Asset Pricing and ML

Semyon Malamud

EPFL

Table of Contents

Mean-Variance Optimization: Unconditional 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

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

- Thus, the excess return on your wealth is

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

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

Mean-Variance Optimization: Unconditional iii

- Intuitively, we like **high return** and **low variance,** hence, we might try to find a **static** portfolio that maximizes

$$\pi \;=\; \arg\max_{\pi} \left(E[\pi' R_{t+1}] \;-\; 0.5 \underbrace{\gamma}_{\textit{risk aversion}} \mathrm{Var}[\pi' R_{t+1}] \right) \tag{5}$$

- The solution is Markowitz

$$\pi \;=\; \gamma^{-1} \mathrm{Var}[R]^{-1}\, E[R]\,. \tag{6}$$

- Alternatively, one could optimize

$$\pi \;=\; \arg\max_{\pi} \left(E[\pi' R_{t+1}] \;-\; 0.5 \underbrace{\gamma}_{\textit{risk aversion}} E[(\pi' R_{t+1})^2] \right) \tag{7}$$

Mean-Variance Optimization: Unconditional iv

and the solution is

$$\begin{aligned} \tilde{\pi} &= \gamma^{-1}(E[R_{t+1}R_{t+1}'])^{-1}\, E[R_{t+1}] \\ &= const \cdot \pi, \qquad const = \frac{1}{1 + E[R_{t+1}]'\mathrm{Var}[R_{t+1}]^{-1}E[R_{t+1}]} \end{aligned} \tag{8}$$

where

$$E[R_{t+1}R_{t+1}'] = \mathrm{Var}[R_{t+1}] + E[R_{t+1}]E[R_{t+1}]' = (E[R_{i,t+1}R_{j,t+1}])_{i,j=1}^{N} \tag{9}$$

Why Are the Two Markowitz Portfolios Proportional? The Sherman-Morrison formula i

The magic behind is the

Lemma (Sherman-Morrison formula)

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

and

$$(A + xx')^{-1}x \;=\; \frac{A^{-1}x}{1 + x'A^{-1}x} \tag{11}$$

Proof[Proof of the Sherman-Morrison formula] Recall that

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

Why Are the Two Markowitz Portfolios Proportional? The Sherman-Morrison formula ii

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

$$\begin{aligned}
&(A+xx')(A^{-1} \; - \; \frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x}) \\
&= \; I \; - \frac{xx'A^{-1}}{1+x'A^{-1}x} + xx'A^{-1} - xx'\frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x} \\
&= \; I \; - \frac{xx'A^{-1}}{1+x'A^{-1}x} + xx'A^{-1} - xx'A^{-1}\frac{x'A^{-1}x}{1+x'A^{-1}x} \; = \; I
\end{aligned} \tag{12}$$

and

$$(A+xx')^{-1}x \; = \; (A^{-1} \; - \; \frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x})x \; = \; \frac{A^{-1}x}{1+x'A^{-1}x} \tag{13}$$

(Very Big) Issues with Markowiz

- Markowitz **assumes that we know the truth! The true**

$$E[R] \;=\; (E[R_{i,t+1}])_{i=1}^{N_t},\ \mathrm{Var}[R] \;=\; (\mathrm{Cov}(R_{i,t+1}, R_{j,t+1}))_{i,j=1}^{N_t} \quad (14)$$

  where N_t is the number of assets (stocks?) available at time t.
- The problem is that:
  - **expected stock returns move a lot over time**: Hence, using **static** portfolio is a **very bad idea**
  - we just **do not have enough data** to estimate $E[R]$ and $\mathrm{Var}[R]$. We can use naive

$$\bar{E}[R] \;=\; \frac{1}{T}\sum_{t=1}^{T} R_t,\ \overline{\mathrm{Var}}[R] \;=\; \frac{1}{T}\sum_{t=1}^{T} \underbrace{\underbrace{(R_t - \bar{E}[R])}_{N\times 1}\underbrace{(R_t - \bar{E}[R])'}_{1\times N}}_{N\times N}$$

Incorporating Conditional Information: The conditional Markowitz i

We would like to incorporate conditional information.

