Random Matrix Theory (RMT)

Semyon Malamud

EPFL

What is Random Matrix Theory (RMT)?

- Mathematical formalism to study high-dimensional random matrices.
- As of today, it is probably the right mathematical language for understanding Machine Learning.
- In high dimensions, stuff tends to “concentrate” and “simplify”

Concentration Phenomena: The Hidden Order in High-Dimensional Reality i

Suppose $X_t \in \mathbb{R}^P$ are i.i.d., with i.i.d. coordinates. Then, $X_{i,t}$ is wild and i.i.d. But,

$$P^{-1}\|X_t\|^2 \;=\; P^{-1}\sum_{i=1}^{P} X_{i,t}^2 \;\approx\; E[X_{i,t}^2] \tag{1}$$

There are more general concentration inequalities of this sort:

Meta-Theorem For good functions f,

$$\lim_{P\to\infty} (f(X_t) - E[f(X_t)]) \;\to\; 0 \tag{2}$$

in probability. This is a Law of Large Numbers in P (and not in T!)

Concentration Phenomena: The Hidden Order in High-Dimensional Reality ii

Click Here if You are Curious: Concentration Inequalities

Where are the traces of concentration in financial data? Where are the i.i.d., high-dimensional observations? Where is the hidden order?

Some Linear Algebra i

- A symmetric matrix $\Psi \in \mathbb{R}^{P \times P}$ admits a spectral decomposition $\Psi = UDU', \; D = \text{diag}(\lambda_i)$.
- Ψ is positive definite if and only if $\lambda_i > 0$.
-
$$\text{tr}(AB) \; = \; \text{tr}(BA) \tag{3}$$
-
$$\text{tr}(A^k) \; = \; \sum_i \lambda_i^k \tag{4}$$

Some Linear Algebra ii

- **Frobenius Norm**

$$\|A\|_2^2 \;=\; \sum_{i,j} A_{i,j}^2 \;=\; \mathrm{tr}(A^2) \;=\; \sum_i \lambda_i^2 \tag{5}$$

- Trace Norm

$$\|A\|_1 \;=\; \sum_i |\lambda_i| \tag{6}$$

- Spectral norm

$$\|A\| \;=\; \max_i |\lambda_i| \;=\; \max_x \|Ax\|/\|x\|$$

- Big and Small Matrices: $A = I/P$: Big or Small?

$$\|A\| = 1/P\,,\; \|A\|_1 \;=\; 1,\; \|A\|_2^2 \;=\; P^{-1}\,. \tag{7}$$

Concentration of Quadratic Forms: Heuristic i

Assumption We have $S_t \in \mathbb{R}^P$ given by $S_t = \Psi_P^{1/2} X_t$, where $X_t \in \mathbb{R}^P$ have i.i.d. coordinates, $X_{i,t}$, with $E[X_{i,t}] = 0, E[X_{i,t}^2] = 1, \ E[X_{i,t}^4] < \infty$.

E.g., $S_t \sim N(0, \Psi_P)$.

Theorem 1 (Pseudo-Theorem)

When P is large,

$$P^{-1} S_t' A S_t \approx P^{-1} \operatorname{tr}(A\Psi)$$

Concentration of Quadratic Forms: Heuristic ii

The intuition behind this lemma is particularly clear for the case $\Psi = I$ (i.e., signals $S_{i,t}$ are i.i.d. across i). Indeed, when P is large, a "law-of-large-numbers-in-P" implies that

$$\begin{aligned} P^{-1} S_t' A S_t &= P^{-1} \sum_{i,j} S_{i,t} S_{j,t} A_{i,j} \\ &= \underbrace{P^{-1} \sum_{i=1}^{P} S_{i,t}^2 A_{i,i}}_{\approx P^{-1} \sum_{i=1}^{P} A_{i,i} \;\; because \;\; E[S_{i,t}^2]=1} \\ &+ \underbrace{P^{-1} \sum_{i \neq j} S_{i,t} S_{j,t} A_{i,j}}_{\approx 0 \;\; because \;\; E[S_{i,t} S_{j,t}]=0} \\ &\approx P^{-1} \sum_{i=1}^{P} A_{i,i} = P^{-1} \operatorname{tr}(A) . \end{aligned} \tag{8}$$

Concentration of Quadratic Forms: Rigorous i

Lemma 2 (Concentration of Quadratic Forms)

Let A be a uniformly bounded matrix and let

$$Y_t \;=\; S_t' A S_t \,. \tag{9}$$

Then,

$$\mathrm{Var}[P^{-1} Y_t] \;\leq\; C \|A\| P^{-1} \tag{10}$$

for some constant $C = C(\|\Psi\|)$. Hence,

$$P^{-1} Y_t \;\approx\; E[P^{-1} Y_t] \;=\; P^{-1} \mathrm{tr}(A\Psi) \,. \tag{11}$$

Proof i

For simplicity, we assume $\Psi = I$ so that $S_t = X_t$. Then,

$$Y_t = \sum_{i,j} X_i X_j A_{i,j} \tag{12}$$

and therefore

$$E[Y_t] = E[X_t' A X_t] = E[\sum_{i,j} X_i X_j A_{i,j}] = \sum_{i,j} E[X_i X_j] A_{i,j} = \sum_{i,j} \delta_{i,j} A_{i,j} \tag{13}$$

and

$$E[Y_t^2] = \sum_{i_1,j_1,i_2,j_2} E[X_{i_1} X_{j_1} A_{i_1,j_1} A_{i_2,j_2} X_{i_2} X_{j_2}]\,. \tag{14}$$

Proof ii

Now, among all fourth-order moments, $E[X_{i_1}X_{j_1}X_{i_2}X_{j_2}]$, the only non-zero moments are those where either all are identical, $i_1 = i_2 = i_3 = i_4$, or when there are exactly two identical pairs. The latter can happen in exactly 3 ways. First, $(i_1 = i_2, j_1 = j_2)$, $(i_1 = j_2, j_1 = i_2)$ give rise to the terms $A^2_{i_1,j_1}$ because, by assumption, A is symmetric, so that $A_{i_1,j_1} = A_{j_1,i_1}$. Second, $(i_1 = j_1, i_2 = j_2)$ gives rise to $A_{i,i}A_{j,j}$.

