Random Matrix Theory and Machine Learning for Finance

Semyon Malamud

EPFL

Table of Contents

Intoduction to Asset Pricing i

- assets $i = 1, \cdots, N$ have prices $P_{i,t}$ and *excess* returns

$$R_{i,t+1} = \frac{P_{i,t+1} + D_{i,t+1}}{P_{i,t}} - \underbrace{R_{f,t}}_{risk\ \ free\ \ rate} \tag{1}$$

- if you invest fraction $\pi_{i,t}$ of your wealth W_t into security i, the rest stays on your bank account and grows at the rate $R_{f,t}$:

$$W_t = \sum_i \underbrace{\pi_{i,t} W_t}_{investment\ \ in\ \ stock\ \ i} + \underbrace{(W_t - \sum_i \pi_{i,t} W_t)}_{bank\ \ account} \tag{2}$$

and then you sell your investments at time t and collect dividends so that

$$\begin{aligned} W_{t+1} &= \sum_i W_t \pi_{i,t} \frac{P_{i,t+1} + D_{t+1}}{P_{i,t}} + (W_t - \sum_i \pi_{i,t} W_t) R_{f,t} \\ &= W_t R_{f,t} + W_t \sum_i \pi_{i,t} R_{i,t+1} \end{aligned} \tag{3}$$

Intoduction to Asset Pricing ii

- Thus, the excess return on your wealth is

$$\frac{W_{t+1}}{W_t} - R_{f,t} = \sum_i \pi_{i,t} R_{i,t+1} = \pi_t' R_{t+1} \tag{4}$$

- Thus, we want π_t that gives good returns. But what is the criterion?
- mean-variance optimization:

$$\pi_t = \arg\max_{\pi_t} \left(E_t[\pi_t' R_{t+1}] - 0.5 \underbrace{\gamma}_{\text{risk aversion}} \mathrm{Var}_t[\pi_t' R_{t+1}] \right) \tag{5}$$

and hence the Mean-Variance Efficient (MVE) portfolio is

$$\underbrace{\pi_t}_{\text{tangency portfolio}} = \gamma^{-1} \underbrace{(\mathrm{Var}_t[R_{t+1}])^{-1}}_{N \times N \text{ covariance matrix}} \underbrace{E_t[R_{t+1}]}_{N \times 1 \text{ expected returns}} \tag{6}$$

Intoduction to Asset Pricing iii

- Similarly,

$$\begin{aligned} \tilde{\pi}_t &= \gamma^{-1}(E_t[R_{t+1}R'_{t+1}])^{-1}\, E_t[R_{t+1}] \\ &= \frac{1}{1 + E_t[R_{t+1}]' \mathrm{Var}_t[R_{t+1}]^{-1} E_t[R_{t+1}]} \pi_t \end{aligned} \tag{7}$$

where

$$E_t[R_{t+1}R'_{t+1}] = \mathrm{Var}_t[R_{t+1}] + E_t[R_{t+1}]E_t[R_{t+1}]' \tag{8}$$

The Sherman-Morrison formula i

The magic behind is the

Lemma (Sherman-Morrison formula)

$$(A + xx')^{-1} = A^{-1} - \frac{A^{-1}xx'A^{-1}}{1 + x'A^{-1}x} \tag{9}$$

and

$$(A + xx')^{-1}x = \frac{A^{-1}x}{1 + x'A^{-1}x} \tag{10}$$

The Sherman-Morrison formula ii

Proof of the Sherman-Morrison formula.

Recall that

$$xx' \;=\; (x_i x_j)_{i,j=1}^{N}$$

is a symmetric, positive, semi-definite, *rank* $-\,1$ matrix (all columns are proportional to x). Then,

$$\begin{aligned}
&(A+xx')(A^{-1} \;-\; \frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x}) \\
&= \; I \;-\frac{xx'A^{-1}}{1+x'A^{-1}x} + xx'A^{-1} - xx'\frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x} \qquad (11)\\
&= \; I \;-\frac{xx'A^{-1}}{1+x'A^{-1}x} + xx'A^{-1} - xx'A^{-1}\frac{x'A^{-1}x}{1+x'A^{-1}x} \;=\; I
\end{aligned}$$

and

$$(A+xx')^{-1}x \;=\; (A^{-1} \;-\; \frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x})x \;=\; \frac{A^{-1}x}{1+x'A^{-1}x} \qquad (12)$$

■

SDF i

- Ok, this is the optimal portfolio, but what does this have to do with asset pricing?
- Intuitively, we expect that

$$P_{i,t} = \underbrace{(R_{f,t})^{-1} E_t[P_{i,t+1} + D_{i,t+1}]}_{\textit{Seems wrong in the data}} \Leftrightarrow E_t[R_{i,t+1}] = 0 \tag{13}$$

because the discount factor $(R_{f,t})^{-1}$ is too naive

- Unless we think beliefs are irrational and wild
- But this actually does hold in the data
- But then,

$$P_{i,t} = \underbrace{(R_{f,t})^{-1} E_t[WrongBeliefs(t, t+1)(P_{i,t+1} + D_{i,t+1})]}_{\textit{Seems ok in the data}} \tag{14}$$

- See Belief Overreaction and Stock Market Puzzles

SDF ii

- We need a smart discount factor (SDF):

$$P_{i,t} \;=\; E_t[\;\underbrace{M_{t,t+1}}_{\textit{stochastic discount factor}}\;(P_{i,t+1}+D_{i,t+1})] \tag{15}$$

- with a bit of algebra, this is equivalent to

$$E_t[R_{i,t+1}\,M_{t,t+1}] \;=\; 0 \tag{16}$$

- SDF is the most important object of asset pricing

Theorem (SDF = Efficient Portfolio)

- *Fix a portfolio q_t. Then,*

$$M_{t+1} \;=\; 1 \;-\; q_t' R_{t+1} \;=\; 1 \;-\; R_{t+1}' q_t \tag{17}$$

is an SDF if and only if $q_t = \tilde{\pi}_t$.

SDF iii

Proof.

$$\begin{aligned} E_t[R_{t+1}\, M_{t,t+1}] &= E_t[R_{t+1}\,(1 - R'_{t+1} q_t)] \\ &= E_t[R_{t+1}] - E_t[R_{t+1} R'_{t+1}]\, q_t = 0 \end{aligned} \tag{18}$$

implies

$$q_t = E_t[R_{t+1} R'_{t+1}]^{-1}\, E_t[R_{t+1}] \tag{19}$$

- Now comes the big question: How do we measure the conditional expectations, $E_t[R_{t+1}]$ and $E_t[R_{t+1} R'_{t+1}]$?
- Once can start with a simple prediction problem: measure $E_t[R_{t+1}]$ by running a regression on observables (economic variables) S_t using past data (time series prediction)
- Interesting already with a single asset: If we new $E_t[R_{t+1}]$ and $E_t[R^2_{t+1}]$ (single asset), we would have

$$M_{t,t+1} = 1 - \tilde{\pi}_t\, R_{t+1}, \text{ where } \tilde{\pi}_t = \frac{E_t[R_{t+1}]}{E_t[R^2_{t+1}]} \in \mathbb{R} \tag{20}$$

is a infeasible timing strategy.

SDF iv

- Let us understand how to use timing strategies to learn about $M_{t,t+1}$. We have

$$\begin{aligned} & E_t[R_{t+1}M_{t+1}] \;=\; 0 \\ \Leftrightarrow\; & \underbrace{E_t[R_{t+1}M_{t+1}] - E_t[R_{t+1}]\,E_t[M_{t+1}]}_{=\mathrm{Cov}_t(R_{t+1},M_{t+1})} \;=\; -E_t[R_{t+1}]\,E_t[M_{t+1}] \\ \Leftrightarrow\; & E_t[R_{t+1}] \;=\; -R_{f,t}\mathrm{Cov}_t(R_{t+1},M_{t+1}) \end{aligned} \tag{21}$$

- Cauchy-Schwarz inequality:

$$|\mathrm{Cov}_t(R_{t+1},M_{t+1})| \;\leq\; STD_t(R_{t+1})\,STD_t(M_{t+1}) \tag{22}$$

implies the conditional Hansen-Jagannathan bound

$$STD_t(M_{t+1}) \;\geq\; R_{f,t}\,SR_t(R_{t+1}) \tag{23}$$

- Suppose we have a timing strategy π_t and define

$$R^{\pi}_{t+1} \;=\; \pi_t \, R_{t+1} \tag{24}$$

- Then,

$$0 \;=\; \pi_t E_t[R_{t+1} M_{t+1}] \;=\; E_t[\pi_t R_{t+1} M_{t+1}] \;=\; E_t[R^{\pi}_{t+1} M_{t+1}] \tag{25}$$

so that

$$0 \;=\; E[E_t[R^{\pi}_{t+1} M_{t+1}]] \underbrace{=}_{\textit{iterated expectations}} E[R^{\pi}_{t+1} M_{t+1}] \tag{26}$$

so that

$$E[R^{\pi}_{t+1}] \;=\; -\frac{\mathrm{Cov}(R^{\pi}_{t+1}, M_{t+1})}{E[M_{t+1}]} \tag{27}$$

- define

$$STD(X) \;=\; \mathrm{Var}[X]^{1/2}, \; SR(X) \;=\; E[X]/STD(X) \tag{28}$$

- Cauchy-Schwarz inequality:

$$|\mathrm{Cov}(R^{\pi}_{t+1}, M_{t+1})| \;\leq\; STD(R^{\pi}_{t+1})\, STD(M_{t+1}) \tag{29}$$

  implies the unconditional Hansen-Jagannathan bound

$$STD(M_{t+1}) \;\geq\; E[R^{-1}_{f,t}]^{-1}\, SR(R^{\pi}_{t+1}) \tag{30}$$

- Thus, any "good" timing strategy provides important information about the STD of the SDF.

- Why do we care? Well, suppose we have some other pricing kernel $\tilde{M}_{t,t+1}$, say, coming from a consumption-based or some other form of a macro or behavioral model. Let Π be the orthogonal projection onto the asset span. First,

$$0 \;=\; E_t[R_{t+1}\tilde{M}_{t,t+1}] \;=\; E_t[R_{t+1}\, M_{t,t+1}] \tag{31}$$

Second,

$$0 \;=\; E_t[R_{t+1}\Pi(\tilde{M}_{t,t+1})] \tag{32}$$

but since both M and $\Pi(\tilde{M})$ are in the asset span, we have

$$\Pi(\tilde{M}) \;=\; M \tag{33}$$

and hence

$$E_t[M_{t,t+1}] \;=\; E_t[\Pi(\tilde{M}_{t,t+1})] \;=\; E_t[\tilde{M}_{t,t+1}]$$

while (since Π is the orthogonal projection),

$$E_t[\tilde{M}_{t,t+1}^2] \;=\; \|\tilde{M}_{t,t+1}\|^2 \;\geq\; \|\Pi(\tilde{M}_{t,t+1})\|^2 \;=\; \|M_{t,t+1}\|^2 \;=\; E_t[M_{t,t+1}^2]$$

and hence (iterated expectations)

$$E[\tilde{M}_{t,t+1}^2] \;\geq\; E[M_{t,t+1}^2]\,,$$

implying that macro pricing kernel is always more volatile than the tradable pricing kernel:

$$\begin{aligned} STD_t(\tilde{M}_{t+1}) \;&\geq\; STD_t(M_{t+1}) \\ STD(\tilde{M}_{t+1}) \;&\geq\; STD(M_{t+1}) \end{aligned} \tag{34}$$

- Thus, macro makes asset pricing puzzles worse

Table of Contents

Statistical Wisdom and Overfitting

Typically, we aim for a trade-off between

- Fit to the training data, e.g.,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_i)-y_i\right)\Big)^2,\ X_i\in R^d,\ y_i\in\mathbb{R}$$

- Complexity of a prediction rule, e.g.,
  - Number of parameters $d\sim n$?
  - Norm of the parameter vector
  - Bandwidth of smoothing kernel
  - ...

This is especially important for nonparametric methods, that is, those for which the number of parameters grows with the sample size.

