Factor Models, Machine Learning, and Asset Pricing

by Stefano Giglio, Bryan Kelly and Dacheng Xiu

Presented by: Firmin Ayivodji

Université de Montréal (UdeM), CIREQ

Econometrics and Machine Learning Reading Group

December 6, 2021

Overview of the paper

- Recent survey of the literature at the intersection of factor models (FM), machine learning (ML) and asset pricing (AP).
- Probably a good starting point for ML and FM in AP.

List of topics covered

- Model specifications
  - Static Factor Models
  - Conditional Factor Models
- Methodologies
  - Measuring Expected Returns
  - Estimating Factors and Exposures
  - Estimating Risk Premia
  - Estimating the SDF
  - Model Specification Tests and Model Comparison
  - Alphas and Multiple Testing
- Asymptotic theory
  - Fixed N, Large T
  - Large N, Large T
  - Large N, Fixed T

Roadmap

1. Introduction
2. Static factor model
3. Conditional factor models
4. Estimating factors and exposures
5. Theory: “Strong” vs. “weak” factors
6. Estimating the SDF and its Loadings
7. Hot topics: ESG investing or Green factors
8. Conclusion

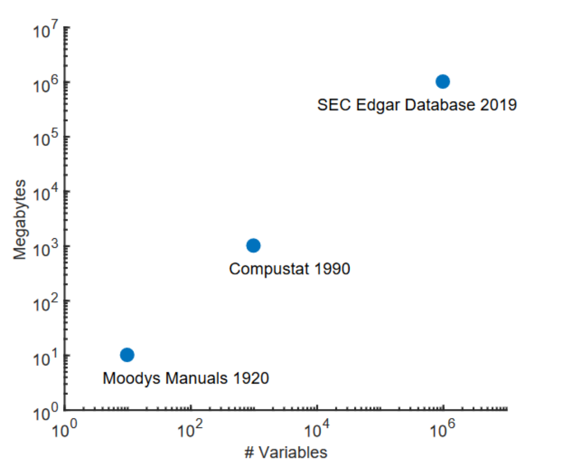
Introduction

Asset Pricing with Risk Factors

- Most important question in finance: Why are prices different for different assets?
- **Fundamental insight**: Arbitrage Pricing Theory: Prices of financial assets should be explained by systematic risk factors.
- **Problem**: "Chaos" in asset pricing factors: Over **300 potential asset pricing factors** published!
- Fundamental question: Which factors are really important in explaining expected returns?

- **Prediction central** to ML and also essential to asset pricing (AP):
  - Forecasting returns
  - Forecasting risk exposures
- ML methods are receiving a lot of attention in **asset pricing**:
  - Model selection in data-rich environments (big-data) for prediction
  - Nonlinear models (Neural Networks, Deep Learning, Trees, etc.)
- ML can be useful: to detect some hidden patterns beyond the documented asset pricing anomalies.
- Blossoming of ML in factor investing has it source at the confluence of three favorable developments: **data availability, computational capacity, and economic groundings.**
- Machine learning (ML) offers potentially useful toolbox for prediction.

Big Data in asset pricing: Example of corporate financial reports data



- Other sources: Textual data from social media, satellite imagery, or credit card.

- The baseline equation in **supervised learning**:

$$\mathbf{y} = g(\mathbf{X}) + \epsilon \tag{1}$$

is translated in financial terms as

$$\mathbf{r}_{t+1,n} = g\left(\mathbf{x}_{t,n}\right) + \epsilon_{t+1,n} \tag{2}$$

where $g\left(\mathbf{x}_{t,n}\right)$ can be viewed as the expected return for time $t+1$ computed at time t, that is, $\mathbb{E}_t\left[r_{t+1,n}\right]$.

- Building **accurate predictions** requires to pay attention to all terms in equation 2.
  - The first step is to gather data and to process it.
  - Second step: the choice of g.
  - Finally, the errors $\epsilon_{t+1,n}$, are often overlooked.

Choice of Factors

1. Statistical approach: use a large set of asset returns to build factors.
    - Factor analysis
    - Principal components
2. Economic approach: based on factors capturing economy-wide systematic risks.

