ledoit and Péché (2011)

[We will use bold symbols for matrices from now on]

- Remember that $\mathbf{X} \in \mathbb{R}^{n \times d}$ and define $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. Let $\mathbf{G} = (\lambda\mathbf{I} + \hat{\mathbf{\Sigma}})^{-1}$ be the resolvent.
- We are interested in computing traces of the form

$$\frac{1}{d}\mathrm{Tr}(\mathbf{G}\mathbf{\Sigma}) \quad \text{and} \quad \frac{1}{d}\mathrm{Tr}(\mathbf{G})$$

Proof of Ledoit and Péché (2011)

[We will use bold symbols for matrices from now on]

- Remember that $\mathbf{X} \in \mathbb{R}^{n \times d}$ and define $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. Let $\mathbf{G} = (\lambda\mathbf{I} + \hat{\mathbf{\Sigma}})^{-1}$ be the resolvent.
- We are interested in computing traces of the form

$$\frac{1}{d}\mathrm{Tr}(\mathbf{G}\mathbf{\Sigma}) \quad \text{and} \quad \frac{1}{d}\mathrm{Tr}(\mathbf{G})$$

- There are many ways to do it; e.g., leave-one-out analysis, free probability, etc. Here, we use a method based on Stein's formula.

Proof of Ledoit and Péché (2011)

Let's first remind the Stein's formula. Hong Hu talked about it briefly last time:

Lemma (Stein's Formula)

Let $X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and $f : \mathbb{R}^d \to \mathbb{R}$ have gradient with at most polynomial growth at infinity. Then for all $i_0 = 1, \ldots, d$:

$$\mathbb{E} X_{i_0} f\left(X_1, \ldots, X_d\right) = \sum_{k=1}^{d} \Sigma_{i_0 k}\, \mathbb{E}\left(\partial_k f\right)\left(X_1, \ldots, X_d\right)$$

Matrix version of Stein's formula

Theorem

For $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ *a good behaving matrix-valued function, we have*

$$\mathbb{E}\left[X^\top \mathbf{F}(X)X\right] = \mathbb{E}\left[\mathrm{Tr}\,\boldsymbol{\Sigma}\mathbf{F}(X)\right] + \mathbb{E}\sum_{k=1}^{d}\left(\boldsymbol{\Sigma}\left(\partial_k \mathbf{F}\right)(X)X\right)_k$$

Theorem

For $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ *a good behaving matrix-valued function, we have*

$$\mathbb{E}\left[X^\top \mathbf{F}(X) X\right] = \mathbb{E}\left[\operatorname{Tr} \boldsymbol{\Sigma} \mathbf{F}(X)\right] + \mathbb{E} \sum_{k=1}^{d} \left(\boldsymbol{\Sigma}\left(\partial_k \mathbf{F}\right)(X) X\right)_k$$

Proof.

$$\begin{aligned}
\mathbb{E} X^\top \mathbf{F}(X) X &= \mathbb{E} \sum_{ij} X_i X_j F(X)_{ij} \\
&= \mathbb{E} \sum_{ijk} \Sigma_{ik} \frac{\partial}{\partial X_k} X_j F(X)_{ij} \\
&= \mathbb{E} \sum_{ijk} \Sigma_{ik} \left[\delta_{j=k} F(X)_{ij} + X_j \left(\partial_k \mathbf{F}\right)(X)_{ij}\right] \\
&= \mathbb{E}\left[\operatorname{Tr} \boldsymbol{\Sigma} \mathbf{F}(X)\right] + \mathbb{E} \sum_{ijk} \Sigma_{ik} X_j \left(\partial_k \mathbf{F}\right)(X)_{ij} \\
&= \mathbb{E}\left[\operatorname{Tr} \boldsymbol{\Sigma} \mathbf{F}(X)\right] + \mathbb{E} \sum_{k} \left[\boldsymbol{\Sigma}\left(\partial_k \mathbf{F}\right)(X) X\right]_k
\end{aligned}$$

Proof of Ledoit and Péché (2011)

- We want to write $\frac{1}{n}\mathrm{Tr}(\mathbf{G}\boldsymbol{\Sigma})$ in terms of $\frac{1}{n}\mathrm{Tr}(\mathbf{G})$.

