Supervised Machine Learning:

Practical Machine Learning Case Studies for Finance

Dr. Claus Huber, CEFA, CFA, FRM

October 2021

Outline

- Macro Forecasting the S&P 500 and the Baa-Spread

- Granger Causality: Testing for Structure in (Alternative) Data

- Sharpe Style Regression Methods for Mutual Funds

- Natural Language Processing for Sentiment Analysis of ESG Company Reports

Typical Areas of Use Cases for AI & ML in Finance

- Financial Monitoring: Cyber Security, Money Laundering, Fraud Detection

- Credit Portfolio Management: Predict loss rates

- Investment Predictions
  - For example, **macro-forecasting**
  - Robo-advisors: derive investor preferences from social media profiles, return forecasts for tactical asset allocation

- Process Automation: Chatbots, automated document reading and storage

- Algorithmic Trading: Reading order books and short-term forecasting of price movements

- **Natural Language Processing / Sentiment Analysis**: Extract current market sentiment from news feed and use for stock trading

- Customer Retention: Predict user behaviour and provide individual product offers based on user demographic data and transaction activity

FitchLearning

Supervised Learning

- Supervised Learning:

$$Y = f(X) + \epsilon$$

- Y is the regressand (the entity we want to predict), X are the regressors (the explaining variables), ϵ is noise

- To predict we need the function f
  - f is unknown and hence needs to be estimated

- In the context of predicting time series:

$$Y_{t+1} = f(X_t) + \epsilon_{t+1}$$
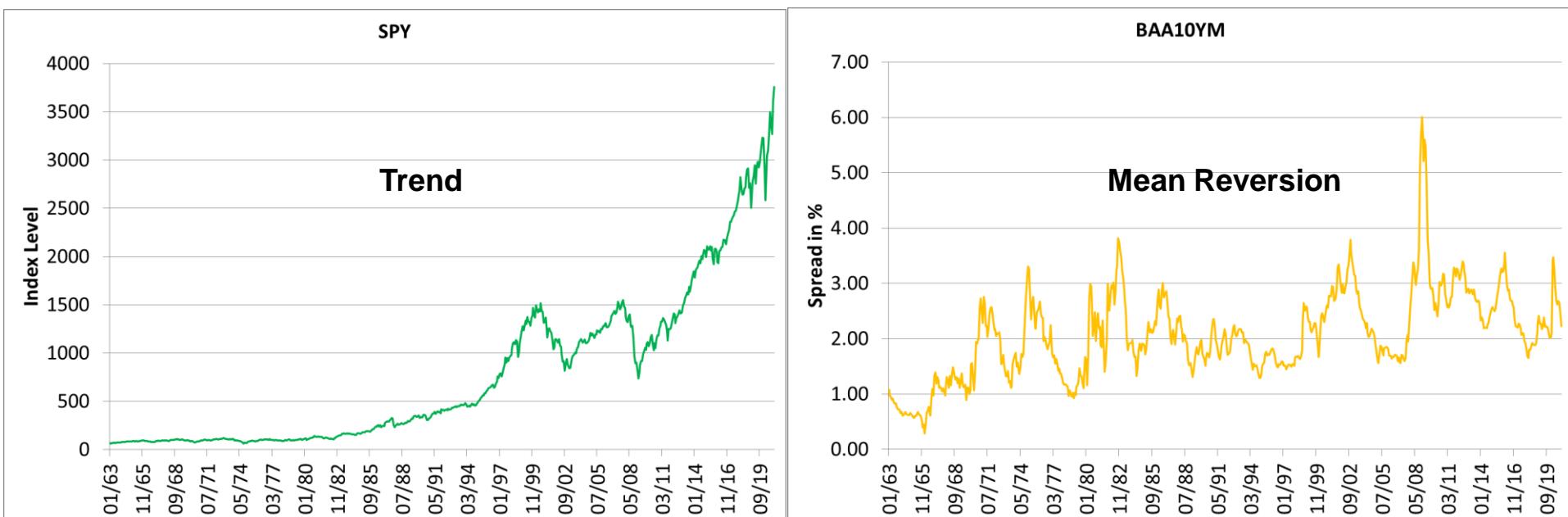
Data in Finance

- What does data in finance look like? What is different to other science disciplines?

  - Low Signal-to-Noise ratio: noisy data, structure in data weak and changing over time

  - Feedback mechanisms: A influencing B, B influencing C, C influencing A, ...

  - Sparse data environment – small data sets compared to, for example, car industry

    – Even high frequency data gives much fewer data points than industrial applications (millions of data points)

  - Time Series:

    - Autocorrelation

    - Trend properties of time series: non-stationarity & co-integration

  - Multi-collinearity

  - Outliers

  - Structural breaks

  - Missing Values

Macro Forecasting

- We forecast 2 time series: S&P500 (source: Yahoo Finance), BAA10YM (Spread BAA – 10Y US Treasuries, source: FRED)

- Monthly data from 1/1963 to 12/2020

- We forecast 1 M

Macro Forecasting: Managing Expectations

- Forecasting GDP and other macro variables useful for asset allocation, e.g., some asset classes or risk premia work well in boom times, others in recession

- Macro forecasting (e.g., unemployment rate, M2 growth rate): ML-based help to improve forecasts (=out-of-sample MSE) by 3-5% on average for the 1 M horizon (Kim / Swanson (2018), Coulombe et al. (2020))
  - The benchmark are AR(1) models → very simple model class!
  - Improvements are larger for longer forecasts horizons (3 M to 24 M) → more structure in the data for longer horizons, less noise?

- Starkest improvements can be expected when variables are pre-processed (e.g., with PCA)

Macro Forecasting

- Correlation matrix (monthly data from 2/63 to 10/20):

- BAA10YM[-1] and SPY[-1] are lagged by 1 M to reflect our forecast horizon

- All variables reflect monthly %-changes
  - For example: if INDPRO rises in one period, PAYEMS also tends to rise in that period (correl = 0.72)

	BAA10YM[-1]	SPY[-1]	US_10Y_1Y	vol_10Y	SPY	vol_SPY	INDPRO	PAYEMS	BAA10YM	BAA	CPI_Core
BAA10YM[-1]	1	-0.08	-0.01	0.25	-0.35	0.26	-0.09	-0.01	0.26	0.23	0.08
SPY[-1]	-0.08	1	0.08	-0.04	0.04	-0.08	0.03	-0.03	0.07	-0.11	-0.14
US_10Y_1Y	-0.01	0.08	1	0	0.05	0.04	-0.15	-0.07	0.24	-0.08	0.05
vol_10Y	0.25	-0.04	0	1	-0.16	0.47	-0.07	-0.05	0.23	0.32	0.13
SPY	-0.35	0.04	0.05	-0.16	1	-0.33	-0.03	-0.06	-0.07	-0.25	-0.12
vol_SPY	0.26	-0.08	0.04	0.47	-0.33	1	-0.13	-0.13	0.41	0.25	-0.01
INDPRO	-0.09	0.03	-0.15	-0.07	-0.03	-0.13	1	0.72	-0.17	0.05	0.02
PAYEMS	-0.01	-0.03	-0.07	-0.05	-0.06	-0.13	0.72	1	-0.07	0.03	0.1
BAA10YM	0.26	0.07	0.24	0.23	-0.07	0.41	-0.17	-0.07	1	0.07	0.07
BAA	0.23	-0.11	-0.08	0.32	-0.25	0.25	0.05	0.03	0.07	1	0.24
CPI_Core	0.08	-0.14	0.05	0.13	-0.12	-0.01	0.02	0.1	0.07	0.24	1

FitchLearning

Knowledge | Skills | Conduct

Benchmark Models: Autoregression

- AR models are simple: only the variable's own history is needed

$$Y_{t+1} = c + b_1 \cdot Y_t + b_2 \cdot Y_{t-1} + \cdots$$
$$+ b_N \cdot Y_{t-N+1} + \epsilon_{t+1}$$

- As a benchmark for our models, we use AR(1): $Y_{t+1} = c + b_1 \cdot Y_t + \epsilon_{t+1}$

- In AR(p) models, p would have to be determined by cross validation or information criteria
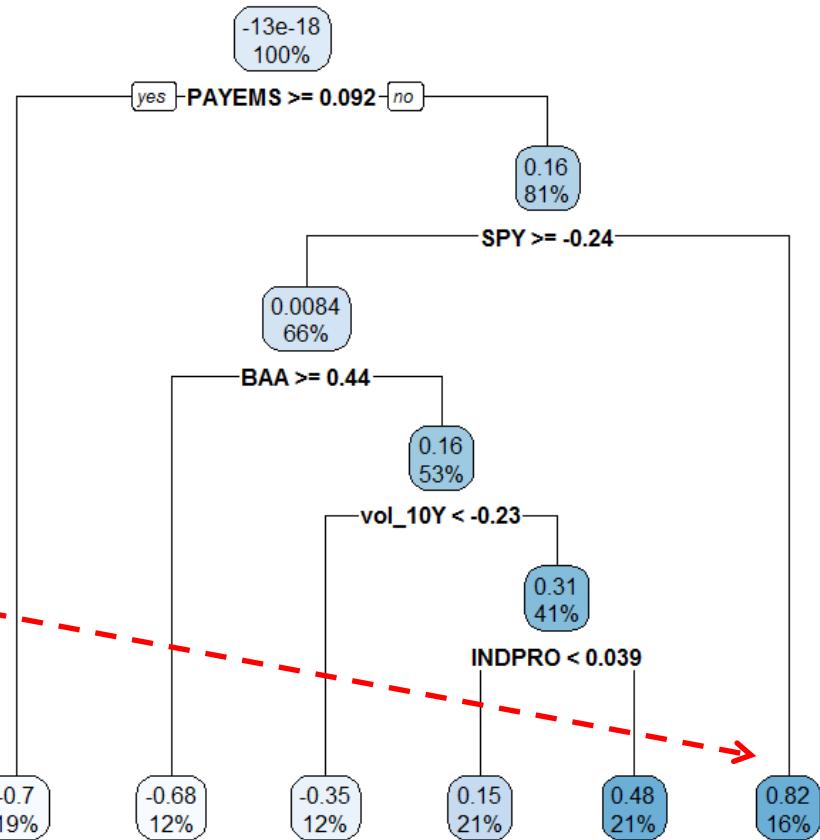
Our ML Methods for Macro Forecasting

- Our ML methods: Linear Regression, Decision Trees, Random Forests, Extreme Gradient Boosting

- Horse Race: Which model turns out to be the best based on out-of-sample (=OOS) forecasts?

- Rolling Regressions & out-of-sample forecasts

- 58 months in-sample, 1 month gap, 1 month OOS → 60 M ≡ 1 RR
  - 1st OOS forecast is for 2/68, last OOS forecast for 12/20

|	1	2	3	...	635
INS start	02/63	03/63	04/63	...	01/16
INS end	12/67	01/68	02/68	...	10/20
OOS	02/68	03/68	04/68	...	12/20

Decision Tree: Predicting SPY
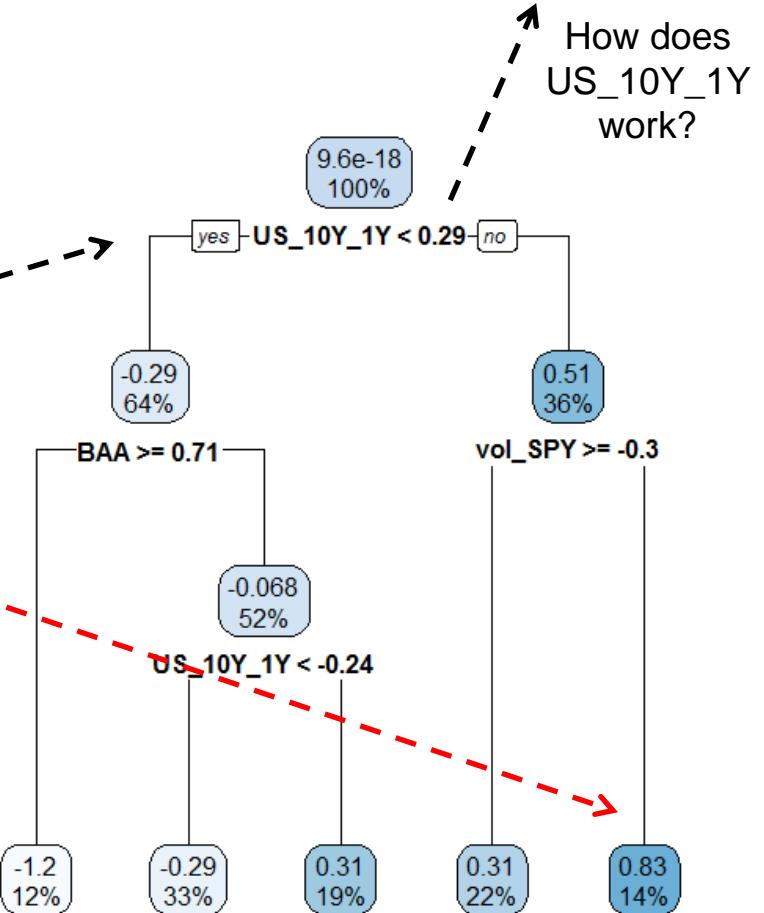
- i = 635 (INS 2016-01 to 2020-10)

- Tree determines 6 final nodes:

  - Strong increase in PAYEMS (non-farm payrolls) → SPY falls strongly

  - SPY rising strongly if SPY fell heavily in previous period (rebound)

FitchLearning

Decision Tree: Predicting SPY

|	US_10Y_1Y				
	10Y	1Y	10Y_1Y	d(10Y_1Y)	Comment
t	3%	2%	1%	NA	YC original
t+1	4%	2%	2%	1%	>0: YC steeper
t+2	2%	2%	0%	-1%	<0: YC flatter
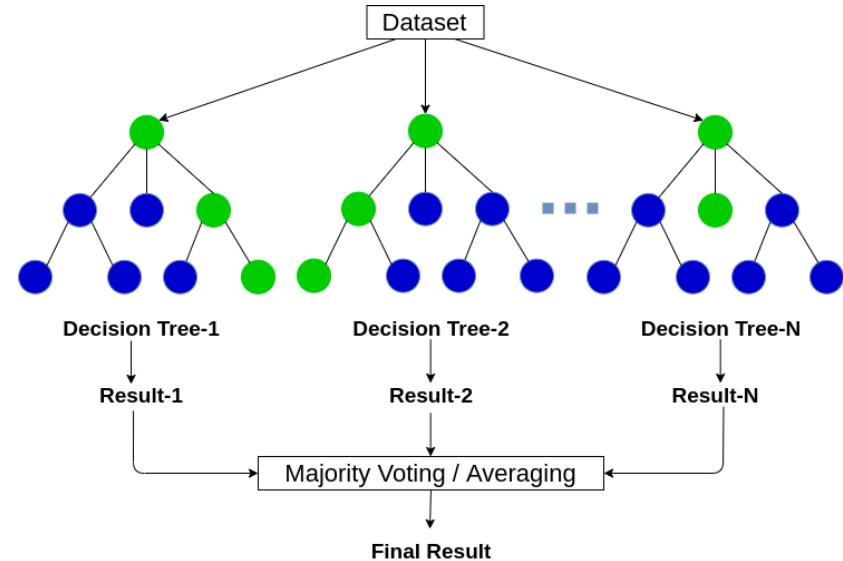
- i = 100 (INS 1971-05 to 1976-02)

- Different features relevant

- During the 1970s, inflation and rising interest rates were dominant

- YC flattening → equities down

- YC steeper and eq vol declining → favourable for equities

How does US_10Y_1Y work?