SoFie 2021: 3 sessions with factor models...

- Specify **macroeconomic and financial market variables** which capture systematic risks in the economy.
- Specify **characteristics of firms** which could explain differential sensitivity to the systematic risks and form portfolios.
- For instance, we have:
    - Expected inflation, industrial production growth,
    - Fama-French factors (**size**: SMB, **value**: HML)
    - Factors based on profitability, investment, momentum, volatility,

Complex today...

- Harvey et al. (2015) catalogue 316 factors in some **300 articles** related to explaining the cross-section equity returns.

Table 1
Factor classification

Risk type		Description	Examples
Common (113)	**Financial** (46)	Proxy for aggregate financial market movement, including market portfolio returns, volatility, squared market returns, among others	Sharpe (1964): market returns; Kraus and Litzenberger (1976): squared market returns
	Macro (40)	Proxy for movement in macroeconomic fundamentals, including consumption, investment, inflation, among others	Breeden (1979): consumption growth; Cochrane (1991): investment returns
	Microstructure (11)	Proxy for aggregate movements in market microstructure or financial market frictions, including liquidity, transaction costs, among others	Pastor and Stambaugh (2003): market liquidity; Lo and Wang (2006): market trading volume
	Behavioral (3)	Proxy for aggregate movements in investor behavior, sentiment or behavior-driven systematic mispricing	Baker and Wurgler (2006): investor sentiment; Hirshleifer and Jiang (2010): market mispricing
	Accounting (8)	Proxy for aggregate movement in firm-level accounting variables, including payout yield, cash flow, among others	Fama and French (1992): size and book-to-market; Da and Warachka (2009): cash flow
	Other (5)	Proxy for aggregate movements that do not fall into the above categories, including momentum, investors' beliefs, among others	Carhart (1997): return momentum; Ozoguz (2009): investors' beliefs
Characteristics (202)	**Financial** (61)	Proxy for firm-level idiosyncratic financial risks, including volatility, extreme returns, among others	Ang et al. (2006): idiosyncratic volatility; Bali, Cakici, and Whitelaw (2011): extreme stock returns
	Microstructure (28)	Proxy for firm-level financial market frictions, including short sale restrictions, transaction costs, among others	Jarrow (1980): short sale restrictions; Mayshar (1981): transaction costs
	Behavioral (3)	Proxy for firm-level behavioral biases, including analyst dispersion, media coverage, among others	Diether, Malloy, and Scherbina (2002): analyst dispersion; Fang and Peress (2009): media coverage
	Accounting (87)	Proxy for firm-level accounting variables, including PE ratio, debt-to-equity ratio, among others	Basu (1977): PE ratio; Bhandari (1988): debt-to-equity ratio
	Other (24)	Proxy for firm-level variables that do not fall into the above categories, including political campaign contributions, ranking-related firm intangibles, among others	Cooper, Gulen, and Ovtchinnikov (2010): political campaign contributions; Edmans (2011): intangibles

The numbers in parentheses represent the number of factors identified. See Table 6 and http://faculty.fuqua.duke.edu/~charvey/Factor-List.xlsx.

- Observe excess returns of N assets over T time periods:
- Matrix notation

$$\underbrace{X}_{T\times N} = \underbrace{F}_{T\times K}\underbrace{\Lambda^{\top}}_{K\times N} + \underbrace{e}_{T\times N} \tag{3}$$

- N assets (large)
- T time-series observation (large)
- K systematic factors (fixed)
- F is are factors
- Λ are loadings
- e idiosyncratic factors
- F, Λ and e are unknown

Static factor model

Static factor model

A static factor model can be written as:

$$r_t = \mathrm{E}\left(r_t\right) + \beta v_t + u_t, \tag{4}$$

where r_t is an $N \times 1$ vector of excess returns, β is an $N \times K$ matrix of factor exposures, v_t is a $K \times 1$ vector of factor innovations, and u_t is an $N \times 1$ vector of idiosyncratic errors.