Statistical Wisdom and Overfitting

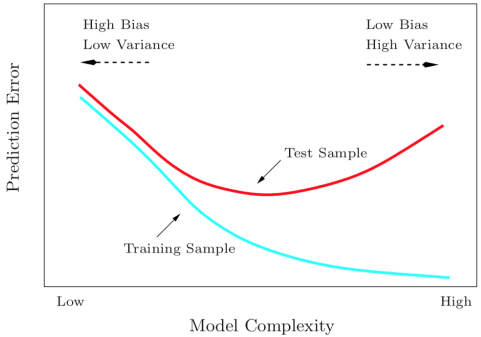
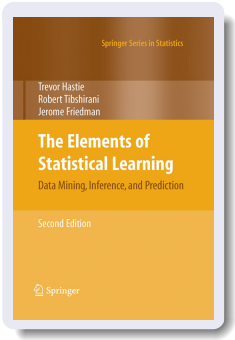


FIGURE 2.11. *Test and training error as a function of model complexity.*

Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In



Statistical Wisdom and Overfitting

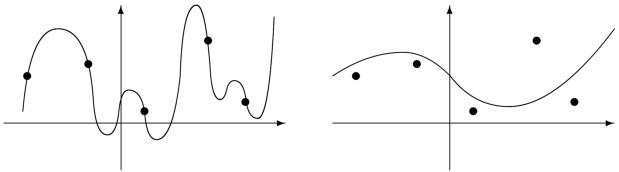
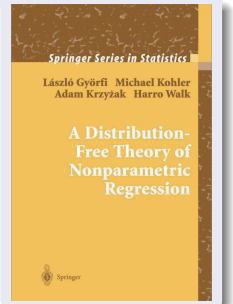
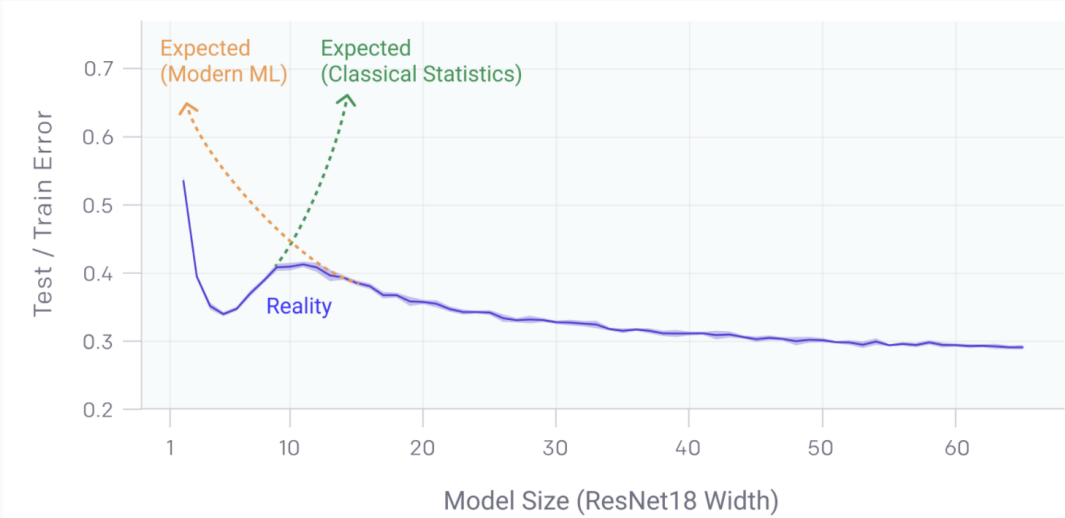
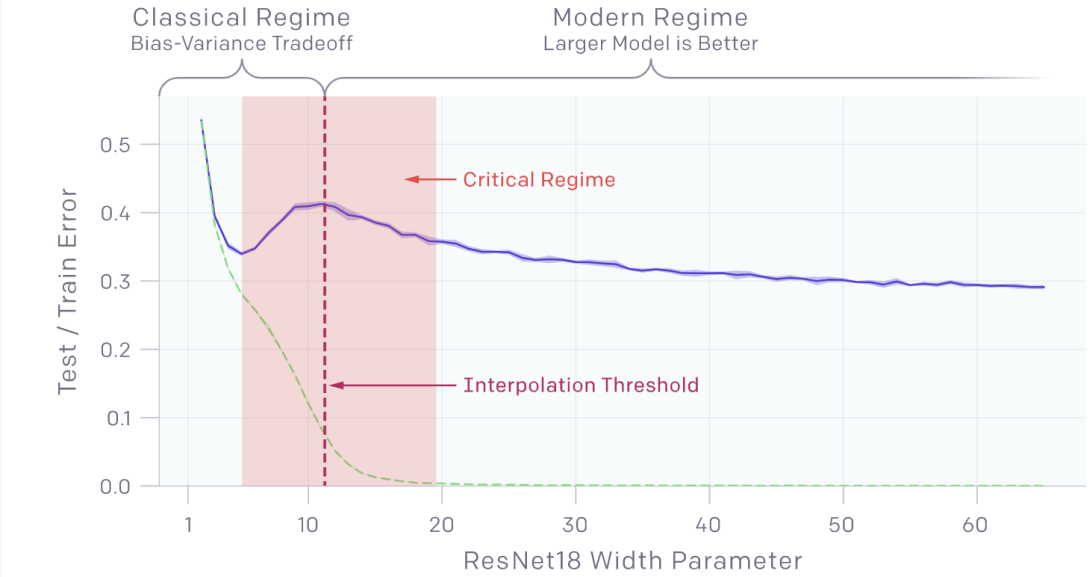


Figure 2.3. The estimate on the right seems to be more reasonable than the estimate on the left, which interpolates the data.

over $\mathcal{F}_n$. Least squares estimates are defined by minimizing the empirical L_2 risk over a general set of functions $\mathcal{F}_n$ (instead of (2.7)). Observe that it doesn't make sense to minimize (2.9) over all (measurable) functions f, because this may lead to a function which interpolates the data and hence is




Expected
(Modern ML)
Expected
(Classical Statistics)
Reality
Test / Train Error
0.7
0.6
0.5
0.4
0.3
0.2
1
10
20
30
40
50
60
Model Size (ResNet18 Width)


Classical Regime
Bias-Variance Tradeoff
Modern Regime
Larger Model is Better
Critical Regime
Interpolation Threshold
Test / Train Error
0.0
0.1
0.2
0.3
0.4
0.5
1
10
20
30
40
50
60
ResNet18 Width Parameter

Virtue of Complexity

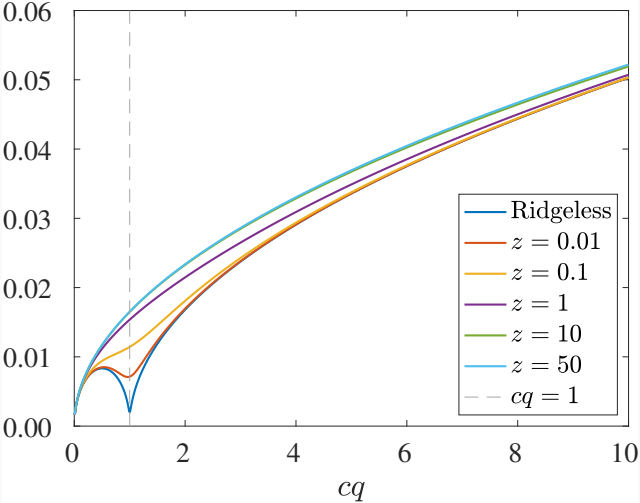


Figure: Expected Out-of-sample Timing Strategy Sharpe Ratio From Mis-specified Models

What Does This Mean For Finance?

- $E_t[R_{t+1}]$, $E_t[R_{t+1}R'_{t+1}]$ and M_{t+1} are not observable and also **not learnable**
- Classical Econometrics:
  - Make unrealistic conditions so that $E_t[R_{t+1}]$, $E_t[R_{t+1}R'_{t+1}]$ become learnable.
  - Thus, assume that we have learned the true pricing kernel
  - In-sample and out-of-sample is the same
  - Enjoy an easy life
- Complexity Econometrics:
  - Be honest and realistic
  - Admit that M_{t+1} is not observable and also **not learnable**
  - Predict the difference between in-sample and out-of-sample behavior
  - Understand the true nature of the data without being able to learn the details of the DGP. Depends crucially on the degree of sparsity
  - Look for learning algorithms and shrinkage:

$$M_{t+1} \;=\; F(S_t,\; R_{t+1},\; Data([1,t])) \tag{35}$$

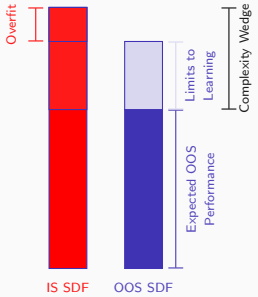
  - **How do we select F when the world is complex?**

Complexity and the SDF: The Wedge

1. "Complexity wedge" = IS Performance – Expected OOS Performance

   "Complexity wedge" = $\underbrace{\text{IS} - \text{True}}_{\text{"Overfit"}} + \underbrace{\text{True} - \text{OOS}}_{\text{"Limits to Learning"}}$

- Quantifiable based on training data
- Can infer the performance of true SDF and how far you are from it but cannot recover it!



Summary

- Pricing Kernel $= 1 -$ *ReturnOnEfficientPortfolio*.
- Conditional Pricing Kernel $= 1 -$ *ReturnOnConditionallyEfficientPortfolio*.
- *ConditionallyEfficientPortfolio* $= \Sigma_t[R]^{-1} E_t[R]$.
- Neither $\Sigma_t[R]$ nor $E_t[R]$ are learnable in finite samples.
- We can only build "feasible" counter-parts
- We need statistical tools to know how far we are from the "truth."
- Every feasible strategy defines a lower bound on the variability of the "true" SDF

Table of Contents

OLS

- Predictors' vector $S_t \in \mathbb{R}^P$
- OLS (linear regression, Carl Friedrich Gauss):

$$\hat{\beta}_{OLS} \;=\; \arg\min_{\beta} \|R - \beta' S\|^2 \;=\; \arg\min_{\beta} \sum_t \left(R_{t+1} - \sum_j S_{j,t}\beta_j \right)^2$$

and

$$\hat{\beta}_{OLS} \;=\; \hat{\Psi}_T^{-1} T^{-1} \sum_t S_t' R_{t+1} \;\in\; \mathbb{R}^P$$

-

$$\hat{\Psi}_T \;=\; T^{-1} \sum_t S_t S_t' \;\in\; \mathbb{R}^{P \times P} \tag{36}$$

is the sample covariance matrix

- it is BLUE (Best **Linear** Unbiased Estimator): If

$$R_{t+1} \;=\; S_t'\beta \;+\; \varepsilon_{t+1} \;=\; \sum_j S_{j,t}\beta_j \;+\; \varepsilon_{t+1} \tag{37}$$

with $E[S_t\varepsilon_{t+1}] = 0$, then

$$\begin{aligned}
\hat{\beta}_{OLS} \;&=\; \hat{\Psi}_T^{-1} T^{-1} \sum_t S_t R_{t+1} \\
&=\; \hat{\Psi}_T^{-1} T^{-1} \sum_t S_t (S_t'\beta \;+\; \varepsilon_{t+1}) \\
&=\; \hat{\Psi}_T^{-1} (\underbrace{T^{-1} \sum_t S_t S_t'}_{=\hat{\Psi}_T} \beta \;+\; T^{-1} \sum_t S_t \varepsilon_{t+1}) \\
&=\; \beta \;+\; \underbrace{\hat{\Psi}_T^{-1} T^{-1} \sum_t S_t \varepsilon_{t+1}}_{noise}
\end{aligned} \tag{38}$$

and hence

$$E[\hat{\beta}_{OLS}] \underbrace{=}_{unbiased} \beta \tag{39}$$

- If $E_t[\varepsilon_{t+1}] = 0$, then

$$E_t[R_{t+1}] \;=\; S_t'\beta \tag{40}$$

and hence, at least intuitively, we expect that

$$E_t[R_{t+1}] \;\approx\; S_t'\hat{\beta}_{OLS} \tag{41}$$

which can make money and tell us about SDF: With $\pi_t \;=\; \hat{\beta}'_{OLS}S_t$,

$$E[R^{\pi}_{t+1}] \;=\; E[\pi_t E_t[R_{t+1}]] \underbrace{\approx}_{kind\ of\ intuitive} E[\pi_t^2]\ is\ big? \tag{42}$$

- All these arguments can be made rigorous when $c = P/T \approx 0$: zero complexity models = under-parametrized models
- However, when $c > 0$, we face the risk of over-fitting

How to Run Regression when $P > T$? Ridge Regression i

$$\hat{\beta} \;=\; \arg\min_{\beta}(T^{-1}\|R - \beta' S\|^2 + z\|\beta\|^2)$$

gives

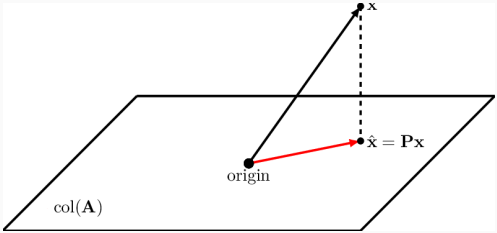
$$\hat{\beta}(z) \;=\; \left(zI + T^{-1}\sum_t S_t S_t'\right)^{-1} T^{-1}\sum_t S_t R_{t+1}$$

When P moves beyond T, there are more parameters than observations and the least squares problem has multiple solutions. A particularly interesting solution invokes the Moore-Penrose pseudo-inverse, $\left(T^{-1}\sum_t S_t S_t'\right)^{+}\frac{1}{T}\sum_t S_t R_{t+1}$. Recall that the Moore-Penrose pseudo-inverse A^+ of a matrix A is defined via $A^+ = (A'A)^{-1}A'$ if $A'A$ is invertible, and $A^+ = A'(AA')^{-1}$ if AA' is invertible. This solution is equivalent to the ridge estimator as the shrinkage parameter approaches zero:

$$\hat{\beta}(0^+) = \lim_{z\to 0+}\left(zI + T^{-1}\sum_t S_t S_t'\right)^{-1}\frac{1}{T}\sum_t S_t R_{t+1}.$$

Orthogonal Projection in Linear Algebra i

Given a vector x and a subspace S of $\mathbb{R}^n$, given by the span of columns of a matrix A (denoted by $col(A)$), the **orthogonal projection** P of x onto S, is the closest vector $\hat{x}$ in S to x.



