Empirical Asset Pricing: S04 Statistical Asset Pricing

Amedeo Andriollo

Finance Department (DRM)
Université Paris Dauphine - PSL.

Structure of the course

- Email: amedeo.andriollo@dauphine.psl.eu
- Grading follows a 30-70 rule: 70% final exam, 30% project/homework.
- Dense slides. References at the end.
- Office hours: feel "free" to (DO) come.
- If you spot any typos/mistake, please let me know: slides are updated regularly.
- This session partially builds on Giglio et al. (2022).

$\trianglerighteq$ Updates will come.

A Quick Warm Up: Previous Session

- After Fama French's series of papers, the gold standard for pricing shifted from CAPM to the FF3 factor model. Benchmark was: MKT, SMB, HML (Fama and French, 1993), then:
  - Jegadeesh and Titman (1993): FF3 + MOM ("Carhart four-factor model").
  - Pástor and Stambaugh (2003): FF3 + LIQ.
  - Fama and French (2015): FF3 + RMW (Robust-Minus-Weak: operating profitability) + CMA (Conservative-Minus-Aggressive, investment).
  - Hou et al. (2015): *"An empirical q-factor model consisting of the market factor, a size factor, an investment factor, and a profitability factor largely summarizes the cross section..."*
  - Stambaugh and Yuan (2017): *"A four-factor model with two "mispricing" factors, in addition to market and size factors..."*

↪ Hundreds of signals/factors competing in explaining the cross-section of expected returns:

- Factor zoo is still growing: *"A war-related factor model derived from textual analysis of media news reports explains the cross section of expected stock returns."* (Hirshleifer et al., 2025).

A Quick Warm Up: What are the PCs?

- Consider the high-dimensional data matrix $X \in \mathbb{R}^{T \times N}$ generated by the latent factor model:

$$X = FB^\top + E$$

where F is a $T \times r$ matrix of latent factors and B is a $N \times r$ matrix of factor loadings, and E is the matrix of idiosyncratic error, such that: $\mathbb{E}[E] = 0_{T \times N}$.

- The Principal Component (PC) estimator $(\widehat{F}, \widehat{B})$ are defined as:

$$(\widehat{F}, \widehat{B}) = \arg\min ||X - FB^\top||_F^2, \qquad T^{-1}F^\top F = I_r, \qquad B^\top B \quad \text{diagonal}$$

- This translates into an eigendecomposition problem:

$$T^{-1}XX^\top = T^{-1}QDQ^\top = T^{-1}Q_r D_r Q_r^\top + T^{-1}Q_{r+1:T} D_{r+1:T} Q_{r+1:T}^\top$$

where:

$$\widehat{F} = \sqrt{T}Q_r, \qquad \widehat{B} = T^{-1}X^\top \widehat{F}, \qquad EE^\top = T^{-1}Q_{r+1:T} D_{r+1:T} Q_{r+1:T}^\top$$

Latent Factors

- APT assumes returns follow a factor structure (or approximate).
- Implicit assumption: factors are observable (econ/finance)

vs. Connor and Korajczyk (1986): factors and their exposures are latent (statistical).

- *"a new class of beta pricing model statistics, based on the recent asymptotic principal components theory of Chamberlain and Rothschild (1982) [...] large numbers of securities can be used in estimating factor returns."*
- The returns follow: $r_t = \mathbb{E}[r_t] + \beta v_t + \epsilon_t$, where v_t are estimated via PCs: *"sample principal components of the excess returns on the n observed assets converge to the realized k market factors as n grows large."*
- A factor model divides the returns into k common sources of randomness and n asset-specific sources of randomness (Connor and Korajczyk, 1995).

↪ If betas and factors are latent: k rotational indeterminacy, $\tilde{\beta} = \beta H, \tilde{v_t} = H^{-1} v_t$, and replacing β and v_t with $\tilde{\beta}$ and $\tilde{v_t}$ yields an observationally equivalent return generating model.

- Extra. a third approach is Rosenberg (1974): observable exposures but latent factors.

XS Regression and Latent Factors (1)

- Two-Pass Cross-sectional Regression (2P): 1) TS regressions of returns on risk factors to get the betas, 2) XS regressions of returns on asset betas to estimate the factor-mimicking returns.
- Connor et al. (2015)'s fixed point result: iterating the 2P leads to the same factor estimates, independent of the choice of the initial factors.

↪ *"starting with the Fama and French (1993) 3-factor model, or a model using three macroeconomic series, or three nonsensical factors (like sunspot numbers) leads to the same final estimates (up to a k-dimensional rotation, where k is the number of factors)"*.

