# ML in Finance, Chapter 4.1-4.3

Deheng Xu

University of Texas at Austin

March 3, 2022

# Overview

1. Asset Pricing with Characteristics-Based Factors
2. Supervised Learning Approach
3. Empirical Analysis

# Asset Pricing with Characteristics-Based Factors

- For any point in time $t$, let $r_t$ denote an $N \times 1$ vector of excess returns for $N$ stocks. Each stock has $K$ characteristics that we collect in the $N \times K$ matrix $X_t$. We use $K$ factor portfolios formed by weighting stocks' returns with their characteristics. The factor returns are $f_t = X_{t-1} r_t$. Then one can always find a price-of-risk vector $b$ such that an SDF

$$M_t = 1 - b'(f_t - Ef_t)$$

satisfies the unconditional pricing equation

$$E(M_t f_t) = 0$$

# Asset Pricing with Characteristics-Based Factors

- It can be solved that

$$b = (\Sigma)^{-1} E(f_t)$$

where $\Sigma = E[(f_t - Ef_t)(f_t - Ef_t)']$.Rewriting this expression as

$$b = (\Sigma\Sigma)^{-1} \Sigma E(f_t)$$

shows that SDF coefficients can be interpreted as the coefficients in a cross-sectional regression of the expected asset returns to be explained by the SDF,which are the K elements of $E[f_t]$, on the K columns of covariances of each factor with the other factors and with itself.

- In practice, this does not work without assuming sparse characteristics.

# Asset Pricing with Characteristics-Based Factors

- Kozak, Nagel, and Santosh (2020) point out that there are no strong economic reasons to expect characteristics-sparsity of the SDF.
- Kozak, Nagel, and Santosh (2018) argue that absence of near-arbitrage (extremely high Sharpe ratios) implies that factors earning substantial risk premia must be major sources of co-movement.
- Furthermore, for typical sets of test assets, returns have a strong factor structure dominated by a small number of PCs with the highest variance. Under these two conditions, an SDF with a small number of these high-variance PCs as factors should explain most of the cross-sectional variation in expected returns.

# Asset Pricing with Characteristics-Based Factors

To empirically test this theoretical result, construct PC factors as follows:

- Let $Q$ be the matrix of eigenvectors of $\Sigma$, $\Lambda$ is the diagonal matrix of eigenvalues ordered in decreasing magnitude. Using the eigenvector as portfolio weights, we obtain the PC factors

$$p_t = Q' f_t$$

- Using all PCs, and with knowledge of population moments, we could express the SDF as

$$M_t = 1 - b_p'(p_t - Ep_t), \qquad b_p = \Lambda^{-1} E[p_t]$$

# Supervised Learning Approach

- Consider a sample with size $T$ and let $\bar{\mu}$, $\bar{\Sigma}$ be sample moments.
- A natural and naive estimator will be

$$\hat{b} = (\bar{\Sigma}\bar{\Sigma})^{-1}\bar{\Sigma}\bar{\mu}$$

- When $K$ is not very small compare to $T$, the estimation is very imprecise because of imprecise estimation of factor return means.

# Shrinkage Estimator

To avoid spurious overfitting, we bring in economically motivated prior beliefs about the factors' expected returns.

- Under the assumption that $\Sigma$ is known, we impose the following prior beliefs about expected returns of the $K$ factor portfolios

$$\mu \sim \mathbf{N}(0, \frac{\kappa^2}{\tau}\Sigma^2)$$

Where $\tau = tr(\Sigma)$, and $\kappa$ is controlling the "scale" of $\mu$.

$$\hat{b} = (\bar{\Sigma}\bar{\Sigma})^{-1}\bar{\Sigma}\bar{\mu}$$

- For PC portfolio, the belief is expressed as

$$\mu_p \sim \mathbf{N}(0, \frac{\kappa^2}{\tau}\Lambda^2)$$

# Shrinkage Estimator

- If we expressed it as Sharpe Ratio, we have

$$\Lambda^{-\frac{1}{2}}\mu_p \sim \mathbf{N}(0, \frac{\kappa^2}{\tau}\Lambda)$$

- In rational expectations models in which cross-sectional differences in expected returns arise from exposure to macroeconomic risk factors, risk premia are typically concentrated in one or a few common factors.
- Kozak, Nagel, and Santosh (2018) show that a similar prediction also arises in plausible behavioral models in which investors have biased beliefs.
- An investor should not expect any assets to have very large or very small portfolio weights.