Orthogonal Projection in Linear Algebra ii

Lemma

Let $\Pi = (AA')^{+}AA'$. *Then,* Π *is the orthogonal projection onto the* $col(A)$.

Proof.

First, Π is symmetric and positive definite. Second, $AA' = U\,\mathrm{diag}(D)U'$ and hence

$$(AA')^{+}AA' = U \lim_{z\to 0+} \mathrm{diag}(D/(D+z))U' \;=\; U\,\mathrm{diag}(\mathbf{1}_{D>0})U'. \tag{43}$$

and hence Π is the projection onto the image of AA', while it kills the kernel of AA'. But $AA'x = 0$ means $x'AA'x = 0$ means $A'x = 0$ and we know

$$\ker A' = (ImA)^{\perp}. \tag{44}$$

■

What does the Ridgeless Regression Do? i

Any solution $\hat{\beta} \in \mathbb{R}^P$ to the system

$$\hat{\beta}' S_t \;=\; R_{t+1},\;\; t = 1, \cdots, T \tag{45}$$

of T equations and P variables is called an interpolator.

Lemma

Suppose that $R \in \mathrm{span}(S)$, so that (45) has a solution. Then, any solution β to (45) can be written down as

$$\beta \;=\; \underbrace{\hat{\beta}(0^+)}_{\in\ \mathrm{span}(S)} \;+\; \underbrace{\beta^{\perp}}_{\textit{orthogonal to}\ \mathrm{span}(S)} \tag{46}$$

In particular, $\hat{\beta}(0^+)$ gives the minimal norm interpolator:

$$\hat{\beta}(0^+) \;=\; \arg\min\{\|\beta\|^2 :\ \beta \textit{ satisfies } (45)\} \tag{47}$$

What does the Ridgeless Regression Do? ii

Proof.

We have

$$\hat{\beta}(0^+) \;=\; (SS')^+ SR \tag{48}$$

Let

$$S'\beta \;=\; R \tag{49}$$

be some other interpolator. Then,

$$SS'\beta \;=\; SR \;\Rightarrow\; (SS')^+ SS'\beta \;=\; (SS')^+ SR \;=\; \hat{\beta}(0^+)\,. \tag{50}$$

Let $\Pi = (SS')^+ SS'$. Then, Π is the orthogonal projection onto the $\mathrm{span}(S)$. Thus,

$$\Pi\beta \;=\; \hat{\beta}(0^+) \tag{51}$$

and the claim follows. ∎

What does the Ridgeless Regression Do? iii

Summarizing, we can formulate the complexity principle

Theorem (heuristic)

$\|\hat{\beta}(0+)\|$ *is monotone decreasing in* P*. Thus, if the small norm of* β *is a good inductive bias, ridgeless regression generalizes better when* P *gets larger.*

The Magic of Ridge: Playing with Dimention

Lemma (Swapping Dimension)

Let $S \in \mathbb{R}^{T \times P}$. Then,

$$(zI + S'S)^{-1}S' = S'(zI + SS')^{-1} \tag{52}$$

Proof.

We have

$$\begin{aligned} S'(zI + SS')^{-1} &= (zI + S'S)^{-1}S' \Leftrightarrow \\ S'(zI + SS') &= (zI + S'S)S' \Leftrightarrow \\ S'z + SS'S &= zS' + SS'S \end{aligned} \tag{53}$$

■

The Big Question We Will Study: Why (and How) Do Big Models Generalize? i

- suppose

$$R_{t+1} = S_t'\beta + \varepsilon_{t+1} \tag{54}$$

- suppose we run a linear regression

$$\hat{\beta}(z) \;=\; \left(zI + T^{-1}\sum_t S_t S_t' \right)^{-1} T^{-1}\sum_t S_t R_{t+1}$$

and build the prediction

$$\hat{\beta}(z)' S_{T+1}\,. \tag{55}$$

- • How does the generalization error

$$MSE \;=\; E[(R_{T+1} - \hat{\beta}(z)' S_T)^2] \tag{56}$$

behave?

The Big Question We Will Study: Why (and How) Do Big Models Generalize? ii

- How does the timing strategy

$$R^{\pi}_{t+1} \;=\; R_{t+1}\,\pi_t,\;\; \pi_t \;=\; \hat{\beta}(z)'S_t \tag{57}$$

  behave for $t \geq T$.

- Answering these questions will require developing some mathematical tools

Experiments with High Dimensional Regressions

Please click on the link:

Understanding High-Dimensional Regressions

Table of Contents

Law of Large Numbers i

Theorem

If X_i are i.i.d. then

$$\frac{1}{n}\sum_i X_i \;\rightarrow\; E[X]$$

as $n \rightarrow \infty$

The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.

Proof of the Weak Law of Large Numbers.

Define

$$Y_i = X_i - E[X_i],\ E[Y_i] = E[X_i - E[X_i]] = E[X_i] - E[X_i] = 0 \tag{58}$$

Then,

$$\begin{aligned} \frac{1}{n}\sum_i X_i - E[X] &= \frac{\sum_i X_i - nE[X]}{n} = \frac{\sum_i (X_i - E[X])}{n} \\ &= \frac{\sum_i Y_i}{n} \end{aligned} \tag{59}$$

and

$$\begin{aligned} E[(\frac{\sum_i Y_i}{n})^2] &= n^{-2}E[(\sum_i Y_i)^2] = n^{-2}(E[\sum_i Y_i^2 + \sum_{i\neq j} Y_i Y_j]) \\ &= n^{-2}(\sum_i E[Y_i^2] + \sum_{i\neq j} \underbrace{E[Y_i]E[Y_j]}_{independence}) \\ &= n^{-2}\sum_i E[Y_i^2] = n^{-1}E[Y_1^2] \to 0 \end{aligned} \tag{60}$$

■

Central Limit Theorem: Can we have large deviations from LLN?

Theorem

$$\text{Prob}\left(\sqrt{n}\left(\frac{1}{n}\sum_i X_i \; - \; E[X]\right) \; < \; x\right) \; \to \; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/(2\sigma^2)}dt$$

$e^{-t^2/(2\sigma^2)}$ **is the Gaussian density**. It is **is very small** for large t.

Table of Contents

Trace i

- $\mathrm{tr}(A) = \sum_i A_{i,i}$
- $\mathrm{tr}(A) = \sum_i \lambda_i(A)$
- $\mathrm{tr}(A) = \mathrm{tr}(A')$

- $$\mathrm{tr}(AB) = \mathrm{tr}(BA) \text{ for all } A,\ B$$

- $$\mathrm{tr}(A^k) = \sum_i \lambda_i(A)^k$$

- $$P^{-1}\,\mathrm{tr}((A - zI)^{-1}) = P^{-1}\sum_i (\lambda_i(A) - z)^{-1}$$

Trace ii

- For any matrix A, the matrix AA' is symmetric and positive semi-definite (p.s.d.):

$$x'AA'x \;=\; \|A'x\|^2 \;\geq\; 0 \tag{61}$$

  By the spectral theorem, $AA' = UDU'$ where $D = \text{diag}(\lambda_1, \cdots, \lambda_P)$ are the eigenvalues. Then,

$$|A| \;=\; (AA')^{1/2} \;=\; UD^{1/2}U' \;=\; U\,\text{diag}(\lambda_1^{1/2}, \cdots, \lambda_P^{1/2})U'$$

  is well-defined.

- **von Schatten norms:**

$$\|A\|_p \;=\; (\text{tr}((AA')^{p/2}))^{1/p} \;=\; (\text{tr}((A'A)^{p/2}))^{1/p} \;=\; (\text{tr}(|A|^p))^{1/p}$$

  For $p = 1$, this is the trace norm:

$$\|A\|_1 \;=\; \text{tr}(|A|)$$

Trace iii

and for $p = 2$ this is the Frobenius norm:

$$\|A\|_2 \;=\; (\mathrm{tr}(|A|^2))^{1/2} \;=\; \left(\sum_{i,j} A_{i,j}^2\right)^{1/2}.$$

- $\|A\|_p^p$ captures different things:

$$P^{-1}\|A\|_p^p \;=\; P^{-1}\sum_i (\lambda_i(|A|))^p \;=\; E[(\lambda(|A|))^p] \tag{62}$$

- Hence, by Hölder inequality,

$$(P^{-1}\|A\|_p^p)^{1/p}$$

is increasing in p, while $\|A\|_p^p$ is decreasing in p. E.g., $A = I_{P\times P}/P$ has

$$\|A\|_\infty \;=\; 1/P \;\to\; 0$$

Trace iv

and

$$\|A\|_1 \;=\; \mathrm{tr}(I/P) \;=\; 1$$

and

$$\|A\|_2 \;=\; (\mathrm{tr}(A^2))^{1/2} = \;(\mathrm{tr}((I/P)^2))^{1/2} \;=\; (P/P^2)^{1/2} \;=\; 1/P^{1/2} \;\to\; 0$$

when $P \to \infty$.

The Big Random Matrix Trick: Concentration of Quadratic Forms i

Lemma

Suppose that $X = (X_i)_{i=1}^P$ *with* X_i *independent of* X_j, $E[X_i] = 0$, $E[X_i X_j] = \delta_{i,j}, E[X_i^4] \leq k$, *and* A_P *are random matrices independent of* X *and such that* $\|A_P\|_2 = o(1)$. *Let also*

$$Y_t \;=\; X_t' A_P X_t \,. \tag{63}$$

Then,

$$\begin{aligned} &(1)\ Y_t = \mathrm{tr}(A_P X_t X_t') \\ &(2)\ \lim_{P\to\infty} E[(Y_t - \mathrm{tr}(A_P))^2 | A_P] = 0 \end{aligned} \tag{64}$$

In particular, If $A_P = B_P/P$ *where* $\|B_P\| \leq K$, *we have* $\|A_P\|_2^2 \leq P\|B_P\|^2/P^2 \leq K$, *and hence*

$$\lim_{P\to\infty} E[(X_t' B_P X_t - \mathrm{tr}(B_P))^2 | B_P]/P^2 \;=\; 0\,. \tag{65}$$

Heuristics: $E[X_t X_t'] = I$ and BIG DATA (high P). Then, the law of large numbers implies

$$X_t X_t' \approx I$$

and hence

$$X_t' A_P X_t = \mathrm{tr}(X_t' A_P X_t) = \mathrm{tr}(A_P X_t X_t') \approx \mathrm{tr}(A_P I) = \mathrm{tr}(A_P)$$

Rigorous Proof of Lemma 9.