Random Forests

- Random Forests generate a large number of trees

- Each of those trees uses a different subset of the training data

- Subsets selected by sampling at random

https://miro.medium.com/max/1482/0*Srg7htj4TOMP5ldX.png

Cross Validation

- There are plenty of hyper parameters in ML models

  - For example, in Random Forests: # trees, depth of trees, …

- Which hyper parameters should I pick?

→ **Cross validation**

- We want to identify a robust model that performs well when predicting unknown data

Cross Validation: Pseudo OOS vs. k-fold

- Example:

- Forecasting 1 period ahead

- 2 rolling regressions: A, B

- Pseudo OOS: only 1 CV per rolling regression
  → can data be used more efficiently?

- k-fold CV:
  - k=2
  - 10 data points, 5 dp in each fold
    - A: [1 .. 5], [6 .. 10]
    - B: [2 .. 6], [7 .. 11]
  - 4 validations per rolling regression: A1 to A4, B1 to B4

- See, for example, Coulombe et al. (2020), Supplementary Material, p. 23ff.

Legend:
- 🟢 **Training**
- 🟡 Cross Validation
- 🔴 **Test**

Pseudo OOS

	...	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...
A			🟢	🟢	🟢	🟢	🟢	🟢	🟢		🟡	🟡		🔴			
B				🟢	🟢	🟢	🟢	🟢	🟢	🟢		🟡	🟡		🔴		

K-Folds

	...	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...
A1			🟢	🟢	🟡	🟢	🟢	🟡	🟢	🟢	🟢	🟢		🔴			
A2			🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟡	🟡	🟢		🔴			
A3			🟢	🟢	🟢	🟡	🟢	🟡	🟢	🟢	🟢			🔴			
A4			🟢	🟡	🟢	🟢	🟢	🟢	🟢	🟢	🟡			🔴			
B1				🟢	🟡	🟢	🟢	🟢	🟢	🟢	🟢	🟡			🔴		
B2				🟡	🟢	🟢	🟢	🟢	🟡	🟢	🟢	🟢			🔴		
B3				🟢	🟢	🟢	🟡	🟢	🟢	🟢	🟢	🟢			🔴		
B4				🟢	🟢	🟢	🟢	🟡	🟢	🟡	🟢	🟢			🔴		

Data Leakage

- Model incorporates information that was not available at the time a prediction is made

→ over-optimistic models if not entirely invalid models:

  - Model might be excellent on training data but useless on unknown data

- Data Leakage often occurs in time series forecasting

  - For example, erroneously including the regressand (= dependent variable) as regressor (= independent variable) with the same time stamp

Data Leakage

- If we assume an AR(p) process in s_reg(t), then s_reg(t) influences s_reg(t+1)

- The same holds in a process with exogenous variables if s_reg(t) influences s_dep(t+1)

Leakage occurs here

Leakage

	INS	INS	INS	INS	OOS	OOS	OOS	OOS		# months
	28/02/1963	...	30/11/1967	31/12/1967	31/01/1968	29/02/1968	31/03/1968	...		2/63-12/67
s_reg(t)	X	X	X	X	X	X	X	X		INS
s_dep(t)	Y	Y	Y	Y	Y	Y	Y	Y		59

No Leakage

|	INS	INS	INS		OOS	OOS	OOS	OOS		# months
	28/02/1963	...	30/11/1967		31/01/1968	29/02/1968	31/03/1968	...		2/63-11/67
s_reg(t)	X	X	X		X	X	X	X		INS
s_dep(t)	Y	Y	Y		Y	Y	Y	Y		58
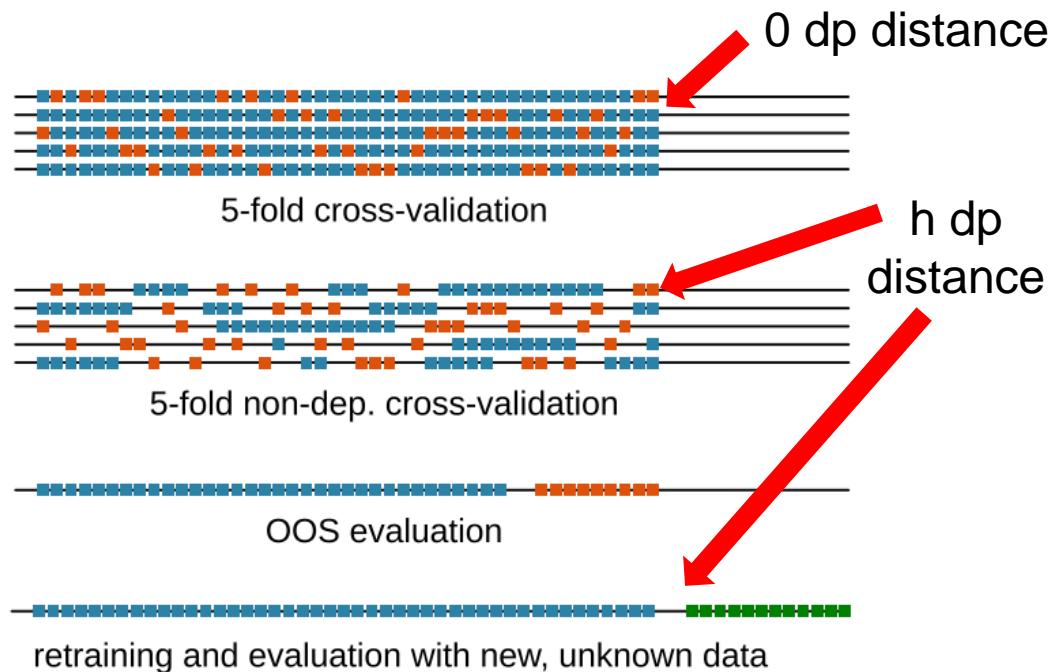
To prevent leakage, avoid overlapping periods between INS and OOS

Cross Validation: Time Series

- Serial correlation
  - Cutting folds would destroy the time series properties

- Splitting the data and holding out an earlier part of the data would mean to train a model with future data to predict the past

- **However: If the residuals of our time series model are serially uncorrelated, then k-fold cross-validation can and should be used over the holdout strategy (Bergmeir et al. (2018))**

- This is a new finding from 2018!

- Earlier, a dominant suggestion was h-block CV (Burman et al (1994))
  - The h observations preceding and following the observation are left out in the test set
  - Training set contains 2h fewer data points → inefficient use of available data

FitchLearning

Cross Validation: Leakage?

- What about leakage if we do k-fold CV for time series?

- My interpretation: there is leakage, but the information gain is higher than the detrimental effect from removing the h rows from the data

- Bergmeir et al. (2018) show that k-fold CV exhibits more favourable performance than OOS evaluation

0 dp distance

h dp distance

5-fold cross-validation

5-fold non-dep. cross-validation

OOS evaluation

retraining and evaluation with new, unknown data
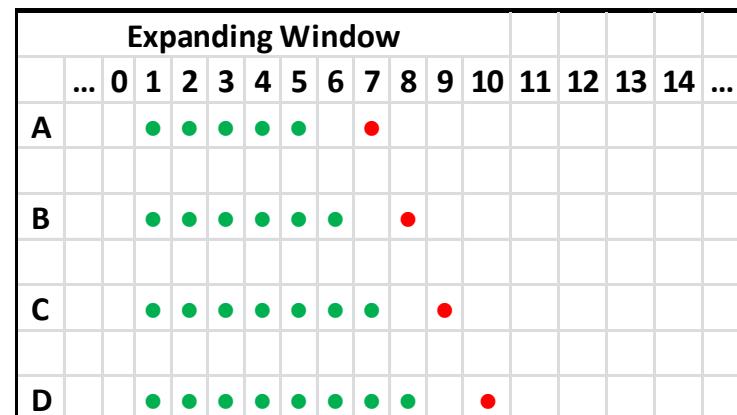
Figure 2: *Training and test sets used for the experiments. The blue and orange dots represent values in the training and test set, respectively. The green dots represent future data not available at the time of model building.*

FitchLearning

Rolling Window Or Expanding Window

Our Approach

- Rolling window vs. expanding window

- Pro rolling window:
  - Can take changes in data into account flexibly
  - Very old and hence possibly irrelevant data is disposed of
  - More robust to issues of model instability

- Con:
  - Rolling time window might be too short to capture relevant information

● Training
● Test

Rolling Window

	...	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...
A			●	●	●	●	●		●								
B				●	●	●	●	●		●							
C						●	●	●	●			●					
D							●	●	●	●	●			●			

Expanding Window

	...	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...
A			●	●	●	●	●		●								
B			●	●	●	●	●	●		●							
C			●	●	●	●	●	●	●			●					
D			●	●	●	●	●	●	●	●			●				

Rolling Window Or Expanding Window

- Kim/Swanson (2018): Expanding window estimation strategies dominate rolling strategies when constructing 1-step ahead forecasts

  - More noise in short-term signals?

- At longer forecast horizons, rolling estimation methods are preferred to expanding windows

- **Despite this finding, we use rolling windows**

  - Rationale: we are interested in 1 M forecasts in a period of unprecedented central bank action

- Papers on model stability: Pesaran and Timmermann (2007), Pesaran et al. (2013), Boot and Pick (2020)

FitchLearning

Pre-Processing: Differencing / Returns

- Nature of data requires different pre-processing

- Table from Kim / Swanson (2018): right column shows required transformation ($\rightarrow$ to achieve stationarity (I(0))

Table 1: Target Forecast Variables *

Series	Abbreviation	Y_{t+h}
Unemployment Rate	UR	$Z_{t+1} - Z_t$
Personal Income Less transfer payments	PI	$\ln(Z_{t+1}/Z_t)$
10-Year Treasury Bond	TB	$Z_{t+1} - Z_t$
Consumer Price Index	CPI	$\ln(Z_{t+1}/Z_t)$
Producer Price Index	PPI	$\ln(Z_{t+1}/Z_t)$
Nonfarm Payroll Employment	NPE	$\ln(Z_{t+1}/Z_t)$
Housing Starts	HS	$\ln(Z_t)$
Industrial Production	IPX	$\ln(Z_{t+1}/Z_t)$
M2	M2	$\ln(Z_{t+1}/Z_t)$
S&P 500 Index	SNP	$\ln(Z_{t+1}/Z_t)$
Gross Domestic Product	GNP	$\ln(Z_{t+1}/Z_t)$

* Notes: Data used in model estimation and prediction construction are monthly U.S. figures for the period 1960:1-2009:5. Data transformations used in prediction experiments are given in the last column of the table. See Section 4 for further details.

FitchLearning
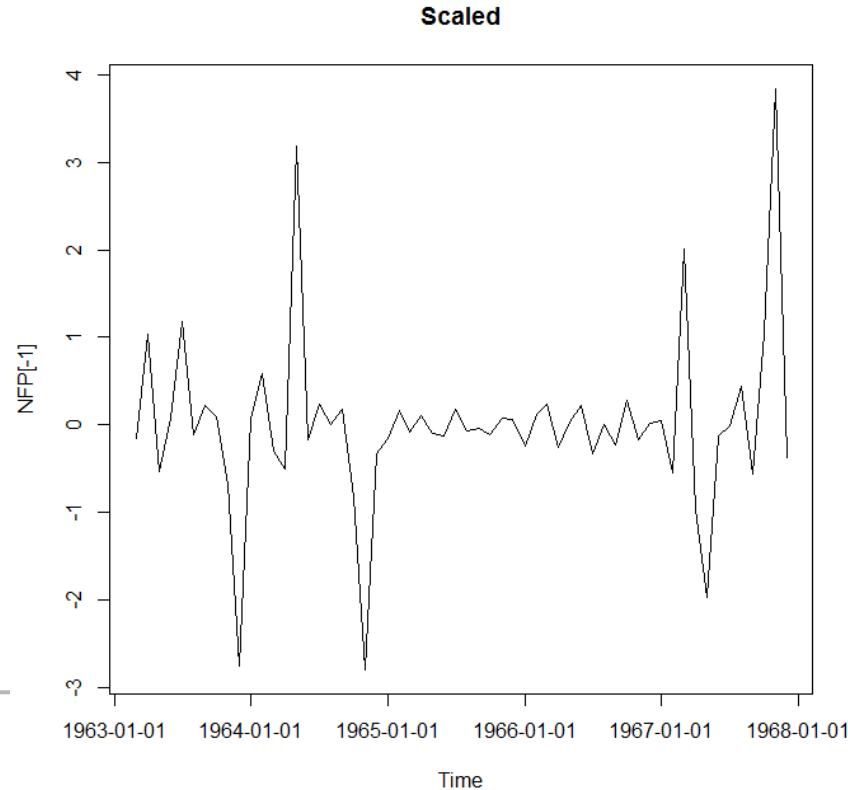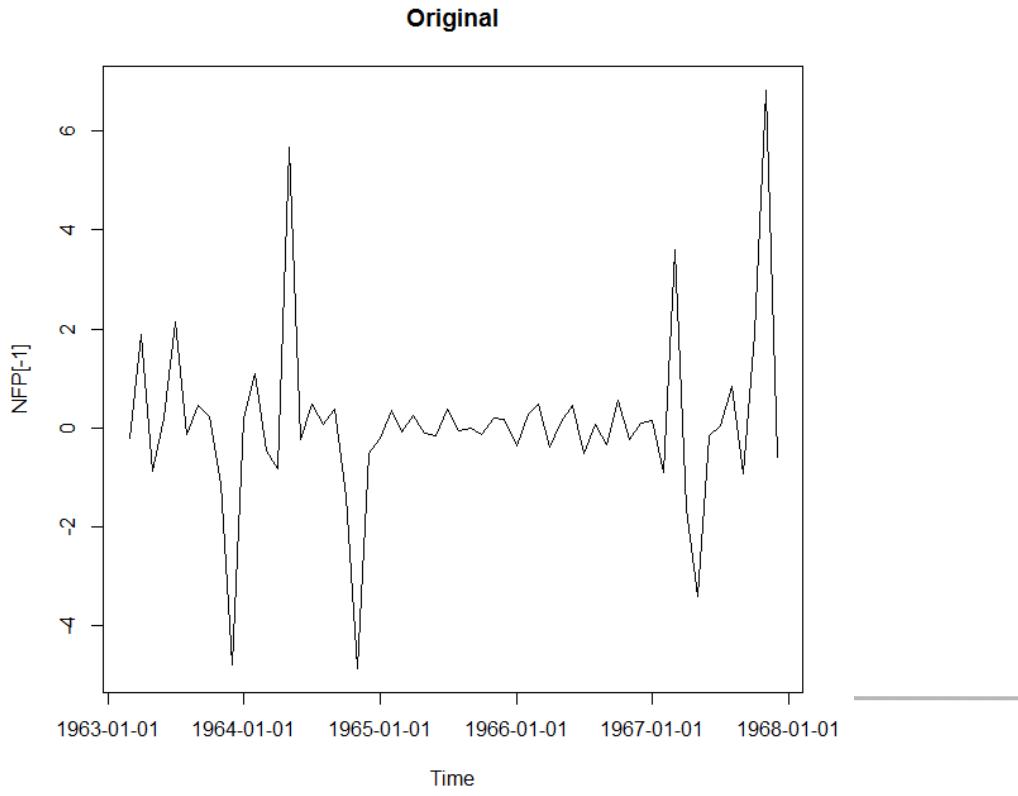
Pre-Processing: Data Transformations