# Shrinkage Estimator

- Combining the prior, belief, we have the following bayesian estimator of $\mu$

$$\hat{b} = (\Sigma + \gamma I_k)^{-1}\bar{\mu}$$

  where $\gamma = \frac{\tau}{\kappa^2 T}$
- Intuitively, the bayesian estimator shrinks the estimates when the variation of the associated factor is low.
- The Bayesian estimator can also be written as

$$\hat{b} = argmin_b\{(\bar{\mu} - \Sigma b)'\Sigma^{-1}(\bar{\mu} - \Sigma b) + \gamma b'b\}$$

  i.e. GLS plus a rigde penalized term.

# Sparsity and Data-Driven Penalty Choice

- We can incorporate a Lasso penalty term in the previous specification

$$\hat{b} = argmin_b\{(\bar{\mu} - \Sigma b)'\Sigma^{-1}(\bar{\mu} - \Sigma b) + \gamma_2 b'b + \gamma_1 \sum_{j=1}^{H} |b_j|\}$$

- To implement the estimation, we need to set the value for penalty terms.
- In this chapter, the authors used k-fold cross validation. That is, the author divide the sample into K-subsample and estimate $\hat{b}$ using $k-1$ subsample and assess its OOS perform using the withheld subsample. This procedure will produce $K$ OOS $R^2$. The author then take average of $R^2$ and pick the $\gamma$ which deliver the largest average $R^2$.

# Data

- 50 characteristics from literature
- 68 financial ratios from WRDS
- 12 portfolios sorted on past monthly returns in month $t-1$ through $t-12$

# Data Normalization

- The authors first perform rank transformation:

$$rc_{s,t}^{i} = \frac{rank(c_{s,t}^{i})}{n_t + 1}$$

- Then, they normalized each rank-transformed characteristics $rc_{s,t}^{i}$ as

$$x_{s,t}^{i} = \frac{(rc_{s,t}^{i} - rc_{t}^{i})}{\sum_{s=1}^{n_t} |rc_{s,t}^{i} - rc_{t}^{i}|}$$

- This ensures that zero-investment long-short portfolios of transformed characteristics $x_{s,t}^{i}$ are insensitive to outliers and allow the authors to keep the absolute amount of long and short positions invested in the characteristic-based strategy (i.e., leverage) fixed.

# Data Normalization

- Finally, the author combines all transformed characteristics $x^i_{s,t}$ for all stocks into a matrix of instruments $X_t$. Interaction with returns, $f_t = X'_{t-1} r_t$, then yields one factor for each characteristic.
- Besides, to ensure the results are not driven by very small illiquid stocks, the authors exclude small-cap stocks with market caps below 0.01% of aggregate stock market capitalization at the time of portfolio formation.
- The authors use daily return data and adjust daily portfolio weights on individual stocks within each month to correspond to a monthly rebalanced buy-and-hold strategy during that month.
- At last, the authors orthogonalize all portfolio returns with respect to the CRSP value-weighted index return using market factor loadings estimated in the full sample.

# Fifty Anomaly Characteristics

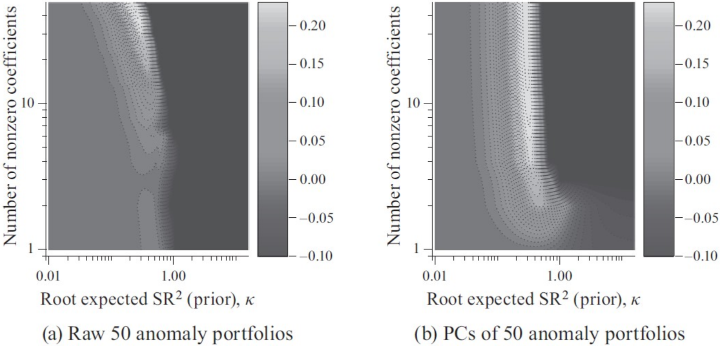



Figure 4.1. **OOS $R^2$ from dual-penalty specification (50 anomaly portfolios).** OOS cross-sectional $R^2$ for families of models that employ both $L^1$ and $L^2$ penalties simultaneously using 50 anomaly portfolios (Panel a) and 50 PCs based on anomaly portfolios (Panel b). We quantify the strength of the $L^2$ penalty by prior root expected squared Sharpe ratio on the $x$-axis. We show the number of retained variables in the SDF, which quantifies the strength of the $L^1$ penalty, on the $y$-axis. Both axes are plotted on logarithmic scale.