Thus,

$$
\begin{aligned}
E[Y_t^2] &= \sum_{i_1,j_1,i_2,j_2} A_{i_1,j_1} A_{i_2,j_2} E[X_{i_1} X_{j_1} X_{i_2} X_{j_2}] \\
&= \sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j,i\neq j} (2A_{i,j}^2 + A_{i,i}A_{j,j}) \\
&= \sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j,i\neq j} 2A_{i,j}^2 \; - \; \sum_i A_{i,i}^2 \; + \; (\sum_i A_{i,i})^2 \\
&= \sum_i A_{i,i}^2 E[X_i^4] - 2\sum_i A_{i,i}^2 + \sum_{i,j} 2A_{i,j}^2 \; - \; \sum_i A_{i,i}^2 \; + \; (\sum_i A_{i,i})^2 \\
&= \sum_i A_{i,i}^2 (E[X_i^4] - 3) \; + \; 2\|A\|_2^2 \; + \; (\mathrm{tr}(A))^2
\end{aligned} \tag{15}
$$

Thus, since $E[Y_t] = \mathrm{tr}(A)$, we have

$$
E[Y_t^2] - E[Y_t]^2 \; = \; \sum_i A_{i,i}^2 (E[X_i^4] - 3) \; + \; 2\|A\|_2^2 \; \leq \; (E[X_i^4] - 1)\|A\|_2^2 \tag{16}
$$

because

$$\sum_{i} A_{i,i}^2 \;\leq\; \sum_{i,j} A_{i,j}^2 \;=\; \|A\|_2^2\,, \tag{17}$$

Correlated Case

Homework Use the uncorrelated case $Y_t = X_t' A_P X_t$ to prove the analogous result for the correlated case $Y_t = S_t' A_P S_t$ with $S_t = \Psi^{1/2} X_t$.

So What Can We Learn from One Observation?

We have

$$P^{-1} S_t' A S_t \;\approx\; P^{-1}\,\mathrm{tr}(\Psi A)\,. \tag{18}$$

So, measuing it for many A gives us everything we need to know about Ψ?

S_t has P dimensions, Ψ has P^2 dimensions??

Beware of multiple testing!!

Sample Covariance Matrix i

- Empirical covariance

$$\hat{\Psi} = \frac{1}{T}\sum_{t=1}^{T} S_t S_t' \in \mathbb{R}^{P\times P} \tag{19}$$

is an unbiased estimator

Howework:

$$E[\hat{\Psi}] = \Psi \,. \tag{20}$$

Howework:

$$E[\hat{\Psi}^2] = \Psi^2 + \textit{bias} \tag{21}$$

Sample Covariance Matrix ii

- Ridge regression

$$\hat{\beta} = (zI + \hat{\Psi})^{-1} \frac{1}{T} \sum_{t=1}^{T} S_t y_t \tag{22}$$

- We want to understand $\hat{\Psi}$
- It is a high-dimensional Random Matrix
- Is there a hidden structure inside it?
- Eigenvalue decomposition

$$\hat{\Psi} = \hat{U} \hat{D} \hat{U}' \tag{23}$$

- Eigenvectors $\hat{U}$ are poorly understood
- Eigenvalues $\hat{D}$ are much better understood

Stieltjes Transform and the Eigenvalue Distribution i

- A key object of RMT is the eigenvalue distribution $H_A(x)$ of a symmetric matrix $A \in \mathbb{R}^{P\times P}$:

$$H_A(x) \;=\; \frac{1}{P}\sum_{i=1}^{P} \mathbf{1}_{x<\lambda_i(A)}, \tag{24}$$

where $\lambda_i(A)$ are the eigenvalues of the matrix A.

- It is encoded in the Stieltjes Transform

$$m_\Psi(-z) \;=\; P^{-1}\mathrm{tr}((zI+\Psi)^{-1})\;,\;\; z>0\,, \tag{25}$$

because

$$P^{-1}\mathrm{tr}((zI+\Psi)^{-1}) \;=\; P^{-1}\sum_{i=1}^{P}((z+\lambda_i(\Psi))^{-1}) \;=\; \int \frac{1}{z+\lambda} dH_\Psi(\lambda)\,. \tag{26}$$

Stieltjes Transform and the Eigenvalue Distribution ii

- Of course, m_Ψ is not observable! We need to work with its sample counterpart

$$\hat{m}(-z) \;=\; P^{-1}\operatorname{tr}((zI+\hat{\Psi})^{-1})\;,\;\; z>0 \tag{27}$$

- It turns out that the key determinant of its behavior is complexity $c = P/T$.
- When $c \to 0$, we have (see below)

$$\hat{m}(-z) \;\approx\; m_\Psi(-z) \tag{28}$$

- When $c > 0$, this is not the case. A striking discovery of the RMT is that there is a universal way of linking $\hat{m}$ to m via a fixed point equation.

Stieltjes Transform and the Eigenvalue Distribution iii

Theorem 3 (Bai and Zhou (2008))

For each $z > 0$,

$$\lim_{T,P\to\infty,P/T\to c} \hat{m}(-z) \;=\; m(-z;c) \tag{29}$$

exists in probability and $m(-z;c)$ *is the unique positive solution to the nonlinear master equation*

$$m(-z;c) = \frac{1}{1 \;-\; c \;+\; c\,z\,m(-z;c)}\, m_\Psi\left(\frac{-z}{1 \;-\; c \;+\; c\,z\,m(-z;c)},\right) . \tag{30}$$

Understanding Marcenko-Pastur

Homework

- Howework: Derive m in closed form when $\Psi = I$ (this is the Marcenko-Pastur Theorem)
- Howework: Derive m in closed form when Ψ has just two eigenvalues λ_1, λ_2. What else matters in addition to λ_1, λ_2?

Click the button to reveal hidden content:

▸ Solving The Master Equation

Table of Contents

1. The Master Theorem of RMT
2. Proof of the Master Theorem
3. Ridge Regression
4. Proof of Bias-Variance Tradeoff
5. Appendix
6. Solving The Fixed Point Equation

Implicit Regularization i

Define the implicit shrinkage

$$Z_*(z; c) = \frac{z}{1 - c + c\,z\,m(-z; c)} \tag{31}$$

Implicit Regularization ii

Theorem 4

We have

$$z\, m(-z; c) \;=\; Z_*(z; c)\, m(-Z_*(z; c)) \tag{32}$$

That is, $(zI + \hat{\Psi})^{-1}$ *behaves as if we are doing* $(Z_* I + \Psi)^{-1}$.