- In matrix form, presume the returns are: $R = BF + \phi$.
  - Suppose we start with some F^0, then the 2PR:
    1. TS reg: $\widehat{B}^0 = R(F^0)'(F^0(F^0)')^{-1}$
    2. XS reg: $F^1 = ((\widehat{B}^0)'\widehat{B}^0)^{-1}(\widehat{B}^0)'R$
  - Then iterating up until convergence: $\widehat{F} = \arg\min \text{trace}[(R - \widehat{B}\widehat{F})'(R - \widehat{B}\widehat{F})]$
- Connor and Korajczyk (1986) proposed to estimate the factors as principal components from the $T \times T$ covariance-variance matrix of returns: $R'R/n$.

XS Regression and Latent Factors (2)

- Connor et al. (2015)'s Theorem 1: (algebraic) equivalence of two estimators.
  - Lemma 1: *"Let H be any matrix of k eigenvectors of $R'R$ and let L be a non singular $k \times k$ orthogonal matrix ($L'L = LL' = I_k$). Then $\widehat{F} = LH$ is a solution to the equation"*:

$$(5) : \widehat{F} = (\widehat{B}'\widehat{B})^{-1}\widehat{B}'R, \qquad (6) : \widehat{B} = R\widehat{F}'(\widehat{F}\widehat{F}')^{-1}$$

    (**HW**: show it by plugging in the solution and by making use of: 1) $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$, 2) $HR'R = HH^{-1}\Lambda H = \Lambda H$).
  - *"Conversely, let $\widehat{F}$ denote any solution to equation (5) and (6). Then $\widehat{F} = LH$ where H is a matrix of k eigenvectors of $R'R$ and L is a nonsingular $k \times k$ matrix.*
- Since $\widehat{B}\widehat{F}$ is orthogonal to $\widehat{\phi}$, then: $\text{trace}[R'R] = \text{trace}[\widehat{F}'\widehat{B}'\widehat{B}\widehat{F}] + \text{trace}[\widehat{\phi}'\widehat{\phi}]$.
- Look at the first part: $\text{trace}[\widehat{F}'\widehat{B}'\widehat{B}\widehat{F}]$
  - Substitute (6): $\text{trace}[\widehat{F}'(R\widehat{F}'(\widehat{F}\widehat{F}')^{-1})'R\widehat{F}'(\widehat{F}\widehat{F}')^{-1}\widehat{F}]$.
  - F is the k eigenvectors of R: $\widehat{F}R'R = \Lambda\widehat{F}$.

  ↪ $\text{trace}[\widehat{F}'(\widehat{F}\widehat{F}')^{-1}\Lambda\widehat{F}\widehat{F}'(\widehat{F}\widehat{F}')^{-1}\widehat{F}] = \text{trace}[\widehat{F}'(\widehat{F}\widehat{F}')^{-1}\Lambda\widehat{F}]$.
  - By property of the trace: $\text{trace}[\widehat{F}'(\widehat{F}\widehat{F}')^{-1}\Lambda\widehat{F}] = \text{trace}[\widehat{F}\widehat{F}'(\widehat{F}\widehat{F}')^{-1}\Lambda] = \text{trace}[\Lambda]$
- *"since trace$(R'R)$ is fixed, minimizing trace$(\widehat{\phi}'\widehat{\phi})$ is equivalent to maximizing trace(Λ)"*.

Towards Giglio and Xiu (2021): Blocks #1-2

- *"Expand the set of test portfolios beyond size-B/M portfolios"* (Lewellen et al., 2010).
- (Managed) Portfolios are helping also moving from conditional beta-pricing to unconditional:
  - In general, if $r_t = \beta_{t-1}\gamma_{t-1} + \beta_{t-1}v_t + u_t$, consider some characteristics $\{c_t\}$:

  ↪ $\tilde{r}_t = (c_{t-1}^\top c_{t-1})^{-1} c_{t-1}^\top r_t = \beta\gamma_{t-1} + \beta v_t + (c_{t-1}^\top c_{t-1})^{-1} c_{t-1}^\top u_t$.

↪ Block #1: as LHS variables, many portfolios (rather than individual assets).

- CS regression can be understood as a mimicking/tracking portfolio:

$$\widehat{\gamma}_t = (\widehat{\beta}^\top \widehat{\beta})^{-1} \widehat{\beta}^\top r_t = (\widehat{\beta}^\top \widehat{\beta})^{-1} \widehat{\beta}^\top (\beta\gamma_{t-1} + \beta v_t + u_t), \qquad \mathbb{E}[\widehat{\gamma}_t] \approx \mathbb{E}[\gamma_{t-1}] = \gamma$$

- FM portfolios track v_t while earning (on average) γ. Thus, to hedge a given variable, we form the portfolio: $g_t = \xi + \eta v_t + \epsilon_t$.