- Each data point x(i) is replaced by the transformed value x'(i) = f(x(i))

- Common transformations:

  - x'(i) = $\frac{x-\mu}{\sigma}$: mean of 0 and variance 1 (demeaning and standardisation)

  - x'(i) = ln(x): log transformation (beware of negative data points)

  - x'(i) = $\frac{x-min}{max-min}$: interval between 0 and 1

- Transformations are useful to:

  - Linearise non-linear data (to process further with, for example, linear regression)

  - Make data series of different orders of magnitude comparable, e.g., interest rates (0.1%) and GDP ($ 21.5 trn)

  - This can be important with variable selection / regularisation models (w/o standardisation large scale variables would be preferred)

Data Pre-Processing: Data Transformations

- US non-farm payrolls

- % changes as in Kim / Swanson (2018)

- Plots show original data (% changes, left) and standardised (right) → scale!
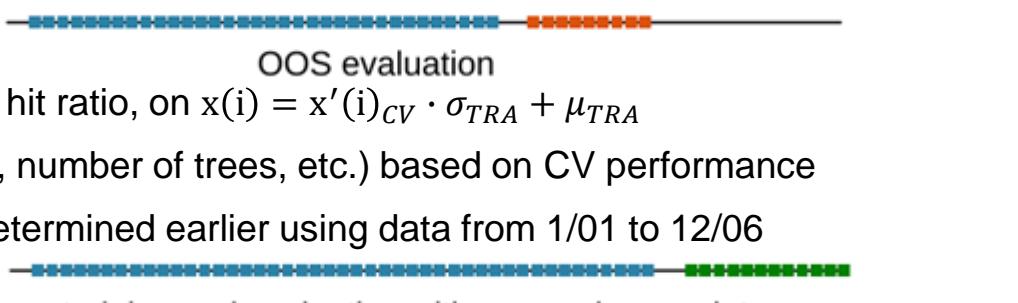
	NFP[-1]	
	original	scaled
Min	-5.28	-2.80
1st Quartile	-0.47	-0.23
Median	-0.02	0.01
Mean	-0.04	0.00
3rd Quartile	0.36	0.22
Max	6.84	3.68
SD	1.87	1.00

Original

Scaled

Data Pre-Processing: Data Transformations for Cross Validation

- x'(i) = $\frac{x-\mu}{\sigma}$: mean of 0 and variance 1

- Mean and variance can only be calculated on known data (=TRA)

- For CV (= "unknown" data), transformations should be based on TRA's parameters

Example: steps to **simulate a forecast** for 2/2007 (OOS evaluation):

- Select data points for TRA (e.g., from 1/2001 to 12/2005) & CV (2/2006 to 12/2006)
- Calculate μ and σ for TRA
- Build model with data from TRA
- For CV, calculate $x(i)'_{CV} = \frac{x-\mu_{TRA}}{\sigma_{TRA}}$, i.e., variables are transformed with parameters from TRA only
- Cross-validate model on $x(i)'_{CV}$:

  OOS evaluation

- Calculate untransformed signals, e.g., hit ratio, on $x(i) = x'(i)_{CV} \cdot \sigma_{TRA} + \mu_{TRA}$
- Determine parameters (e.g., variables, number of trees, etc.) based on CV performance
- Re-estimate model with parameters determined earlier using data from 1/01 to 12/06
- Do not use 1/2007, predict 2/2007:

  retraining and evaluation with new, unknown data

Performance Measurement

- MSE or RMSE are often used in the literature

$$\text{RMSE}= \sqrt{\frac{1}{\#OOS}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$$
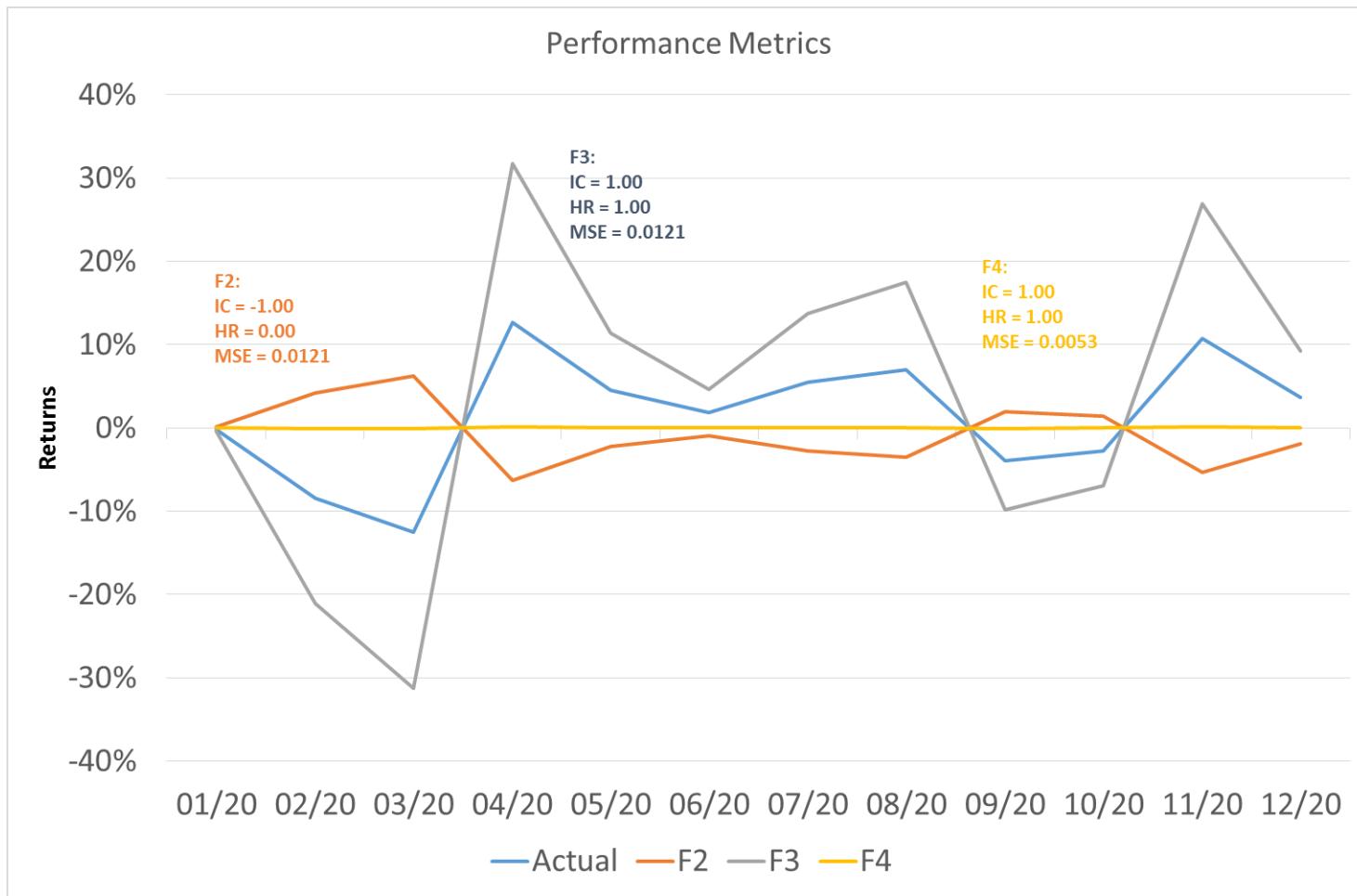
- y(t) measures level of the variable

- Information Coefficient (IC): correl bw. actual and predicted values

- Hit Ratio: [# forecasts with correct sign] / [# all forecasts]

	SPX	-1*Actual				-0.5*Actual				Multiplier to match SPX with MSE(F2): 2.5015				0.01*Actual				close to 0			
	Actual	F1	e2	ED	Direction	F2	e2	ED	Direction	F3	e2	ED	Direction	F4	e2	ED	Direction	F5	e2	ED	Direction
31/01/2020	-0.2%	0.2%	0.0000	0.0033	0	0.1%	0.0000	0.0024	0	-0.4%	0.0000	0.0024	1	0.0%	0.0000	0.0016	1	0.1%	0.0000	0.0026	0
29/02/2020	-8.4%	8.4%	0.0283	0.1682	0	4.2%	0.0159	0.1262	0	-21.0%	0.0159	0.1263	1	-0.1%	0.0069	0.0833	1	0.1%	0.0072	0.0851	0
31/03/2020	-12.5%	12.5%	0.0626	0.2502	0	6.3%	0.0352	0.1877	0	-31.3%	0.0353	0.1879	1	-0.1%	0.0153	0.1239	1	0.1%	0.0159	0.1261	0
30/04/2020	12.7%	-12.7%	0.0644	0.2537	0	-6.3%	0.0362	0.1903	0	31.7%	0.0363	0.1905	1	0.1%	0.0158	0.1256	1	0.1%	0.0158	0.1258	1
31/05/2020	4.5%	-4.5%	0.0082	0.0906	0	-2.3%	0.0046	0.0679	0	11.3%	0.0046	0.0680	1	0.0%	0.0020	0.0448	1	0.1%	0.0020	0.0443	1
30/06/2020	1.8%	-1.8%	0.0014	0.0368	0	-0.9%	0.0008	0.0276	0	4.6%	0.0008	0.0276	1	0.0%	0.0003	0.0182	1	0.1%	0.0003	0.0174	1
31/07/2020	5.5%	-5.5%	0.0121	0.1102	0	-2.8%	0.0068	0.0827	0	13.8%	0.0068	0.0827	1	0.1%	0.0030	0.0546	1	0.1%	0.0029	0.0541	1
31/08/2020	7.0%	-7.0%	0.0196	0.1401	0	-3.5%	0.0110	0.1051	0	17.5%	0.0111	0.1052	1	0.1%	0.0048	0.0694	1	0.1%	0.0048	0.0691	1
30/09/2020	-3.9%	3.9%	0.0062	0.0785	0	2.0%	0.0035	0.0588	0	-9.8%	0.0035	0.0589	1	0.0%	0.0015	0.0388	1	0.1%	0.0016	0.0402	0
31/10/2020	-2.8%	2.8%	0.0031	0.0553	0	1.4%	0.0017	0.0415	0	-6.9%	0.0017	0.0415	1	0.0%	0.0008	0.0274	1	0.1%	0.0008	0.0287	0
30/11/2020	10.8%	-10.8%	0.0463	0.2151	0	-5.4%	0.0260	0.1613	0	26.9%	0.0261	0.1615	1	0.1%	0.0113	0.1065	1	0.0%	0.0116	0.1076	0
31/12/2020	3.7%	-3.7%	0.0055	0.0742	0	-1.9%	0.0031	0.0557	0	9.3%	0.0031	0.0557	1	0.0%	0.0014	0.0368	1	0.0%	0.0014	0.0372	0
MSE:			0.0215	1.4762			0.0121	1.1071			0.0121	1.1083			0.0053	0.7307			0.0054	0.7383	
IC:					-1.00				-1.00				1.00				1.00				-0.36
HR:					0.00				0.00				1.00				1.00				0.42