Furthermore,

$$Z_* \;=\; z \;+\; cZ_* \int \frac{x dH(x)}{x + Z_*} \tag{33}$$

so that

$$Z_* \;\in\; [z,\; z + c]\,. \tag{34}$$

Implicit Regularization iii

Formally, in finite samples,

$$\begin{aligned} Z_* &= z + cZ_* \int \frac{xdH(x)}{x+Z_*} \\ &= z + \lim cZ_* P^{-1} \sum_i \lambda_i/(\lambda_i + Z_*) \\ &\approx z + Z_* \frac{P}{T} P^{-1} \operatorname{tr}(\Psi(\Psi + Z_*)^{-1}) \\ &= z + Z_* T^{-1} \operatorname{tr}(\Psi(\Psi + Z_*)^{-1}) \end{aligned} \tag{35}$$

The Master Theorem of RMT

Theorem 5

We have

- $$\beta' z(zI + \hat{\Psi})^{-1}\beta \;\to\; \beta' Z_*\,(Z_* I + \Psi)^{-1}\beta \tag{36}$$

- $$P^{-1}\,\mathrm{tr}(\underbrace{A}_{P\times P}\, z(zI + \hat{\Psi})^{-1}) \;\to\; P^{-1}\,\mathrm{tr}(\underbrace{A}_{P\times P}\, Z_*\,(Z_* I + \Psi)^{-1}) \tag{37}$$

for any bounded A!!!

Table of Contents

1 The Master Theorem of RMT

2 Proof of the Master Theorem

3 Ridge Regression

4 Proof of Bias-Variance Tradeoff

5 Appendix

6 Solving The Fixed Point Equation

Sherman-Morrison i

Lemma 6 (Sherman-Morrison Formula)

Suppose $A \in \mathbb{R}^{n\times n}$ *is an invertible square matrix and* $u,\ v \in \mathbb{R}^P$ *are column vectors. Then* $A + uv'$ *is invertible if* $1 + v'A^{-1}u \neq 0$. *In this case,*

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u} \tag{38}$$

and

$$(A + uv')^{-1}u = A^{-1}u\,\frac{1}{1 + v'A^{-1}u} \tag{39}$$

Homework: Prove Sherman-Morrison.

Concentration of Stieltjes Transform

Lemma 7

We have

$$P^{-1}\operatorname{tr}(Q_P(zI+\hat{\Psi}_T)^{-1}) \;\approx\; P^{-1}E[\operatorname{tr}(Q_P(zI+\hat{\Psi}_T)^{-1})] \tag{40}$$

almost surely for any sequence of uniformly bounded matrices Q_P.

What does this mean, and why is this striking?

- $\hat{\Psi}_T$ is **very random**
- $(zI+\hat{\Psi}_T)^{-1}$ is **very random (but bounded)**
  Homework: Prove that $\|(zI+\hat{\Psi}_T)^{-1}\| \;\le\; z^{-1}$.
- But $P^{-1}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1})$ is not random
- $P^{-1}\operatorname{tr}(Q_P(zI+\hat{\Psi}_T)^{-1})$ is also not random for any Q

[Proof of Lemma 7] The proof follows by the same arguments as in Bai and Zhou (2008). Let $\Psi_{T,t} = \frac{1}{T}\sum_{\tau\neq t} S_\tau S_\tau'$. By the Sherman-Morrison formula

$$(zI+\hat{\Psi}_T)^{-1} = (zI+\hat{\Psi}_{T,t})^{-1} \; - \; \frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_tS_t'(zI+\hat{\Psi}_{T,t})^{-1}\frac{1}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}. \tag{41}$$

Let E_t denote the conditional expectation given $S_1,\cdots,S_t$. Let also

$$q_T(z) = \frac{1}{P}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1}Q_P)\,.$$

With this notation, since $\hat{\Psi}_{T,t}$ is independent of S_t, we have

$$\begin{aligned} E_t[\frac{1}{P}\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1}Q_P)] \; &= \; E[\frac{1}{P}\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1}Q_P)|S_1,\cdots,S_{t-1},S_t] \\ = \; E[\frac{1}{P}\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1}Q_P)|S_1,\cdots,S_{t-1}] \; &= \; E_{t-1}[\frac{1}{P}\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1}Q_P)]\,. \end{aligned} \tag{42}$$

Formally, we can rewrite this as

$$(E_t - E_{t-1})[\frac{1}{P}\operatorname{tr}((zI + \hat{\Psi}_{T,t})^{-1} Q_P)] \;=\; 0\,. \tag{43}$$

Therefore,

$$E[q_T(z)] - q_T(z) = E_0[q_T(z)] - E_T[q_T(z)] \underbrace{=}_{telescope\ \ sum} \sum_{t=1}^{T}(E_{t-1}[q_T(z)] - E_t[q_T(z)])$$

$$= \sum_{t=1}^{T}(E_{t-1} - E_t)[q_T(z)]$$

$$= \sum_{t=1}^{T}(E_{t-1} - E_t)[q_T(z)] - \underbrace{(E_{t-1} - E_t)[\frac{1}{P}\operatorname{tr}((zI + \hat{\Psi}_{T,t})^{-1}Q_P)]}_{\color{red}{=0:\ \ we\ \ are\ \ subtracting\ \ zero}}$$

$$= \frac{1}{P}\sum_{t=1}^{T}(E_{t-1} - E_t)[\underbrace{\operatorname{tr}((zI + \hat{\Psi}_T)^{-1}Q_P)}_{=q_T} - \operatorname{tr}((zI + \hat{\Psi}_{T,t})^{-1}Q_P)]$$

$$= -\frac{1}{P}\sum_{t=1}^{T}(E_{t-1} - E_t)[\gamma_t]\,.$$

Let

$$\delta_t \;=\; -\frac{1}{P}(E_{t-1} - E_t)[\gamma_t] \;=\; E_{t-1}[q_T(z)] - E_t[q_T(z)] \tag{45}$$

be the martingale differences for the martingale $M_t \;=\; E_t[q_T(z)]$, where we have used (50) and defined

$$\begin{aligned}\gamma_t &= \operatorname{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t(1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}Q_P\right)\\ &\underbrace{=}_{cyclicality\;\;of\;\;trace} \frac{S_t'(zI+\hat{\Psi}_{T,t})^{-1}Q_P\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t}{(1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}\end{aligned} \tag{46}$$

We will need the following

Homework

Lemma 8

$$|x'ABAx| \;\leq\; \|A^{1/2}BA^{1/2}\|\, x'Ax \tag{47}$$

for any positive definite A.