(1):

$$\begin{aligned} &X_t' A X_t \in R \Rightarrow X_t' A X_t = \mathrm{tr}(X_t' A X_t) \\ &\mathrm{tr}(AB) = \mathrm{tr}(BA) \Rightarrow \mathrm{tr}(X_t' A X_t) = \mathrm{tr}(A X_t X_t') \end{aligned} \tag{66}$$

(2): Define $Y_t = X_t' A_P X_t$. We have

$$E[Y_t] = E[\mathrm{tr}(A_P(X_t X_t'))|A_P] = \mathrm{tr}(A_P E[X_t X_t']) = \mathrm{tr}(A_P),$$

and hence

$$E[(Y_t - \mathrm{tr}(A_P))^2 | A_P] \;=\; \mathrm{Var}[Y_t | A_P] \;=\; E[Y_t^2 | A_P] - E[Y_t | A_P]^2 \tag{67}$$

and hence it suffices to prove that

$$E[Y_t^2 | A_P] \;-(\mathrm{tr}(A_P))^2 \;\to\; 0 \tag{68}$$

For simplicity, we assume from now on that A_P is deterministic, and write $A_P = (A_{i,j})_{i,j=1}^P$. Then,

$$Y_t \;=\; \sum_{i,j} X_i X_j A_{i,j} \tag{69}$$

and therefore

$$Y_t^2 \;=\; \sum_{i_1,j_1,i_2,j_2} X_{i_1} X_{j_1} A_{i_1,j_1} A_{i_2,j_2} X_{i_2} X_{j_2} \tag{70}$$

Now we compute the expectation:

$$
\begin{aligned}
E[Y_t^2] &= \sum_{i_1,j_1,i_2,j_2} A_{i_1,j_1} A_{i_2,j_2} E[X_{i_1} X_{j_1} X_{i_2} X_{j_2}] \\
&= (\sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j} (A_{i,j}^2 + 2A_{i,i}A_{j,j}) E[X_i^2 X_j^2] \\
&\underbrace{\leq}_{E[X_i^4] \leq k} (\sum_i k A_{i,i}^2 + \sum_{i,j} (A_{i,j}^2 + 2A_{i,i}A_{j,j})) \\
&= ((k-1) \sum_i A_{i,i}^2 + \sum_{i,j} A_{i,j}^2 + \mathrm{tr}(A)^2)
\end{aligned}
\tag{71}
$$

We have

$$
\sum_i A_{i,i}^2 \leq \sum_{i,j} A_{i,j}^2 = \|A\|_2^2 , \tag{72}
$$

and therefore

$$
|E[Y_t^2] - \mathrm{tr}(A)^2| \leq k \|A_P\|_2^2 , \tag{73}
$$

and the proof is complete.

Lemma

Suppose that $S_t = \Psi^{1/2} X_t$, *where* Ψ *is a symmetric, P-dimensional positive semi-definite (p.s.d.) matrix. Then,*

$$\lim(P^{-1} S_t' A_P S_t - P^{-1} \operatorname{tr}(\Psi A_P)) \ \to \ 0 \tag{74}$$

in L_2 *and hence probability. More generally, if the Frobenius norm* $\|A_P\|_2$ *is uniformly bounded, then*

$$\lim(S_t' A_P S_t - \operatorname{tr}(\Psi A_P)) \ \to \ 0 \tag{75}$$

in L_2 *and hence in probability.*

Proof.

We have

$$E[X_t X_t'] \;=\; I \tag{76}$$

and hence

$$\begin{aligned} E[S_t S_t'] \;&=\; E[\Psi^{1/2} X_t X_t' \Psi^{1/2}] \\ &=\; \Psi^{1/2} E[X_t X_t'] \Psi^{1/2} \;=\; \Psi^{1/2} I \Psi^{1/2} \;=\; \Psi \end{aligned} \tag{77}$$

Let

$$\tilde{A} \;=\; \Psi^{1/2} A \Psi^{1/2}\,. \tag{78}$$

Then, we have

$$\begin{aligned} &Y_t \;=\; P^{-1} S_t' A S_t \;=\; P^{-1} (\Psi^{1/2} X_t)' A \Psi^{1/2} X_t \;=\; P^{-1} X_t' \tilde{A} X_t \\ &P^{-1} \operatorname{tr}(\Psi A) \;=\; P^{-1} \operatorname{tr}(\Psi^{1/2} \Psi^{1/2} A) \;=\; P^{-1} \operatorname{tr}(\Psi^{1/2} A \Psi^{1/2}) \;=\; P^{-1} \operatorname{tr}(\tilde{A}) \\ &E[Y_t] \;=\; P^{-1} E[X_t' \tilde{A} X_t] \;=\; P^{-1} \operatorname{tr}(\tilde{A}) \end{aligned} \tag{79}$$

and the claim follows from Lemma 9. ∎

The Magic of Big Data

- I observe $S_t \in \mathbb{R}^P$
- I want to learn $\Psi \in \mathbb{R}^{P \times P}$
- Obviously I cannot because $P << 0.5P(P+1)$
- But actually, I can (can I?) because I can evaluate

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

on many matrices A_P.

Assumptions

Assumptions:

- **Linear predictive structure**

$$R_{t+1} = S_t'\beta + \varepsilon_{t+1} \tag{80}$$

  with ε_{t+1} i.i.d., $E[\varepsilon_{t+1}] = 0$ and S_t a P-vector of predictor variables.
- **Noise size** $E[\varepsilon_{t+1}^2] = \sigma^2$. We normalize $\sigma^2 = 1$.
- **Signal covariance structure**
- **Structure of the** β **vector**. We assume $\beta = \beta_P$ is random, $\beta = (\beta_i)_{i=1}^P \in \mathbb{R}^P$, independent of S and R, and satisfies $E[\beta] = 0$, and $E[\beta\beta'] = P^{-1}b_* I$ and $E[\beta_i^4] < KP^{-2}$ for some K.
- That is, all features are equally informative

Lemma

$$\lim_{P\to\infty} \beta' X \beta \;=\; b_* \lim_{P\to\infty} \mathrm{tr}(X)/P$$

for any bounded matrix X.

Proof Formally, the claim follows directly from Lemma 9. But here is a non-rigorous heuristic proof. We have $E[\beta\beta'] = b_* I/P$ and hence, by the law of large numbers, $\beta\beta' \sim b_* I/P$ and, hence,

$$\beta' X \beta \;=\; \mathrm{tr}(\beta' X \beta) \;=\; \mathrm{tr}(X\beta\beta') \;\approx\; \mathrm{tr}(X b_* I/P) \;=\; b_* \,\mathrm{tr}(X)/P\,. \tag{81}$$

In particular,

$$\beta' \Psi \beta \;\to\; b_* \psi_{*,1},\; \psi_{*,1} \;=\; \lim_{P\to\infty} \mathrm{tr}(\Psi)/P\,. \tag{82}$$

Infeasible MSE and R2 i

If we know the true β, we get $\pi_t = S_t'\beta,\ R_{t+1} - S_t'\beta = \varepsilon_{t+1}$ and hence

$$\mathit{InfeasibleMSE} \ = \ E[\varepsilon_{t+1}^2] \ = \ 1$$

and

$$\begin{aligned}
E[R_{t+1}^2] \ &= \ E[(S_t'\beta + \varepsilon_{t+1})^2] \ = \ E[(S_t'\beta)^2 + 2(S_t'\beta)\varepsilon_{t+1} + \varepsilon_{t+1}^2] \\
&= \ E[(S_t'\beta)^2] \ + \ 2\underbrace{E[(S_t'\beta)]E[\varepsilon_{t+1}]}_{\textit{by indepdendence}} \ + \ E[\varepsilon_{t+1}^2] \\
&= \ E[\beta' S_t S_t'\beta] \ + \ \underbrace{0}_{E[\varepsilon_{t+1}]=0} \ + 1 \\
&= \ \beta' E[S_t S_t']\beta \ + \ 1 \\
&= \ \beta' \Psi \beta \ + \ 1 \\
&\to \ b_* \psi_{*,1} \ + \ 1 \, .
\end{aligned} \tag{83}$$

Infeasible MSE and R2 ii

Thus,

$$InfeasibleR^2 = \frac{\beta'\Psi\beta}{\beta'\Psi\beta + 1} \to \frac{b_*\psi_{*,1}}{b_*\psi_{*,1} + 1}.$$

Table of Contents

Computing Expected Return i

Let

$$\hat{\pi}_t \;=\; S_t'\hat{\beta}(z)\,.$$

Then,

$$MSE(\hat{\beta}) = E\left[\left(R_{t+1} - S_t'\hat{\beta}\right)^2 |\hat{\beta}\right] = E[R_{t+1}^2] - 2\underbrace{E[\hat{\pi}_t R_{t+1}|\hat{\beta}]}_{\substack{\text{Timing}\\ \text{Expected Return}}} + \underbrace{E[\hat{\pi}_t^2|\hat{\beta}]}_{\substack{\text{Timing}\\ \text{Leverage}}}. \tag{84}$$

and

$$R^2 \;=\; 1 - \frac{MSE}{E[R^2]}$$

The second moment of the managed return is

$$E[(\hat{\pi}_t R_{t+1})^2] \;=\; E[(\hat{\pi}_t)^2 R_{t+1}^2] \;=\; E[(S_t'\hat{\beta}(z))^2\,(S_t'\beta + \varepsilon_{t+1})^2] \tag{85}$$

Computing Expected Return ii

Recall

$$\begin{aligned}\hat{\beta}(z) &= (zI + \hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_t R_{t+1} \\ &= (zI + \hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_t(S_t'\beta + \varepsilon_{t+1}) \\ &= (zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta + q_T)\end{aligned} \tag{86}$$

where

$$q_T = \frac{1}{T}\sum_t S_t\varepsilon_{t+1} \tag{87}$$

Hence, for $t > T$,

$$
\begin{aligned}
& E[R_{t+1}\hat{\pi}_t|\mathcal{F}_T] \\
&= E[(\beta' S_t + \underbrace{\varepsilon_{t+1}}_{E[\varepsilon_{t+1}]=0})S_t'(zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \ + \ q_T)|\mathcal{F}_T] \\
&= \beta' E[\underbrace{S_t S_t'}_{E[S_t S_t']=\Psi}(zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \ + \ q_T)|\mathcal{F}_T] \\
&= \beta' E[\Psi(zI + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta \ + \ q_T)|\mathcal{F}_T] \\
&= \underbrace{\beta'\Psi(zI + \hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta}_{making\ \ money} \ + \ \underbrace{\beta'\Psi(zI + \hat{\Psi}_T)^{-1}q_T}_{noise}
\end{aligned}
\tag{88}
$$

Lemma

The first and second moments of q_T conditional on all Signals S (both in-sample and out-of-sample) are equal to

$$E[q_T|S] = 0 \tag{89}$$

and

$$E[q_T q_T'|S] = \frac{\sigma^2}{T}\hat{\Psi}_T \tag{90}$$

respectively.

Proof of Lemma 12 For the first moment, we have

$$E[q_T|S] \;=\; \frac{1}{T}\sum_{t=1}^{T} E[S_t\varepsilon_{t+1}|S] \;=\; \frac{1}{T}\sum_{t=1}^{T} S_t E[\varepsilon_{t+1}|S] \;=\; 0\,. \tag{91}$$

Understanding Noise ii

For the second moment, we have

$$
\begin{aligned}
E[q_T q_T'|S] &= \frac{1}{T^2} E[\sum_{t=1}^{T} S_t \varepsilon_{t+1} \sum_{t_1=1}^{T} \varepsilon_{t_1+1} S_{t_1}'|S] \\
&= \frac{1}{T^2} E[\sum_{t,t_1=1}^{T} S_t \varepsilon_{t+1} \varepsilon_{t_1+1}' S_{t_1}'|S] \\
&= \frac{1}{T^2} E[\sum_{t=t_1} S_t S_{t_1}' \varepsilon_{t+1}^2|S] + T^{-2} \sum_{t \neq t_1} E[S_t S_{t_1}' \varepsilon_{t+1} \varepsilon_{t_1+1}'] \\
&= \frac{1}{T^2} \sum_t S_t S_t' E[\varepsilon_{t+1}^2|S] + T^{-2} \sum_{t \neq t_1} S_t S_{t_1}' \underbrace{E[\varepsilon_{t+1}] E[\varepsilon_{t_1+1}]}_{=0} \\
&= \frac{\sigma^2}{T^2} \sum_t S_t S_t' \\
&= \frac{\sigma^2}{T} \hat{\Psi}_T \,.
\end{aligned}
\tag{92}
$$

Understanding Noise iii

Now we can prove the following Lemma:

Lemma

$q_T \to 0$ weakly in probability in the sense that $a_T' q_T \to 0$ for any uniformly bounded sequence of vectors a_T.

Proof of Lemma 13 We must show that for every vector a with $\|a\| = 1$, $a' q_T \approx 0$. We have

$$\begin{aligned} E[(a' q_T)^2] &= E[a' q_T q_T' a] \\ &= a' E[q_T q_T'] a \\ &\underbrace{=}_{\text{iterated expectations}} a' E[E[q_T q_T' | S]] a \\ &\underbrace{\approx}_{\text{Lemma 12}} \frac{\sigma^2}{T} a' E[\hat{\Psi}_T] a = \frac{\sigma^2}{T} a' \Psi a \leq \frac{\sigma^2}{T} \|\Psi\| \to 0 \end{aligned} \tag{93}$$

Understanding Noise iv

It is important to note that

$$E[\|q_T\|^2] = E[q_T' q_T] = E[\mathrm{tr}(q_T q_T')] = \mathrm{tr}\, E[q_T q_T'] = \frac{\sigma^2}{T} E[\mathrm{tr}(\hat{\Psi})] = \frac{\sigma^2}{T} \mathrm{tr}(\Psi) \to \sigma^2 c \psi_{*,1} \neq 0. \tag{94}$$

Thus, q_T does not converge to zero strongly, but only weakly.