- mean-variance optimization:

$$\pi_t = \arg\max_{\pi_t} \left(E_t[\pi_t' R_{t+1}] - 0.5 \underbrace{\gamma}_{risk\ aversion} \mathrm{Var}_t[\pi_t' R_{t+1}] \right) \quad (15)$$

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

$$\underbrace{\pi_t}_{conditional\ tangency\ portfolio} = \gamma^{-1} \underbrace{(\mathrm{Var}_t[R_{t+1}])^{-1}}_{N\times N\ covariance\ matrix} \underbrace{E_t[R_{t+1}]}_{N\times 1\ expected\ returns} \quad (16)$$

Incorporating Conditional Information: The conditional Markowitz ii

- Similarly,

$$\tilde{\pi}_t = \arg\max_{\pi} \left(E_t[\pi' R_{t+1}] - 0.5 \underbrace{\gamma}_{risk\ aversion} E_t[(\pi' R_{t+1})^2] \right) \tag{17}$$

is given by

$$\begin{aligned} \tilde{\pi}_t &= \gamma^{-1}(E_t[R_{t+1}R'_{t+1}])^{-1} E_t[R_{t+1}] \\ &= \frac{1}{1 + E_t[R_{t+1}]'\mathrm{Var}_t[R_{t+1}]^{-1}E_t[R_{t+1}]}\pi_t \end{aligned} \tag{18}$$

where

$$E_t[R_{t+1}R'_{t+1}] = \mathrm{Var}_t[R_{t+1}] + E_t[R_{t+1}]E_t[R_{t+1}]' \tag{19}$$

Incorporating Conditional Information: The conditional expectation i

- We would need

$$
\begin{aligned}
E_t[R_{t+1}] &= \arg \min_{F:\mathbb{R}^P \to \mathbb{R}^N} E[\|R_{t+1} - F(S_t)\|^2] \\
E_t[R_{t+1}R'_{t+1}] &= \arg \min_{G:\mathbb{R}^P \to \mathbb{R}^{N\times N}} E[\|R_{t+1}R'_{t+1} - G(S_t)\|^2]
\end{aligned} \tag{20}
$$

- The reality is that **we still cannot compute $E[\cdot]$** because we do not have enough data. So, we will still be doing

$$
E_t[X_{t+1}] = \arg \min_F \frac{1}{T} \sum_t |X_{t+1} - F(S_t)|^2 \tag{21}
$$

Table of Contents

Intoduction to Asset Pricing i

- I promised Asset Pricing, but we did Markowitz instead. Why?
- Intuitively, we expect that

$$P_{i,t} \; = \; \underbrace{(R_{f,t})^{-1}\, E_t[P_{i,t+1} + D_{i,t+1}]}_{\textit{Definitely wrong in the data}} \tag{22}$$

  because the **discount factor** $(R_{f,t})^{-1}$ is too naive
- **We need a smart discount factor (SDF)**:

$$P_{i,t} \; = \; E_t[\qquad \underbrace{M_{t,t+1}}_{\textit{stochastic discount factor}} \qquad (P_{i,t+1} + D_{i,t+1})] \tag{23}$$

Intoduction to Asset Pricing ii

- with a bit of algebra, this is equivalent to

$$E_t[R_{i,t+1}\, M_{t,t+1}] \;=\; 0 \tag{24}$$

and

$$\underbrace{E_t[M_{t,t+1}]}_{scale\ \ of\ \ the\ \ SDF} \;=\; R_{f,t}^{-1}$$

- By direct calculation,

$$M_{t+1} \;=\; 1 \;-\; \tilde{\pi}_t' R_{t+1} \tag{25}$$

does the job:

$$\begin{aligned} E_t[R_{t+1} M_{t,t+1}] \;&=\; E_t[R_{t+1}\,(1 \;-\; R_{t+1}'\tilde{\pi}_t)] \\ =\; E_t[R_{t+1}] \;-\; E_t[R_{t+1} R_{t+1}']\,\tilde{\pi}_t \;&=\; 0 \end{aligned} \tag{26}$$

Intoduction to Asset Pricing iii

implies

$$\tilde{\pi}_t = E_t[R_{t+1} R'_{t+1}]^{-1} E_t[R_{t+1}] \tag{27}$$

We now state

Theorem

Nothing Has Alpha Against $\tilde{\pi}'_t R_{t+1}$

Implications for Testing “If we have found a new, useful strategy”

- If we have found the true, ultimate $\tilde{\pi}_t$, nothing has alpha against it: If you have some other portfolio ξ_t and run the regression

$$\xi_t' R_{t+1} \;=\; \alpha \;+\; \beta \tilde{\pi}_t' R_{t+1} \;+\; \varepsilon_{t+1}\,.$$

- If α is not significant, reject the strategy ξ_t, move somewhere else.
- if α is significant, the $\tilde{\pi}_t$ is not efficient and you should try combining it with $\tilde{\pi}_t$

Testing Conditional Efficiency

- We cannnot compute $E_t[\cdot]$
- Instead, we can build instruments Z_t and test that

$$E_t[M_{t+1}R_{t+1}] \; = \; 0 \; \Leftrightarrow \; E[Z_t\, M_{t+1}R_{t+1}] \; = \; 0$$

  for all instruments!
- Thus, we need to build infinitely many Z_t thought machine learning and then test

$$\frac{1}{T}\sum_t Z_t\, M_{t+1}R_{t+1} \; \approx \; 0$$

  Complexity is always there!