# Fifty Anomaly Characteristics

For 50 portfolios based on anaomaly characteristics

- Unregularized models (top-right corner) demonstrate extremely poor performance with OOS $R^2$ substantially below zero. Hence, substantial regularization is needed to get good OOS performance.
- To attain the maximum OOS $R^2$, the data calls for substantial $L^2$-shrinkage but essentially no sparsity. Imposing sparsity (i.e., moving down in the plot) leads to a major deterioration in OOS $R^2$.

# Fifty Anomaly Characteristics

For PCs of the anaomaly portfolio returns as basis asset

- A relatively sparse SDF with only four PCs, for example, does quite well in terms of OOS $R^2$, and with ten PCs we get close to the maximum OOS $R^2$.

# Fifty Anomaly Characteristics

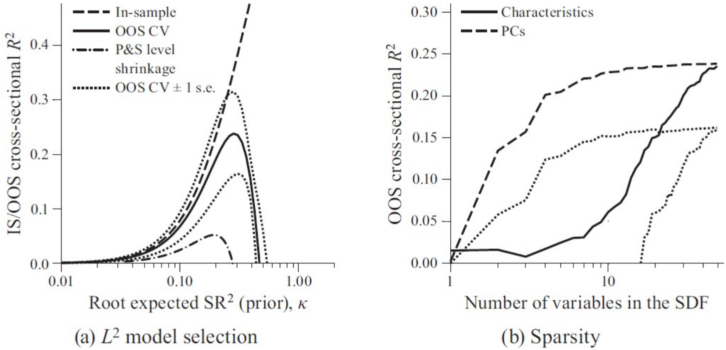



Figure 4.2. $L^2$ **model selection and sparsity (50 anomaly portfolios).** Panel (a) plots the in-sample cross-sectional $R^2$ (dashed), OOS cross-sectional $R^2$ based on cross-validation (solid), and OOS cross-sectional $R^2$ based on the proportional shrinkage (dash-dot) from Pástor and Stambaugh (2000). In Panel (b), we show the maximum OOS cross-sectional $R^2$ attained by a model with $n$ factors (on the $x$-axis) across all possible values of $L^2$ shrinkage, for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines in Panel (b) depict $-1$ s.e. bounds of the CV estimator.

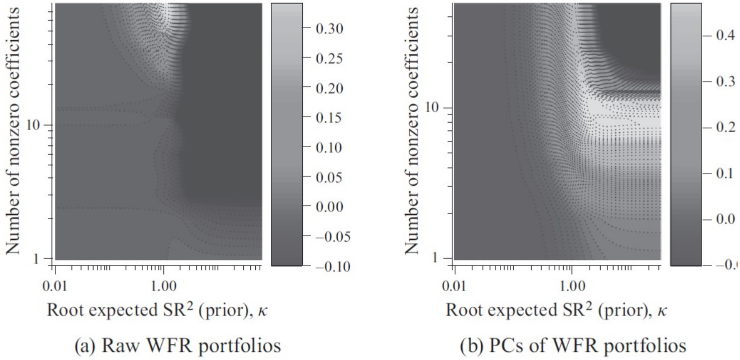



Figure 4.3. **OOS $R^2$ from dual-penalty specification (WFR portfolios).** OOS cross-sectional $R^2$ for families of models that employ both $L^1$ and $L^2$ penalties simultaneously using 80 WFR portfolios (Panel a) and 80 PCs based on WFR portfolios (Panel b). We quantify the strength of the $L^2$ penalty by prior root expected squared Sharpe ratio on the $x$-axis. We show the number of retained variables in the SDF, which quantifies the strength of the $L^1$ penalty, on the $y$-axis. Both axes are plotted on logarithmic scale.

# WRDS Financial Ratio (WFR)

- Unregularized models (top-right corner) demonstrate extremely poor performance with OOS $R^2$ substantially below zero. Hence, substantial regularization is needed to get good OOS performance.
- $L^2$-penalty-only based models (top edge of a plot) perform well for both raw portfolio returns and PCs.
- $L1$-penalty-only "lasso" based models (right edge of the plot) work poorly for raw portfolio returns in the left-hand figure.

# WRDS Financial Ratio (WFR)

- As can be seen toward the right edge of the right-hand side figure, a PC-sparse SDF not only does quite well in terms of OOS $R^2$, but it does so even without much L2-shrinkage.
- A potential explanation of this finding is that the data mining and publication bias toward in-sample significant factors may play a bigger role in the anomalies data set, which is based on published anomalies, than in the WFR data set.
- As a consequence, stronger shrinkage of SDF coefficients toward zero may be needed in the anomalies data set to prevent these biases from impairing OOS performance, while there is less need for shrinkage in the WFR data set because in- and out-of-sample returns are not so different.