The expected return can be decomposed as:

$$\mathrm{E}\left(r_t\right) = \alpha + \beta\gamma,$$

where γ is a $K \times 1$ vector of risk premia and α is an $N \times 1$ vector of pricing errors.

$$f_t = \mu + v_t$$

We can rewrite equation (4) by:

$$r_t = \alpha + \beta f_t + u_t,$$

1. Factors f_t are **known and observable** (eg: industrial production growth)
2. Factor exposures are **observable** but factors are **latent**
3. All factors and their exposures are assumed **latent**

Conditional factor models

The conditional factor model can be specified as:

$$\tilde{r}_t = \alpha_{t-1} + \beta_{t-1}\gamma_{t-1} + \beta_{t-1}v_t + \tilde{u}_t,$$

where $\tilde{r}_t$ and $\tilde{u}_t$ are $M \times 1$ vectors of excess returns and idiosyncratic errors of individual stocks.

- Obviously, the **right-hand side contains too many degrees of freedom** and the model cannot be identified without additional restrictions.
- Rosenberg (1974) imposes that $\beta_{t-1} = b_{t-1}\beta$, where b_{t-1} is an $M \times N$ matrix of observable characteristics and β is an $N \times K$ vector of parameters.

Consequently, the model becomes

$$\tilde{r}_t = b_{t-1}\tilde{f}_t + \tilde{\varepsilon}_t$$

where $\tilde{f}_t := \beta\left(\gamma_{t-1} + v_t\right)$ is a new $N \times 1$ vector of latent factors, and $\tilde{\varepsilon}_t := \alpha_{t-1} + \tilde{u}_t$

Conditional factor models

- Kelly et al. (2019) suggest a new modeling approach know as instrumented principal components analysis (IPCA).

$$\tilde{r}_t = b_{t-1}\beta f_t + \tilde{\varepsilon}_t \tag{5}$$

  where β and $\{f_t\}$ have $N \times K$ and $K \times T$ unknown parameters, respectively.

- If we project b_{t-1} on both sides of **equation 5** at each t, we obtain

$$r_t := \left(b_{t-1}^{\mathrm{T}} b_{t-1}\right)^{-1} b_{t-1}^{\mathrm{T}} \tilde{r}_t = \beta f_t + u_t, \quad \text{where} \quad u_t := \left(b_{t-1}^{\mathrm{T}} b_{t-1}\right)^{-1} b_{t-1}^{\mathrm{T}} \tilde{\varepsilon}_t$$

- $\left(b_{t-1}^{\mathrm{T}} b_{t-1}\right)^{-1} b_{t-1}^{\mathrm{T}}$ can be interpreted as portfolio weights for characteristics-sorted portfolio returns.
- IPCA incorporates stock-level and portfolio-level asset pricing in a single specification.
- Gu et al. (2021) extend the IPCA to a nonlinear setting using a conditional autoencoder model, augmented with additional explanatory variables.

ML for asset pricing (Gu et al. 2020)

Determine the expected return of an asset:

$$\begin{aligned} r_{i,t+1} &= \mathrm{E}_t\left(r_{i,t+1}\right) + \epsilon_{i,t+1} \\ \mathrm{E}_t\left(r_{i,t+1}\right) &= g^{\star}\left(z_{i,t}\right) \end{aligned} \tag{6}$$

Functional form of g, the paper explores...

- Linear Models
  - **Ordinary Least Squares [OLS] (standard)**
  - OLS + Elastic Net (ENet) + Huber's Loss (H)
- Dimension Reduction
  - Partial Least Squares [PLS]
  - Principal Component Regression [PCR]
- Generalized Linear Models (GLM)
  - Series Regression + Group Lasso
- Regression Trees
  - Random Forest (RF)
  - Gradient Boosted Regression Trees [GBRT]
- Neural Networks (with layers K) [NNK]

ML for asset pricing (Gu et al. 2020): The Features

- 94 **characteristics**
- 8 **macroeconomic variables** $+1$ constant
- 74 **industry dummies** (first two digits of SIC codes)
- Treasury-bill rate to proxy for the risk-free rate
- The characteristic and macroeconomic variables are combined into $\mathbf{z}$ as,

  $$z_{i,t} = x_t \otimes c_{i,t}$$

  $c_{i,t}$ is a $P_c \times 1$ matrix of characteristics for each stock i,
  x_t is a $P_x \times 1$ vector of macroe conomic predictors.
  $\#$ of covariates: $94 \times (8+1) + 74 = 920$
- 30 000 stocks
- From 1957 to 2016 (almost 60 years).