Proof of Ledoit and Péché (2011)

- We want to write $\frac{1}{n}\mathrm{Tr}(\mathbf{G}\boldsymbol{\Sigma})$ in terms of $\frac{1}{n}\mathrm{Tr}(\mathbf{G})$.
- Note that we have

$$\frac{1}{n}\mathrm{Tr}(\mathbf{G}\hat{\boldsymbol{\Sigma}}) = \frac{1}{n}\mathrm{Tr}\left(\mathbf{G}(\hat{\boldsymbol{\Sigma}} + \lambda I - \lambda I)\right) = \gamma - \frac{\lambda}{n}\mathrm{Tr}(\mathbf{G}).$$

Proof of Ledoit and Péché (2011)

- We want to write $\frac{1}{n}\mathrm{Tr}(\mathbf{G}\boldsymbol{\Sigma})$ in terms of $\frac{1}{n}\mathrm{Tr}(\mathbf{G})$.
- Note that we have

$$\frac{1}{n}\mathrm{Tr}(\mathbf{G}\hat{\boldsymbol{\Sigma}}) = \frac{1}{n}\mathrm{Tr}\left(\mathbf{G}(\hat{\boldsymbol{\Sigma}} + \lambda I - \lambda I)\right) = \gamma - \frac{\lambda}{n}\mathrm{Tr}(\mathbf{G}).$$

- We will start with $\frac{1}{n}\mathrm{Tr}(\mathbf{G}\hat{\boldsymbol{\Sigma}})$.

Proof of Ledoit and Péché (2011)

$$\mathbb{E}\left[X^\top \mathbf{F}(X)X\right] = \mathbb{E}\left[\operatorname{Tr} \mathbf{\Sigma}\mathbf{F}(X)\right] + \mathbb{E}\sum_{k=1}^{d} \left(\mathbf{\Sigma}\left(\partial_k \mathbf{F}\right)(X)X\right)_k$$

Proof of Ledoit and Péché (2011)

$$\mathbb{E}\left[X^\top \mathbf{F}(X)X\right] = \mathbb{E}\left[\operatorname{Tr} \boldsymbol{\Sigma}\mathbf{F}(X)\right] + \mathbb{E}\sum_{k=1}^{d}\left(\boldsymbol{\Sigma}\left(\partial_k \mathbf{F}\right)(X)X\right)_k$$

$$\begin{aligned}
\mathbb{E}\operatorname{Tr}\mathbf{G}\hat{\boldsymbol{\Sigma}} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\mathbf{G}X(i)X(i)^\top\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X(i)^\top \mathbf{G}X(i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\operatorname{Tr}\mathbf{G}\boldsymbol{\Sigma}\right] + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\left[e_k^\top \boldsymbol{\Sigma}\left(\frac{\partial}{\partial X(i)_k}\mathbf{G}\right)X(i)\right]
\end{aligned}$$

Proof of Ledoit and Péché (2011)

$$\mathbb{E}\left[X^\top \mathbf{F}(X)X\right] = \mathbb{E}\left[\operatorname{Tr} \boldsymbol{\Sigma}\mathbf{F}(X)\right] + \mathbb{E}\sum_{k=1}^{d}\left(\boldsymbol{\Sigma}\left(\partial_k \mathbf{F}\right)(X)X\right)_k$$

$$\begin{aligned}
\mathbb{E}\operatorname{Tr}\mathbf{G}\hat{\boldsymbol{\Sigma}} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\mathbf{G}X(i)X(i)^\top\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X(i)^\top \mathbf{G}X(i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\operatorname{Tr}\mathbf{G}\boldsymbol{\Sigma}\right] + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\left[e_k^\top \boldsymbol{\Sigma}\left(\frac{\partial}{\partial X(i)_k}\mathbf{G}\right)X(i)\right]
\end{aligned}$$

