The Cross-Section of Stock Returns: Rank-Based, Linear, and Non-Linear Models

Matéo Molinaro

M2 272 Economic & Financial Engineering, Dauphine-PSL University

08 July 2025

Summary

Introduction and motivations (1)

- Initial Research Project at Candriam: Mid-Term Abnormal Trends Detection
  - price-based
  - non-price based
  - alternative data-based
- Conclusions on price-based (Country and Sector Indices) approach:
  - Price-based leads to poor results (except for momentum and Hamed-Rao Mann-Kendall test)
  - Length of the outperformance ("abnormality") is short (1-3 months)
  - Market timing is more difficult that stock selection
- Project Update: price-based market timing (time-series) –> fundamental data-based stock selection (cross-section)

Introduction and motivations (2)

- Asset Pricing: Understanding the cross-section of expected asset (equity) returns
- Foundation: Capital Asset Pricing Model (**CAPM**) of **Sharpe (1964)** and **Lintner (1965)**
- Academic to Industry Transfer 1: **first index fund** (Bogle, Vanguard 500 fund, 1976)
- Questioning (**CAPM, EMH**):
  - Market risk is not the only driver of excess returns –> Arbitrage Pricing Theory **APT (Ross, 1976)**.
  - Discovering of anomalies –> **Value** (Basu, 1977) and **Size** (Banz, 1981)
- Development:
  - multi-factor models: **Fama and French (1993) three-factor model**, and **five-factor extension (Fama and French, 2015)**, incorporating **size, value, profitability**, and **investment** factors.
- Racing: **"factor zoo" Cochrane** (2011).

Introduction and motivations (3)

- Academic to Industry Transfer 2: **Smart Beta** (Research Affiliates, WisdomTree) circa 2011
- Recent Advances:
  - **ML –> Empirical Asset Pricing via Machine Learning** (Gu, Kelly and Xiu, 2019)
  - **DL –> Autoencoder Asset Pricing Models** (Gu, Kelly and Xiu, 2021)
  - ML and DL –> model **complex** and **non-linear relationships** between firm characteristics and expected returns.
- Aim of our study: Empirically assess, in the US (Russell 1000), which type of models lead to the best economic advantage (SR) in constructing factor portfolios
  - **Rank-based** (univariate sorts, sequential filtering)
  - **Linear-Models** (LASSO, Ridge, EN)
  - **Non-Linear Models** (RF, NN)

Data and Feature Engineering (1)

- Fundamental data comes from Bloomberg (BQL and API) for the Russell 1000 (RIY) Index.
- It consists of two sets of fundamental variables: i) "raw fields" that can be directlty retrieved from Bloomberg or with little transformations (change, ratios, volatility) and ii) aggregated scores (mean of ranks).
- Details on the computations for the transformed variables are left in the appendices.
- On the next slides, you can find the list of all the variables.

Data and Feature Engineering (2)

Raw Fields	Aggregated Scores
FREE_CASH_FLOW_MARGIN	QUALITY_SCORE
ROC_WACC_RATIO	GROWTH_SCORE
RETURN_COM_EQY	MOM_SCORE
NET_DEBT_TO_EBITDA	LOW_VOL_SCORE
CUR_MKT_CAP	SENTIMENT_SCORE
PX_TO_CASH_FLOW	VALUE_SCORE
PE_RATIO	TOTAL_SCORE
PX_TO_BOOK_RATIO	TOTAL_SCORE_MAHALANOBIS
HEADLINE_EV_TO_SALES	
HEADLINE_DVD_YIELD	
CF_ISSUE_COM_STOCK	
FCFE_PAYOUT	
OCF_OVER_TOT_ASSETS	
SALES_PER_SHARE	
FREE_CASH_FLOW_EQUITY_PER_SHARE	
FCFE_VARIABILITY	
ROA_CHANGE	
CUR_RATIO_CHANGE	
ASSET_TURNOVER_CHANGE	
LT_DEBT_TO_TOT_ASSET_CHANGE	
GROSS_MARGIN_CHANGE	
SALES_PER_SHARE_CHANGE	
FREE_CASH_FLOW_EQUITY_PER_SHARE_CH	
TARGET_PRICE_CHANGE	
EPS_CHANGE	
CONSENSUS_UPSIDE	

Table: Raw financial ratios and aggregated scores.