From Non-Tradable to Tradable SDFs i

- What about asset pricing theory?
- the SDF

$$\widetilde{M}_{t+1} = \underbrace{\frac{e^{-\rho}U'(C_{t+1})}{U'(C_t)}}_{=IMRS}$$

comes from the Euler equation (things get more complex with Epstein-Zin preferences, expectations, sentiments, etc)

$$E_t[\underbrace{\frac{e^{-\rho}U'(C_{t+1})}{U'(C_t)}}_{=IMRS}(R_{t+1}+R_{f,t})] \;=\; 1 \;\Leftrightarrow\; E_t[\widetilde{M}_{t+1}R_{t+1}] \;=\; 0 \tag{28}$$

because

$$R_{f,t} \;=\; E_t[\widetilde{M}_{t+1}]^{-1}\,. \tag{29}$$

From Non-Tradable to Tradable SDFs ii

- When markets are complete,

$$M_{t+1} \text{ proportional to } \widetilde{M}_{t+1}$$

and the proportionality constant can be pinned down by

$$\widetilde{M}_{t+1} = \frac{M_{t+1}}{R_{f,t} E_t[M_{t+1}]} \tag{30}$$

- The normalization ensures that

$$E_t[\widetilde{M}_{t+1}] = R_{f,t}^{-1} \tag{31}$$

From Non-Tradable to Tradable SDFs iii

- In general, we need to **project**

$$\underbrace{\frac{M_{t+1}}{R_{f,t}E_t[M_{t+1}]}}_{unique\ \ tradable} = Proj_t(\widetilde{M}_{t+1}) = \arg\min_{a,\pi} E_t[(\widetilde{M}_{t+1} - a(1 - \pi' R_{t+1}))^2] \tag{32}$$

- solution is

$$\pi = \tilde{\pi}_t, \ a = \frac{1}{R_{f,t}E_t[M_{t+1}]}$$

Table of Contents

Panel Datasets: Leveraging the Power of Big Data i

- Now comes the big question: **How do we measure the conditional** expectations, $E_t[R_{t+1}]$ and $E_t[R_{t+1}R'_{t+1}]$?
- Running prediction models **per stock** is infeasible due to insufficient data:

$$E_t[R_{i,t+1}] \underbrace{=}_{bad\ \ idea} g_i(X_{i,t})$$

- use panel data

$$E_t[R_{i,t+1}] \underbrace{=}_{good\ \ idea} g(X_{i,t})$$

- **panel** means **same function** g for all stocks.
- **non-linear** g means machine learning
- What about the covariance matrix? How do we model a **time-varying** covariance structure?

Panel Datasets: Leveraging the Power of Big Data ii

- Typically, we assume a factor structure:

$$R_{t+1} = \underbrace{S}_{factor\ \ exposures} \underbrace{F_{t+1}}_{factors} + \varepsilon_{t+1}$$

- In reality, factor exposures are time-varying:

$$R_{t+1} = S_t F_{t+1} + \varepsilon_{t+1}$$

- If $\mathrm{Cov}_t[F_{t+1}] = \Sigma_F$ and

$$\mathrm{Cov}_t[\varepsilon_{t+1}] = \mathrm{diag}(\sigma_{i,t}^2), \ \sigma_{i,t} = \textit{idiosyncratic volatility}$$

so that the Conditional covariance matrix is given by

$$E_t[R_{t+1}R'_{t+1}] = S'_t \Sigma_F S_t + \mathrm{diag}(\sigma_{i,t}^2)$$

- Equivalently:

$$E_t[R_{i,t+1}R_{j,t+1}] \;=\; \underbrace{S_{i,t}'\,\Sigma_F\,S_{j,t}}_{systematic\;\;covariance} \;+\; \underbrace{\delta_{i,j}\sigma_{i,t}^2}_{idiosyncratic\;\;variance}$$

  where Σ_F and $\sigma_{i,t}$ are to be estimated.

- **Can we avoid computing the conditional covariance matrix?**

Managed Portfolios and Rich Conditional Factor Structures i

- Suppose

$$R_{i,t+1} = \underbrace{S'_{i,t}}_{conditional\ \ betas} \cdot \underbrace{\tilde{F}_{t+1}}_{latent\ \ factors} + \varepsilon_{i,t+1}$$

-

$$E_t[\tilde{F}_{t+1}] = \underbrace{\lambda_F}_{latent\ \ factor\ \ risk\ \ premia}, \; E_t[\tilde{F}_{t+1}\tilde{F}'_{t+1}] = \underbrace{\Sigma_F}_{latent\ \ factor\ \ cov}$$