Let

$$q_* \;=\; \sup_P \|Q_P\|.$$

Then,

$$|\gamma_t| \;=\; \frac{|S_t'(zI+\hat{\Psi}_{T,t})^{-1}Q_P\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t|}{(1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)} \tag{48}$$

Using (47) with $x=S_t,\; A=T^{-1}(zI+\hat{\Psi}_{T,t})^{-1},\; B=Q_P$, we get

$$|\gamma_t| \;=\; \frac{|x'ABAx|}{1+x'Ax} \;\leq\; \|A^{1/2}BA^{1/2}\|\frac{|x'Ax|}{1+x'Ax} \;\leq\; \|A^{1/2}BA^{1/2}\| \;\leq\; \|A\|\,\|B\| \;\leq\; z^{-1}q_*.$$

Thus, the margingale differences satisfy

$$|\delta_t| \;=\; |\frac{1}{P}(E_{t-1} - E_t)[\gamma_t]| \;\le\; P^{-1}(E_{t-1}[|\gamma_t|] \;+\; E_t[|\gamma_t|]) \;\le\; 2P^{-1}z^{-1}q_*.$$

We first prove a weaker form of our result.

Proposition 1

$E[(E[q_T(z)] - q_T(z))^2] \;\le\; P^{-2}T(2z^{-1}q_*)^2$ and, hence, $E[q_T(z)] - q_T(z) \to 0$ in probability when $P^{-2}T \to 0$.

The claim follows directly from the Ito isometry

$$E[(q_T - E[q_T])^2] \;=\; E[\sum_t \delta_t^2]$$

Homework: Prove this.

It turns out, however, that a more powerful result holds.

Theorem 9 (Burkholder-Davis-Gundy Inequality)

For any $q > 2$, where exists a $K_q > 0$ such that

$$E[(q_T - E[q_T])^q] \;\leq\; K_q E\left[\left(\sum_t \delta_t^2\right)^{q/2}\right].$$

Thus,

$$E[(q_T - E[q_T])^q] \;\leq\; K_q P^{-q} T^{q/2} (2z^{-1} q_*)^q \tag{49}$$

Almost sure convergence follows with $q > 2$ from the following lemma.

Lemma 10

Suppose that

$$E[|X_T|^q] \;\le\; T^{-\alpha}$$

for some $\alpha > 1$ and some $q > 0$. Then, $X_T \to 0$ almost surely.

Proof.

It is known that if

$$\sum_{T=1}^{\infty} Prob(|X_T| > \varepsilon) \;<\; \infty$$

for any $\varepsilon > 0$, then $X_T \to 0$ almost surely. In our case, the Chebyshev inequality implies that

$$Prob(|X_T| > \varepsilon) \;\le\; \varepsilon^{-q} E[|X_T|^q] \;\le\; \varepsilon^{-q} T^{-\alpha}$$

and convergence follows because $\alpha > 1$. ∎

The proof of the Lemma 7 is complete.

The ξ function i

- 99% of proofs in RMT use Sherman-Morrison:
- Let $\Psi_{T,t} = \frac{1}{T}\sum_{\tau\neq t} S_\tau S_\tau'$. By the Sherman-Morrison formula

$$(zI + \hat{\Psi}_T)^{-1} = (zI + \hat{\Psi}_{T,t})^{-1} - \frac{1}{T}(zI + \hat{\Psi}_{T,t})^{-1} S_t S_t' (zI + \hat{\Psi}_{T,t})^{-1} \frac{1}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t}. \tag{50}$$

- The quantity

$$(T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t \tag{51}$$

appears everywhere.

The ξ function ii

- Concentration of Quadratic Forms implies

$$
\begin{aligned}
T^{-1} S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t &= cP^{-1}\,\mathrm{tr}(S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t) \\
&= cP^{-1}\,\mathrm{tr}((zI+\hat{\Psi}_{T,t})^{-1}\underbrace{S_tS_t'}_{LLN}) \approx cP^{-1}\,\mathrm{tr}((zI+\hat{\Psi}_{T,t})^{-1}E[S_tS_t']) \\
&= cP^{-1}\,\mathrm{tr}((zI+\hat{\Psi}_{T,t})^{-1}\Psi)
\end{aligned}
\tag{52}
$$

- Question:

$$
\lim T^{-1}\,\mathrm{tr}((zI+\underbrace{\hat{\Psi}_{T,t}}_{known})^{-1}\underbrace{\Psi}_{unknown}) \tag{53}
$$

The ξ function Characterization i

Proposition 2

We have

$$\lim_{T\to\infty} \frac{1}{T} \mathrm{tr}((zI + \hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{54}$$

almost surely, where

$$\xi(z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}.$$

Intuition:

First, for large T,

$$\lim T^{-1}\mathrm{tr}((zI + \hat{\Psi}_{T,t})^{-1}\Psi) \;=\; \lim cP^{-1}\mathrm{tr}((zI + \hat{\Psi}_T)^{-1}\Psi) \tag{55}$$

The ξ function Characterization ii

Second, $\Psi \approx \hat{\Psi}$ and hence,

$$\begin{aligned} P^{-1}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1}\Psi) &\approx P^{-1}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T) \\ &= P^{-1}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1}(-zI+zI+\hat{\Psi}_T)) \\ &= P^{-1}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)) - z\,P^{-1}\operatorname{tr}((zI+\hat{\Psi}_T)^{-1}) \\ &= 1 - z\,\hat{m}(-z) \to 1 - z\,m(-z;c)\,. \end{aligned} \tag{56}$$

But this is wrong! The right expression is

$$\frac{1-zm(-z;c)}{c^{-1}-1+zm(-z;c)} \tag{57}$$

The ξ function Characterization iii

- Let

$$\hat{\Psi}_{T,t} = \frac{1}{T} \sum_{\tau \neq t} S_\tau S_\tau' \,. \tag{58}$$

Then, by the Sherman-Morrison formula (50),

$$\begin{aligned}
(zI + \hat{\Psi}_T)^{-1} S_t &= (zI + \hat{\Psi}_{T,t})^{-1} S_t \\
&- \frac{1}{T} (zI + \hat{\Psi}_{T,t})^{-1} S_t S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t \frac{1}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} \\
&= (zI + \hat{\Psi}_{T,t})^{-1} S_t \frac{1}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} \,.
\end{aligned} \tag{59}$$

The ξ function Characterization iv

- By concentration,

$$P^{-1}\, S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \;-\; P^{-1}\, \mathrm{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1}) \;\to\; 0 \tag{60}$$

in probability. At the same time, by Lemma 7,

$$P^{-1}\, \mathrm{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1}) \;-\; E[P^{-1}\, \mathrm{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1})] \;\to\; 0$$

almost surely. Thus,

$$P^{-1}\, S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \;-\; E[P^{-1}\, \mathrm{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1})] \;\to\; 0 \tag{61}$$

is probability and

$$P^{-1}\, \mathrm{tr}\, E[(zI + \hat{\Psi}_T)^{-1}] \;\to\; m(-z;c) \tag{62}$$

The ξ function Characterization v

- Now, we have

$$
\begin{aligned}
1 &= P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)] \\
&= P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}]z \;+\; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&= \; z\hat{m}(-z) \;+\; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_t S_t'] \\
&= \; \{symmetry\;\;across\;\;t\} \;=\; z\hat{m}(-z,c) \;+\; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}S_t S_t'] \\
&= \; \{using\;\;Sherman-Morrison\;(59)\} \\
&= \; z\hat{m}(-z) \;+\; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}S_t'] \\
&= \; z\hat{m}(-z) \;+\; E[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}].
\end{aligned}
\tag{63}
$$