If $g^\star(z_{i,t})$ is linear, this reduces to the beta-pricing representation

$$\mathrm{E}_t\left(r_{i,t+1}\right) = \beta_{i,t}'\gamma_t$$

With the betas being the risk exposure and the gammas the dynamic risk premiums.

$$\beta_{i,t} = \theta_1 c_{i,t} \quad \gamma_t = \theta_2 x_t$$

$$g^\star\left(z_{i,t}\right) = \mathrm{E}_t\left(r_{i,t+1}\right) = \beta_{i,t}'\gamma_t = c_{i,t}'\theta_1'\theta_2 x_t = \left(x_t \otimes c_{i,t}\right)' \operatorname{vec}\left(\theta_1'\theta_2\right) =: z_{i,t}'\theta$$

ML for asset pricing (Gu et al. 2020) - training objective

The training objective:

$$\mathcal{L}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(r_{i,t+1} - g\left(z_{i,t};\theta\right)\right)^2$$

Heavy tails can cause issues with the least square objective, hence also consider the Huber loss

$$\mathcal{L}_H(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H\left(r_{i,t+1} - g\left(z_{i,t};\theta\right),\xi\right)$$

where

$$H(x;\xi) = \begin{cases} x^2, & \text{if} \quad |x| \leq \xi; \\ 2\xi|x| - \xi^2, & \text{if} \quad |x| > \xi. \end{cases}$$

The hyper-parameter ξ is determined by the model performance in a validation sample.

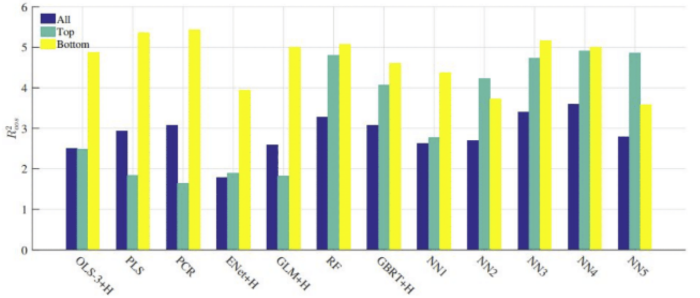
Performance Evaluation Out of sample R-squared

$$R^2_{\mathrm{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}\left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}},$$

$\mathcal{T}_3$ denotes the test set

Table 2: Annual Out-of-sample Stock-level Prediction Performance (Percentage R^2_{oos})

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	-34.86	2.50	2.93	3.08	1.78	2.60	3.28	3.09	2.64	2.70	3.40	3.60	2.79
Top	-54.86	2.48	1.84	1.64	1.90	1.82	4.80	4.07	2.77	4.24	4.73	4.91	4.86
Bottom	-19.22	4.88	5.36	5.44	3.94	5.00	5.08	4.61	4.37	3.72	5.17	5.01	3.58



Estimating factors and exposures

1. If the factors are know, we can estimate factor exposures via **asset-by-asset TSR** as:
$$\text{TSR}: \quad \widehat{\beta} = \bar{R}\bar{F}^\top \left(\bar{F}\bar{F}^\top\right)^{-1}$$
2. If the factors are latent but exposures are observable, we can estimate factors by CSR at each time point as:
$$\text{CSR}: \quad \widehat{F} = \left(\beta^\top \beta\right)^{-1} \beta^\top R$$
3. In case, the factors loadings can be proxied by firm characteristics, we obtain for each t:
$$\widehat{\tilde{f}}_t = \left(b_{t-1}^\top b_{t-1}\right)^{-1} b_{t-1}^\top \tilde{r}_t$$
4. If factors and exposures are latent, we can use PCA or its variants to estimate them.