Final Expression For the Expected Return i

$$
\begin{aligned}
E[R_{t+1}\hat{\pi}_t|\mathcal{F}_T] \;&\approx\; \beta'\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T\beta \\
&\underbrace{\approx}_{\beta' A\beta \approx P^{-1}b_*\,\mathrm{tr}(A)\;\forall\; A} \quad b_* P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T) \\
&= b_* P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T+zI-zI)) \qquad (95) \\
&= b_* P^{-1}\,\mathrm{tr}(\Psi((zI+\hat{\Psi}_T)^{-1}(\hat{\Psi}_T+zI)-z(zI+\hat{\Psi}_T)^{-1})) \\
&= \underbrace{b_* P^{-1}\,\mathrm{tr}(\Psi)}_{\textit{infeasible return}} - z\,b_* P^{-1}\,\mathrm{tr}(\Psi(zI+\hat{\Psi}_T)^{-1})
\end{aligned}
$$

Final Expression For the Expected Return ii

We have

$$P^{-1}\,\mathrm{tr}(\hat{\Psi}) \;\approx\; P^{-1}T^{-1}\,\mathrm{tr}\Big(\sum_t S_t S_t'\Big) \;=\; P^{-1}\sum_t \mathrm{tr}(S_t S_t') \;=\; P^{-1}T^{-1}\sum_t \mathrm{tr}(\Psi^{1/2}X_t X_t'\Psi^{1/2})$$
$$=\; P^{-1}T^{-1}\sum_t \mathrm{tr}(X_t X_t'\Psi^{1/2}\Psi^{1/2}) \;=\; P^{-1}T^{-1}\sum_t \mathrm{tr}(X_t X_t'\Psi) \;=\; T^{-1}\sum_t (P^{-1}\,\mathrm{tr}(X_t X_t'\Psi))$$
$$\underbrace{\approx}_{LLN} E[P^{-1}\,\mathrm{tr}(X_t X_t'\Psi)] \;=\; P^{-1}\,\mathrm{tr}(\underbrace{E[X_t X_t']}_{=I}\Psi) \;=\; P^{-1}\,\mathrm{tr}(\Psi) \tag{96}$$

and hence we may assume that $\mathrm{tr}(\Psi)$ is observable. Thus, we have

$$E[R_{t+1}\hat{\pi}_t|\mathcal{F}_T] \;\approx\; b_*\,P^{-1}\,\mathrm{tr}(\hat{\Psi}) \;-\; z\,b_*\,c^{-1}\xi(z)\,, \tag{97}$$

where

$$\underbrace{\xi(z)}_{\text{the most important function}} \;=\; T^{-1}\,\mathrm{tr}\Big(\underbrace{\Psi}_{\text{unobservable}}\,(zI+\underbrace{\hat{\Psi}_T}_{\text{observable}})^{-1}\Big) \tag{98}$$

Final Expression For the Expected Return iii

Intuitively, we expect that

$$\Psi \approx \hat{\Psi}_T \; ?? \tag{99}$$

It turns out this intuition is entirely wrong.

$$\Psi \not\approx \hat{\Psi}_T \, . \tag{100}$$

It order to compute $\xi(z)$, we need random matrix theory. It will allow us to compute traces of quantities that involve $(zI + \hat{\Psi}_T)^{-1}$.

Table of Contents

Eigenvalue Distributions and the Stieltjes Transform i

We will use $\lambda_k(\Psi),\ k = 1, \cdots, P$, to denote the eigenvalues of an arbitrary matrix Ψ. In the limit as $P \to \infty$, the spectral distribution F^{Ψ} of the eigenvalues of Ψ,

$$F^{\Psi}(x) = \frac{1}{P} \sum_{k=1}^{P} \mathbf{1}_{\lambda_k(\Psi) \leq x} \tag{101}$$

converges to a non-random probability distribution H supported on $(0, +\infty)$.[1] Furthermore, Ψ is uniformly bounded as $P \to \infty$. We will use

$$\psi_{*,k} = \lim_{P \to \infty} P^{-1} \operatorname{tr}(\Psi^k)\ ,\ k \geq 1$$

to denote asymptotic moments of the eigenvalues of Ψ.

For any symmetric matrix Ψ, a convenient matrix identity states

$$\frac{1}{P} \operatorname{tr}\left((\Psi - zI)^{-1}\right) = \frac{1}{P} \sum_{i=1}^{P} (\lambda_i(\Psi) - z)^{-1} \ = \ \int \frac{1}{x - z} dF^{\Psi}(x)\ ,\ z < 0\,.$$

where $\lambda_i(\Psi)$ are the eigenvalues of Ψ. From this identity, we immediately see the fundamental connection between ridge regularization and the distribution of eigenvalues for Ψ. The right-side quantity is the *Stieltjes transform* of the eigenvalue distribution of Ψ, denoted F^{Ψ}. And

$$\begin{aligned} & \lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\Psi - zI)^{-1}\right) \\ = \; & \lim \int \frac{1}{x-z} \underbrace{dF^{\Psi}(x)}_{\to dH(x)} \\ = \; & \int \frac{1}{x-z} dH(x) \\ \equiv \; & m_{\Psi}(z) \end{aligned} \tag{102}$$

[1] If 0 is in support of H, then Ψ is strictly degenerate, meaning that some signals are redundant.

Empirical Eigenvalues and Stieltjes Transforms i

The empirical counterpart to the unobservable $m_\Psi(z)$ is

$$\hat{m}_T(z) \;=\; \frac{1}{P}\,\mathrm{tr}\left((\hat{\Psi}_T - zI)^{-1}\right),$$

where

$$\hat{\Psi}_T \;=\; T^{-1}\sum_t S_t S_t' \;\in\; \mathbb{R}^{P\times P} \tag{103}$$

Observations:

- Ψ has P eigenvalues $\lambda_1(\Psi) > \cdots > \lambda_P(\Psi)$ and eigenvectors $\hat{U}_T$: $\Psi = U\,\mathrm{diag}(\lambda(\Psi))U'$
- $\hat{\Psi}_T$ has P eigenvalues $\lambda_1(\hat{\Psi}_T) > \cdots > \lambda_P(\hat{\Psi}_T)$ and eigenvectors $\hat{U}_T$: $\hat{\Psi}_T = \hat{U}_T\,\mathrm{diag}(\hat{\lambda}(\Psi))\hat{U}_T'$

Empirical Eigenvalues and Stieltjes Transforms ii

- if $P > T$, then $P - T$ of these eigenvalues are zero:

$$\hat{\Psi}_T \;=\; T^{-1} S' S \tag{104}$$

  has the same non-zero eigenvalues as $T^{-1} S S'$
- $\hat{U}_T$ (empirical eigenvectors) does not generally converge
- $\hat{\lambda}$ does converge in some sense. Namely, the "histogram" converges (although nearby eigenvalues can sort of interchange with each other a bit)
- We can capture eigenvalues through their distribution,

$$F_{\hat{\Psi}_T}(x) \;=\; \frac{1}{P} \sum_{i=1}^{P} \mathbf{1}_{\hat{\lambda}_i < x} \tag{105}$$

Empirical Eigenvalues and Stieltjes Transforms iii

- Or we can do it through the Stieltjes transform

$$\hat{m}_T(z) \;=\; \frac{1}{P}\,\mathrm{tr}\left((\hat{\Psi}_T - zI)^{-1}\right) \;=\; \frac{1}{P}\sum_{i=1}^{P}\frac{1}{\hat{\lambda}_i - z} \;=\; \int \frac{1}{x-z}\,dF_{\hat{\Psi}_T}(x)$$

- How can we recover all eigenvalues from $\hat{m}_T(z)$? Well, these are the *singularities (poles)*: $\hat{m}_T(\hat{\lambda}_i) = \infty$
- What about moments? We have (assuming big z) that

$$\begin{aligned}(\hat{\Psi}_T - zI)^{-1} \;&=\; (z(z^{-1}\hat{\Psi}_T - I))^{-1} \\ &=\; z^{-1}(z^{-1}\hat{\Psi}_T - I)^{-1} \;=\; -z^{-1}\sum_{k=0}^{\infty}(z^{-1}\hat{\Psi}_T)^k\end{aligned} \tag{106}$$

Empirical Eigenvalues and Stieltjes Transforms iv

Thus,

$$\hat{m}_T(z) = \frac{1}{P}\operatorname{tr}\left((\hat{\Psi}_T - zI)^{-1}\right) = -z^{-1}\sum_{k=0}^{\infty} z^{-k} P^{-1}\operatorname{tr}((\hat{\Psi}_T)^k)$$

$$= -z^{-1}\sum_{k=0}^{\infty} z^{-k} P^{-1}\operatorname{tr}((\hat{\Psi}_T)^k) \tag{107}$$

$$= -z^{-1}\sum_{k=0}^{\infty} z^{-k}\hat{\psi}_{*,k}$$

where

$$\hat{\psi}_{*,k} = P^{-1}\operatorname{tr}((\hat{\Psi}_T)^k) \tag{108}$$

are the empirical eigenvalue moments

- We have

$$
\begin{aligned}
E[\hat{\psi}_{*,1}] &= P^{-1} E[\mathrm{tr}(\hat{\Psi}_T)] \\
&= P^{-1} \,\mathrm{tr}\, E[T^{-1} \sum_t S_t S_t'] \\
&= P^{-1} T^{-1} \sum_t \mathrm{tr}\, E[S_t S_t'] \\
&\underbrace{=}_{\text{summetry across } t} P^{-1} \,\mathrm{tr}\, \underbrace{E[S_t S_t']}_{=\Psi} \\
&= P^{-1} \,\mathrm{tr}\, \Psi \;\to\; \psi_{*,1}
\end{aligned}
\tag{109}
$$

Empirical Eigenvalues and Stieltjes Transforms vi

- But already in the second moment, things get tricky:

$$
\begin{aligned}
\hat{\psi}_{*,2} &= P^{-1}\,\mathrm{tr}(\hat{\Psi}_T^2) \\
&= P^{-1}\,\mathrm{tr}\,E[(T^{-1}\sum_t S_t S_t')^2] \\
&= P^{-1}T^{-2}\sum_{t_1,t_2}\mathrm{tr}\,E[S_{t_1}S_{t_1}'S_{t_2}S_{t_2}'] \\
&= P^{-1}T^{-2}\sum_{t_1=t_2}\mathrm{tr}\,E[S_{t_1}S_{t_1}'S_{t_2}S_{t_2}'] \;+\; P^{-1}T^{-2}\sum_{t_1\neq t_2}\mathrm{tr}\,E[S_{t_1}S_{t_1}'S_{t_2}S_{t_2}'] \\
&= P^{-1}T^{-2}\sum_t \mathrm{tr}\,E[(S_tS_t')^2] \;+\; P^{-1}T^{-2}\sum_{t_1\neq t_2}\mathrm{tr}(\underbrace{E[S_{t_1}S_{t_1}']\,E[S_{t_2}S_{t_2}'] \;=\; \Psi^2}_{independence}) \\
&\underbrace{=}_{summetry\ across\ t_1,t_2} P^{-1}T^{-2}T\,\mathrm{tr}\,E[(S_tS_t')^2] \;+\; P^{-1}\frac{T(T-1)}{T^2}\mathrm{tr}\,E[\Psi^2] \\
&\approx \frac{1}{PT}\mathrm{tr}\,E[(S_tS_t')^2] + \psi_{*,2}
\end{aligned}
\tag{110}
$$

where

$$
\begin{aligned}
& \operatorname{tr} E[(S_t S_t')^2] \\
= \; & \operatorname{tr} E[S_t S_t' S_t S_t'] \\
= \; & \operatorname{tr} E[\|S_t\|^2 S_t S_t'] \\
= \; & \operatorname{tr} E[\|S_t\|^2 S_t' S_t] \\
= \; & E[\|S_t\|^4] \\
= \; & E[(\sum_{i=1}^{P} S_i^2)^2] \; = \; \sum_{i=1}^{P} E[S_i^4] \; + \; \sum_{i,j} E[S_i^2 S_j^2]
\end{aligned}
\tag{111}
$$

This grows at the order of P^2 and does not converge to zero!