Data and Feature Engineering (3)

Raw Fields	Aggregated Scores
MOMENTUM_3M	
MOMENTUM_6M	
MOMENTUM_12M	
BETA	
VOLATILITY_1Y	
BUYBACK_YIELD	
EARNINGS_REVISION	
EARNINGS_SURPRISE	
NORMALIZED_ROA	
IDIOSYNCRATIC_VOLATILITY_1Y	
ALPHA	
NEW_EQ_ISSUANCE_LTM	
PFS	
IDIO_SHARPE	

Table: Raw financial ratios and aggregated scores.

- We then, computed cross-sectional z-scores and winsorized them at the 2nd and 98th percentiles.
- We also put a negative sign on the following ratios because economically higher means "bad" for the companies:

Data and Feature Engineering (4)

Table: List of Z-Scored Winsorized Financial Variables with negative sign

Variable Name
NET_DEBT_TO_EBITDA_ZSWINSO
CUR_MKT_CAP_ZSWINSO
PX_TO_CASH_FLOW_ZSWINSO
PE_RATIO_ZSWINSO
PX_TO_BOOK_RATIO_ZSWINSO
HEADLINE_EV_TO_SALES_ZSWINSO
CF_ISSUE_COM_STOCK_ZSWINSO
FCFE_VARIABILITY_ZSWINSO
LT_DEBT_TO_TOT_ASSET_CHANGE_ZSWINSO
BETA_ZSWINSO
VOLATILITY_1Y_ZSWINSO
IDIOSYNCRATIC_VOLATILITY_1Y_ZSWINSO

Ranked-Based Strategies - Univariate Sorts

- The Ranked-Based Long-Only (LO) strategies consist in going long the top n% of the ranked stocks according to a variable. This gives us the signals.
- The portfolios are constructed with an Equally-Weighting (EW) Scheme. Other Weighting Scheme (WS) may be of interest but this is beyond the scope of this study.
- We also performed "grid search" for the strategies' parameters such as the n%, AllIndustries or BestInIndustries and rebalancing frequencies but to avoid overhelming the presentation we'll just compare one strategy setting across all models: top 10% (deciles), selection of the stocks across the entire universe (AllIndustries) and rebalanced monthly assuming no TCs.

Rank-Based Strategies - Univariate Sorts

- We run the backtests for univariate sorts on the 46 (we dropped NEW_EQ_ISSUANCE_LTM as this is a binary variable and BUYBACK_YIELD due to poor coverage) fundamental ratios. Start date is 31-01-2015, end date 31-05-2025.
- All results (performance metrics) are available in this **GitHub repository**. To keep the presentation concise, we focus on the top 10 portfolios ranked by two metrics: the Information Ratio (IR) and the Excess Sharpe Ratio (eSR).

Rank-Based Strategies - Univariate Sorts - Results IR

top10	ir_EW
SALES_PER_SHARE_CHANGE_ZSWINSO_D8_AllIndustries_monthly	1.21
FREE_CASH_FLOW_MARGIN_ZSWINSO_D2_AllIndustries_monthly	1.20
FREE_CASH_FLOW_MARGIN_ZSWINSO_D1_AllIndustries_monthly	1.16
GROWTH_SCORE_ZSWINSO_D7_AllIndustries_monthly	1.10
FREE_CASH_FLOW_MARGIN_ZSWINSO_P5_AllIndustries_monthly	1.04
FCFE_PAYOUT_ZSWINSO_D5_AllIndustries_monthly	0.94
SALES_PER_SHARE_CHANGE_ZSWINSO_D6_AllIndustries_monthly	0.90
QUALITY_SCORE_ZSWINSO_P5_AllIndustries_monthly	0.87
FREE_CASH_FLOW_EQUITY_PER_SHARE_CHANGE_ZSWINSO_D7_AllIndustries_monthly	0.86
QUALITY_SCORE_ZSWINSO_D1_AllIndustries_monthly	0.85

Figure: Top 10 factors by IR (against Russell 1000 EW)

- The top 10 performing "factors" are mainly ratios related to profitability/quality (and growth in a lesser extent).