Estimating factors and exposures

- A limitation of PCA is that it only applies to static factor models.
- It also lacks the flexibility to incorporate other data beyond returns.
- To address both issues, Kelly et al. (2019) estimate the conditional factor model, by solving the optimization problem:

$$\min_{\beta,\{f_t\}} \sum_{t=2}^{T} \left\|\tilde{r}_t - b_{t-1}\beta f_t\right\|^2 \tag{7}$$

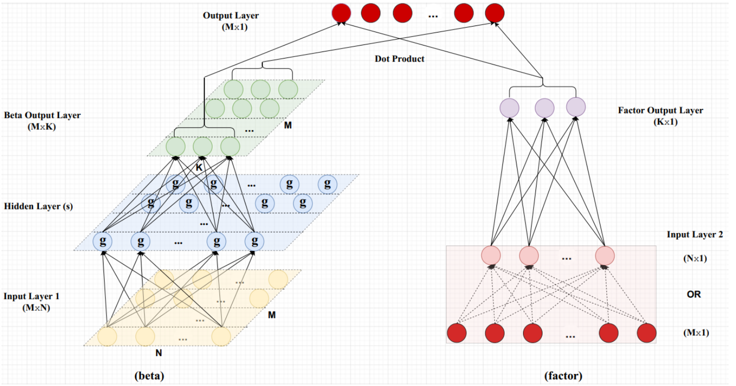
- From the first order condition, we get that for $t = 1, 2, \ldots, T-1$:

$$\begin{aligned} \widehat{f}_t &= \left(\widehat{\beta}^\top b_{t-1}^\top b_{t-1}\widehat{\beta}\right)^{-1} \widehat{\beta}^\top b_{t-1}^\top \tilde{r}_t \\ \mathbf{vec}\left(\widehat{\beta}^\top\right) &= \left(\sum_{t=2}^{T} b_{t-1}^\top b_{t-1} \otimes \widehat{f}_t \widehat{f}_t^\top\right)^{-1} \left(\sum_{t=2}^{T} \left(b_{t-1} \otimes \widehat{f}_t^\top\right)^\top \tilde{r}_t\right) \end{aligned} \tag{8}$$

- Given **conditional betas**, factors are estimated from cross section regressions of returns on betas.
- The authors recommend an **iterative algorithm** to update $\widehat{\beta}$ and $\widehat{f}_t$ until convergence.

Autoencoder learning - architecture

- IPCA assumes the factor exposures are a linear function of the covariates.
- Existing literature suggests their relationship might be **nonlinear**.
- **Autoencoder** allows betas to depend on stock characteristics



- On the left side of the network, factor loadings are a nonlinear function of covariates (e.g., firm characteristics)

$$b_{i,t-1}^{(0)} = b_{i,t-1},$$
$$b_{i,t-1}^{(l)} = g\left(b^{(l-1)} + W^{(l-1)} b_{i,t-1}^{(l-1)}\right), \quad l = 1, \ldots, L_\beta,$$
$$\beta_{i,t-1} = b^{\left(L_\beta\right)} + W^{\left(L_\beta\right)} b_{i,t-1}^{\left(L_\beta\right)}.$$

- On the right side of the network models factors as portfolios of individual stock returns.

$$r_t^{(0)} = \left(b_{t-1}^\top b_{t-1}\right)^{-1} b_{t-1}^\top r_t,$$
$$r_t^{(l)} = \widetilde{g}\left(\widetilde{b}^{(l-1)} + \widetilde{W}^{(l-1)} r_t^{(l-1)}\right), \quad l = 1, \ldots, L_f,$$
$$f_t = \widetilde{b}^{\left(L_f\right)} + \widetilde{W}^{\left(L_f\right)} r_t^{\left(L_f\right)}.$$

- The **structure of the neural network** is summarized below.