The ξ function Characterization vi

Now, $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \leq \ c\|\Psi\|z^{-1}$ and hence is uniformly bounded. Let us pick a subsequence of T converging to infinity and such that $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \to q$ for some $q>0$. By (60),

$$\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t} \ \to \ \frac{c^{-1}q}{1+q}$$

in probability, and this sequence is uniformly bounded. Hence,

$$E[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+(T)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}] \ \to \ \frac{c^{-1}q}{1+q}$$

and we get

$$1-zm(-z,c)=\frac{c^{-1}q}{1+q}.$$

The ξ function Characterization vii

Thus, the limit of $\xi(z; c) = E[T^{-1}\,\mathrm{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1})]$ is independent of the subsequence of T and satisfies the required equation.

The proof of Proposition 3 is complete.

Marcenko-Pastur

$$\xi(z;c) = \lim T^{-1}\operatorname{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1}) \;=\; \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}, \tag{64}$$

$$m(-z;c) \;=\; \lim P^{-1}\operatorname{tr}((zI + \hat{\Psi}_{T,t})^{-1})$$

For $\Psi = \sigma^2 I$, we get

$$\xi(z;c) \;=\; c\sigma^2 m(-z;c)\,. \tag{65}$$

This gives a quadratic equation for m:

$$\sigma^2 m(-z;c) \;=\; \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}\,. \tag{66}$$

Proposition 3

We have

$$\lim_{T\to\infty} \frac{1}{T} \operatorname{tr}((zI + \hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{67}$$

almost surely, where

$$\xi(z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}.$$

Similarly,

$$\lim_{T\to\infty} \frac{1}{T} \operatorname{tr}((zI + \hat{\Psi})^{-2}\Psi) \;\to\; -\xi'(z;c) \tag{68}$$

almost surely, where

$$\xi'(z;c) = \frac{d}{dz}\left(\frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}\right). \tag{69}$$

Table of Contents

1 The Master Theorem of RMT

2 Proof of the Master Theorem

3 Ridge Regression

4 Proof of Bias-Variance Tradeoff

5 Appendix

6 Solving The Fixed Point Equation

Data Generating Process i

Assumption 1

There exists a vector $\beta \in \mathbb{R}^P$ *such that*

$$d_{t+1} \;=\; \beta' S_t \;+\; \varepsilon_{t+1}\,, t = 0, \cdots, t+1\,, \tag{70}$$

where $E[\varepsilon_{t+1}] = 0,\ E[\varepsilon_{t+1}^2] = \sigma_\varepsilon^2,\ E[\varepsilon_{t+1}^4] < \infty$ *are i.i.d., and* $S_t \;=\; \Psi^{1/2} X_t$, *where* $X_t = (X_{i,t})$ *where* $E[X_{i,t}] = 0,\ E[X_{i,t}^2] = 1,\ E[X_{i,t}^4] < \infty$ *are i.i.d., and* $\Psi = E[S_t S_t']$ *is p.s.d. and bounded.*

Below, we frequently use the convenient matrix notation $d = (d_\tau)_{\tau=1}^{T}$, *and* $S = (S_\tau)_{\tau=0}^{T-1} \in \mathbb{R}^{T \times P}$.

Data Generating Process ii

By (70), the total variance of d_{T+1} admits the standard decomposition

$$\mathrm{Var}[d_{T+1}] \;=\; \underbrace{\beta'\Psi\beta}_{explained\ \ variance} \;+\; \underbrace{\sigma_\varepsilon^2}_{irreducible\ \ noise} \quad . \tag{71}$$

An econometrician knowing the true β would then achieve the *infeasible* R^2 given by

$$R^2_{infeasible} \;=\; 1 \;-\; \frac{\sigma_\varepsilon^2}{\beta'\Psi\beta \;+\; \sigma_\varepsilon^2} \,. \tag{72}$$

Ridge Estimator Decomposition i

$$
\begin{aligned}
\hat{\beta}(z) &= (zI + \hat{\Psi})^{-1}\frac{S'd}{T} \\
&= (zI + \hat{\Psi})^{-1}\frac{S'(S\beta+\varepsilon)}{T} \\
&= \underbrace{(zI + \hat{\Psi})^{-1}\hat{\Psi}\beta}_{information} + \underbrace{(zI + \hat{\Psi})^{-1}\frac{S'\varepsilon}{T}}_{noise} \\
\hat{\pi}_T(z) &= \hat{\beta}_T(z)'S_T
\end{aligned}
\tag{73}
$$

be the ridge estimator.

Ridge Estimator Decomposition ii

Our goal is to characterize the out-of-sample behavior:

$$\begin{aligned} E_T\left[(d_{T+1} - \hat{\pi}_T(z))^2\right] &= E_T\left[(\beta' S_T + \varepsilon_{T+1} - \hat{\beta}(z)' S_T)^2\right] \\ &= E_T\left[\varepsilon_{T+1}^2 + (\beta' S_T - \hat{\beta}(z)' S_T)^2\right] \\ &= \sigma_\varepsilon^2 + E_T\left[(\beta' S_T - \hat{\beta}(z)' S_T)^2\right], \end{aligned} \tag{74}$$

Note that $\beta' S_T - \hat{\beta}(z)' S_T = (\beta - \hat{\beta}(z))' S_T$. Taking expectations and using $E_T\left[S_T S_T'\right] = \Psi$, we obtain

$$E_T\left[((\beta - \hat{\beta}(z))' S_T)^2\right] = (\beta - \hat{\beta}(z))' \Psi (\beta - \hat{\beta}(z)). \tag{75}$$

Ridge Estimator Decomposition iii

Thus, the out-of-sample prediction error becomes

$$E_T\left[(d_{T+1} - \hat{\pi}_T(z))^2\right] = \sigma_\varepsilon^2 + (\beta - \hat{\beta}(z))'\Psi(\beta - \hat{\beta}(z)). \tag{76}$$