Discussion i

- In traditional finite P statistics, we would have a convergence between the sample covariance $\hat{\Psi}$ and the true covariance Ψ as $T \to \infty$. One might be tempted to think that $\lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\hat{\Psi} - zI)^{-1}\right)$ and $\lim_{P\to\infty} \frac{1}{P} \operatorname{tr}\left((\Psi - zI)^{-1}\right)$ also converge as $T \to \infty$.
- **But this is not the case!**
- The limiting eigenvalue distributions of $\hat{\Psi}$ and Ψ remain divergent in the limit as $T \to \infty$ if $P/T \to c > 0$. Here we see a first glimpse of the complexity of machine learning and how we can understand it with random matrix theory.
- When signals are i.i.d. with $\Psi = I$ and $m_\Psi(z) = (1-z)^{-1}$, Marcenko and Pastur (1967) showed that

$$m(-z; c) = \frac{-((1-c)+z) + \sqrt{((1-c)+z)^2 + 4cz}}{2cz}. \tag{112}$$

Discussion ii

- While the eigenvalue distributions of the sample and true covariance matrices do not coincide, there is a precise non-linear way they relate to each other. In particular, when $P > T$, the matrix $\hat{\Psi}$ has $P - T$ zero eigenvalues and therefore, $P^{-1} \operatorname{tr}\left((zI + \hat{\Psi})^{-1}\right)$ contains a singular part, $P^{-1}(P - T)z^{-1} = (1 - c^{-1})z^{-1}$.
- Quite remarkably, if we constrain ourselves to linear ridge regression estimators, all asymptotic expressions depend only on $m(z; c)$ and do not require m_Ψ. This is crucial because m_Ψ is not observable and also not recoverable in finite samples!

Marcenko-Pastur for the Spectral Density

Suppose $E[S] = 0,\ E[SS'] = \sigma^2 I$.

Theorem

$\hat{\Psi} = \frac{1}{T}\sum_t S_t S_t'$, *Suppose that* $T, P \to \infty$ *and* $P/T \to c \in (0, \infty)$. *Then, eigenvalues are distributed like*

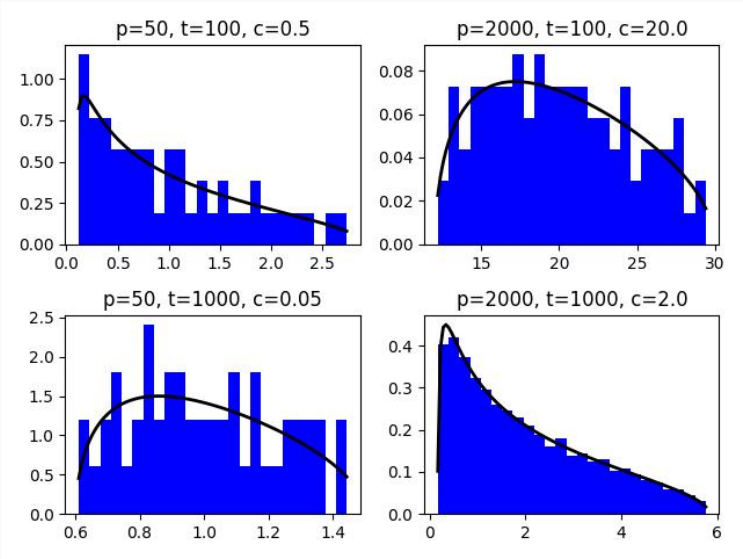
$$\mu(A) = \begin{cases} (1 - c^{-1})\mathbf{1}_{0 \in A} + \nu(A), & c > 1 \\ \nu(A),\ 0 \le c \le 1 \end{cases}$$

and

$$d\nu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{cx}\mathbf{1}_{x \in [\lambda_-, \lambda_+]}dx$$

$$\lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2$$

Note: If $P > T$ then the fraction $(P - T)/P = 1 - c^{-1}$ of eigenvalues are zero.


p=50, t=100, c=0.5
p=2000, t=100, c=20.0
p=50, t=1000, c=0.05
p=2000, t=1000, c=2.0

Marcenko-Pastur for the cumulative distribution

Same assumptions and definitions as before:

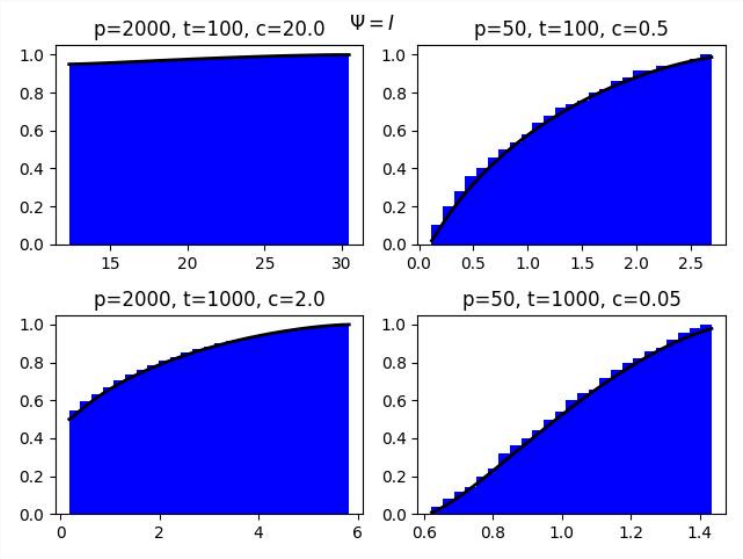
Theorem

$$F_\lambda(x) = \begin{cases} \frac{c-1}{c}\mathbf{1}_{x\in[0,\lambda_-)} + \left(\frac{c-1}{2c} + F(x)\right)\mathbf{1}_{x\in[\lambda_-,\lambda_+)} + \mathbf{1}_{x\in[\lambda_+,\infty)}, & \text{if } c > 1 \\ F(x)\mathbf{1}_{x\in[\lambda_-,\lambda_+)} + \mathbf{1}_{x\in[\lambda_+,\infty)}, & \text{if } 0, \end{cases}$$

$$F(x) = \frac{1}{2\pi c}\left(\pi c + \sigma^{-2}\sqrt{(\lambda_+ - x)(x-\lambda_-)}\right.$$

$$\left. - (1+c)\arctan\frac{r(x)^2-1}{2r(x)} + (1-c)\arctan\frac{c - r(x)^2 - \lambda_+}{2\sigma^2(1-c)r(x)}\right)$$

$$r(x) = \sqrt{\frac{\lambda_+ - x}{x-\lambda_-}}$$


$\Psi = I$
p=2000, t=100, c=20.0
p=50, t=100, c=0.5
p=2000, t=1000, c=2.0
p=50, t=1000, c=0.05

Marcenko-Pastur for the Stieltjes Transform

If $E[S] = 0,\ E[SS'] = I$, then

$$
\begin{aligned}
& P^{-1}\,\mathrm{tr}((zI + \hat{\Psi})^{-1}) \ \to \\
& (1 - c^{-1})z^{-1}\mathbf{1}_{c>1} \ + \ \int_{\lambda_-}^{\lambda_+} \frac{1}{2\pi(z+x)} \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{cx} \mathbf{1}_{x \in [\lambda_-, \lambda_+]} dx \\
& = \ \frac{-((1-c)+z) + \sqrt{((1-c)+z)^2 + 4cz}}{2cz} \\
& = \ \frac{2}{(1-c)+z+\sqrt{((1-c)+z)^2 + 4cz}}
\end{aligned}
\tag{113}
$$

Please click on the link:

Understanding High-Dimensional Covariance Matrices and the Marcenko-Pastur Theorem

Table of Contents

- Recall that $c = P/T > 0$
- **Naive approximation:**

$$\begin{aligned}\xi(z) &= \frac{1}{T}\operatorname{tr}((zI+\hat{\Psi})^{-1}\Psi) = c\,\frac{1}{P}\operatorname{tr}((zI+\hat{\Psi})^{-1}\Psi)\\ &\approx P^{-1}\operatorname{tr}((zI+\hat{\Psi})^{-1}\hat{\Psi}) = P^{-1}\operatorname{tr}((zI+\hat{\Psi})^{-1}(\hat{\Psi}+zI-zI))\\ &= P^{-1}\operatorname{tr}(I-z(zI+\hat{\Psi})^{-1}) = 1 - zP^{-1}\operatorname{tr}((zI+\hat{\Psi})^{-1}) = 1 - z\,m(-z;c)\,.\end{aligned} \tag{114}$$

- Strikingly, we will show that this naive approximation is wrong, and the correct formula is

$$\xi(z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}. \tag{115}$$

Technical Results ii

Proposition

Suppose that

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

exists. Then, if Ψ is uniformly bounded, we have

$$\lim_{T\to\infty} \frac{1}{T}\operatorname{tr}((zI+\hat{\Psi})^{-1}\Psi) \;\to\; \xi(z;c) \tag{116}$$

almost surely, where

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

To prove this formula, we will use **Leave-One-Out Decomposition:**

$$\hat{\Psi}_{t,T} \;=\; \sum_{\tau\neq t} S_\tau S_\tau' \tag{117}$$

Technical Results iii

gives

$$\hat{\Psi}_T \;=\; \hat{\Psi}_{t,T} \;+\; \frac{1}{T} S_t S_t' \tag{118}$$

and hence, by the Sherman-Morrison formula,

$$(zI + \hat{\Psi})^{-1} S_t \;=\; \frac{(zI + \hat{\Psi}_{T,t})^{-1} S_t}{1 \;+\; T^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} \;\approx\; \frac{(zI + \hat{\Psi}_{T,t})^{-1} S_t}{1 \;+\; \xi(z)} \tag{119}$$

Proposition

For Ψ uniformly bounded, we have

$$\lim_{T\to\infty} \frac{1}{T} S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \;=\; \lim_{T\to\infty} \frac{1}{T} \operatorname{tr}((zI + \hat{\Psi})^{-1}\Psi) \;=\; \xi(z;c) \tag{120}$$

in probability, where

$$\xi(z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}. \tag{121}$$

The proof is based on several steps, the first one being:

Lemma

[The Law of Large Numbers of Inverse Covariance] Suppose that

$$\hat{\Psi}_T \;=\; \frac{1}{T}\sum_{t=1}^{T} S_t S_t', \tag{122}$$

where $S_t \in \mathbb{R}^P$ are independent random vectors. Define

$$q_T(z) \;=\; P^{-1}\,\mathrm{tr}(Q_P(zI + \hat{\Psi}_T)^{-1}) \tag{123}$$

for some sequence of uniformly bounded matrices Q_P. Then,

$$q_T \;-\; E[q_T] \;\to\; 0$$

almost surely.

Technical Results vi

- By Lemma 10,

$$P^{-1}\, S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \; - P^{-1} \operatorname{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1}) \; \to \; 0 \tag{124}$$

in probability. At the same time, by Lemma 20,

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

almost surely. Thus,

$$P^{-1}\, S_t'(zI + \hat{\Psi}_{T,t})^{-1} S_t \; - E[P^{-1} \operatorname{tr}(\Psi(zI + \hat{\Psi}_{T,t})^{-1})] \; \to \; 0 \tag{125}$$

in probability.

-

$$P^{-1} \operatorname{tr} E[(zI + \hat{\Psi}_T)^{-1}] \; \to \; m(-z; c) \tag{126}$$

- Now, we have

$$
\begin{aligned}
1 &= P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)] = P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}]z \\
&+ \; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&= z\hat{m}(-z,c) \; + \; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_t S_t'] \\
&= \{\textit{symmetry across } t\} \; = \; z\hat{m}(-z,c) \; + \; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_T)^{-1}S_t S_t'] \\
&= \{\textit{using Sherman} - \textit{Morrison (143)}\} \\
&= z\hat{m}(-z,c) \; + \; P^{-1}\,\mathrm{tr}\,E[(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}S_t'] \\
&= z\hat{m}(-z,c) \; + \; E[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}]
\end{aligned}
\tag{127}
$$

Now, $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \leq \|\Psi\| z^{-1}$ and hence is uniformly bounded. Let us pick a sub-sequence of T converging to infinity and such that

$$E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \to q \tag{128}$$

for some $q > 0$. By (124),

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

in probability and this sequence is uniformly bounded. Hence,

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

and we get

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

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

The proof is complete.

Table of Contents

The Fixed Point Equation i

Theorem

Suppose that $\Psi = I$. *Then,*

$$m(-z;c) = \frac{\sqrt{(1-c-z)^2+4cz}-(1-c-z)}{2cz} \tag{129}$$

Proof The proof follows from Proposition 1. Indeed, placing $\Psi = I$ in the definition of $\xi(z;c)$ gives us:

$$\begin{aligned} E[T^{-1}\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1})] &= \left(\frac{T}{P}\right)^{-1} P^{-1}E[(\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1})] \\ &= cP^{-1}E[\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1})] \\ &= cm(-z;c) = \xi(z;c). \end{aligned} \tag{130}$$

The Fixed Point Equation ii

Plugging the previous expression into the closed form solution of $\xi(z;c)$ provides us with:

$$\xi(z;c) = cm(-z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)} \tag{131}$$

We can solve for $m(-z;c)$ as follows:

$$\begin{aligned} & m(-z;c) = \frac{1 - zm(-z;c)}{1 - c + czm(-z;c)} \\ & \Rightarrow (1 - c + z)m(-z;c) + czm(-z;c)^2 - 1 = 0 \\ & \Rightarrow m(-z;c) = \frac{\pm\sqrt{(1-c-z)^2 + 4cz} - (1 - c - z)}{2cz}. \end{aligned} \tag{132}$$

By definition, $m(-z;c)$ is positive. Therefore, the positive root is the true solution, concluding the proof.