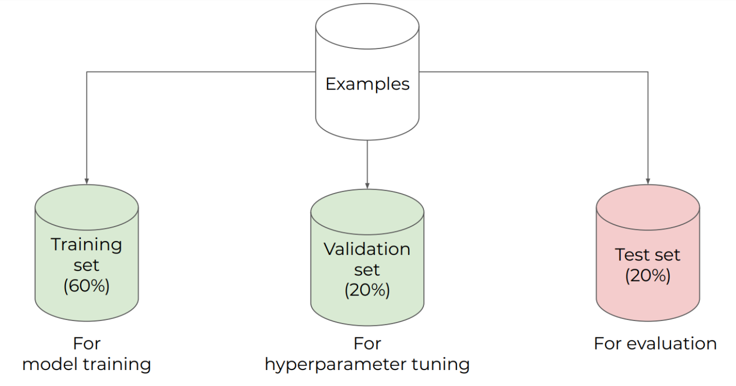
$$\left.\begin{array}{rcl} \text{returns}\,(\mathbf{r}_t) & \xrightarrow{NN_1} & \text{factors}\;\; (\mathbf{f}_t = NN_1\,(\mathbf{r}_t)) \\ \text{characteristics}\,(\mathbf{x}_{t-1}) & \xrightarrow{NN_2} & \text{loadings}\;\; (\boldsymbol{\beta}_{t-1} = NN_2\,(\mathbf{x}_{t-1})) \end{array}\right\} \rightarrow \text{returns}\,(r_t)$$

- Gu et al. (2021) define the **estimation objective** to be

$$\mathcal{L}(\theta;\cdot) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\tilde{r}_{i,t} - \beta'_{i,t-1}f_t\right\|^2 + \phi(\theta;\cdot)$$

Where θ summarizes the weight parameters in the loading and factor networks , $\phi(\theta)$ is a penalty function, such as lasso (or l_1) penalization, which takes the form $\phi(\theta;\lambda) = \lambda\sum_j |\theta_j|$.

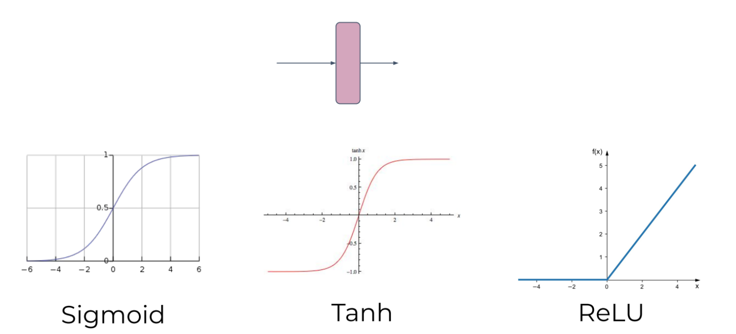
How to choose the best hyperparameters?



- Cross-validation
- Early stopping

Most commonly used activation functions:

- Sigmoid: $\sigma(z) = \frac{1}{1+\exp(-z)}$
- Tanh: $\mathbf{tanh}(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$
- ReLU (Rectified Linear Unit): $\mathbf{ReLU}(z) = \max(0, z)$



Theory: “Strong” vs. “weak” factors

- Key property: Factor strength
  1. **Strong factors**: High variance ($\Leftrightarrow$ affect many assets)
  2. **Weak factors**: Low variance ($\Leftrightarrow$ affect some assets)
- Examples:
  - Stronger factor: Market
  - Weaker factors: Long/short portfolios
- Standard PCA methods assume that all factors are **strong**
- But, PCA can fail to identify weak factors with large risk p
- **Consequence**: PCA captures TS variance but not XS
- Alternative: RP-PCA (Lettau-Pelger, 2020ab)

Key idea

- Apply PCA to a covariance matrix with overweighted mean

$$\frac{1}{T}RR^\top + \gamma \bar{r}.\bar{r}^\top \quad \gamma = \text{ risk-premium weight}$$

- Eigenvectors of largest eigenvalues estimate loadings $\hat{\Lambda}$.
- $\hat{F}$ estimator for factors: $\hat{F} = \frac{1}{N} R \hat{\Lambda} = R \left(\hat{\Lambda}^\top \hat{\Lambda}\right)^{-1} \hat{\Lambda}^\top$.