What is $\frac{\partial \mathbf{G}}{\partial X(i)_k} = \frac{\partial}{\partial X(i)_k}\left[(\lambda \mathbf{I} + \hat{\boldsymbol{\Sigma}})^{-1}\right]$?

$$\frac{\partial\,\mathbf{G}}{\partial X(i)_k} = \frac{1}{n}\left[\mathbf{G}\left(e_k X(i)^\top + X(i)e_k^\top\right)\mathbf{G}\right]$$

Proof of Ledoit and Péché (2011)

$$
\begin{aligned}
&\mathbb{E}\,\mathrm{Tr}\left[\mathbf{G}\hat{\mathbf{\Sigma}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\left[e_k^\top \mathbf{\Sigma}\left(\mathbf{G}\left(e_k X(i)^\top + X(i)e_k^\top\right)\mathbf{G}\right)X(i)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\left[e_k^\top \mathbf{\Sigma}\mathbf{G} e_k X(i)^\top \mathbf{G} X(i) + e_k^\top \mathbf{\Sigma}\mathbf{G} X(i) e_k^\top \mathbf{G} X(i)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\mathrm{Tr}(\mathbf{\Sigma}\mathbf{G})X(i)^\top \mathbf{G}X(i) + X(i)^\top \mathbf{G}\mathbf{\Sigma}\mathbf{G}X(i)\right) \\
&= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}\mathbf{G})\,\mathrm{Tr}(\mathbf{G}\hat{\mathbf{\Sigma}})\right] + \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(\mathbf{G}\mathbf{\Sigma}\mathbf{G}\hat{\mathbf{\Sigma}})\right]
\end{aligned}
$$

Proof of Ledoit and Péché (2011)

$$
\begin{aligned}
&\mathbb{E}\,\mathrm{Tr}\left[\mathbf{G}\hat{\mathbf{\Sigma}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\left[e_k^\top \mathbf{\Sigma}\left(\mathbf{G}\left(e_k X(i)^\top + X(i)e_k^\top\right)\mathbf{G}\right)X(i)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\left[e_k^\top \mathbf{\Sigma}\mathbf{G}e_k X(i)^\top \mathbf{G}X(i) + e_k^\top \mathbf{\Sigma}\mathbf{G}X(i)e_k^\top \mathbf{G}X(i)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\mathrm{Tr}(\mathbf{\Sigma}\mathbf{G})X(i)^\top \mathbf{G}X(i) + X(i)^\top \mathbf{G}\mathbf{\Sigma}\mathbf{G}X(i)\right) \\
&= \mathbb{E}\left[\mathrm{Tr}\,\mathbf{G}\mathbf{\Sigma}\right] + \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}\mathbf{G})\,\mathrm{Tr}(\mathbf{G}\hat{\mathbf{\Sigma}})\right] + \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(\mathbf{G}\mathbf{\Sigma}\mathbf{G}\hat{\mathbf{\Sigma}})\right]
\end{aligned}
$$

Dividing both sides by n and by Gaussian Lipschitz concentration inequality, we have

$$
\frac{1}{n}\,\mathrm{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right) = \frac{1}{n}\,\mathrm{Tr}(\mathbf{G}\mathbf{\Sigma}) + \frac{1}{n}\,\mathrm{Tr}(\mathbf{\Sigma}\mathbf{G})\cdot\frac{1}{n}\,\mathrm{Tr}(\mathbf{G}\hat{\mathbf{\Sigma}}) + o(1).
$$

Proof of Ledoit and Péché (2011)