Knowledge | Skills | Conduct
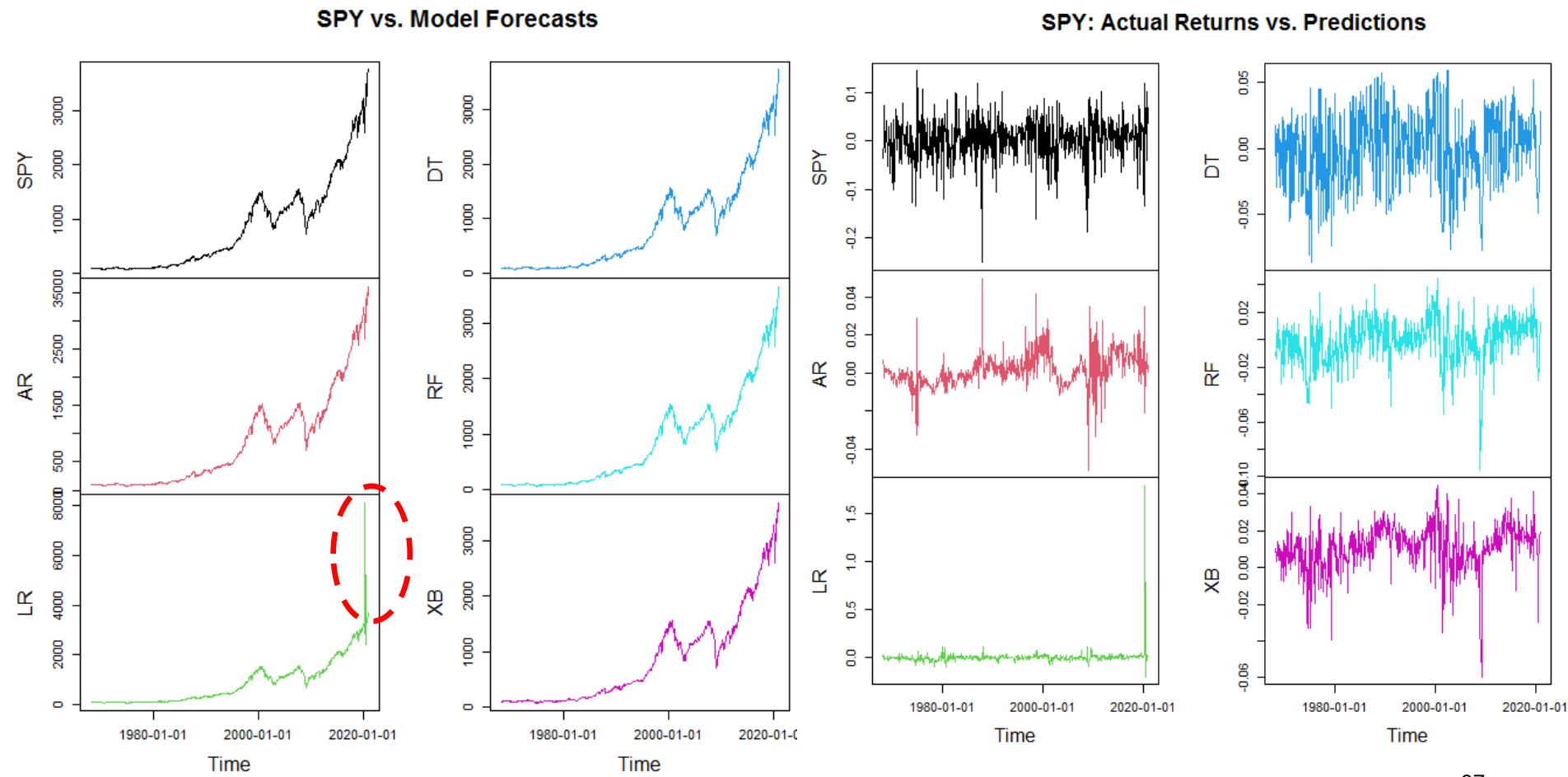
Performance Measurement, Loss Functions

- Detailed numerical example on previous slide

- F2: low MSE, but dismal IC = -1 & HR = 0

- F3: same MSE as F2, but accurate directionality (IC = 1, HR = 1)

- F4: time series almost does not budge, low MSE, IC = 1 & HR = 1

Performance Metrics

F2:
IC = -1.00
HR = 0.00
MSE = 0.0121

F3:
IC = 1.00
HR = 1.00
MSE = 0.0121

F4:
IC = 1.00
HR = 1.00
MSE = 0.0053

Returns

— Actual — F2 — F3 — F4

Empirical Results

- Rolling Regressions, 58 data points INS to predict, 1 to prevent leakage, 1 OOS forecast

- Plots of SPY vs. OOS forecasts

SPY vs. Model Forecasts

SPY: Actual Returns vs. Predictions

Empirical Results: Information Coefficients, Hit Ratios and MSFE for Select Periods (All OOS)

- BAA10YM can be more accurately forecasted than SPY

- Simple models, like AR & LR prove to be hard-to-beat benchmarks
 - LR is the winner for BAA10YM, followed by RF
  - LR has highest IC for SPY, but XB with highest Hit Ratio

			SPY					BAA10YM				
			AR	LR	DT	RF	XB	AR	LR	DT	RF	XB
29/02/1968	31/12/2020	Total										
		Hit Ratio	0.51	0.52	0.52	0.54	0.57	0.54	0.62	0.57	0.61	0.53
		Information Coefficient	0.02	0.09	0.06	0.06	0.04	0.22	0.38	0.22	0.27	0.22
		MSFE	0.0443	0.0481	0.0508	0.0457	0.0458	0.0019	0.0020	0.0021	0.0019	0.0019
31/01/1973	31/12/1982	High Inflation										
		Hit Ratio	0.59	0.60	0.59	0.56	0.51	0.45	0.62	0.58	0.62	0.53
		Information Coefficient	-0.07	0.14	0.13	0.07	0.01	-0.04	0.36	0.25	0.21	0.20
		MSFE	0.0488	0.0538	0.0538	0.0496	0.0504	0.0027	0.0026	0.0027	0.0026	0.0026
28/02/2009	31/12/2020	Post GFC										
		Hit Ratio	0.58	0.51	0.55	0.55	0.67	0.58	0.62	0.58	0.59	0.51
		Information Coefficient	0.00	-0.03	0.05	0.10	0.08	0.24	0.35	0.32	0.34	0.29
		MSFE	0.0372	0.0398	0.0405	0.0363	0.0370	0.0012	0.0016	0.0015	0.0012	0.0013
31/01/2020	31/12/2020	Covid										
		Hit Ratio	0.50	0.50	0.42	0.25	0.50	0.75	0.50	0.75	0.58	0.42
		Information Coefficient	-0.08	0.14	-0.10	-0.32	-0.29	0.16	0.13	0.35	0.44	0.34
		MSFE	0.0744	0.5132	0.0843	0.0811	0.0768	0.0041	0.0251	0.0039	0.0039	0.0040
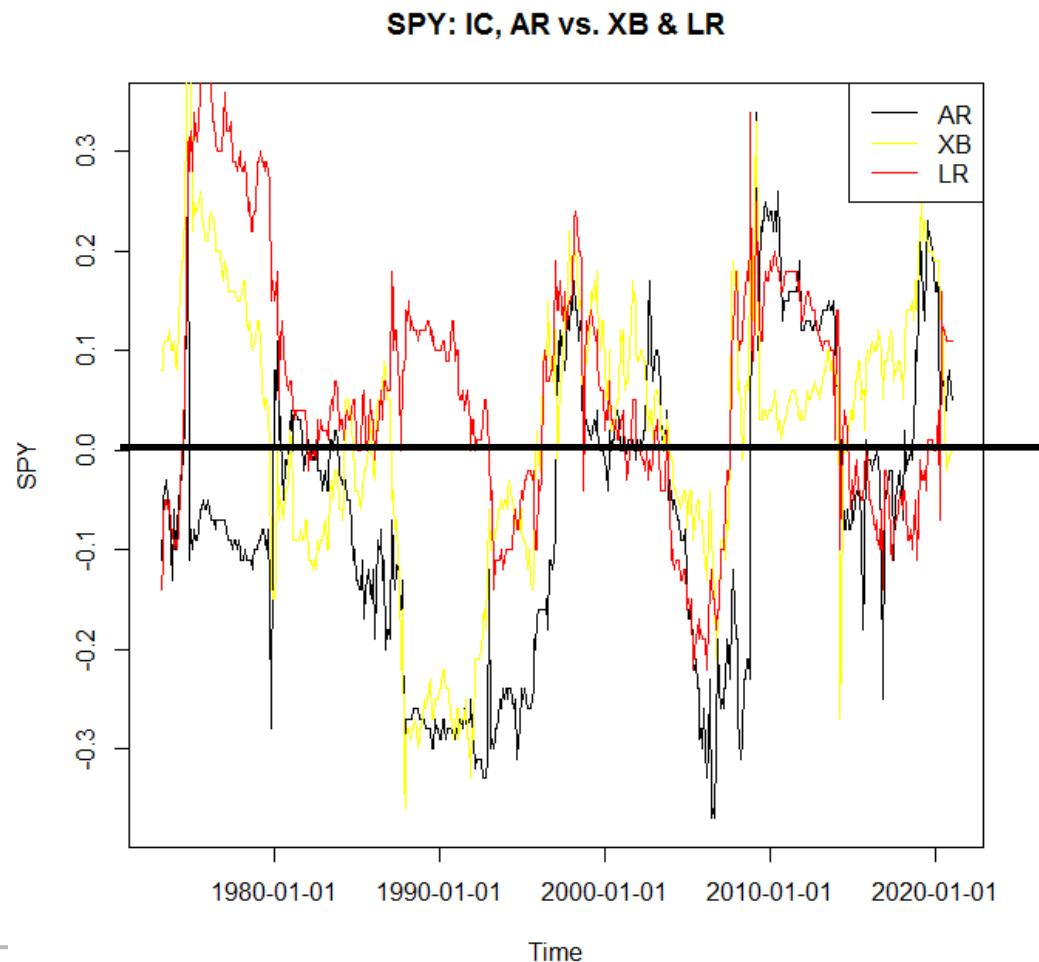
Post-GFC ML models outperform AR, LR

During Covid ML outperforms AR, LR

FitchLearning

Knowledge | Skills | Conduct

28

Empirical Results SPY: Information Coefficients over Time for LR (Red) and XB (Yellow)

- Rolling IC over 5 Y

- Models shown here are complementary to each other

- For example:
  - pre-1980 AR: IC<0, but LR & XB IC > 0
  - Around 1990: LR best performer
  - Post-GFC: XB most successful

→ Model Ensembles!?

SPY: IC, AR vs. XB & LR

FitchLearning

Interpreting the Output SPY: Linear Regression

- Regression output: Adj. R2 = 13%

- Most relevant variables are vol_10Y, INDPRO, PAYEMS, BAA10YM

- Vol_SPY and CPI_Core also contribute

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	0.0000	0.1222	0.0000	1.0000	
x_reg[, s_reg]US_10Y_1Y	0.0568	0.1456	0.3900	0.6983	
x_reg[, s_reg]vol_10Y	0.7657	0.2915	2.6260	0.0116	*
x_reg[, s_reg]SPY	0.0020	0.1916	0.0100	0.9919	
x_reg[, s_reg]vol_SPY	-0.7217	0.3614	-1.9970	0.0515	.
x_reg[, s_reg]INDPRO	1.0340	0.4217	2.4510	0.0179	*
x_reg[, s_reg]PAYEMS	-0.7483	0.3625	-2.0640	0.0444	*
x_reg[, s_reg]BAA10YM	0.5790	0.2369	2.4440	0.0182	*
x_reg[, s_reg]BAA	-0.2278	0.1771	-1.2860	0.2046	
x_reg[, s_reg]CPI_Core	-0.3309	0.1742	-1.9000	0.0635	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 0.9309 on 48 degrees of freedom					
Multiple R-squared: 0.2703, Adjusted R-squared: 0.1335					
F-statistic: 1.976 on 9 and 48 DF, p-value: 0.06322					

Period i=635
(2016-01 to
2020-10)

FitchLearning

Interpreting the Output SPY: Random Forest
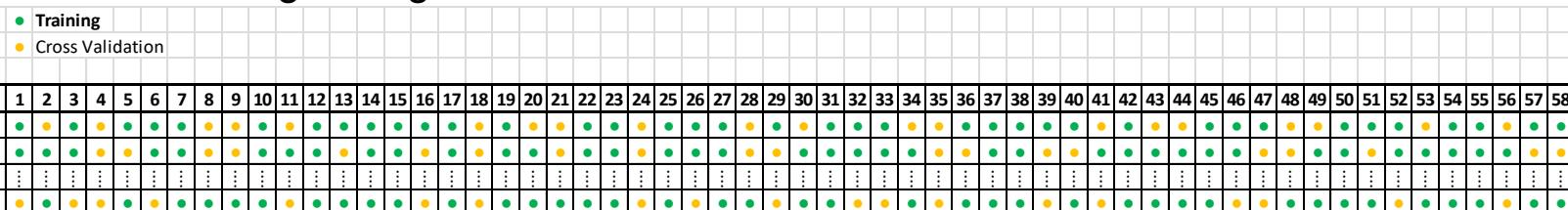
- R: importance(x.RF, scale = TRUE)

- %IncMSE measures the increase in MSE when a given variable is excluded from the model
  - Basis is mean decrease of accuracy in predictions on out of bag samples

- PAYEMS is the most important variable, followed by INDPRO

Var#	s_reg	%IncMSE	Importance Rank
1	US_10Y_1Y	0.63	6
2	vol_10Y	1.64	3
3	SPY	-0.36	7
4	vol_SPY	-1.00	9
5	INDPRO	1.83	2
6	PAYEMS	3.23	1
7	BAA10YM	0.80	5
8	BAA	1.03	4
9	CPI_Core	-0.56	8

Period i=635
(2016-01 to
2020-10)

FitchLearning

Random Forest

- How does R function randomForest perform cross validation?
  - see also complementary R code
  - ntree=100 (grow 100 forests)
  - mtry=3 (use 3 variables in each forest, e.g., vol_SPY, vol_10Y, BAA): select randomly 3 variables out of the 9 variables in our universe for each of the 100 RF grown

- 58 data points are available in our rolling window for training & validation
  - R randomForest uses ca. 1/3 of data points for out-of-bag valuation (here: 19 data points), 2/3 for training (here: 39 data points)

- Example of arranging randomly selected variables and data points for training & CV (= out-of-bag valuation):
  - Green dots → training, orange dots → cross validation

● Training
● Cross Validation

# RF	Variables used	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58
1	vol_SPY, vol_10Y, BAA																																																										
2	INDPRO, BAA, CPI_Core																																																										
⋮	⋮																																																										
100	PAYEMS, INDPRO, US_10Y-1Y																																																										

R Code: "CQF Macro Forecasting.R"

- 3 files belong to this use case:
  - "CQF Macro Forecasting.R": R code
  - "Macro_Data_US_1963.csv": input file economic variables
  - "Macro_OOS Periods.csv": settings file

- Several R libraries need to be installed:
  - timeSeries, rpart, rpart.plot, randomForest, xgboost, car

- Running all Rolling Regressions takes ca. 5 min
  - Rolling Regressions run in a large loop

- User can specify individual periods, run ML for those periods separately and study results
  - In R code: ENTRY POINT, EXIT POINT

- Ca. 400 lines of code
  - Focus is on education, not a production system!