- Thus,

$$E_t[R_{t+1}] = S_t \lambda_F$$

and

$$E_t[R_{t+1}R'_{t+1}] = S_t \Sigma_F S'_t + \Sigma_\varepsilon$$

Managed Portfolios and Rich Conditional Factor Structures ii

- $$M_{t+1} \; = \; 1 - \tilde{\pi}_t' R_{t+1} = \; 1 \; - \; W(S_t)' R_{t+1} \,, \tag{33}$$

  where $\tilde{\pi}_t = E_t[R_{t+1} R_{t+1}']^{-1} E_t[R_{t+1}]$ and, hence,

  $$W(S_t) \; = \; \underbrace{(S_t \Sigma_{F,t} S_t' + \Sigma_\varepsilon)^{-1}}_{conditional \;\; covariance} \qquad \underbrace{S_t \lambda_F}_{conditional \;\; expectation} \tag{34}$$

- Define **managed portfolios**

  $$F_{t+1} \; = \; S_t' R_{t+1}. \tag{35}$$

  and the unconditionally efficient portfolio

  $$\lambda = E[F_{t+1} F_{t+1}']^{-1} E[F_{t+1}] \tag{36}$$

Managed Portfolios and Rich Conditional Factor Structures iii

- By construction,

$$M^{F}{}_{t+1} = 1 - \lambda' F_{t+1} \tag{37}$$

prices factors unconditionally:

$$E[M^{F}{}_{t+1} F_{t+1}] \;=\; 0 \tag{38}$$

- However,

$$E_t[M^{F}{}_{t+1} R_{t+1}] \;\neq\; 0$$

because

$$\lambda' S_t' R_{t+1} \;\neq\; \lambda_F{}' \, S_t' \, \Sigma_t^{-1} \, R_{t+1},$$

with

$$\Sigma_t \;=\; (S_t \Sigma_{F,t} S_t' + \Sigma_\varepsilon)$$

Click on this link to know more:

APT or "AIPT"? The Surprising Dominance of Large Factor Models

Theorem

Suppose that in the limit, as $P \to \infty$, the vector of latent risk premia λ_F satisfies

$$\lambda_F' A \lambda_F \;\to\; 0 \tag{39}$$

for any symmetric, positive definite A with uniformly bounded trace. Let

$$M^F{}_{t+1} = 1 - \lambda' F_{t+1}, \tag{40}$$

be the factor approximation for the SDF with λ. Then, $M^F{}_{t+1}$ converges to M_{t+1} and the Sharpe ratio of $\lambda' F_{t+1}$ converges to that of $W(S_t)' R_{t+1}$ as $P \to \infty$. In particular,

$$E_t[M^F{}_{t+1} R_{t+1}] \;\to\; 0$$

Sources of Complexity i

- We now know: If

$$R_{t+1} \ = \ \underbrace{S_t}_{N_t \times P \ \ signals} \ \ \underbrace{\tilde{F}_{t+1}}_{P \times 1 \ \ latent \ \ factors} \ + \ \underbrace{\varepsilon_{t+1}}_{residuals} \tag{41}$$

  then we build

$$F_{t+1} \ = \ S_t' R_{t+1} \ = \ (S_t' S_t)\tilde{F}_{t+1} \ + \ (S_t' \varepsilon_{t+1}) \tag{42}$$

- But where do S_t come from?
- Suppose

$$R_{i,t+1} = \beta(X_{i,t})' G_{t+1} + u_{i,t+1}, \tag{43}$$

- $$\beta(X_{i,t}) \approx \sum_{p=1}^{P} \xi_p \, S_{i,t,p} \; = \; \xi' \underbrace{S_{i,t}}_{P\times 1}, \tag{44}$$

  where

  $$S_{i,t} = (\sigma(\omega_p' X_{i,t}))_{p=1}^{P}. \tag{45}$$

- This gives

$$\begin{aligned} &\underbrace{R_{t+1}}_{N\times 1} \approx \underbrace{S_t}_{N\times P} \underbrace{\tilde{F}_{t+1}}_{P\times 1} + u_{t+1}, \; \textit{with} \\ &\tilde{F}_{t+1} = \underbrace{\xi}_{P\times 1} G_{t+1}, \; \nu \; = \; E[\tilde{F}_{t+1}] = \xi \, E[G_{t+1}]. \end{aligned} \tag{46}$$

- If β is highly non-linear, we need to go for a high-dimensional S_t

Sources of Complexity iii

- The true SDF return is

$$(\beta_t \beta_t' + \Sigma_u)^{-1} \beta_t E[G_{t+1}] \underbrace{=}_{Sherman-Morrison} \Sigma_u^{-1} \beta_t E[G_{t+1}] \frac{1}{1 + \beta_t' \Sigma_u^{-1} \beta_t} \tag{47}$$

In high dimensions, $\beta_t' \Sigma_u^{-1} \beta_t \approx const$. Furthermore, if β_t are sufficiently complex, $\Sigma_u^{-1} \beta_t \approx const \beta_t$. Thus, we end up with

$$\pi_t \sim \beta_t = S_t \xi \tag{48}$$

and the SDF is

$$\pi_t' R_{t+1} = \underbrace{\xi'}_{factor\ \ weights} \underbrace{S_t' R_{t+1}}_{F_{t+1}} . \tag{49}$$