Next, substitute the expression for $\hat{\beta}(z)$: $\hat{\beta}(z) = (zI + \hat{\Psi})^{-1}(\hat{\Psi}\beta + \frac{S'\varepsilon}{T})$. We can express β as

$$\beta = (zI + \hat{\Psi})^{-1}(zI + \hat{\Psi})\beta = (zI + \hat{\Psi})^{-1}\left(z\beta + \hat{\Psi}\beta\right).$$

Subtracting $\hat{\beta}(z)$ from β, we have:

$$\begin{aligned} \beta - \hat{\beta}(z) &= (zI + \hat{\Psi})^{-1}\left(z\beta + \hat{\Psi}\beta\right) - (zI + \hat{\Psi})^{-1}\left(\hat{\Psi}\beta + \frac{S'\varepsilon}{T}\right) \\ &= (zI + \hat{\Psi})^{-1}\left(z\beta - \frac{S'\varepsilon}{T}\right). \end{aligned} \tag{77}$$

Ridge Estimator Decomposition iv

Plugging (77) into (76) yields

$$(\beta - \hat{\beta}(z))'\Psi(\beta - \hat{\beta}(z)) = \left(z\beta - \frac{S'\varepsilon}{T}\right)'(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}\left(z\beta - \frac{S'\varepsilon}{T}\right). \tag{78}$$

Opening the brackets, we obtain

$$\begin{aligned}(\beta - \hat{\beta}(z))'\Psi(\beta - \hat{\beta}(z)) &= \left(z\beta - \frac{S'\varepsilon}{T}\right)'(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}\left(z\beta - \frac{S'\varepsilon}{T}\right) \\ &= z^2\,\beta'(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}\beta \\ &\quad - \frac{2z}{T}\,\beta'(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}S'\varepsilon \\ &\quad + \frac{1}{T^2}\,\varepsilon' S(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}S'\varepsilon.\end{aligned} \tag{79}$$

Ridge Estimator Decomposition v

Thus, the overall out-of-sample prediction error consists of three main terms in addition to the irreducible error, σ_ε^2. Each term corresponds to (1) the squared bias, (2) the cross-term between the signal and the noise, and (3) the variance term due to overfitting noise.

Proposition[Bias-Variance Tradeoff] We have

$$(\beta - \hat{\beta}(z))'\Psi(\beta - \hat{\beta}(z)) \;=\; \hat{\mathcal{B}}(z) \;-\; \hat{\mathcal{I}}(z) \;+\; \hat{\mathcal{V}}(z) \tag{80}$$

where

$$\begin{aligned} \hat{\mathcal{B}}(z) \;&=\; \underbrace{z^2\,\beta'(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\beta}_{bias} \;\geq\; 0 \\ \hat{\mathcal{V}}(z) \;&=\; \underbrace{\frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'\varepsilon}_{variance} \;\geq\; 0, \end{aligned} \tag{81}$$

while

$$\hat{\mathcal{I}}(z) \;=\; \underbrace{\frac{2z}{T}\,\beta'(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'\varepsilon}_{interaction} \tag{82}$$

satisfies

$$E[\hat{\mathcal{I}}(z)^2] \;\leq\; T^{-1}\, 4z^{-1}\sigma_\varepsilon^2 \,\|\beta\|^2 \,\|\Psi\|^2 \tag{83}$$

and, hence, is negligible for $T \to \infty$, irrespective of z and P.

Proposition shows how ridge regularization leads to the well known bias-variance tradeoff. However, the nature of this tradeoff changes drastically depending on whether we are in the *classical regime*, with $P < T$, or the *modern regime*, with $P > T$. In the classical regime, $\hat{\Psi}$ is typically non-degenerate and, hence, the bias term in (80) vanishes when $z \to 0$ because $\hat{\beta}(0)$ is the unbiased OLS estimator. At the same time, the variance term in (80) tends to be larger for small z. When $z \to \infty$, $\hat{\mathcal{V}}(z)$ vanishes, while the bias $\hat{\mathcal{B}}(z)$ converges to $\beta' \Psi \beta$. By contrast, in the over-parametrized regime when $P > T$, $\hat{\Psi} \in \mathbb{R}^{P \times P}$ is degenerate $(\text{rank}(\hat{\Psi}) \leq T)$ and, hence, the bias does not vanish even when $z \to 0$.

Table of Contents

1 The Master Theorem of RMT

2 Proof of the Master Theorem

3 Ridge Regression

4 Proof of Bias-Variance Tradeoff

5 Appendix

6 Solving The Fixed Point Equation

Interaction is Negligible

$$
\begin{aligned}
E[\hat{\mathcal{I}}(z)^2] &= \frac{4z^2}{T^2}\sigma_\varepsilon^2 E[\beta'(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'S(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\beta] \\
&= \frac{4z^2}{T}E[\beta'(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\beta] \\
&\le \|\beta\|^2\frac{4z^2}{T}z^{-1}\|\Psi\|z^{-1}\|\Psi\|z^{-1} \\
&= \|\beta\|^2\frac{4z^{-1}\|\Psi\|^2}{T}
\end{aligned}
\tag{84}
$$

Proof.

Lemma 11

Suppose that $u = (u_i)_{i=1}^{P}$ *where* u_i *are i.i.d., with* $E[u_i] = 0,\ E[u_i^2] = \sigma^2,\ E[u_i^4] < \infty$. *Suppose also* A_P *is a sequence of symmetric random matrices that is independent of* u *and is such that* $E[(P^{-1}\operatorname{tr}(A_P))^2] < K$ *and* $\lim P^{-2}E[\operatorname{tr}(A_P^2)] = 0$. *Then,*

$$\frac{1}{P}u'Au \ - \ P^{-1}\sigma^2\operatorname{tr}(A) \ \to \ 0 \tag{85}$$

in L_2 *and, hence, in probability.*

Lemma 12

Let $z \in \mathbb{R}$ be such that the matrices

$$zI_N + \frac{1}{T}S'S \quad \text{and} \quad zI_T + \frac{1}{T}SS'$$

are invertible. Then, the following identity holds:

$$S\left(zI_N + \frac{1}{T}S'S\right)^{-1} = \left(zI_T + \frac{1}{T}SS'\right)^{-1}S.$$

As a consequence, we have

$$\begin{aligned} S\left(zI_N + \frac{1}{T}S'S\right)^{-2} &= S\left(zI_N + \frac{1}{T}S'S\right)^{-1}\left(zI_N + \frac{1}{T}S'S\right)^{-1} \\ &= \left(zI_T + \frac{1}{T}SS'\right)^{-1}S\left(zI_N + \frac{1}{T}S'S\right)^{-1} \\ &= \left(zI_T + \frac{1}{T}SS'\right)^{-2}S. \end{aligned} \tag{86}$$

Lemma 13

If two symmetric matrices A, B satisfy $A \le B$ in the sense of positive definite order (i.e., $B - A \ge 0$), then

$$C'AC \;\le\; C'BC$$

for any matrix C. In particular, since $A \le \|A\| I$, we have

$$C'AC \;\le\; \|A\|\, C'C\,.$$

We can now use these lemmas to prove the following results.