- All data are from public sources
  - "Macro_Data_US_1963.csv": FRED and US Fed

FitchLearning

R Code: "CQF Macro Forecasting.R"

- "Macro_OOS Periods.csv": specify OOS period and display results in df_results

  - For example, row 1 shows IC, HR etc. for period 2/1968 to 12/2019
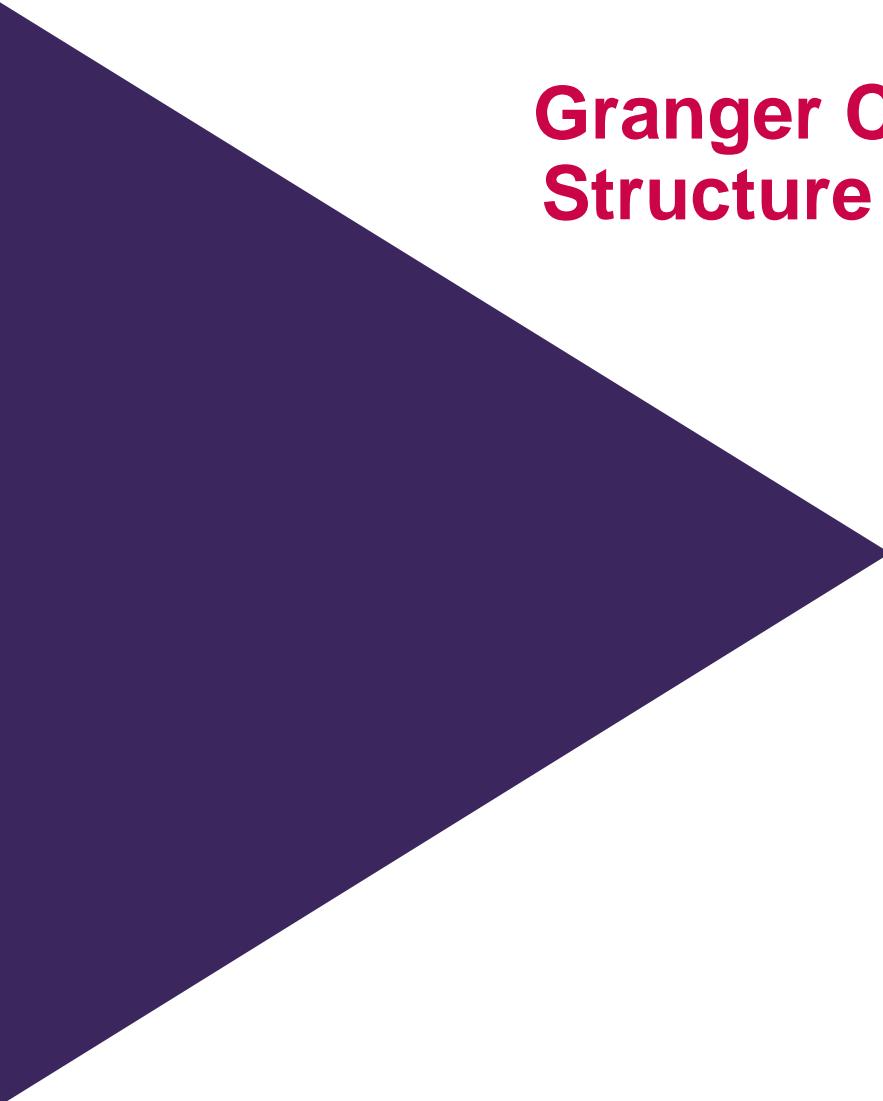
Macro_OOS Periods.csv		
Start	End	Comment
29/02/1968	31/12/2019	Total2019
29/02/1968	31/12/2020	Total2020
31/01/1973	31/12/1982	High Inflation
⋮	⋮	⋮

- R code generates output file "Results SPY.csv":

	Start	End	Comment	AR_IC	LR_IC	DT_IC	RF_IC	XB_IC	AR_HR	LR_HR	DT_HR	RF_HR	XB_HR	AR_RMSFE	LR_RMSFE	DT_RMSFE	RF_RMSFE	XB_RMSFE
1	29/02/1968	31/12/2019	Total2019	0.02	0.09	0.06	0.1	0.06	0.51	0.52	0.52	0.55	0.57	0.0443	0.0481	0.0508	0.0452	0.0455
2	NA	NA	Relative to AR:1	0	0.07	0.04	0.08	0.04	0	0.01	0.01	0.04	0.06	1.0000	1.0850	1.1464	1.0205	1.0257
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

IC for AR(1): 0.02

IC for LR, **relative** to AR(1): 0.09-0.02 = 0.07

FitchLearning

Knowledge | Skills | Conduct

Granger Causality: Testing for Structure in (Alternative) Data

Alternative Data: Features

- Alternative economic data collected from new media, e.g., smart phones, Google Trends, satellite data, …

- Can offer visibility into activity that traditional measures miss

- Shorter time lags, especially for areas where information is scarce or unreliable

  - US government shutdown (2019), Covid lockdown (2020)

- Potentially enhanced transparency for area or economies where as yet data is not plenty (e.g., Emerging Markets)

- Higher granularity (specific firms, industries, demographic groups, geographic regions)

- But short histories, collection systems that are prone to change and inconsistent samplings of the population

Alternative Data: Testing for Structure

- Vendor of Alternative Data

  - Platform offers several hundred data sets for download

  - Data are pre-processed and ready to consume in Excel format. API under construction

  - User does not have to collect the raw data and/or pre-process it

  - Data are sold to many parties: no exclusivity

  - Data history typically short (max. back to 2010)

- Do Alternative Data add any value to our investment process?

  - Universe of several hundred alternative economic indicators, e.g., activity in European Chemical industry

- How can we find out whether / which variables contain useful information? → Granger Causality Test

Granger Causality

- Granger causality is based on the intuition that a cause X helps to predict a target variable Y:

- $Y_t = a_0 + a_1 \cdot Y_{t-1} + \cdots + a_p \cdot Y_{t-p} + b_1 \cdot X_{t-1} + \cdots + b_p \cdot X_{t-p} + \epsilon_t$

- H0: $b_1 = b_2 = \ldots = b_p = 0$

- If H0 can be rejected, X_{t-1} "Granger-causes" Y_t

- Various applications, also outside of Finance: which methods cause other variables?
  - Climate change: which measures taken cause climate change to slow?
  - Covid vaccination: which measures help?

- Interesting for vendors of Alternative Data: how can they prove that their data offer value?

FitchLearning

Knowledge | Skills | Conduct

Granger Causality Test in R

- "CQF Granger Causality Test.R"
  - 17 lines of code

- Grangertest (R package lmtest): Bivariate Granger causality testing

- Currently, the methods for the generic function grangertest only perform tests for Granger causality in **bivariate** series

- The test is simply a test comparing 2 models:
  - Unrestricted model: $Y \rightarrow ARX(p)$
  - Restricted model: $Y \rightarrow AR(p)$

- For more elaborate methods, see Lawrence et al. (2020), Marinazzo et al. (2008)

Granger Causality Test in R: Simple Example

- Test if BAA spread contributes to predicting SPY (monthly data)
  - Here: 2 lags

- Model 1 is unrestricted (SPY lags & BAA lags)

- Model 2 is restricted (SPY lags only)

- 2 periods:
  1. 2/63 to 11/67
  2. 12/15 to 9/20

- Granger causality varies depending on time period

	Granger causality test				
i = 1	**2/1963 - 11/1967**				
	Model 1: x_reg[, "SPY"] ~ Lags(x_reg[, "SPY"], 1:2) + Lags(x_reg[, "BAA"], 1:2)				
	Model 2: x_reg[, "SPY"] ~ Lags(x_reg[, "SPY"], 1:2)				
	Res.Df	Df	F	Pr(>F)	
1	51	NA	NA	NA	
2	53	-2	5.3526	0.0078	**
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
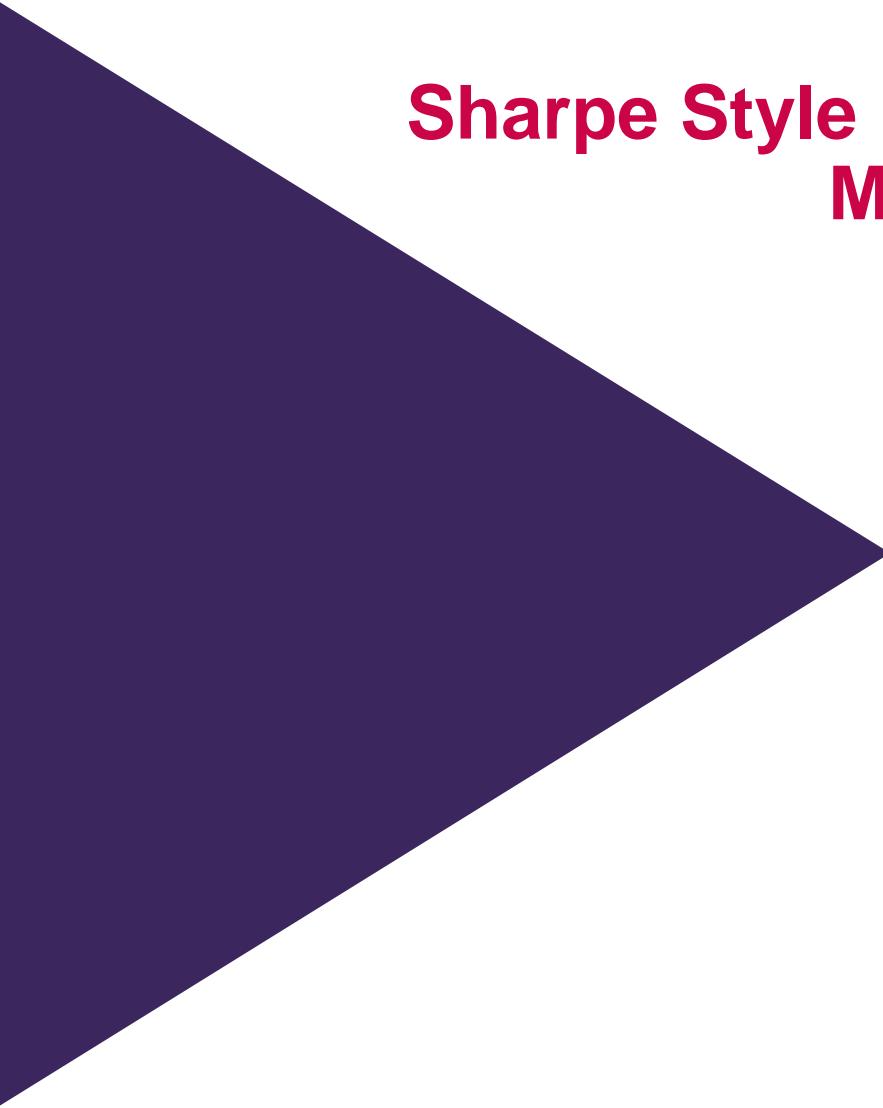
p-value significant @ 1% level:
Reject the Null of **no** Granger Causality,
BAA spread helps to forecast SPY

	Granger causality test				
i = 635	**12/2015 - 9/2020**				
	Model 1: x_reg[, "SPY"] ~ Lags(x_reg[, "SPY"], 1:2) + Lags(x_reg[, "BAA"], 1:2)				
	Model 2: x_reg[, "SPY"] ~ Lags(x_reg[, "SPY"], 1:2)				
	Res.Df	Df	F	Pr(>F)	
1	51	NA	NA	NA	
2	53	-2	0.0251	0.9752	

p-value insignificant,
BAA spread does not help to forecast SPY

FitchLearning

Granger Causality: Summary of Our Alternative Data Tests

- Our test results with several hundreds of Alt. Data series showed merely no significant Granger causality

- Even if Alt. Data does not fit to our **quantitative** investment process (monthly rebalancing, 1 M horizon), they might fit to other ones

- Corporate-specific alt. data probably more suited for **fundamental** analysis

- Do Alt Data add value in the presence of structural breaks (e.g., Covid)?

- From a macro perspective, Alternative Data probably most useful in area with little data available, e.g., Emerging Markets, during a government shutdown (e.g., US 2019)
  - Individual states, regions, cities, industries, firms, stocks, etc.

Sharpe Style Regression Methods for Mutual Funds

Sharpe Style Regression

- Sharpe Style Analysis (1992) useful for identifying a fund's exposures or risk factors:

$$R_{i,t} = b_{i,1}F_{1,t} + b_{i,2}F_{2,t} + \cdots + b_{i,n}F_{n,t} + \epsilon_{i,t}$$

- Which model can **explain** (NOT: predict) most of the managers' return variance?

- Sharpe (1992): styles *F* represented by asset classes

- Commercial applications, e.g., for Morningstar ratings: Kaplan (2003), also Hardy (2003)

Why Sharpe Style Regression?

- Copy cat funds: copy the most successful (hedge) fund managers just by looking at their returns
  - CAIA endowment investable portfolio (Kazemi / Wilkens (2017, 2020): replicate endowment funds with simple ETFs

- Performance attribution: which sector / asset class mostly drive returns of a manager?
  - Determine exposures of funds to risk factors (e.g., Fama / French SMB, HML, …)

- Basis for peer group analyses
  - Compare equity & bond weights across (competitor) funds and over time

- Do managers follow the style they claim to follow?
  - For example, a fund claims superior selection skills, but actually loads high on momentum or value

Sharpe Style Regression

SPY	SPDR S&P 500 ETF Trust (SPY)
IWM	iShares Russell 2000 ETF (IWM)
QQQ	Invesco QQQ Trust (QQQ)
EZU	iShares MSCI Eurozone ETF (EZU)
RWO	SPDR Dow Jones Global Real Estate ETF (RWO)
VWO	Vanguard FTSE Emerging Markets Index Fund ETF Shares (VWO)
DBC	Invesco DB Commodity Index Tracking Fund (DBC)
HYG	iShares iBoxx $ High Yield Corporate Bond ETF (HYG)
VLUE	iShares MSCI USA Value Factor ETF (VLUE)
GOVT	iShares U.S. Treasury Bond ETF (GOVT)
LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF

- Why is Sharpe Style Analysis an ML problem?