Complexity in the Cross Section: A Brief History i

- Most academic attempts to build an SDF assume

$$M^{\star}_{t+1} = 1 - \sum_{i=1}^{N} w(X_{i,t}) R_{i,t+1} \tag{50}$$

- Cross-sectional asset pricing is about $w_t = w(X_t)$
  - Explains differences in average returns
  - Defines the MVE portfolio
- Why does cross-section literature rarely start here? Because w must be estimated
  - This is a high-dimensional (*complex*) problem
  - We know: In-sample tangency portfolio behaves horribly out-of-sample
  - Why? Complexity ($n/T \not\to 0$) $\to$ LLN doesn't apply $\to$ IS and OOS diverge

Complexity in the Cross Section: A Brief History ii

- Standard solution: Restrict w
  - E.g., Fama-French: $w_{i,t} = b_0 + b_1 \text{Size}_{i,t} + b_2 \text{Value}_{i,t}$ (Brandt et al. 2007 generalize):

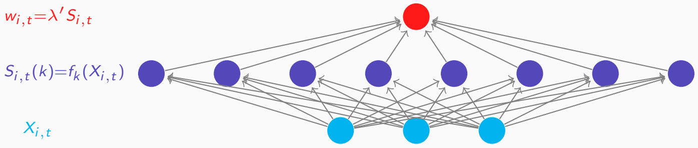
$$\begin{aligned} \sum_{i=1}^{N} w(X_{i,t}) R_{i,t+1} &= \sum_{i=1}^{N} (b_0 + b_1 \text{Size}_{i,t} + b_2 \text{Value}_{i,t}) R_{i,t+1} \\ &= b_0 \sum_{i=1}^{N} R_{i,t+1} + b_1 \sum_{i=1}^{N} \text{Size}_{i,t} R_{i,t+1} + b_2 \sum_{i=1}^{N} \text{Value}_{i,t} R_{i,t+1} \\ &= b_0 MKT_{t+1} + b_1 SMB_{t+1} + b_2 HML_{t+1} . \end{aligned} \tag{51}$$

  - Reduces parameters, implies factor model: $M_{t+1} = 1 - b_0 MKT - b_1 SMB - b_2 HML$
  - "Shrinking the cross-section" Kozak et al. (2020) — use a few PCs of anomaly factors

Complexity in the Cross Section: Machine Learning Perspective i

Rather than restricting $w(X_t)$....

- ...expand parameterization, saturate with conditioning information
- For example, approximate w with neural network: $w(X_{i,t}) \approx \lambda' S_{i,t}$
- $P \times 1$ vector $S_{i,t}$ is known nonlinear function of original predictors $X_{i,t}$



Complexity in the Cross Section: Machine Learning Perspective ii

- Implies that empirical SDF is a high-dimensional factor model

$$\sum_{i=1}^{N} w(X_{i,t})R_{i,t+1} = \sum_{i=1}^{N}(\sum_{k}\lambda_k \underbrace{S_{i,t}(k))}_{S_{i,t}(k)=f_k(X_{i,t})} R_{i,t+1} = \sum_{k}\lambda_k \underbrace{\sum_{i=1}^{N} S_{i,t}(k)R_{i,t+1}}_{F_{k,t+1}} \tag{52}$$

$$M^{\star}_{t+1} \approx M_{t+1} = 1 - \lambda' S_t' R_{t+1} = 1 - \lambda' F_{t+1}$$

Complexity in the Cross Section: Machine Learning Perspective i

The Objective:

- Maximize out-of-sample Sharpe ratio (equivalently, minimize out-of-sample pricing errors) of SDF

The Choice:

- Fix T data points. Decide on "complexity" (number of factors P) to use in approximating model

The Tradeoff:

Complexity in the Cross Section: Machine Learning Perspective ii

- Simple SDF ($P << T$) has low variance (thanks to parsimony) but is a poor approximator of w
- Complex SDF ($P > T$) is a good approximator but may behave poorly (and requires shrinkage)

- Which P should the analyst opt for? Does the benefit of more factors justify their cost?