Lemma 14

The matrix

$$A_T \; = \; T^{-1} S (zI + \hat{\Psi})^{-1} \Psi (zI + \hat{\Psi})^{-1} S' \tag{87}$$

is positive definite and uniformly bounded.

[Proof of Lemma 14] We have by Lemma 13 that

$$0 \;\leq\; A_T \;\leq\; \|\Psi\|\, T^{-1} S (zI + \hat{\Psi})^{-1} (zI + \hat{\Psi})^{-1} S' \tag{88}$$

By (86),

$$\|\Psi\|\, T^{-1} S (zI + \hat{\Psi})^{-1} (zI + \hat{\Psi})^{-1} S' \;=\; \|\Psi\| (zI + SS'/T)^{-2} SS'/T\,. \tag{89}$$

The matrix SS'/T is symmetric and positive definite. Hence, by the spectral theorem,

$$\|(zI + SS'/T)^{-2} SS'/T\| \;\leq\; \max_{\lambda} (z + \lambda)^{-2} \lambda \;\leq\; z^{-1}\,. \tag{90}$$

Lemma 15

We have

$$\frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'\varepsilon \;-\; \frac{\sigma_\varepsilon^2}{T}\,\mathrm{tr}\big(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\big) \;\to\; 0 \qquad (91)$$

in probability, as $T \to \infty$.

Proof i

Let

$$A \;=\; T^{-1}S(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'\,. \tag{92}$$

By Lemma 14, A is a random, uniformly bounded matrix that is independent of ε. Hence, by Lemma 11,

$$\frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'\varepsilon \;=\; T^{-1}\varepsilon' A\varepsilon \tag{93}$$

satisfies

$$T^{-1}\varepsilon' A\varepsilon \;-\; T^{-1}\sigma_\varepsilon^2\,\mathrm{tr}(A) \;\to\; 0\,. \tag{94}$$

Proof ii

Now, by the commutativity of the trace, $\mathrm{tr}(CD) = \mathrm{tr}(DC)$ for any matrices C, D and, hence,

$$\begin{aligned} \mathrm{tr}(A) &= \mathrm{tr}(T^{-1}S(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}S') \\ &= \mathrm{tr}(T^{-1}S'S(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}) \\ &= \mathrm{tr}(\hat{\Psi}(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1})\,. \end{aligned} \tag{95}$$

The Characterization of Variance i

Proposition 4

The quantity in Lemma 15 satisfies

$$\lim_{T\to\infty} \frac{\sigma_\varepsilon^2}{T} \operatorname{tr}\left(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\right) = \sigma_\varepsilon^2\left(\xi(z;c) + z\,\xi'(z;c)\right),$$

almost surely, where $\xi(z;c)$ and $\xi'(z;c)$ are defined in Proposition 3.

[Proof] We begin by noting that

$$\hat{\Psi}(zI+\hat{\Psi})^{-1} = I - z(zI+\hat{\Psi})^{-1}.$$

The Characterization of Variance ii

Multiplying both sides on the right by $\Psi(zI + \hat{\Psi})^{-1}$ gives

$$\hat{\Psi}(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1} = \Psi(zI + \hat{\Psi})^{-1} - z(zI + \hat{\Psi})^{-1}\Psi(zI + \hat{\Psi})^{-1}.$$

Taking the trace and dividing by T, we obtain

$$\frac{1}{T}\operatorname{tr}\Big(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\Big) = \frac{1}{T}\operatorname{tr}\Big(\Psi(zI+\hat{\Psi})^{-1}\Big) - z\,\frac{1}{T}\operatorname{tr}\Big((zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\Big).$$

By Proposition 3 we have that

$$\lim_{T\to\infty}\frac{1}{T}\operatorname{tr}\Big((zI + \hat{\Psi})^{-1}\Psi\Big) = \xi(z;c)$$

and

$$\lim_{T\to\infty}\frac{1}{T}\operatorname{tr}\Big((zI + \hat{\Psi})^{-2}\Psi\Big) = -\xi'(z;c).$$

The Characterization of Variance iii

Substituting these limits into the previous expression yields

$$\lim_{T\to\infty} \frac{1}{T} \operatorname{tr}\left(\hat{\Psi}(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\right) = \xi(z;c) + z\,\xi'(z;c).$$

Multiplying through by σ_ε^2 completes the proof.

Convergence of Derivatives

How did we get convergence of derivatives?

What about the Bias? i

$$\hat{\mathcal{B}}(z) = \underbrace{z^2\,\beta'(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\beta}_{bias} \geq 0$$

$$\hat{\mathcal{V}}(z) = \underbrace{\frac{1}{T^2}\,\varepsilon' S(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}S'\varepsilon}_{variance} \geq 0, \tag{96}$$

What about the Bias? ii

In general, the expression for $\hat{\mathcal{B}}(z)$ is complex. However, in the case when β is itself random, $\beta \sim N(0, \sigma_\beta^2/P)$, we get

$$\begin{aligned}
\hat{\mathcal{B}}(z) &= z^2\beta'(zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}\beta \\
&\approx \sigma_\beta^2 z^2 P^{-1}\,\mathrm{tr}((zI+\hat{\Psi})^{-1}\Psi(zI+\hat{\Psi})^{-1}) \\
&= \sigma_\beta^2 z^2 P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi})^{-2}) \\
&= -\sigma_\beta^2 z^2 \frac{d}{dz}(P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi})^{-1})) \\
&= -\frac{\sigma_\beta^2 z^2}{c}\frac{d}{dz}(T^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi})^{-1})) \\
&\approx -\frac{\sigma_\beta^2 z^2}{c}\frac{d}{dz}\xi(z;c)\,.
\end{aligned} \tag{97}$$

Table of Contents

1 The Master Theorem of RMT

2 Proof of the Master Theorem

3 Ridge Regression

4 Proof of Bias-Variance Tradeoff

5 Appendix

6 Solving The Fixed Point Equation

Table of Contents

1 The Master Theorem of RMT

2 Proof of the Master Theorem

3 Ridge Regression

4 Proof of Bias-Variance Tradeoff

5 Appendix

6 Solving The Fixed Point Equation

Solving The Fixed Point Equation i

For $P > T$, we have $c > 1$ and

$$\hat{m}(z) = P^{-1}\operatorname{tr}((\hat{\Psi} - zI)^{-1}) = -P^{-1}(P-T)z^{-1} + \textit{stuff from nonzero eigenvalues}$$
$$= -c^{-1}(c-1)z^{-1} + \textit{stuff from nonzero eigenvalues}\,. \tag{98}$$