→ Find significant variables, i.e., risk factors, that help to explain a manager's returns

- Regularisation: 11 highly correlated features

- Monthly returns 2015 to 2018, download from Yahoo Finance

2015-2018	SPY	IWM	QQQ	EZU	RWO	VWO	DBC	HYG	VLUE	GOVT	LQD
SPY	1.0	0.8	0.9	0.7	0.6	0.6	0.3	0.7	0.9	-0.3	0.1
IWM	0.8	1.0	0.7	0.5	0.5	0.4	0.3	0.6	0.9	-0.4	0.0
QQQ	0.9	0.7	1.0	0.7	0.5	0.6	0.2	0.6	0.8	-0.2	0.1
EZU	0.7	0.5	0.7	1.0	0.5	0.7	0.2	0.7	0.6	-0.2	0.3
RWO	0.6	0.5	0.5	0.5	1.0	0.5	0.0	0.5	0.5	0.3	0.6
VWO	0.6	0.4	0.6	0.7	0.5	1.0	0.4	0.6	0.5	-0.1	0.4
DBC	0.3	0.3	0.2	0.2	0.0	0.4	1.0	0.5	0.4	-0.4	0.0
HYG	0.7	0.6	0.6	0.7	0.5	0.6	0.5	1.0	0.7	-0.1	0.4
VLUE	0.9	0.9	0.8	0.6	0.5	0.5	0.4	0.7	1.0	-0.5	0.0
GOVT	-0.3	-0.4	-0.2	-0.2	0.3	-0.1	-0.4	-0.1	-0.5	1.0	0.7
LQD	0.1	0.0	0.1	0.3	0.6	0.4	0.0	0.4	0.0	0.7	1.0

Sharpe Style Regression: Machine Learning Methods

1) Linear Regression: $min\{\|y - Fb\|_2^2\}$

- Short positions allowed (b_i <0), leverage allowed (b_i >1, $\sum b_i > 1$)

2) Constrained Regression: like LR, but b_i >0, $\sum b_i = 1$

- Long only, no leverage

3) LASSO: $min\{\|y - Fb\|_2^2 + \lambda_1\|b\|_1\}$

4) Ridge Regression: $min\{\|y - Fb\|_2^2 + \lambda_2\|b\|_2^2\}$

5) Elastic Net: $min\{\|y - Fb\|_2^2 + \lambda_1\|b\|_1 + \lambda_2\|b\|_2^2\}$

- Elastic Net is a blend of LASSO and Ridge Regression: $\lambda_1 + \lambda_2$= 1

- Task is to find b (and λ_1 and λ_2 where applicable)

  - All models without intercept
  - Penalty terms → variables need to be centred & scaled

Empirical Results: All 636 Mutual Funds

- Averages over 636 funds, INS: 2015-2018 (48 M), OOS: 2019 (12 M)

| 636 funds	short	INS: 2015-2018					OOS: 2019			
		Avg. R2	avg. vol	avg. return	avg. ret/vol	DW	Avg. R2	avg. vol	avg. return	avg. ret/vol
Fund	-	NA	9.8%	0.3%	0.08	NA	NA	9.3%	15.7%	1.82
Constrained Regression	CR	0.72	10.3%	3.3%	0.35	1.87	0.52	11.8%	22.3%	2.00
Linear Regression	LR	0.81	11.1%	2.6%	0.26	1.88	0.53	12.6%	19.8%	1.66
LASSO	LA	0.77	10.0%	3.0%	0.32	1.86	0.66	11.3%	19.6%	1.80
Ridge Regression	RR	0.77	10.0%	3.1%	0.33	1.83	0.66	11.4%	20.1%	1.80
Elastic Net	EN	0.81	11.1%	2.6%	0.26	1.87	0.53	12.6%	19.8%	1.66

- CR produces lowest R2(INS) @ 0.72, others are higher between 0.77 and 0.82

- Highest R2(OOS) by LA & RR @ 0.66, but highest return/vol by CR @ 2.00
  - Fund vol is 9.3% → replicants exceed funds' vol (between 11.3% and 12.6%)
  - LA & RR also produce lowest vol of the 5 ML methods

FitchLearning
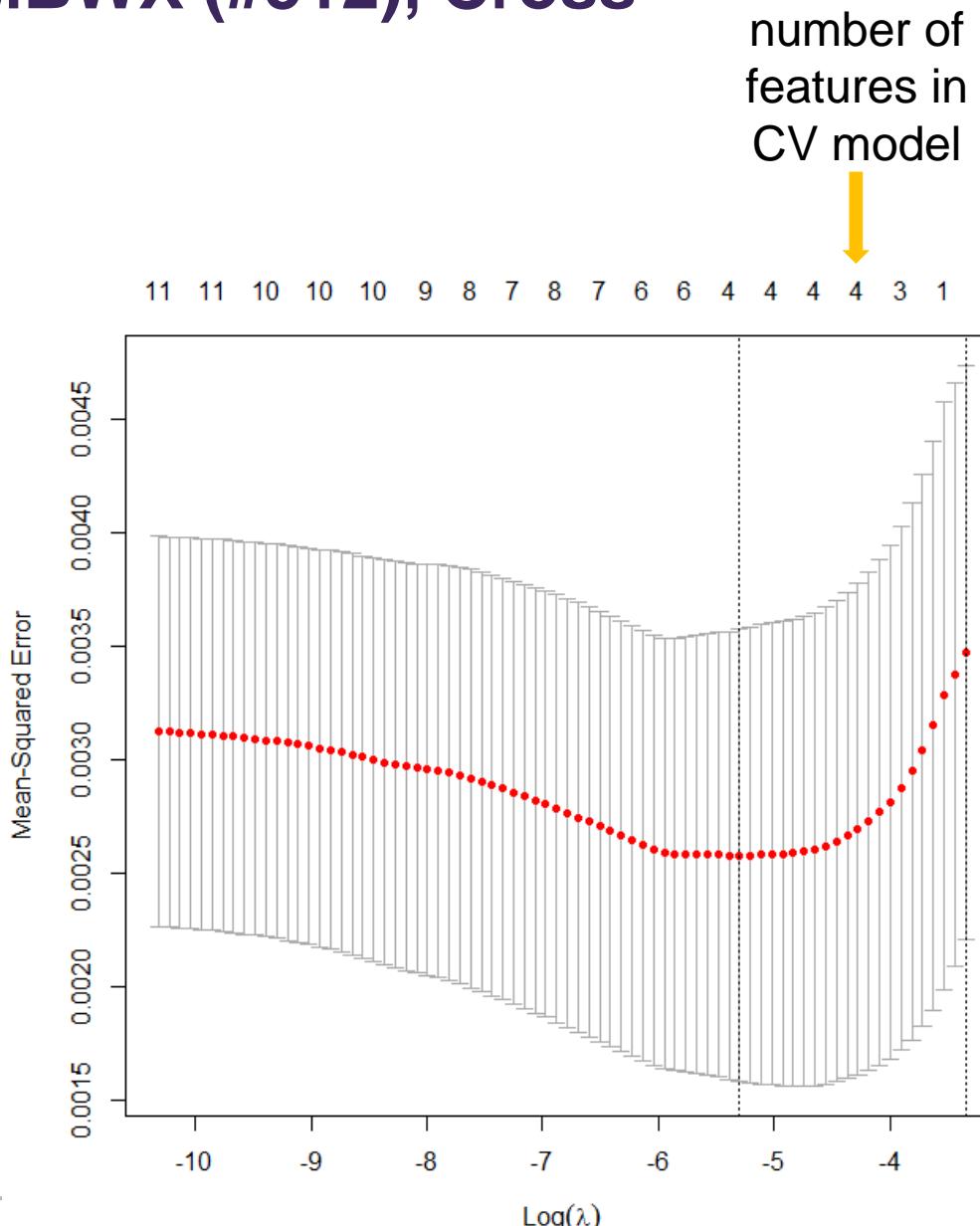
Example: 1 fund, UMBWX (#512)

- From Fidelity Fund Research:

  - https://fundresearch.fidelity.com/mutual-funds/view-all/14214L171#

- The fund normally pursues its objectives by investing in a diversified portfolio consisting primarily of equity securities of established companies either located outside the United States or whose primary business is carried on outside the United States. The equity securities in which the fund invests include common stocks, depositary receipts, preferred stocks, convertible securities, and warrants and other rights. It normally invests at least 80% of its net assets in equity securities.
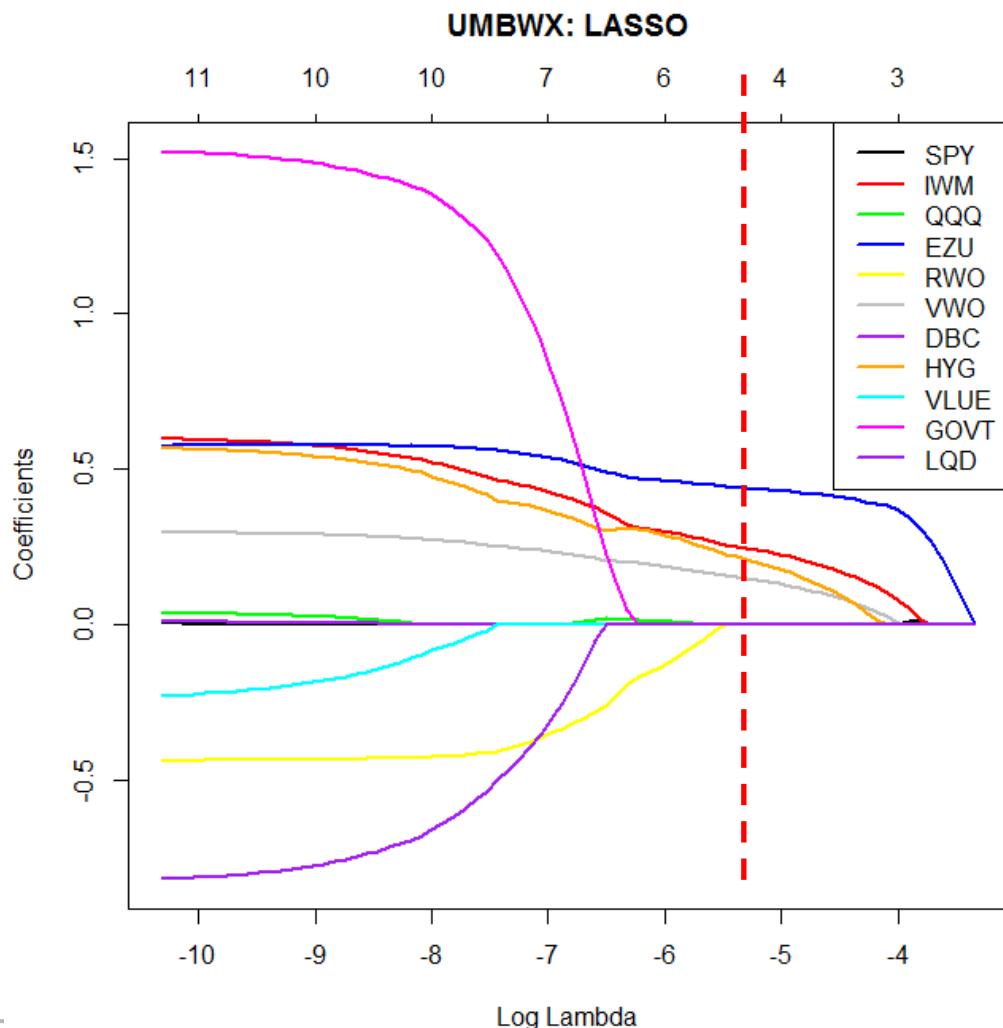
Example: 1 fund, UMBWX (#512), Cross Validation

number of features in CV model

- LASSO

- INS: 2015-2018 (48 M)

- Min. MSE @ log(lambda) = -5.298 or $\lambda_1 = 0.0050$

- Re-estimate LASSO with whole INS and this lambda as shrinkage factor

FitchLearning

Example: 1 fund, UMBWX (#512), Re-Estimation with Whole In-Sample Period (2015-2018)

- Red dashed line marks the LASSO with the lowest MSE based on cross validation

- This solution holds positive weights for 4 assets

UMBWX: LASSO

Legend:
- SPY (black)
- IWM (red)
- QQQ (green)
- EZU (blue)
- RWO (yellow)
- VWO (grey)
- DBC (purple)
- HYG (orange)
- VLUE (cyan)
- GOVT (magenta)
- LQD (purple)

FitchLearning

Example: 1 fund, UMBWX

- CR, LA, RR show most plausible weights ~100%

- LR, EN show shorts in RWO, VLUE, LQD

- UMBWX's Vol(INS) = 20%, replicants' vol 12 – 14%
  - Vol(OOS) at similar levels 14-16%

- Ret/Vol higher than fund for all models

- Highest R2(INS) @ 45% by LR, EN, R2(OOS) by CR, LA @ 79%

	Russell 2000			Eurozone Equities	Global Real Estate	EM Equities		USD High Yield			$ Corp Bonds	
	SPY	IWM	QQQ	EZU	RWO	VWO	DBC	HYG	VLUE	GOVT	LQD	Sum
UMBWX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CR	0%	30%	1%	47%	0%	20%	2%	0%	0%	0%	0%	100%
LA	0%	25%	0%	44%	0%	15%	0%	22%	0%	0%	0%	107%
RR	11%	14%	10%	23%	-5%	15%	4%	37%	6%	-3%	-2%	109%
LR	4%	61%	4%	57%	-44%	30%	1%	58%	-27%	153%	-83%	213%
EN	1%	60%	4%	57%	-44%	30%	1%	57%	-23%	152%	-82%	212%
Avg	3%	38%	4%	46%	-18%	22%	2%	35%	-9%	60%	-33%	148%

	INS: 2015-2018				OOS: 2019			
	R2	Return	Vol	Ret/Vol	R2	Return	Vol	Ret/Vol
UMBWX	NA	-14.3%	20.4%	-0.70	NA	20.7%	14.4%	1.43
CR	0.39	2.7%	12.4%	0.21	0.79	23.5%	14.6%	1.61
LA	0.39	2.9%	11.5%	0.25	0.79	23.2%	13.4%	1.72
RR	0.37	4.3%	10.8%	0.40	0.74	24.9%	13.6%	1.84
LR	0.45	4.5%	14.1%	0.32	0.60	27.0%	15.7%	1.72
EN	0.45	4.4%	14.0%	0.32	0.61	27.0%	15.8%	1.71
Avg	0.42	3.8%	12.4%	0.31	0.73	25.1%	14.5%	1.73

FitchLearning

Knowledge | Skills | Conduct

Example: 1 fund, UMBWX

- Why are allocation weights for EN and LR almost identical, while they differ strongly for EN vs. LA & RR?