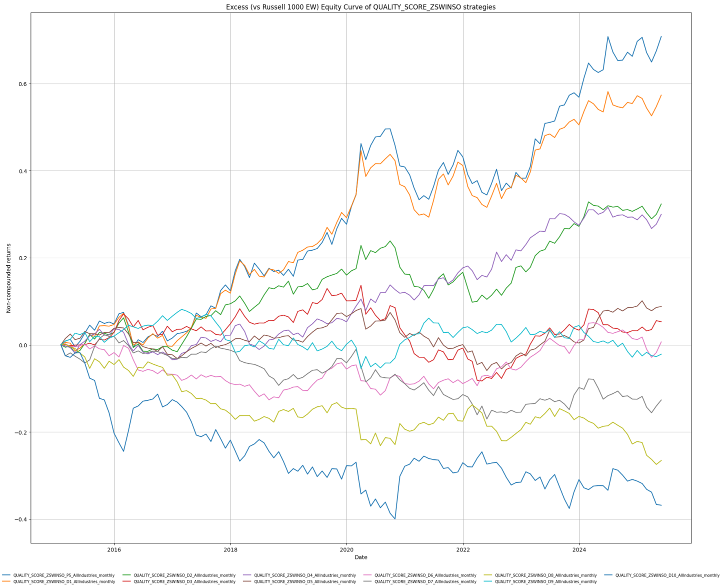
Rank-Based Strategies - Univariate Sorts - Results eSR

	top_excess_ann_sr_EW
QUALITY_SCORE_ZSWINSO_P5_AllIndustries_monthly	0.49
QUALITY_SCORE_ZSWINSO_D1_AllIndustries_monthly	0.45
FREE_CASH_FLOW_MARGIN_ZSWINSO_D2_AllIndustries_monthly	0.42
FREE_CASH_FLOW_MARGIN_ZSWINSO_D1_AllIndustries_monthly	0.41
ROC_WACC_RATIO_ZSWINSO_P5_AllIndustries_monthly	0.38
IDIO_SHARPE_ZSWINSO_D2_AllIndustries_monthly	0.37
TOTAL_SCORE_ZSWINSO_D9_AllIndustries_monthly	0.35
FREE_CASH_FLOW_MARGIN_ZSWINSO_P5_AllIndustries_monthly	0.35
OCF_OVER_TOT_ASSETS_ZSWINSO_D1_AllIndustries_monthly	0.34
ROC_WACC_RATIO_ZSWINSO_D1_AllIndustries_monthly	0.34

Figure: Top 10 factors by eSR (against Russell 1000 EW)

- The top performing factor is the QUALITY_SCORE, followed by profitability metrics.

Rank-Based Strategies - Univariate Sorts - excess EC



Rank-Based Strategies - Sequential Filtering

- In the previous subsection, we presented results from univariate sort portfolios.
- In this subsection, we present results for sequential filtering. We refer to sequential filtering as a process where we first rank the top n% stocks, then we further rank these top n% stocks on another ratio and so on. By doing that, we avoid averaging ratios which can result in a stock being selected if it has a strong ratio that can offset another low ratio. In the sequential filtering, the stocks being selected must pass all the filters.

Rank-Based Strategies - Sequential Filtering

- Sequential Filtering (SF) slightly allows to outperform univariate sorts: 0.53 (1.03) e(SR) vs 0.49 (0.97) e(SR).

Linear Models Strategies - Setup (1)

$$r_{i,t+1} - rf_{t+1} = \mathbb{E}_t\left(r_{i,t+1} - rf_{t+1}\right) + \varepsilon_{i,t+1} \tag{1}$$

$$r_{i,t+1} - rf_{t+1} = g^*(z_{i,t}) + \varepsilon_{i,t+1} \tag{2}$$

$$r_{i,t+1} - rf_{t+1} = \alpha + \sum_{k=1}^{K} \beta_k \, z_{i,t}^{(k)} + \varepsilon_{i,t+1} \tag{3}$$

- Functional form: equations (1) and (2) describe an asset's excess return without assuming any specific functional form.
- Linearity: equation (3) is specific to the linear models category.
- Objective: find parameters that maximize the out-of-sample (OOS) explanatory power.
- Portfolio construction: we take the top n% (or deciles) of the predicted returns and equally-weight them.

Linear Models Strategies - Setup (2)

- Sample Splitting: Expanding Training, Validation and Test
- For 10Y of data we used 3Y of training initially (increasing monthly), 2Y of validation and 1 month of test
- Models are re-estimated each month but hyperparameters are only tuned once, at the beginning
- Remark: as of the date of writting the report, we have data available from 2014-12-31. Hence, with the training-validation-testing period, we have first strategy return on 2020-02-29. This makes the model-free and model approaches not comparable. We still report the results.

Non-Linear Models Strategies - Setup (1)

$$r_{i,t+1} - rf_{t+1} = \mathbb{E}_t \left(r_{i,t+1} - rf_{t+1} \right) + \varepsilon_{i,t+1} \tag{4}$$

$$r_{i,t+1} - rf_{t+1} = g^*(z_{i,t}) + \varepsilon_{i,t+1} \tag{5}$$

- Functional form: equation (5) cannot be more detailed as in ML we do not know the explicit form of g*.
- Linearity: hence, equation (5) allows for non-linearity in the functional learned.
- Objective: find parameters that maximize the out-of-sample (OOS) explanatory power.
- Sample splitting: the same as the one for linear models.
- Portfolio construction: we take the top n% (or deciles) of the predicted returns and equally-weight them.