The Marcenko-Pastur Master Equation i

Theorem

For any $c > 0$ and $z < 0$, the distribution of eigenvalues of $\hat{\Psi}$ in the limit as $P, T \to \infty$, $P/T \to c$ converges to a distribution whose Stieltjes transform, $m(z; c)$, is the unique positive solution to the equation

$$\begin{aligned} m(z;c) &= \frac{1}{1 \; - \; c \; - \; c\, z\, m(z;c)} \; m_{\Psi}\left(\frac{z}{1 \; - \; c \; - \; c\, z\, m(z;c)}\right) \\ &= \int \frac{1}{\lambda(1 \; - \; c \; - \; c\, z\, m(z;c)) - z} dF^{\Psi}(\lambda). \end{aligned} \tag{133}$$

Implicit Regularization

- Stieltjes transforms

$$\begin{aligned} m(-z) &= P^{-1}\,\mathrm{tr}((zI+\Psi)^{-1}) = P^{-1}\sum_i (z+\lambda_i(\Psi))^{-1} \\ m(-z;c) &\approx P^{-1}\,\mathrm{tr}\left(\left(zI+\hat{\Psi}_T\right)^{-1}\right) = P^{-1}\sum_i (z+\lambda_i(\hat{\Psi}_T))^{-1} \end{aligned} \tag{134}$$

-

$$\xi(z;c) = \frac{1-zm(-z;c)}{c^{-1}-1+zm(-z;c)}\ ,\ 1+\xi(z;c) = \frac{1}{1-c-c\,z\,m(z;c)} \tag{135}$$

- The implicit shrinkage function

$$Z_*(z;c) = z\,(1+\xi(z;c)) \tag{136}$$

Theorem When $P \to \infty,\ P/T \to c$:

$$m(-z;c) = \frac{Z_*(z;c)}{z}\,m(-Z_*(z;c)) \tag{137}$$

Properties of the Implicit Shrinkage Function

Theorem

$Z^*(z;c)$ is monotone increasing in z and c. In the ridgeless limit as $z \to 0$, we have

$$Z^*(z;c) \;\to\; \begin{cases} 0, & c < 1 \\ 1/\tilde{m}(c), & c > 1 \end{cases} \tag{138}$$

where $\tilde{m}(c) > 0$ is the unique positive solution to

$$c - 1 \;=\; \frac{\int \frac{dH(x)}{\tilde{m}(1+\tilde{m}\,x)}}{\int \frac{x dH(x)}{1+\tilde{m}\,x}}, \tag{139}$$

and H is the limiting eigenvalue distribution of Ψ.

Recovering the Eigenvalue Density from The Stieltjes Transform

All we then need is to use the Stieltjes Transform inversion formula: If

$$m(-z) \;=\; \int \frac{1}{x-z}\mu(dx)\,,$$

then, if a, b are continuity points of μ, then

$$\mu([a,b]) \;=\; \frac{1}{\pi}\lim_{y\downarrow 0}\int_a^b \Im(m(x+iy))dx$$

If $\mu(x)$ admits a density $f(x)$, then

$$f(x) \;=\; \frac{1}{\pi}\lim_{y\downarrow 0}\Im(m(x+iy))\,.$$

If $\mu(\{x\}) > 0$, then

$$\mu(\{x\}) \;=\; -\frac{1}{\pi}\lim_{y\downarrow 0}(iym(x+iy))$$

Some Links

Please click on the link to get some Useful Data. Please download them and then upload them to your google drive

and now we continue our experiments:

Understanding High-Dimensional Covariance Matrices and the Marcenko-Pastur Theorem

Summary

- **Theorem:** Big Data Behaves as if it chooses a bigger ridge penalty
- This implicit regularization is what makes it generalize better (=work better OOS)
- A lot of Information is lost: Parts of spectral information (low eigenvalues, eigenvectors) are entirely lost in Big Data and cannot be recovered. Implicit regularization takes care of it.
- Can can predict the gap (complexity wedge) using only the distribution of eigenvalues.

Table of Contents

Lemma

Suppose that

$$\hat{\Psi}_T \;=\; \frac{1}{T}\sum_{t=1}^{T} S_t S_t', \tag{140}$$

where $S_t \in \mathbb{R}^P$ *are independent random vectors. Define*

$$q_T(z) \;=\; P^{-1}\,\mathrm{tr}(Q_P(zI + \hat{\Psi}_T)^{-1}) \tag{141}$$

for some sequence of uniformly bounded matrices Q_P*. Then, there exists a constant* K_q *such that*

$$E[|q_T - E[q_T]|^q] \;\leq\; z^{-q} K_q \frac{T^{q/2}}{P^q} \tag{142}$$

The Magic of Big Data: Deterministic Stuff (aka concentration) ii

In particular, if $T^{1/2}/P \to 0$, then $q_T \to 0$ in L_2 and also in L_q for any $q > 2$. If $T^{1/2}/P = o(1/T^a)$ for some $a > 0$ then the convergence is almost sure.

Proof of Lemma 20 i

Let $\Psi_{T,t} = \frac{1}{T} \sum_{\tau \neq t} S_\tau S_\tau'$. By the Sherman-Morrison formula,

$$(zI + \hat{\Psi}_T)^{-1} = (zI + \hat{\Psi}_{T,t})^{-1} \; - \; \frac{\frac{1}{T}(zI + \hat{\Psi}_{T,t})^{-1} S_t S_t' (zI + \hat{\Psi}_{T,t})^{-1}}{1 + (T)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t} \tag{143}$$

Let E_τ denote the conditional expectation given $S_{\tau+1}, \cdots, S_T$. That is,

$$\begin{aligned} E_t[X] &= E[X | S_T, \cdots, S_{t+1}] \\ E_{t-1}[X] &= E[X | S_T, \cdots, S_{t+1}, S_t] \\ (E_{t-1} - E_t)[X] &= E_{t-1}[X] - E_t[X] \end{aligned} \tag{144}$$

Let also

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

Proof of Lemma 20 ii

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

$$\begin{aligned} E_{t-1}[\frac{1}{P}\operatorname{tr}(zI+\Psi_{T,t})^{-1}Q_P] &= E[\frac{1}{P}\operatorname{tr}(zI+\Psi_{T,t})^{-1}Q_P|S_T,\cdots,S_{t+1},S_t] \\ &= E[\frac{1}{P}\operatorname{tr}(zI+\Psi_{T,t})^{-1}Q_P|S_T,\cdots,S_{t+1}] \\ &= E_t[\frac{1}{P}\operatorname{tr}(zI+\Psi_{T,t})^{-1}Q_P] \end{aligned} \tag{145}$$

and therefore, we have

$$(E_{t-1}-E_t)[\frac{1}{P}\operatorname{tr}(zI+\Psi_{T,t})^{-1}Q_P]=0$$

Proof of Lemma 20 iii

and therefore

$$
\begin{aligned}
E[q_T(z)] - q_T(z) \;&=\; E_0[q_T(z)] - E_T[q_T(z)] = \sum_{t=1}^{T}(E_{t-1}[q_T(z)] - E_t[q_T(z)]) \\
&= \sum_{t=1}^{T}(E_{t-1} - E_t)[q_T(z)] \\
&= \sum_{t=1}^{T}(E_{t-1} - E_t)[q_T(z) - \frac{1}{P}\,\mathrm{tr}(zI + \Psi_{T,t})^{-1}Q_P] \qquad (146) \\
&= \frac{1}{P}\sum_{t=1}^{T}(E_{t-1} - E_t)[\mathrm{tr}(zI + \hat{\Psi}_T)^{-1}Q_P - \mathrm{tr}(zI + \hat{\Psi}_{T,t})^{-1}Q_P] \\
&= -\frac{1}{P}\sum_{\tau=1}^{T}(E_{t-1} - E_t)[\gamma_t]\,,
\end{aligned}
$$

where we have used (143) and defined

$$\gamma_t = \mathrm{tr}\left(\frac{1}{T}(zI + \hat{\Psi}_{T,t})^{-1} S_t (1 + \frac{1}{T} S_t' (zI + \hat{\Psi}_{T,t})^{-1} S_t)^{-1} S_t' (zI + \hat{\Psi}_{T,t})^{-1} Q_P\right) \tag{147}$$

We will need the following known properties of the trace:

Lemma

If A, B are symmetric positive semi-definite, then

$$\mathrm{tr}(AB) \leq \mathrm{tr}(A)\|B\| \tag{148}$$

and

$$\mathrm{tr}(A^{1/2} B A^{1/2}) \leq \mathrm{tr}(B)\,\|A\| \tag{149}$$

Proof of Lemma 20 v

Thus, with

$$A \;=\; (zI + \hat{\Psi}_{T,t})^{-1},\; A^{1/2} = (zI + \hat{\Psi}_{T,t})^{-1/2}, \tag{150}$$

we have

$$\|A\| \leq\; z^{-1},\; \|A^{1/2}\| \;\leq\; z^{-1/2} \tag{151}$$

and we get

Proof of Lemma 20 vi

$$
\begin{aligned}
|\gamma_t| &= (1+\frac{1}{T}S_t'AS_t)^{-1}\,\mathrm{tr}\left(\frac{1}{T}AS_tS_t'AQ_P\right)\\
&= (1+\frac{1}{T}S_t'AS_t)^{-1}\,\mathrm{tr}\left(\frac{1}{T}A^{\frac{1}{2}}S_tS_t'A^{\frac{1}{2}}A^{\frac{1}{2}}Q_PA^{\frac{1}{2}}\right)\\
&\leq \|A^{\frac{1}{2}}Q_PA^{\frac{1}{2}}\|(1+\frac{1}{T}S_t'AS_t)^{-1}\,\mathrm{tr}\left(\frac{1}{T}A^{\frac{1}{2}}S_tS_t'A^{\frac{1}{2}}\right)\\
&\leq \|A^{\frac{1}{2}}Q_PA^{\frac{1}{2}}\|(1+\frac{1}{T}S_t'AS_t)^{-1}\frac{1}{T}S_t'AS_t\\
&\leq \|A^{\frac{1}{2}}Q_PA^{\frac{1}{2}}\|(1+B)^{-1}B \leq \|A^{\frac{1}{2}}Q_PA^{\frac{1}{2}}\| \;\leq\; \|Q_P\|\,\|A\| \leq\; z^{-1}\|Q_P\|
\end{aligned}
\tag{152}
$$

where

$$
B=\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t \;\in\; \mathbb{R}_+
$$

Proof of Lemma 20 vii

Thus,

$$\frac{1}{P}(E_{t-1} - E_t)[\mathrm{tr}(zI + \hat{\Psi}_T)^{-1} Q_P] = \frac{1}{P}(E_{t-1} - E_t)[\gamma_t]$$

forms a bounded martingale difference sequence:

$$\begin{aligned} |(E_{t-1} - E_t)[q_T]| &= |(E_{t-1} - E_t)[\gamma_t]| \\ &\leq |E_{t-1}[\gamma_t]| + |E_t[\gamma_t]| \underbrace{\leq}_{(152)} 2z^{-1}\|Q_P\| \end{aligned} \tag{153}$$

The Burkholder-Davis-Gundy inequality i

Let $\{\mathcal{F}_t\}$ be a filtration, X_T an $\mathcal{F}_t$-measurable random variable, and

$$\xi_t \;=\; E_t[X_T] \;-\; E_{t-1}[X_T] \tag{154}$$

be the corresponding Martingale difference sequence so that

$$X_T - E_0[X_T] \;=\; \sum_{t=1}^{T} \xi_t \,. \tag{155}$$

Then, for any $q \geq 2$, we have

$$E[|X_T - E_0[X_T]|^q] \;\leq\; K_q E\left[\left(\sum_{t=1}^{T} \xi_t^2\right)^{q/2}\right] \tag{156}$$

The Burkholder-Davis-Gundy inequality ii

for some constant $K_q > 0$. For $q = 2$, it is an identity:

$$E[|X_T - E_0[X_T]|^2] \;=\; E\left[\left(\sum_{t=1}^{T} \xi_t^2\right)\right]$$

Using this, we get

$$E[|q_T(z) - E[q_T(z)]|^q] \;\leq\; K_q P^{-q} E\left[\left(\sum_{t=1}^{T} \underbrace{|(E_{t-1} - E_t)[\gamma_t]|^2}_{\leq (2z^{-1}\|Q_P\|)^2}\right)^{q/2}\right] \tag{157}$$

$$\leq\; K_q P^{-q} (T(2z^{-1}\|Q_P\|)^2)^{q/2} \;=\; K_q (2\|Q_P\|/z)^q P^{-q/2} \left(\frac{P}{T}\right)^{-q/2}.$$

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

The Burkholder-Davis-Gundy inequality iii

Lemma

Suppose that

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

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

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

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

The proof of Lemma 20 is complete.