  - Given that EN is a blend of LA & RR, we would expect EN to show allocations between LA & RR, but not as close to LR

- To estimate LA, RR & EN, R glmnet fits a generalized linear model (GLM) via penalised maximum likelihood:

RR Penalty + LA Penalty

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i l_i(y_i, \beta_0 + \beta^T x_i) + \lambda \left[\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right]$$

- λ controls the overall strength of the penalty

- Here are the model parameters as estimated for UMBWX:
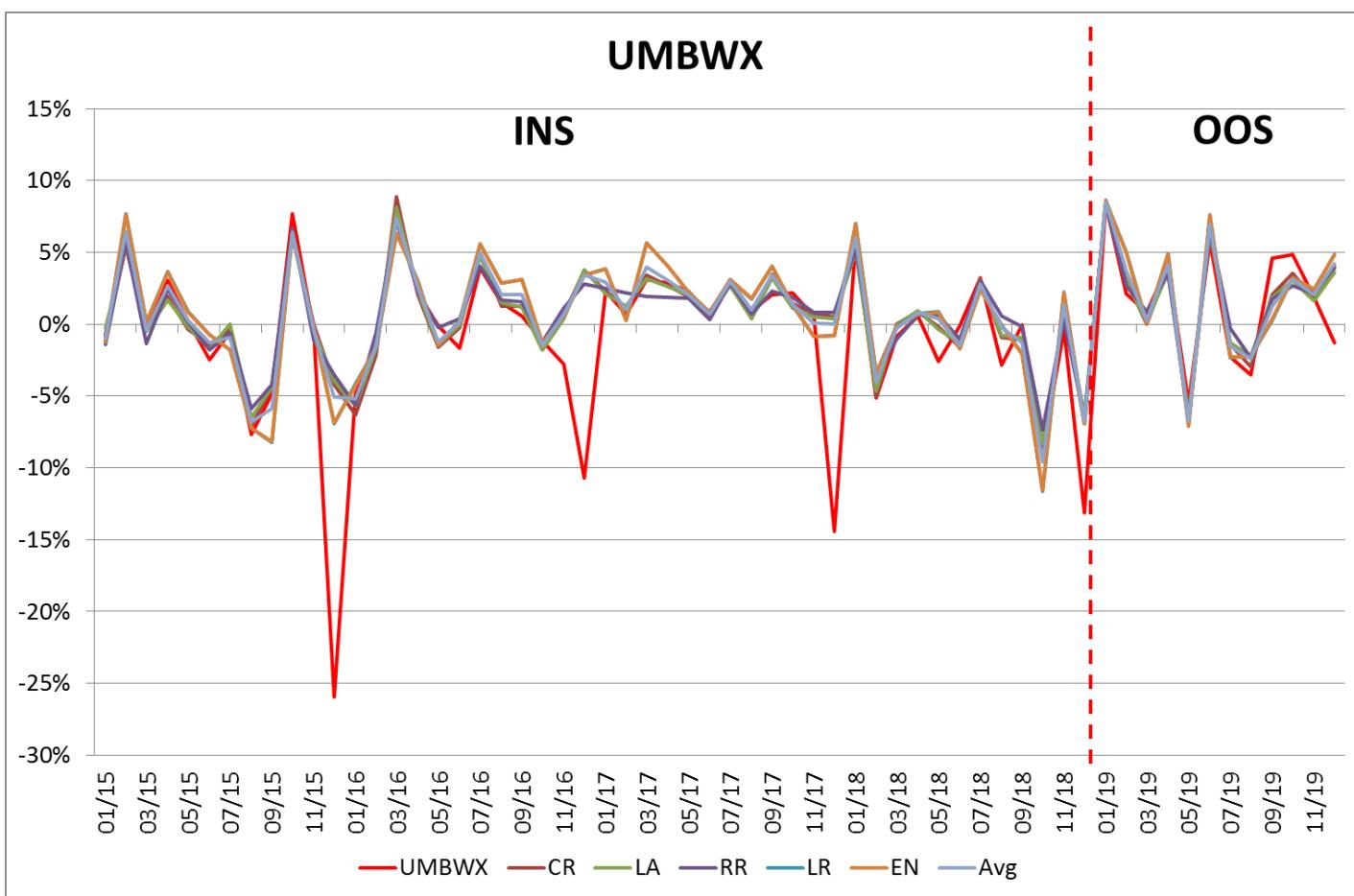
Model	Lambda	Alpha	Alpha why?
LA	0.0054882	1.00	by definition
RR	0.0631009	0.00	by definition
EN	0.0000316	0.35	based on CV

λ(EN) is close to 0 → this pushes the penalty term to almost 0 and hence gives results close to LR
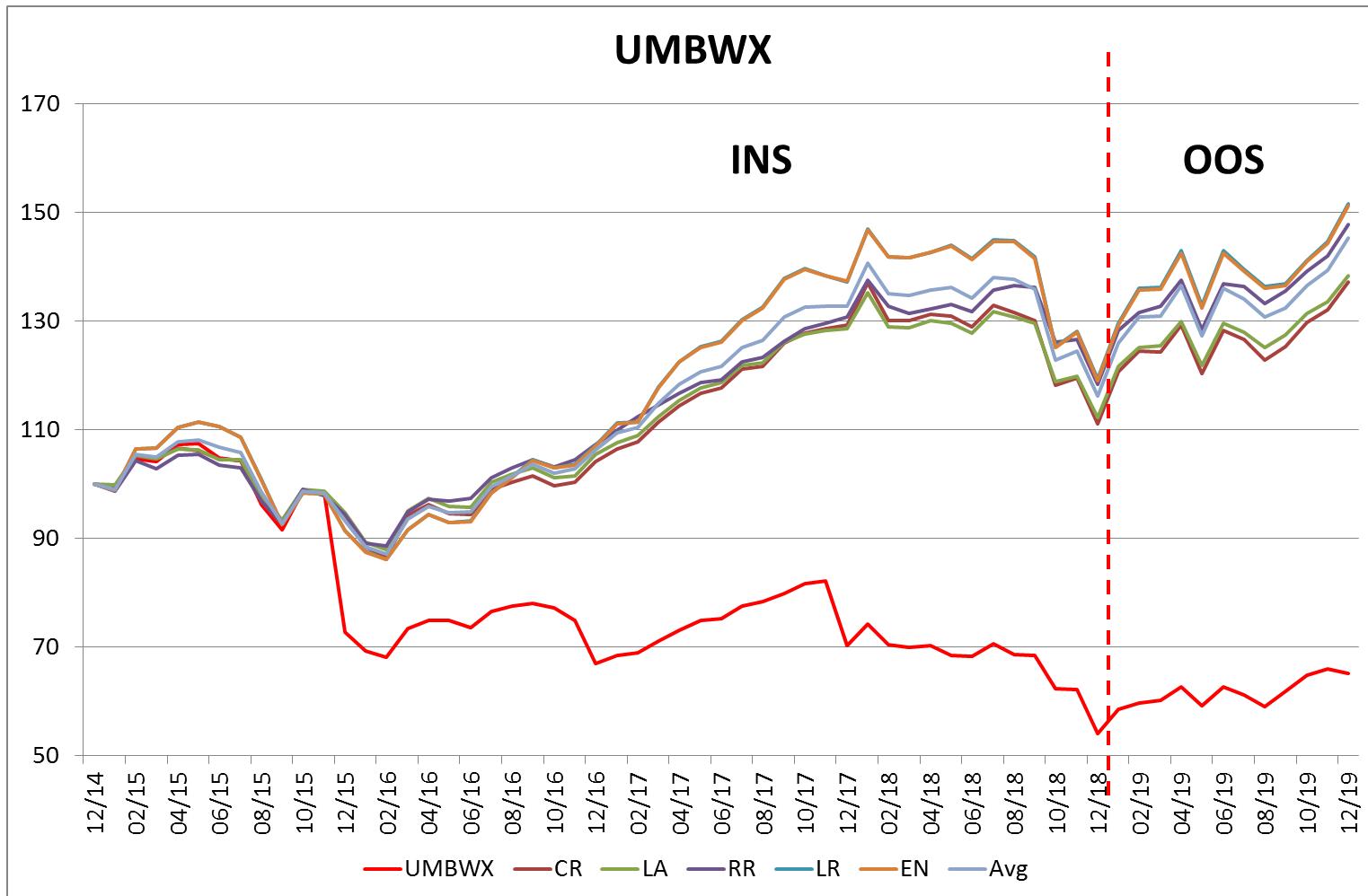
# Example: 1 fund, UMBWX (#512)

- Some returns cannot be explained by the models, e.g., 12/15, 12/16, 12/17

- Those months lead to a low R2 of 39-45%…

UMBWX

INS   OOS

Legend: UMBWX, CR, LA, RR, LR, EN, Avg

Example: 1 fund, UMBWX (#512)

- … and to underperformance of the fund vs. the models

Other Interesting Funds And Their Models

- **AHLPX (#91)**
  - Managed Futures fund (frequently changing long & short positions), hard to explain with standard risk factors

- **INDAX (#525)**
  - Fund invests in Indian equites
  - Models show how exposure can be replicated with standard risk factors, like US equities

- **BLPIX (#16)**
  - Strategy can well be explained by models, R2(INS & OOS) 90 – 100%
  - Models outperform fund

R Code: "CQF Sharpe Style Regression.R"

- 2 files belong to this use case
  - "CQF Sharpe Style Regression.R", "Data Dump All.csv"

- R code loops over all 636 funds
  - To analyse individual funds, run code until "# 1st STOP point" and follow instructions there

- Run time for all funds ca. 5 min
  - Ca. 200 lines of code

- Sometimes the ML algorithms do not converge and throw an error
  - Restart can help, but no guarantee!
  - If you only want to check a few funds, change variable x_funds to 5 or 10 (originally set to 636)

- Output files:
  - "Coeff_LA.csv": Estimated coefficients for LASSO. For each fund 1 column
  - Dto. for Ridge Regression (RR), Linear Regression (LR), Elastic Net (EN)

Natural Language Processing for Sentiment Analysis of ESG Company Reports

Knowledge | Skills | Conduct

Text Mining ESG Reports

- Large companies publish extensive ESG reports (100+ pages)

- Our self-developed tool uses the ESG definitions of the Sustainability Accounting Standards Board (SASB) to analyse the reports:

  - On which ESG areas (e.g., environment, human capital) has each company focused?

  - Are there differences between companies? Has the focus shifted over time?

  - Is the general tone of the ESG report positive or negative? Does a company report with a particularly positive or negative sentiment in specific ESG areas?

- Agenda:

1. Assign words & sentences to the 5 SASB categories

2. Sentiment Analysis with Vader

1) Assign words & sentences to the 5 SASB categories: SASB Materiality Map

- https://materiality.sasb.org/

Dimension	General Issue Category	Consumer Goods	Extractives & Minerals Processing	Financials
		Click to expand	Click to expand	Click to expand
Environment	GHG Emissions			
	Air Quality			
	Energy Management			
	Water & Wastewater Management			
	Waste & Hazardous Materials Management			
	Ecological Impacts			
Social Capital	Human Rights & Community Relations			
	Customer Privacy			
	Data Security			
	Access & Affordability			
	Product Quality & Safety			
	Customer Welfare			
	Selling Practices & Product Labeling			
Human Capital	Labor Practices			
	Employee Health & Safety			
	Employee Engagement, Diversity & Inclusion			
Business Model & Innovation	Product Design & Lifecycle Management			
	Business Model Resilience			
	Supply Chain Management			
	Materials Sourcing & Efficiency			
	Physical Impacts of Climate Change			
Leadership & Governance	Business Ethics			
	Competitive Behavior			
	Management of the Legal & Regulatory Environment			
	Critical Incident Risk Management			
	Systemic Risk Management			

1
2
3
4
5

Fit

SASB Materiality Map: Dimensions

- 5 dimensions on Layer 1 and 26 sub-dimensions on Layer 2

- We focus on the 5 dimensions of Layer 1

- More granular analysis of Layer 2 would also be possible

- See file "CQF ESG.csv"

Layer 1	Layer 2	Layer 3 (Examples)
Environment (Dim. 1)	GHG Emissions	greenhouse, gas
	Air Quality	air, factories
	Energy Management	grid, reliance
	Water & Wastewater Management	water, wastewater
	Waste & Hazardous Materials Management	hazardous, waste
	Ecological Impacts	ecosystems, biodiversity
Social Capital (Dim. 2)	Human Rights & Community Relations	relationship, businesses
	Customer Privacy	secondary, purposes
	Data Security	collection, retention
	Access & Affordability	groups, needs
	Product Quality & Safety	unintended, products
	Customer Welfare	welfare, health
	Selling Practices & Product Labeling	social, transparency
Human Capital (Dim .3)	Labor Practices	labor, laws
	Employee Health & Safety	workplace, injuries
	Employee Engagement, Diversity & Inclusion	hiring, promotion
Business Model & Innovation (Dim . 4)	Product Design & Lifecycle Management	lifecycle, packaging
	Business Model Resilience	long-term, model
	Supply Chain Management	human, rights
	Materials Sourcing & Efficiency	supply, chains
	Physical Impacts of Climate Change	physical, climate
Leadership & Governance (Dim . 5)	Business Ethics	fraud, corruption
	Competitive Behavior	monopolies, excessive
	Management of the Legal & Regulatory Environment	conflicting, corporate
	Critical Incident Risk Management	scenario, planning
	Systemic Risk Management	large-scale, weakening

FitchLearning

Text Mining ESG Company Reports

- SASB provides verbal explanations of each of its ESG dimensions

- From those explanations, we extract keywords (319) that are specific for each dimension
  - For example, GHG emissions (layer 2): keyword "greenhouse"
  - "greenhouse" belongs to layer 1, Environment

- We screen the ESG report for "greenhouse" and count the number of occurrences, e.g., 35

- Assuming 3500 terms in the report, "greenhouse" would contribute 1% to the SASB-relevant content
  - By analogy for the other 318 keywords