# WRDS Financial Ratio (WFR)

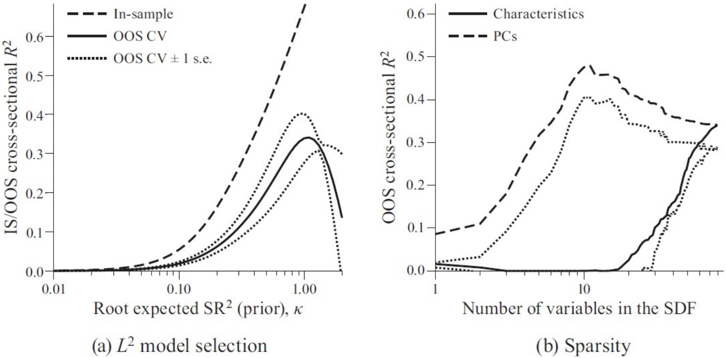
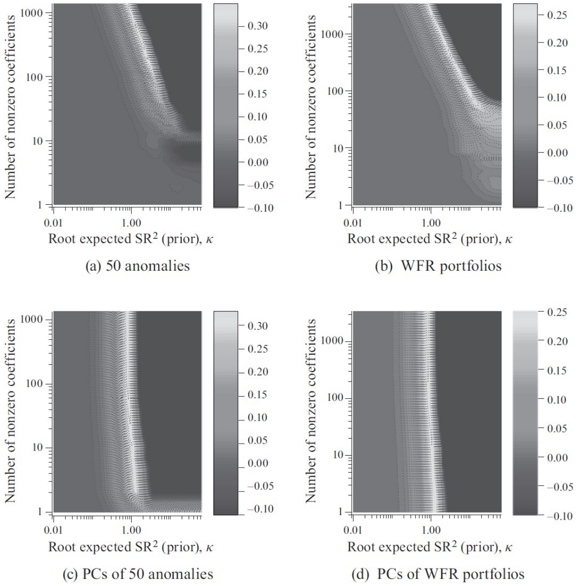



Figure 4.4. $L^2$ **model selection and sparsity (WFR portfolios).** Panel (a) plots the in-sample cross-sectional $R^2$ (dashed) and OOS cross-sectional $R^2$ based on cross-validation (solid). In Panel (b), we show the maximum OOS cross-sectional $R^2$ attained by a model with $n$ factors (on the $x$-axis) across all possible values of $L^2$ shrinkage for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines in Panel (b) depict $-1$ s.e. bounds of the CV estimator.

# Interactions

- 50 anomaly and 80 WFR raw characteristics with additional ones constructed as second and third powers and linear first-order interactions of the raw characteristics
- Renomalized the weight to accommodate interaction and higher order.
- Switch to 2-fold cross-validation because 3-fold cross-validation is not stable anymore.

# Interactions



# Interactions

- Strengthening the L1-penalty to the point that only around 100 of the characteristics and their powers and interactions remain in the SDF (out of 1,375 and 3,400, respectively) does not reduce the OOS $R^2$ as long as one picks the L2-penalty optimal for this level of sparsity. As before, an L1-penalty-only approach leads to poor OOS performance.
- Very few PCs,or even just one, suffice to obtain substantial OOS explanatory power.
- Adding more PCs does not hurt as long as substantial L2 shrinkage is imposed, but it does not improve OOS performance much either.

# Interactions

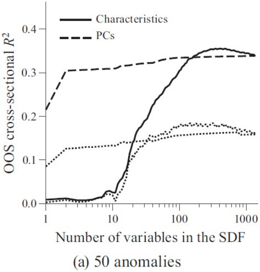



Figure 4.6. $L^1$ **sparsity of models with interactions.** We show the maximum OOS cross-sectional $R^2$ attained by a model with $n$ factors (on the $x$-axis) across all possible values of $L^2$ shrinkage for models based on interactions of original characteristics portfolios (solid) and PCs (dashed). Panel (a) focuses on the SDF constructed from PCs of interactions of 50 anomaly portfolios. Panel (b) shows coefficient estimates corresponding to PCs based on interactions of WFR portfolios. Dotted lines depict −1 s.e. bounds of the CV estimator.