Answer:

- Use the largest factor model (largest P) that you can compute

Implementation i

- Build a bunch of features (random features if you want a shallow model; deep features (output layer) if you want a deep model).
- Call them $S_{i,t}(k) = f_k(X_{i,t};\theta_k),\ k = 1,\cdots,P$
- Build the factors

$$F_{t+1}(k) \ = \ \sum_{i=1}^{N_t} S_{i,t}(k) R_{i,t+1} \tag{53}$$

- Take the vector of factors $F_{t+1} = (F_{t+1}(k))_{k=1}^{P}$ and minimize

$$\min_{\lambda} \frac{1}{T}\sum_{t=1}^{T}(1-\lambda' F_{t+1})^2 \ + \ z\,\|\lambda\|^2 \tag{54}$$

This objective is known as the **Maximal Sharpe Ratio Regression (MSRR)**. For a deep model, you need to minimize this objective using GD

Implementation ii

- Why MSRR? Well,

$$\frac{1}{T}\sum_{t=1}^{T}(1-\lambda' F_{t+1})^2 \;\approx\; E[(1-\lambda' F_{t+1})^2] \\ =\; 1-2E[\lambda' F_{t+1}]+E[(\lambda' F_{t+1})^2] \;=\; 1 \;-\; 2E[U(\lambda' F_{t+1})], \tag{55}$$

where

$$U(x) \;=\; x-0.5x^2$$

- Now, $\tilde{\pi}_t = E_t[R_{t+1}R'_{t+1}]^{-1}E_t[R_{t+1}]$ solves

$$\max_{\pi} E_t[U(\pi_t' R_{t+1})] \tag{56}$$

Implementation iii

It is conditionally efficient for a quadratic utility. By the law of iterated expectations,

$$E[E_t[U(\pi_t' R_{t+1})]] \ = \ E[U(\pi_t' R_{t+1})]$$

and dynamic consistency gives

$$\max_{all \ \ policies \ \ \pi_t} E[U(\pi_t' R_{t+1})] \ = \ E[\max_{\pi} E_t[U(\pi_t' R_{t+1})]$$

- Thus, MSRR looks for conditional policies that maximize unconditional utility and hence, by consistency, are conditionally optimal.

RMT i

- $$\hat{\lambda}(z) = \left(zI + \frac{1}{T}\sum_{t=1}^{T} F_t F_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} F_t \underbrace{\not\approx}_{Complexity!} \lambda_*(z) \quad (57)$$

  where

  $$\lambda_*(z) = (zI + E[FF'])^{-1} E[F] \quad (58)$$

RMT ii

- Leave-One-Out (LOO):

$$\hat{\Psi} = \frac{1}{T}\sum_{\tau=1}^{T} F_\tau F_\tau'$$

$$\hat{\Psi}_{T,t} = \frac{1}{T}\sum_{\tau\neq t}^{T} F_\tau F_\tau'$$

$$(zI+\hat{\Psi})^{-1}F_t = \left(zI+\hat{\Psi}_{T,t}\right)^{-1} F_t \frac{1}{1+T^{-1}F_t'\left(zI+\hat{\Psi}_{T,t}\right)^{-1}F_t} \tag{59}$$

- Define the Stieltjes Transform

$$\hat{m}(-z) \;=\; P^{-1}\,\mathrm{tr}((zI + \hat{\Psi})^{-1}) \tag{60}$$

and

$$\hat{Z}_*(z;c) \;=\; \frac{z}{1 - c + cz\hat{m}(-z)}\,. \tag{61}$$

and

$$\hat{\xi}(z;c) \;=\; -1 \;+\; \frac{1}{1 - c + cz\hat{m}(-z)}\,. \tag{62}$$

We have

$$1/Z_*(z;c) \;=\; \lim T^{-1}\,\mathrm{tr}((zI + \underbrace{FF'/T}_{T\times T})^{-1})\,. \tag{63}$$

RMT iv

- **Lemma**

$$T^{-1} F_t' \left(zI + \hat{\Psi}_{T,t} \right)^{-1} F_t \approx \hat{\xi}(z;c) \tag{64}$$

- Implicit Regularization

$$E[\hat{\lambda}(z)' F_{T+1}] \approx \frac{Z_*(z)}{z} E[\lambda_*(Z_*(z))' F_{T+1}], \tag{65}$$

where

$$Z_*(z) > z. \tag{66}$$

- In fact,

$$\begin{aligned} E_T[\hat{\lambda}(z)' F_{T+1}] &= \frac{Z_*(z)}{z} E_T[\lambda_*(Z_*(z))' F_{T+1}] \\ = \frac{Z_*(z)}{z} \lambda_*(Z_*(z))' E[F] &= \frac{Z_*(z)}{z} E[F]' (Z_*(z) I + E[FF'])^{-1} E[F] \end{aligned} \tag{67}$$

The RMT Master Theorem

Theorem

$$P^{-1} z \,\mathrm{tr}(A_P(zI + \underbrace{\hat{\Psi}}_{random})^{-1}) \; - P^{-1} Z_* \,\mathrm{tr}(A_P(Z_* I + \underbrace{\Psi}_{deterministic})^{-1}) \;\to\; 0 \tag{68}$$

almost surely.