- This way we can analyse which keywords of which of the 5 dimensions contribute most / least to the ESG report

Text Mining: Stem Words

- Original sentence (Barclays ESG Report 2019, p. 22):

  - Our financing volume is tracked and screened using Barclays' Impact Eligibility Framework, which provides clear social and environmental inclusion criteria to track and categorise financing volumes…

- After treatment (stem words, removal of punctuation & stopwords, …):

  - Our financ volum is track and screen use Barclays' Impact Eligibl framework which provid clear environment and social inclus criteria to track and categoris financ volum…

FitchLearning
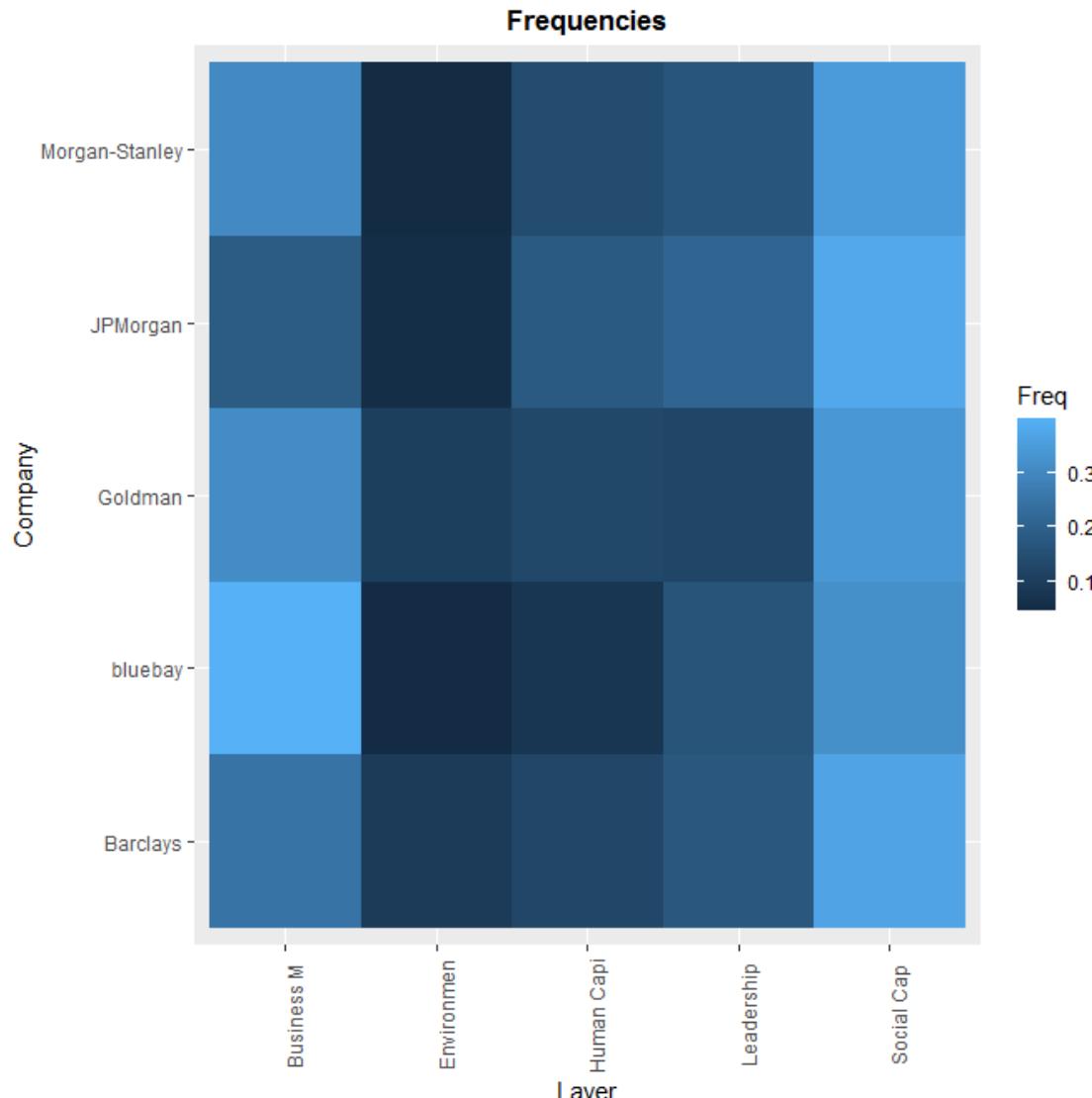
Terms in the Document And Distribution Across SASB Dimensions

- Count each term in each sentence with at least 1 SASB dimension

- Example: the whole doc contains 15 SASB terms

  - For example, SASB dimension 1 holds 3 / 15 = 20% of the terms in the report

  - Sentence 2 does not contribute to SASB relevance

Sentence #	#terms					#terms(i)
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	
1	2	0	3	0	0	5
2	0	0	0	0	0	0
3	1	0	0	0	0	1
4	0	7	0	0	2	9
Sum	**3**	**7**	**3**	**0**	**2**	**15**
	20%	47%	20%	0%	13%	
Assigned to SASB Dim.						

Empirical Results: Term Frequencies

- There are differences in the frequencies of SASB dimensions for reports of companies

  - Social Capital (e.g., Customer Privacy) is the most discussed dimension, followed by Business Model & Innovation
  - Environment discussed least
  - Bluebay talks most about Bus. Model & Inn., followed by Goldman, JP Morgan least

- See "CQF ESG.R"

  - Run time ca. 5 min

FitchLearning

2) Sentiment Analysis with Vader

- VADER: Valence Aware Dictionary and sEntiment Reasoner

- Sentiment Analysis with Vader (Hutto / Gilbert (2014))
  - Specifically designed to process Social Media including emoticons [☺, ☹]

- Simple application in R: output is value in [-1, 1], i.e., not only binary classification good / bad, but a value between -1 and +1
  - -1: very bad sentiment
  - +1: excellent sentiment
  - 0: neutral

See "CQF Vader Example.R":

String	Vader Compound
The food here is great.	0.625
The food here is great!	0.659
The food here is great!!	0.689
The food here is great, but the service is horrible.	-0.494

VADER: Valence Aware Dictionary and sEntiment Reasoner

- Vader is a lexicon and rule-based sentiment analysis tool

- Specifically designed to sentiment expressed in social media

- Works well on texts from other domains

- More examples for text sentiment analysis in Kalamara et al. (2020)

Training Vader

- Humans were hired via Amazon Mechanical Turk to provide a sentiment of several thousand Twitter tweets, movie reviews and newspaper articles

| ROFL | **Description:** Rolling On Floor Laughing |

- [-1] Slightly Negative   - [-2] Moderately Negative   - [-3] Very Negative   - [-4] Extremely Negative
- [0] Neutral (or Neither, N/A)
- [1] Slightly Positive   - [2] Moderately Positive   - [3] Very Positive   - [1] Extremely Positive

Figure 2: Example of the interface implemented for acquiring valid point estimates of sentiment valence (intensity) for each context-free candidate feature comprising the VADER sentiment lexicon. A similar UI was used for all rating activities described in sections 3.1-3.4.

Source: Hutto / Gilbert (2014)

FitchLearning

Valence of the Document

- We derive a valence score for each sentence with at least 1 SASB term in the ESG reports

- How to weigh the valence of one sentence with multiple terms
  - Each term is given a V = 1 / (sum of terms in this sentence)
  - If there is only 1 term (j = 1) in sentence i: $V(i, j) = 1 / 1 = 1$
  - The more terms are in the sentence, the less weight for $V(i, j)$

- The $V(i, j)$ are then assigned to the SASB dimensions, see next slide
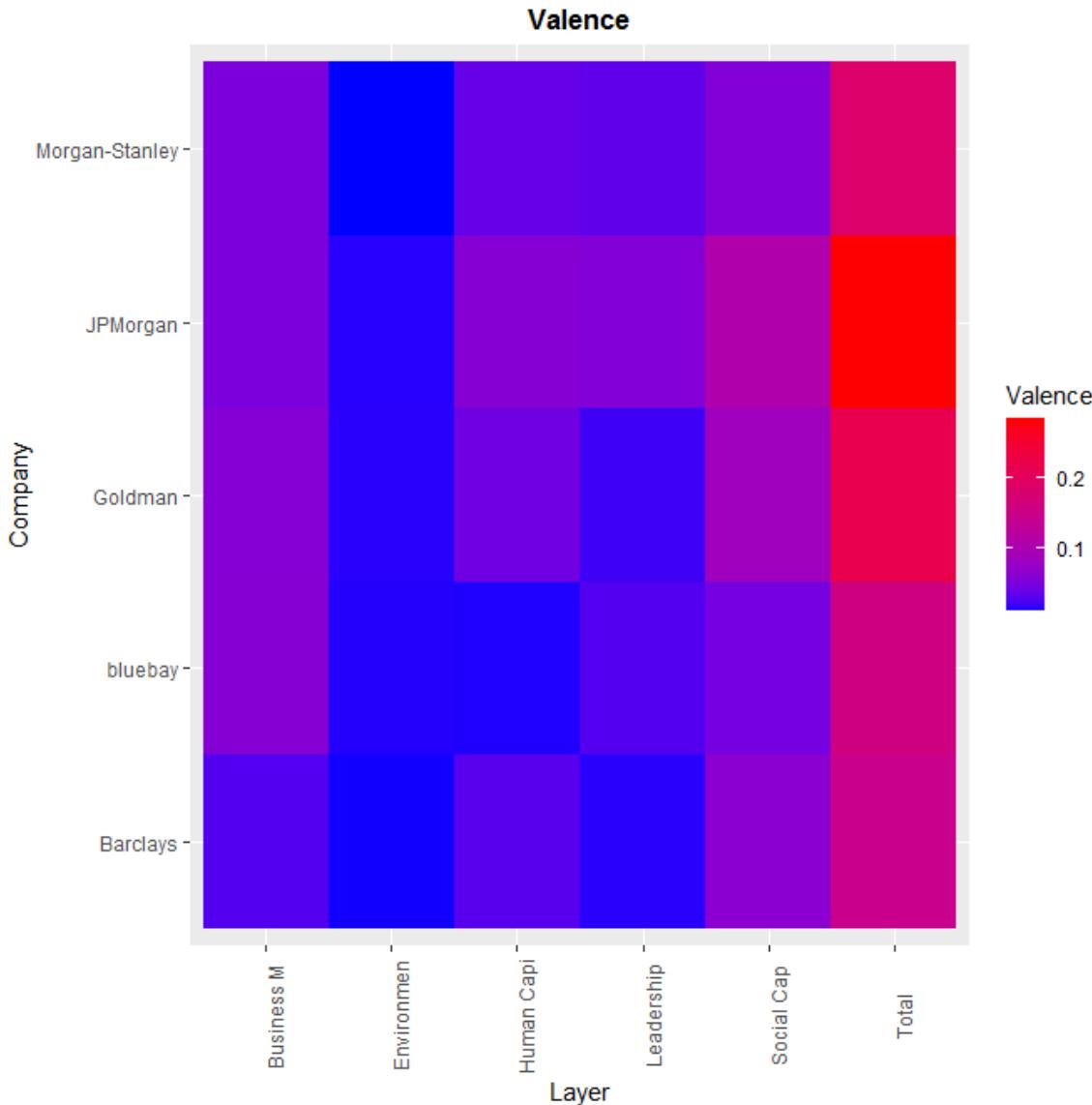
Valence of Document And Distribution Across SASB Dimensions

- To determine the SASB-relevant V, we only calculate the V(i, j) with at least 1 SASB-relevant term

  - In our example, there are 3 SASB-relevant sentences in the table with V = 0.20

  - Dim 1 contributes 0.13 to this

  - Dim 2 and 5 contribute negatively

  - Next slide shows those values as charts for several reports

			#relevant sentences:			3
Sentence #	**Dim 1**	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**	**V**
1	0.07	0	0.10	0	0	0.17
2	0	0	0	0	0	0
3	0.07	0	0	0	0	0.07
4	0	-0.03	0	0	-0.01	-0.03
Sum	**0.13**	**-0.03**	**0.10**	**0**	**-0.01**	**0.20**
						SASB relevant V

Empirical Results: Valence Scores

- JP Morgan achieves highest Total Valence Score, followed by Goldman

- The last 2 Valence scorers are Barclays and Bluebay

- Dimension Social Capital is most favourably discussed, Environment least

References

- Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191:192–213.

- Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120, 70-83.

- Boot, T. and Pick, A. (2020). Does Modeling a Structural Break Improve Forecast Accuracy? Journal of Econometrics, 215(1):35–59.

- Burman, P., Chow, E., & Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, *81*(2), 351-358.

- Carriero, A., Galvão, A. B., and Kapetanios, G. (2019). A Comprehensive Evaluation of Macroeconomic Forecasting Methods. International Journal of Forecasting, 35(4):1226 – 1239.

- Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting?. arXiv preprint arXiv:2008.12477.

- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.

- Hardy, R. S. (2003). Style analysis: A ten-year retrospective and commentary. The Handbook of Equity Style Management, 109-128.

FitchLearning

References

- Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).

- Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making text count: economic forecasting using newspaper text.

- Kaplan, P. (2003). Holdings-based and returns-based style models. Chicago, IL: Morningstar, Inc.

- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), 1-21.

- Kazemi, H., Wilkens, K. (2020): The CAIA Endowment Investable Index, Q1 2020 AIAR Vol. 9 Issue 1, 82-83.

- Kazemi, H., Wilkens, K. (2020): A Simple Approach To The Management of Endowments, Q1 2017 Vol. 6 Issue 1, 1-3.

- Kim, H. H. and Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. International Journal of Forecasting, 34(2):339–354.

FitchLearning

Knowledge | Skills | Conduct

References

- Lawrence, A. R., Kaiser, M., Sampaio, R., & Sipos, M. (2020). Data Generating Process to Evaluate Causal Discovery Techniques for Time Series Data.

- Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008). Kernel-Granger causality and the analysis of dynamical networks. Physical review E, 77(5), 056215.McCracken, M. W., and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics, 34(4), 574-589.

- Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. Journal of Econometrics, 137(1), 134-161.

- Pesaran, M. H., Pick, A., & Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. Journal of Econometrics, 177(2), 134-152.

- Sharpe, W. F. (1992). Asset allocation: Management style and performance measurement. Journal of Portfolio Management, 18(2), 7-19.

FitchLearning