↪ The hedging portfolio return is: (**HW**: using the SDF definition)

$$\widehat{\eta\gamma}_t = \widehat{\eta}(\widehat{\beta}^\top \widehat{\beta})^{-1} \widehat{\beta}^\top r_t$$

↪ Block #2: (weighted) FM portfolios constitute the hedging portfolio for a given variable.

Towards Giglio and Xiu (2021): Blocks #3-4

- Standard settings: the number of assets is fixed (compared to T), and the CS uses the $\widehat{\beta}$ as observed regressors (Shanken correction).

↪ Block #3: when $n, T \to \infty$, no need to correct the (asymp.) variance of RP.

- Another class of portfolios: maximally correlated mimicking portfolio (MP), using an asset return basis y_t (= *"portfolios with the maximum correlations with the [...] state variables"*. Breeden, 1979) with associated weights: $w_g = (\Sigma^y)^{-1}\text{Cov}(y_t, g_t)$. Which y-basis?
  1. All (excess) returns: $y_t = r_t$.

     $\implies$ $w_g = (\Sigma^r)^{-1}\beta\Sigma^v\eta'$, and so: $\gamma_g = w_g'\mathbb{E}[r_t] = \eta\Sigma^v\beta^\top(\Sigma^r)^{-1}\beta\gamma$
  2. The FM portfolios: $y_t = (\beta^\top\beta)^{-1}\beta^\top r_t$.

     $\implies$ $w_g = (\beta^\top\Sigma^r\beta)^{-1}\beta^\top\beta\Sigma^v\eta'$, and so: $\gamma_g = w_g'\mathbb{E}[y_t] = \eta\Sigma^v\beta^\top(\beta(\beta^\top\Sigma^r\beta)^{-1}\beta^\top)\beta\gamma$

↪ #4 block: Under appropriate assumptions, as $n \to \infty$, we have:

$$\beta(\beta^\top\Sigma^r\beta)^{-1}\beta^\top \approx (\Sigma^r)^{-1}, \quad \Sigma^v\beta^\top(\Sigma^r)^{-1}\beta \approx I_k \implies \gamma_g \approx \eta\lambda$$

meaning that there is a RP equivalence between:
a) all-returns max. corr. portfolio and b) FM hedging portfolio.

Giglio Xiu (2021)

- Recap:
  - #1 block: unconditional pricing with N portfolios.
  - #2 block: RP inference does not need corrections when $n, T \to \infty$.
  - #3 block: (weighted) FM portfolios constitute the hedging portfolio for a given variable.
  - #4 block: RP equivalence between all-returns max. corr and the FM hedging portfolios.
  - ...when the pricing factors f_t are given:
  - ↪ #5 block: The p pricing factors are the first p PCs: Statistical Factors.
- The Three-pass (3P) procedure:
  1. Extract the first p PCs from demeaned excess returns: $\widehat{v}_t$. TS regression: $\widehat{\beta}$.
  2. CS regression of mean returns of PCs onto $\widehat{\beta}$: $\widehat{\gamma}$.
  3. Mimicking g_t onto the PCs: $\widehat{\eta}$.
- The (estimated) risk premium of g_t with respect to the excess returns r_t is: $\widehat{\eta}\widehat{\gamma}$.

Post Giglio Xiu (2021): Statistical Factors

- Giglio Xiu (2021)'s *Copernican revolution*:
  - From a small to large panel of returns: to summarize info in a statistical way.
  - From observed economically motivated factors to unobservable statistical factors.

↪ New solutions, new issues: the pricing factors are statistically identified.

1. How the pricing information (i.e., "signal") is diluted in the large panel of returns.
2. How the "algorithm" extracts the information from the returns.
3. What is the economic interpretation of such factors?

- Puzzles: empirically, they find that SMB and HML are not priced.
- Can we conclude that are not pricing (any part of) the cross-section? Rather, we ask: "Are the PCs the latent pricing/SDF factors?"

↪ New gold rush: instead of new factors, we have new algorithms.

Empirical AP with Machine Learning (1)

- From empirical asset pricing to financial machine learning: Gu et al. (2020)
  - *"The fundamental goal of asset pricing is to understand the behavior of risk premiums.[...] risk premiums are notoriously difficult to measure [...] Our research highlights gains that can be achieved in prediction [...] This helps resolve the problem of* **risk premium measurement**, *which then facilitates more reliable investigation into economic mechanisms of asset pricing."*
  - *"Two main research agendas have monopolized modern empirical asset pricing research. The first seeks to describe [...] differences in expected returns across assets. The second focuses on dynamics [...].* **Measurement of an asset's risk premium is fundamentally a problem of prediction**—*the risk premium is the conditional expectation of a future realized excess return.* **Machine learning**, *whose methods are largely specialized for prediction tasks, is thus* **ideally suited** *to the problem of risk premium measurement."*
  - *"With an emphasis on variable selection and* **dimension reduction** *techniques [...] further complicating the problem is ambiguity about the* **functional forms** *through which the high-dimensional predictor sets enter into risk premiums".*

↪ From AP theoretical (economic) models to machines' (statistical) guidance.