Estimating the SDF and its Loadings

In the setup of 4, an SDF can be written as:

$$m_t = 1 - b^\top v_t$$

where $b = \Sigma_v^{-1}\gamma$ and Σ_v is the covariance matrix of factor innovations.

- The SDF is **central** to the field of asset pricing
- In the absence of arbitrage, covariances with the SDF unilaterally explain cross-sectional difference in**expected returns**.

1. GMM estimator:
$$\min_{b,\mu} \widehat{g}_T(b,\mu)^\top \widehat{W} \widehat{g}_T(b,\mu)$$
where the sample moments are given by
$$\widehat{g}_T(b,\mu) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T r_t\left(1 - b^\top (f_t - \mu)\right) \\ \frac{1}{T}\sum_{t=1}^T f_t - \mu \end{pmatrix}_{(N+K)\times 1}.$$
2. Penalized regressions: Kozak et al. (2020) consider an SDF represented in terms of a set of tradable test asset returns:
$$m_t = 1 - \underline{b}^\top \left(r_t - \mathrm{E}\left(r_t\right)\right)$$
where $\underline{b}$ satisfies $\mathrm{E}\left(r_t\right) = \boldsymbol{\Sigma}\underline{b}$, and $\boldsymbol{\Sigma}$ is the covariance matrix of r_t.
3. To estimate the SDF, they suggest solving an optimization problem, which amounts to a regression of $\bar{r}$ onto $\widehat{\boldsymbol{\Sigma}}$:
$$\underline{b} = \arg\min_{\underline{b}} \left\{(\bar{r} - \widehat{\Sigma}\underline{b})^\top \widehat{\Sigma}^{-1}(\bar{r} - \widehat{\Sigma}\underline{b}) + p_\lambda(\underline{b})\right\}, \qquad (9)$$
with which the estimated **pricing kernel** is given by
$$\widehat{m}_t = 1 - \widehat{b}^\top \left(r_t - \bar{r}\right).$$

- The objective function in 9 appears to require the inverse of the sample covariance matrix $\widehat{\Sigma}^{-1}$, which is not well-defined when $N > T$.
- Instead, it is equivalent to optimizing a different form of 9 :

$$\widehat{b} = \arg\min_{\underline{b}} \left\{ \underline{b}^\top \widehat{\Sigma} \underline{b} - 2\underline{b}^\top \bar{r} + \underline{b}^\top \widehat{\Sigma} \underline{b} + p_\lambda(\underline{b}) \right\},$$

which avoids calculating $\widehat{\Sigma}^{-1}$.
- $p_\lambda(\underline{b})$ is a **penalty term** (such as lasso, ridge, elastic-net, etc.)
- Other papers consider:
    - Deep Learning SDF (Cong et al., 2021, Chen et al., 2021).
    - Double Machine Learning (in the spirit of Chernozhukov et al., 2018).

Hot topics: ESG investing or Green factors

Hot topics: ESG investing or Green factors

- More and more, researchers study the financial impact of climate change
- favorable: ESG investing works (Kempf and Osthoff (2007), Cheema-Fox et al. (2020)), can work (Nagy, Kassam, and Lee (2016), Alessandrini and Jondeau (2020)).
- unfavorable: Ethical investing is not profitable according to Adler and Kritzman (2008) and Blitz and Swinkels (2020).
- mixed: ESG investing may be beneficial globally but not locally (Chakrabarti and Sen (2020)).
- On top of these contradicting results, several articles point towards complexities in the measurement of ESG.
- I end this short section by noting that of course ESG criteria can directly be integrated into ML model, as is for instance done in Franco et al. (2020).

Conclusion

Conclusion

- Machine learning methods can yield great improvement on forecasts as compared to the status quo traditional models (i.e. OLS).
- This is due to their ability to handle a large number of predictors without overfitting and being able to exploit non-linearities that may exist.
- There is a lot more topics in the paper (but not much details).
- Machine learning is **no magic wand** .