We proved that

$$\frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right) = \frac{1}{n}\operatorname{Tr}(\mathbf{G}\mathbf{\Sigma}) + \frac{1}{n}\operatorname{Tr}(\mathbf{\Sigma}\mathbf{G}) \cdot \frac{1}{n}\operatorname{Tr}(\mathbf{G}\hat{\mathbf{\Sigma}}) + o(1).$$

Proof of Ledoit and Péché (2011)

We proved that

$$\frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right) = \frac{1}{n}\operatorname{Tr}(\mathbf{G}\mathbf{\Sigma}) + \frac{1}{n}\operatorname{Tr}(\mathbf{\Sigma}\mathbf{G}) \cdot \frac{1}{n}\operatorname{Tr}(\mathbf{G}\hat{\mathbf{\Sigma}}) + o(1).$$

This implies that

$$\frac{1}{n}\operatorname{Tr}(\mathbf{G}\mathbf{\Sigma}) \approx \frac{\frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right)}{1 + \frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right)} + o(1).$$

Proof of Ledoit and Péché (2011)

We proved that

$$\frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right) = \frac{1}{n}\operatorname{Tr}(\mathbf{G}\mathbf{\Sigma}) + \frac{1}{n}\operatorname{Tr}(\mathbf{\Sigma}\mathbf{G}) \cdot \frac{1}{n}\operatorname{Tr}(\mathbf{G}\hat{\mathbf{\Sigma}}) + o(1).$$

This implies that

$$\frac{1}{n}\operatorname{Tr}(\mathbf{G}\mathbf{\Sigma}) \approx \frac{\frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right)}{1 + \frac{1}{n}\operatorname{Tr}\left(\mathbf{G}\hat{\mathbf{\Sigma}}\right)} + o(1).$$

In other words, we have

$$\begin{aligned}\frac{1}{n}\operatorname{Tr}\left((\lambda I + \hat{\mathbf{\Sigma}})^{-1}\mathbf{\Sigma}\right) &\approx \frac{\gamma - \frac{\lambda\gamma}{d}\operatorname{Tr}\left((\lambda I + \hat{\mathbf{\Sigma}})^{-1}\right)}{1 + \gamma - \frac{\lambda\gamma}{d}\operatorname{Tr}\left((\lambda I + \hat{\mathbf{\Sigma}})^{-1}\right)} + o(1) \\ &= \frac{1}{\gamma}\left(\frac{1}{\lambda v_{\hat{\mathbf{\Sigma}}}(-\lambda)} - 1\right) + o(1) \to \frac{1}{\gamma}\left(\frac{1}{\lambda v_{\mathbf{\Sigma}}(-\lambda)} - 1\right) \quad \square\end{aligned}$$

More General Results

Fixed β

- Eigenvalue Decomposition: $\Sigma \rightsquigarrow (s_1, v_1), \ldots, (s_d, v_d)$.
- Assume that

$$\widehat{H}_n(s) := \frac{1}{d} \sum_{i=1}^{d} 1_{\{s \geq s_i\}} \to H(s)$$

$$\widehat{G}_n(s) = \frac{1}{\|\beta\|_2^2} \sum_{i=1}^{d} \langle \beta, v_i \rangle^2 1_{\{s \geq s_i\}} \to G(s)$$

- Hastie, Montanari, Rosset, and Tibshirani (2020) derive the asumptotic risk in this case.

References

References

- **Dobriban and Wager** (2016)
  High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification
  Annals of Statistics.
- **Hastie, Montanari, Rosset, and Tibshirani** (2020)
  Surprises in High-Dimensional Ridgeless Least Squares Interpolation
  Annals of Statistics.
- **Ledoit and Péché** (2011)
  Eigenvectors of some large sample covariance matrix ensembles
  Probability Theory and Related Fields.
- **Benaych-Georges** (2022)
  A short proof of Ledoit-Péché's RIE formula for covariance matrices
  Technical Report.

Thank You!