Solving The Fixed Point Equation ii

Lemma 16 (Get rid of zero eigenvalues)

Let $z < 0$ and $c > 0$. Define

$$\tilde{m}(z;c) = cm(z;c) - (1-c)z^{-1} \tag{99}$$

Then we have,

$$\tilde{m}(z;c) > 0 \tag{100}$$

[**Proof of Lemma**] We have, for $z < 0$,

$$\tilde{m}(z;c) = \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1}\,\mathrm{tr}((\hat{\Psi} - zI)^{-1}) - (1-c)z^{-1} \tag{101}$$

Solving The Fixed Point Equation iii

When $c < 1$, we have $-(1-c)z^{-1} > 0$ and hence the result immediately follows. When $c \geq 1$, we have

$$\begin{aligned}\tilde{m}(z;c) &= \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1}\operatorname{tr}((\hat{\Psi} - zI)^{-1}) - (1-c)z^{-1} \\ &= \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1}\sum_{i=1}^{P}\frac{1}{\lambda_i(\hat{\Psi}) - z} - (1-c)z^{-1}\end{aligned} \tag{102}$$

Now, note that when $c \geq 1$, we have $P > T$; hence, $\hat{\Psi}$ has $P - T$ zero eigenvalues. Let us sort the eigenvalues in decreasing order $\lambda_1 \geq \ldots \geq \lambda_P \geq 0$, where eigenvalue from

Solving The Fixed Point Equation iv

index $i = T + 1$ to $i = P$ are zero. So,

$$\begin{aligned}
&\lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1} \sum_{i=1}^{P} \frac{1}{\lambda_i(\hat{\Psi}) - z} - (1-c)z^{-1} \\
&= \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1} \sum_{i=1}^{T} \frac{1}{\lambda_i(\hat{\Psi}) - z} + cP^{-1} \sum_{i=T+1}^{P} \frac{1}{\underbrace{\lambda_i(\hat{\Psi})}_{=0} - z} - (1-c)z^{-1} \qquad (103) \\
&= \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1} \sum_{i=1}^{T} \frac{1}{\lambda_i(\hat{\Psi}) - z} - cP^{-1}(P-T)\frac{1}{z} - (1-c)z^{-1}
\end{aligned}$$

Solving The Fixed Point Equation v

$$
\begin{aligned}
&= \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1}\sum_{i=1}^{T}\frac{1}{\lambda_i(\hat{\Psi})-z} - c\frac{1}{z} + \frac{1}{z} - (1-c)z^{-1} \\
&= \lim_{P\to\infty, T\to\infty, P/T\to c} cP^{-1}\sum_{i=1}^{T}\frac{1}{\lambda_i(\hat{\Psi})-z} > 0\,.
\end{aligned} \tag{104}
$$

Hence,

$$\tilde{m}(z;c) > 0 \tag{105}$$

Homework: Let $z < 0$ and $c > 0$. Prove that

$$\tilde{m}'(z;c) = cm'(z;c) + (1-c)z^{-2} \tag{106}$$

Solving The Fixed Point Equation vi

satisfies

$$\tilde{m}'(z; c) > 0 \tag{107}$$

Deriving a Clean Fixed Point Equation i

For $z < 0$, $m(z;c)$ is the unique positive solution to the nonlinear master equation

$$m(z;c) = \frac{1}{1 - c - c\,z\,m(z;c)}\, m\left(\frac{z}{1 - c - c\,z\,m(z;c)}\right), \tag{108}$$

where

$$m(z) = \int \frac{dH(x)}{x - z}. \tag{109}$$

Substituting

$$\begin{aligned} \tilde{m}(z;c) &= -(1-c)z^{-1} + cm(z;c) \\ \Leftrightarrow\ z\tilde{m}(z;c) &= -(1-c) + czm(z;c) \end{aligned} \tag{110}$$

Deriving a Clean Fixed Point Equation ii

into the Master equation, we get

$$\begin{aligned} zm(z;c) &= \frac{z}{1 \; - \; c \; - \; c\,z\,m(z;c)}\, m\left(\frac{z}{1 \; - \; c \; - \; c\,z\,m(z;c)}\right) \\ &= \; -\frac{z}{z\tilde{m}(z;c)} m(-1/\tilde{m}(z;c)) \; = \; -\frac{1}{\tilde{m}(z;c)} \int \frac{dH(x)}{x + 1/\tilde{m}(z;c)}, \end{aligned} \tag{111}$$

that is

$$zm(z;c) \;=\; -\int \frac{dH(x)}{\tilde{m}(z;c)x + 1}\,. \tag{112}$$

Rewriting

$$zm(z;c) \;=\; c^{-1} z\tilde{m}(z;c) + c^{-1}(1 - c) \tag{113}$$

Deriving a Clean Fixed Point Equation iii

and substituting gives

$$c^{-1} z \tilde{m}(z;c) + c^{-1}(1-c) \;=\; -\int \frac{dH(x)}{\tilde{m}(z;c)x + 1}, \tag{114}$$

which can be rewritten as

$$c - 1 \;-\; z\tilde{m} - c \int \frac{dH(x)}{(1+\tilde{m}x)} \;=\; 0\,. \tag{115}$$

We can also rewrite it as an equation for

$$Z_*(z;c) \;=\; 1/\tilde{m}\,: \tag{116}$$

Deriving a Clean Fixed Point Equation iv

$$\begin{aligned} 0 &= c - 1 - z\tilde{m} - c\int \frac{dH(x)}{(1+\tilde{m}x)} \\ &= -1 - z\tilde{m} + c(1 - \int \frac{dH(x)}{(1+\tilde{m}x)}) \\ &= -1 - z\tilde{m} + c\tilde{m}\int \frac{xdH(x)}{(1+\tilde{m}x)} \\ &= -1 - z\tilde{m} + c\int \frac{xdH(x)}{Z_* + x} \end{aligned} \tag{117}$$

That is,

$$1 = -z\tilde{m} + c\int \frac{xdH(x)}{Z_* + x} \Leftrightarrow Z_* = -z + cZ_*\int \frac{xdH(x)}{x + Z_*} \tag{118}$$

Go Back

Bai, Zhidong and Wang Zhou, “Large sample covariance matrices without independence structures in columns,” *Statistica Sinica*, 2008, pp. 425–442.