Linear and NL Models Strategies - Results IR

top10	ir_EW
NN_D10_AllIndustries_monthly	0.46
Lasso_D2_AllIndustries_monthly	0.40
Lasso_P5_AllIndustries_monthly	0.40
RF_D2_AllIndustries_monthly	0.39
RF_P5_AllIndustries_monthly	0.39
RF_D7_AllIndustries_monthly	0.37
ElasticNet_D6_AllIndustries_monthly	0.37
ElasticNet_D10_AllIndustries_monthly	0.34
NN_P5_AllIndustries_monthly	0.30
Lasso_D1_AllIndustries_monthly	0.30

Figure: Top 10 models strategies by IR (against Russell 1000 EW)

- The top performing strategy is the decile 10 (smallest predicted excess returns) of a Neural Network. It contrasts with our desire to play the largest predicted returns: we expect to see D1 or P5 first. With this expectation, LASSO becomes the top strategy, followed by RF.

Linear and NL Models Strategies - Results eSR

	excess_ann_sr_EW
Lasso_P5_AllIndustries_monthly	0.20
RF_P5_AllIndustries_monthly	0.18
Lasso_D2_AllIndustries_monthly	0.18
RF_D2_AllIndustries_monthly	0.17
ElasticNet_D6_AllIndustries_monthly	0.16
Lasso_D1_AllIndustries_monthly	0.14
NN_P5_AllIndustries_monthly	0.13
ElasticNet_P5_AllIndustries_monthly	0.12
ElasticNet_D3_AllIndustries_monthly	0.12
Ridge_D2_AllIndustries_monthly	0.12

Figure: Top 10 models strategies by eSR (against Russell 1000 EW)

- The top performing strategy is based on LASSO, closely followed by Random Forest.

Linear and NL Models Strategies - Results overview

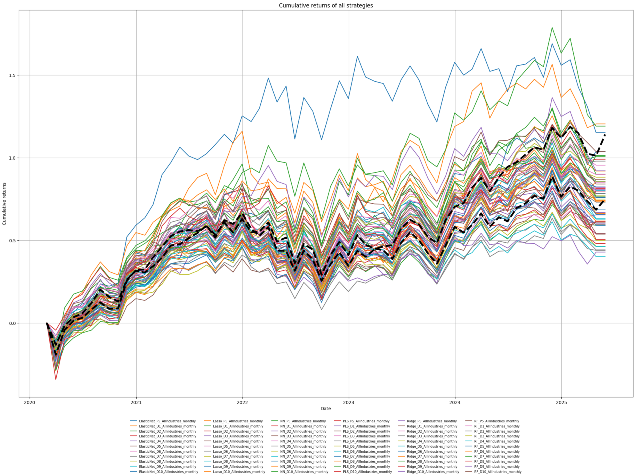


Figure: All model-strategies cumulative returns (dashed black lines are Russell 1000 EW and VW)

Conclusion

- In this project, we studied 3 approaches for constructing portfolios:
  - Model-free (univariate sorts, sequential filtering)
  - Linear Models (LASSO, Ridge, ElasticNet)
  - Non-Linear Models (RF, NN)
- From the results we have at this moment, we can say that the model-free approach offer simplicity, interpretability and good economic performance.
- Cautious: the studied period for the model approach is short and not the same as the model-free approach. Moreover, we tested expanding window with one initial hyperparameter tuning. Rolling window and/or hyperparameter re-tuning could lead to differend results.

Conclusion

- Additional results: we tried to explore some classification models based on some percentiles but we do not find better results. We also remarked that the models have large negative R2 but that it can lead to positive strategy performance: indeed, as we sort the predicted excess returns, what's important is the spearman (rank) correlation between our predictions and target variables. We looked at that and LASSO was at 1.87% (which is still low but explain the positive results).
- Discussion: we also tried to custom the loss of LASSO by replacing the MSE part by the pairwise logistic loss (because it focuses on the rank ability of the models. We cannot directly use spearman correlation in the loss as the rank function is not differentiable). It does not lead to better results but we find that this is a field to explore.
- Next Steps: explore other rank-based loss (so that it will approximately maximize the spearman correlation between our predicted and target values).

Conclusion

End

Matéo Molinaro