Empirical AP with Machine Learning (2)

- Back to individual asset excess return, $r_{i,t}$: 30K stocks over 60y (1957-2016).
- Excess returns as functions of the (large number) predictors, $z_{i,t}$:

$$r_{i,t+1} = \mathbb{E}_t[r_{i,t+1}] + \epsilon_{i,t+1}, \qquad \mathbb{E}_t[r_{i,t+1}] = g(z_{i,t})$$

- 8 macro predictors ($dp, bm, ..$), 74 industry dummies (SIC codes), 94 stock-level (forward-looking) characteristics: # predictors = 920 ($= 94 \times (8+1) + 74$).
- Methodology:
  - Algorithms/Models: Linear (and Robust) regression, Penalized regression, Principal Component Regression, Partial Least Squares, Regression trees, Neural Nets.
  - Hedging against overfitting:
    - Sample (60y) spliting: 18y of training (1957 - 1974), 12y of validation (1975 - 1986), and the remaining (1987 - 2016) for out-of-sample testing.
    - *"We refit once every year as most of our signals are updated once per year. Each time we refit, we increase the training sample by 1 year. We maintain the same size of the validation sample, but roll it forward to include the most recent 12 months"*.
  - *"Hyperparameter tuning amounts to searching for a degree of model complexity that tends to produce reliable out-of-sample performance"* (Gu et al., 2020).

Empirical AP with Machine Learning (3)

- Out-of-sample performance:
  - Statistical: Out-of-sample R^2, Diebold Mariano test, Variable importance.
- *"These improved predictions are only measurements. The measurements do not tell us about economic mechanisms or equilibria. Machine learning methods on their own do not identify deep fundamental associations among asset prices and conditioning variables"*.
- However, *"We can assess the economic contribution [...] via its contribution to risk-adjusted portfolio return performance"* $\implies$ in terms of Sharpe Ratios (SRs):

  Intuition from APT, Max SR portfolio is the MVE portfolio (and is the min variance SDF).
  - Predictive R^2 to SR: active vs. buy-and-hold investor (Campbell and Thompson, 2008).
  - ML portfolios: portfolios built based on machine learning forecasts.

  $\hookrightarrow$ Translating in SR gains: horse race to the highest SR.

Empirical AP with Machine Learning (4)

- Improving our empirical understanding of asset returns:
  - Nonlinear over linear models: *"allowing for interactions among the baseline predictors is a crucial aspect [...] in the expected return function"*.
  - *"Shallow learning outperforms deeper learning"*: small signal-to-noise ratio.
  - Best model: Neural Nets. Most successful predictors are:
    - Price trends (mean): momentum, short-term reversal,..
    - Volatility (variance): idiosyncratic volatility, mkt beta,..
    - Liquidity (costs): dollar volume, bid-ask spread,..
- All that predicts is not gold:
  - DB test: NNs are not statistically outperforming all the others (see Table 3).
  - Some ML portfolio earns annualized SR of 2.45 vs. 0.98 (S&P500).
  - All NNs attach more importance on the long leg.
  - Neural Nets are non-linear and non-convex (stochastic gradient descent):
    ↪ Non-transparent tuning: mini-batches, early-stopping, batch normalization,..

↪ First attempt to understand AI models that empirically understand asset returns.

Thanks for your attention! Good luck with the preparation!

Bibliography I

Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of financial Economics*, 7(3):265–296.

Chamberlain, G. and Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets.

Connor, G. and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of financial economics*, 15(3):373–394.

Connor, G. and Korajczyk, R. A. (1995). The arbitrage pricing theory and multifactor models of asset returns. *Handbooks in operations research and management science*, 9:87–144.

Connor, G., Korajczyk, R. A., and Uhlaner, R. T. (2015). A synthesis of two factor estimation methods. *Journal of Financial and Quantitative Analysis*, 50(4):825–842.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14(1):337–368.

Giglio, S. and Xiu, D. (2021). Asset pricing with omitted factors. *Journal of Political Economy*, 129(7):1947–1990.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hirshleifer, D., Mai, D., and Pukthuanthong, K. (2025). War discourse and the cross section of expected stock returns. *The Journal of Finance*, 80(6):3589–3637.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of financial studies*, 28(3):650–705.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91.

Lewellen, J., Nagel, S., and Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial economics*, 96(2):175–194.

Pástor, L. and Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political economy*, 111(3):642–685.

Rosenberg, B. (1974). Extra-market components of covariance in security returns. *Journal of Financial and quantitative analysis*, 9(2):263–274.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *The review of financial studies*, 30(4):1270–1315.