Similarly, for any sequence of uniformly bounded vectors β, we have

$$z\beta'(zI + \underbrace{\hat{\Psi}}_{random})^{-1}\beta \; - Z_*\beta'(Z_* I + \underbrace{\Psi}_{deterministic})^{-1}\beta \;\to\; 0 \tag{69}$$

The Expected Return Calculation i

Proof.

Let $E[F]=\mu,\ E[FF']=\Psi$; everything is i.i.d. across t. Then,

$$
\begin{aligned}
E[\hat{\lambda}(z)'F_{T+1}] \;&=\; E[\hat{\lambda}(z)'\mu] \;=\; E[\frac{1}{T}\sum_{t=1}^{T}F_t'\left(zI+\frac{1}{T}\sum_{t=1}^{T}F_tF_t'\right)^{-1}]\mu \\
&\underbrace{=}_{symmetry}\; E[F_t'\left(zI+\frac{1}{T}\sum_{t=1}^{T}F_tF_t'\right)^{-1}]\mu \\
&=\; E\left[F_t'(zI+\Psi_{T,t})^{-1}\frac{1}{1+T^{-1}F_t'\left(zI+\hat{\Psi}_{T,t}\right)^{-1}F_t}\right]\mu \\
&\underbrace{\approx}_{F_t\ \ is\ \ independent}\; \mu'E[\left(zI+\hat{\Psi}_{T,t}\right)^{-1}]\mu(1+\xi(z;c))^{-1}
\end{aligned}
\tag{70}
$$

The Expected Return Calculation ii

Proof.

where

$$\hat{\Psi}_{T,t} \;=\; \frac{1}{T}\sum_{\tau=1}^{T} F_\tau F_\tau' \;-\; F_t F_t'$$

where we have used that

$$T^{-1} F_t' \left(zI + \hat{\Psi}_{T,t}\right)^{-1} F_t \;\approx\; \xi(z;c) \tag{71}$$

The claim follows now from the Master Theorem:

$$z\,\mu' \left(zI + \hat{\Psi}_{T,t}\right)^{-1} \mu \;\approx\; Z_*\left(Z_* I + \Psi\right)^{-1} \tag{72}$$

LLG

The Limits-to-Learning Gap (LLG)

$$\mathcal{L}(z;c) \;=\; \underbrace{\frac{d}{dz}Z_*(z;c) \;-\; 1}_{LLG} \;=\; \lim \frac{T^{-1}\,\mathrm{tr}((zI+FF'/T)^{-2})}{(T^{-1}\,\mathrm{tr}((zI+FF'/T)^{-1}))^2} - 1 \quad (73)$$

is always in $[0, T-1]$.

Theorem

$$\underbrace{\mu'\Sigma^{-1}\mu}_{infeasible\ \ SR} \;\;\geq\;\; (1+\mathcal{L}(z;c))\;\underbrace{SR^2_{OOS}(\hat{\lambda}(z))}_{feasible\ \ OOS\ \ SR} \quad (74)$$

Table of Contents

Empirical Analysis

- Analyze empirical analogues to theoretical comparative statics
- Study conventional setting with conventional data
  - Forecast target is monthly return of US stocks from CRSP 1963–2021
  - Conditioning info (X_t) is 130 stock characteristics from Jensen, Kelly, and Pedersen (2022)
- Out-of-sample performance metrics are:
  - SDF Sharpe ratio
  - Mean squared pricing errors (factors as test assets)

Empirical Analysis i

Random Fourier Features

- Empirical model: $M_{t+1} = 1 - \lambda' S_t' R_{t+1}$
- Need framework to smoothly transition from low to high complexity
- Adopt ML method known as "random Fourier features" (RFF)
  - Let $X_{i,t}$ be 130×1 predictors. RFF converts $X_{i,t}$ into

$$S_{\ell,i,t} = \sin(\gamma_\ell' X_{i,t}), \quad \gamma_\ell \sim iidN(0, \gamma I)$$

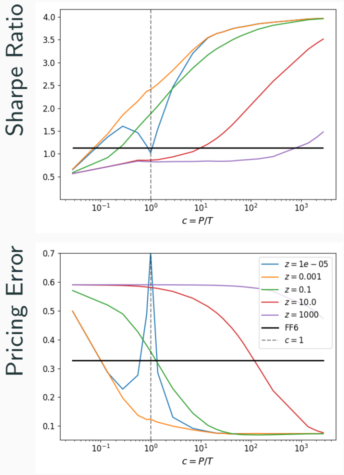
  - $S_{\ell,i,t}$: Random lin-combo of $X_{i,t}$ fed through non-linear activation
  - **we then rank the random features in the cross-section**
- For fixed inputs can create an arbitrarily large (or small) feature set
  - Low-dim model (say $P = 1$) draw a single random weight
  - High-dim model (say $P = 10{,}000$) draw many weights
- In fact, RFF is a two-layer neural network with fixed weights (γ) in the first layer and optimized weights (λ) in the second layer