Table of Contents

Theorem

Suppose that $\Psi = I$. *Then,*

$$m(-z;c) = \frac{\sqrt{(1-c-z)^2+4cz}-(1-c-z)}{2cz} \tag{158}$$

Proof The proof follows from Proposition 1. Indeed, placing $\Psi = I$ in the definition of $\xi(z;c)$ gives us:

$$\begin{aligned} E[T^{-1}\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1})] &= \left(\frac{T}{P}\right)^{-1} P^{-1}E[(\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1})] \\ &= cP^{-1}E[\operatorname{tr}((zI+\hat{\Psi}_{T,t})^{-1})] \\ &= cm(-z;c) = \xi(z;c). \end{aligned} \tag{159}$$

The Fixed Point Equation ii

Plugging the previous expression into the closed form solution of $\xi(z;c)$ provides us with:

$$\xi(z;c) = cm(-z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)} \tag{160}$$

We can solve for $m(-z;c)$ as follows:

$$\begin{aligned} m(-z;c) &= \frac{1 - zm(-z;c)}{1 - c + czm(-z;c)} \\ &\Rightarrow (1 - c + z)m(-z;c) + czm(-z;c)^2 - 1 = 0 \\ &\Rightarrow m(-z;c) = \frac{\pm\sqrt{(1-c-z)^2 + 4cz} - (1 - c - z)}{2cz}. \end{aligned} \tag{161}$$

By definition, $m(-z;c)$ is positive. Therefore, the positive root is the true solution, concluding the proof.

Recall the notion of weak convergence.

The Fixed Point Equation iii

Definition

Given a sequence of measures $d\mu_P(x)$ supported on a bounded interval $[a, b]$, we say that $d\mu_P(x)$ weakly converges to $d\mu(x)$ if

$$\lim_{P\to\infty} \int f(x) d\mu_P(x) \;=\; \int f(x) d\mu(x) \tag{162}$$

for any continuous function $f(x)$ on $[a, b]$.

Assumption The matrices Ψ_P are uniformly bounded, and the eigenvalue distributions of Ψ_P weakly converge to a limit distribution $dF^{\Psi_P}(\lambda)$ as $P \to \infty$. By Definition 24, the limits

$$\begin{aligned} m_\Psi(z) \;&=\; \int \frac{1}{x - z} F^{\Psi}(x) \;=\; \lim_{P\to\infty} P^{-1} \operatorname{tr}((\Psi_P - zI)^{-1}) \\ &= \lim_{P\to\infty} P^{-1} \sum_i (\lambda_i(\Psi_P) - z)^{-1} \end{aligned} \tag{163}$$

The Fixed Point Equation iv

and

$$\psi_{*,k} \;=\; \lim_{P\to\infty} P^{-1}\,\mathrm{tr}(\Psi^k) \;=\; \int \lambda^k dF_\Psi(\lambda) \tag{164}$$

exist.

Lemma

Let

$$\xi_k(z) \;=\; \lim P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi})^{-1}]\,. \tag{165}$$

If the limits ξ_k and ξ_{k+1} exist, then

$$\psi_{*,k} \;=\; z\xi_k(z) + (1-c+czm(-z;c))\xi_{k+1}(z) \tag{166}$$

Since Ψ is uniformly bounded, so is Ψ^k. Therefore, due to Lemma 20 we will have

$$\xi_k(z) \;=\; \lim P^{-1}\,\mathrm{tr}\,\Psi^k(zI+\hat{\Psi})^{-1} \;=\; \lim P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi})^{-1}]\,. \tag{167}$$

The Fixed Point Equation v

Now, the following algebraic identity gives us

$$
\begin{aligned}
P^{-1}\,\mathrm{tr}(\Psi^k) &= P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)] \\
&= zP^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi}_T)^{-1}] + P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&\approx z\,\xi_k(z) + P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_t S_t'] \\
&= \{\textit{symmetry across } t\} = z\xi_k(z) + P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi}_T)^{-1}S_t S_t'] \\
&= \{\textit{by Sherman} - \textit{Morrison } (143)\} \\
&= z\xi_k(z) + P^{-1}\,\mathrm{tr}\,E[\Psi^k(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}S_t'] \\
&= z\xi_k(z) + P^{-1}\,\mathrm{tr}\,E[\frac{S_t'\Psi^k(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}],
\end{aligned}
\tag{168}
$$

where $\approx$ signifies that the difference between the left- and the right-hand sides converge to zero in probability.

The Fixed Point Equation vi

By Lemmas 10 and 20, we have

$$P^{-1}S_t'\Psi^k(zI+\hat{\Psi}_{T,t})^{-1}S_t \to P^{-1}E[\mathrm{tr}(\Psi^{k+1}(zI+\hat{\Psi}_{T,t})^{-1})] \;=\; \xi_{k+1}(z). \tag{169}$$

Using (169) and Proposition 2 together with (167) gives us

$$\mathrm{tr}\,E[\frac{(P)^{-1}S_t'\Psi^k(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}] \to \frac{\xi_{k+1}(z)}{1+\xi(z;c)}. \tag{170}$$

By (121),

$$1+\xi(z;c) = 1+\frac{1-zm(-z;c)}{c^{-1}-1+zm(-z;c)} \;=\; \frac{1}{1-c+czm(-z;c)}\,. \tag{171}$$

Plugging (170) into (168), we get

$$P^{-1}\,\mathrm{tr}(\Psi^k) \approx z\xi_k(z)+\frac{\xi_{k+1}(z)}{1+\xi(z;c)} = z\xi_k(z)+(1-c+czm(-z;c))\xi_{k+1}(z)\,. \tag{172}$$

The proof is complete.

The identity (171) has important implications for the behavior of m that will play a key role in our analysis below. The following is true.

Lemma

We have

$$\Re\left(-\frac{z}{1 - c - c\,z\,m(-z;c)}\right) < -\Re z \tag{173}$$

for any $z \in \mathbb{C}$ *with* $\Re z > 0$. *In particular,*

$$\left|-\frac{z}{1 - c - c\,z\,m(-z;c)}\right| > \Re z \tag{174}$$

when $\Re z > 0$.

The Fixed Point Equation viii

Proof We have

$$\begin{aligned} z\xi(z;c) &= \lim P^{-1}E[\mathrm{tr}(\Psi z(zI+\hat{\Psi})^{-1})] \\ &= \lim P^{-1}E[\mathrm{tr}(\Psi(zI+\hat{\Psi}-\hat{\Psi})(zI+\hat{\Psi})^{-1})] \\ &= \psi_{*,1} - \lim P^{-1}E[\mathrm{tr}(\Psi\hat{\Psi}(zI+\hat{\Psi})^{-1})] \end{aligned} \tag{175}$$

Let $\hat{\Psi} = UDU'$ be the spectral representation of $\hat{\Psi}$. Then,

$$\mathrm{tr}(\Psi\hat{\Psi}(zI+\hat{\Psi})^{-1}) = \mathrm{tr}(\Psi UD(zI+D)^{-1}U') = \mathrm{tr}(U'\Psi UD(zI+D)^{-1}) = \sum_i \alpha_i\lambda_i/(z+\lambda_i)\,, \tag{176}$$

where $\alpha_i > 0$ are the diagonal elements of $U'\Psi U$ and $\alpha_i > 0$ are the eigenvalues of $\hat{\Psi}$. Thus, for any $z = a + ib \in \mathbb{C}$ with a positive real part $a > 0$ and $i = \sqrt{-1}$, we have that

$$1/(z+\lambda_i) = \frac{1}{a+\lambda_i+ib} = \frac{a+\lambda_i-ib}{(a+\lambda_i)^2+b^2}$$

also has a positive real part, and hence

$$\Re(\sum_i \alpha_i \lambda_i/(z+\lambda_i)) \;=\; \sum_i \alpha_i \frac{\lambda_i(a+\lambda_i)}{(a+\lambda_i)^2+b^2} \;\leq\; \sum_i \alpha_i = \mathrm{tr}(U'\Psi U) = \mathrm{tr}(\Psi)$$

Thus, for any z with $\Re(z) > 0$, we have that $z\xi(z;c)$ has a positive real part. Therefore, by (171), we have

$$x \;=\; -\frac{z}{1 \;-\; c \;-\; c\, z\, m(-z;c)} \;=\; -z - z\xi(z) \tag{177}$$

has a negative imaginary part for $\Re z > 0$. ∎

The Marcenko-Pastur Master Equation i

Theorem

For any $c > 0$ and $z < 0$, the distribution of eigenvalues of $\hat{\Psi}$ in the limit as $P, T \to \infty$, $P/T \to c$ converges to a distribution whose Stieltjes transform, $m(z; c)$, is the unique positive solution to the equation

$$\begin{aligned} m(z;c) &= \frac{1}{1 \; - \; c \; - \; c\, z\, m(z;c)}\, m_{\Psi}\left(\frac{z}{1 \; - \; c \; - \; c\, z\, m(z;c)}\right) \\ &= \int \frac{1}{\lambda(1 \; - \; c \; - \; c\, z\, m(z;c)) - z} dF^{\Psi}(\lambda). \end{aligned} \tag{178}$$

Proof Denote $A = 1 - c + czm(-z; c)$. We have

$$(-xI + \Psi)^{-1} \;=\; -x^{-1} \sum_{k=0}^{\infty} x^{-k} \Psi^k \,, \tag{179}$$

and therefore

$$m_\Psi(x) \;=\; P^{-1} \operatorname{tr}((-xI + \Psi)^{-1}) \;=\; -x^{-1} \sum_{k=0}^{\infty} x^{-k} \psi_{*,k} \;=\; \int \frac{1}{\lambda - x} dF^\Psi(\lambda) \,. \tag{180}$$

We rewrite (180), with the recursive relationship of Lemma 25, assuming that $|x| > 1$: In this case, uniform boundedness implies

$$
\begin{aligned}
m_\Psi(x) &= \lim\ -x^{-1}\sum_{k=0}^{\infty} x^{-k} P^{-1}\,\mathrm{tr}(\Psi^k) \\
&= -x^{-1}\sum_{k=0}^{\infty} x^{-k}\lim P^{-1}\,\mathrm{tr}(\Psi^k) \\
&\approx -x^{-1}\sum_{k=0}^{\infty} x^{-k}(z\xi_k(z) + A\xi_{k+1}(z)) \\
&= -x^{-1}\left(z\sum_{k=0}^{\infty} x^{-k}\xi_k(z) + A\sum_{k=0}^{\infty} x^{-k}\xi_{k+1}(z))\right) \\
&= -x^{-1}\left(zm(-z;c) + z\sum_{k=0}^{\infty} x^{-k-1}\xi_{k+1}(z) + Ax\sum_{k=0}^{\infty} x^{-k-1}\xi_{k+1}(z))\right) \\
&= -x^{-1}\left(zm(-z;c) + (z+Ax)\sum_{k=0}^{\infty} x^{-k-1}\xi_{k+1}(z)\right),
\end{aligned}
\tag{181}
$$

Let now $x = -\frac{z}{A} = -\frac{z}{1-c+czm(-z;c)}$, and suppose that $\Re z > 1$. Then, by Lemma 26, $|x^{-1}| = |x|^{-1} < 1$ and therefore (181) holds. Thus,

$$\begin{aligned} & m_\Psi(-\frac{z}{1-c+czm(-z;c)}) \\ = & -\left(\frac{-z}{1-c+czm(-z;c)}\right)^{-1}\left(zm(-z;c)+(z+-z)\sum_{k=0}^{\infty}x^{-k-1}\xi_{k+1}(z)\right) \\ = & (1-c+czm(-z;c))m(-z;c) \\ \Rightarrow & m(-z;c) = \frac{1}{1-c+czm(-z;c)}m_\Psi(-\frac{z}{1-c+czm(-z;c)})\,. \end{aligned} \tag{182}$$

By replacing z with $-z$ we arrive at the desired equation, concluding the proof for $\Re z > 1$. However, the left- and right-hand sides of this identity are analytic functions for z with $\Re z > 0$ by Lemma 26 and the fact that $m_\Psi(x)$ is analytic for $\Re x < 0$. If two analytic functions coincide on an open set, they coincide everywhere. The proof is complete.