Empirical Analysis

Training and Testing

- We estimate out-of-sample SDF with:
  - i. Thirty-year rolling training window ($T = 360$)
  - ii. Various shrinkage levels, $\log_{10}(z) = -12, ..., 3$
  - iii. Various complexity levels $P = 10^2, ..., 10^6$
- For each level of complexity $c = P/T$, we plot
  - i. Out-of-sample Sharpe ratio of the kernels and
  - ii. Pricing errors on 10^6 "complex" factors: $F_{t+1} = S_t' R_{t+1}$
- Also report Sharpe ratio and pricing errors of FF6 to benchmark our results

Out-of-sample SDF Performance

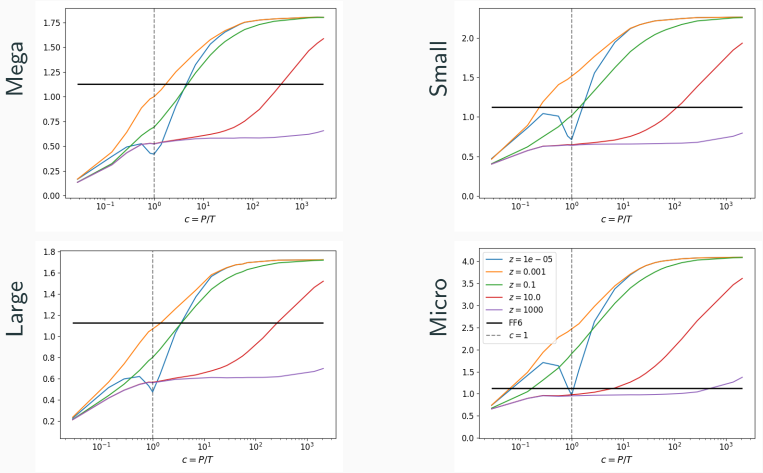


Main Empirical Result

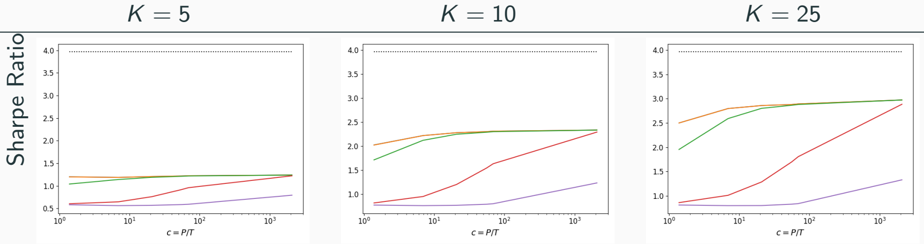
- OOS behavior of ML-based SDF closely matches theory
- High complexity models
  - Improve over simple models by a factor of 3 or more
  - Dominate popular benchmarks like FF6

SDF Performance in Restricted Samples: Sharpe Ratio

Market Capitalization Subsamples



What About “Shrinking” With PCA?



Beyond Own-Signal Portfolios i

All portfolio strategies we have used so far use own-signal weights:

$$\pi_{i,t} \;=\; w(S_{i,t}) \;=\; \sum_k \lambda_k f_k(X_{i,t})$$

where λ_k are estimated through Markowitz.

Beyond Own-Signal Portfolios ii

In Artificial Intelligence Pricing Models, we show how to build strategies that use other stocks' information. The insight is simple: Instead of

$$\pi_t = S_t \lambda,$$

we do

$$\pi_t = \underbrace{A_t \, S_t \lambda}_{one \ \ transformer \ \ block}$$

where

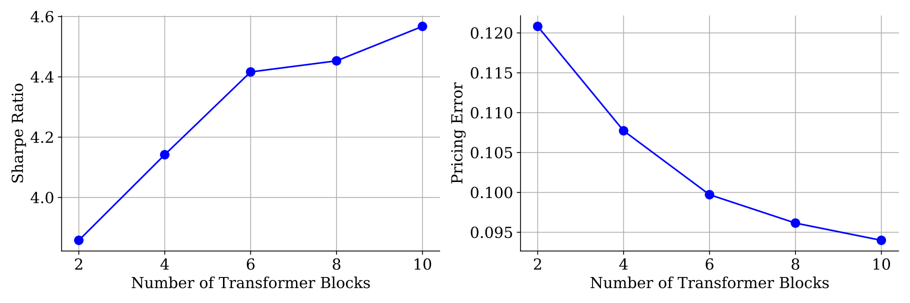
$$A_t = F(S_t M S_t')$$

is the attention matrix, and F is a non-linear transformation.

You can repeat this trick many times, making the attention **deeper**.

Beyond Own-Signal Portfolios iii

Figure: Virtue of complexity for K-block transformer portfolios.



Experiments with Managed Portfolios

Managed Portfolios Notebook