CaseStudy_QuantitativeResearcher_TaiQuoc

Overview

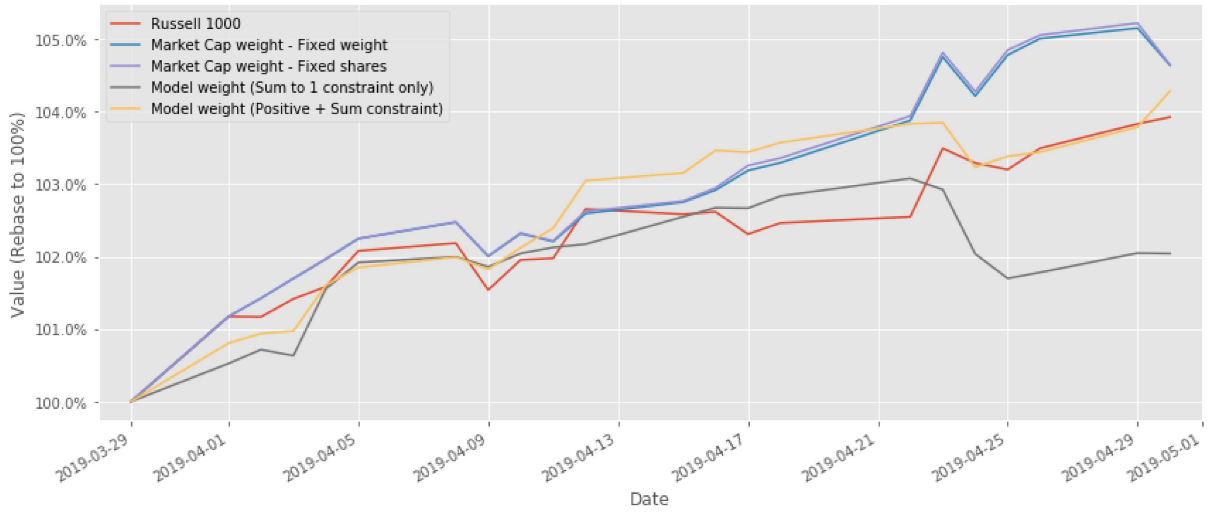
Since the portfolio constituents are already selected, the main focus is on weight optimization

Multiple aspects of index replication and model estimation are taken into account such as:

- Data quality
- Tracking error measurements: RMSE, Standard Deviation, Cumulative Return
- Portfolio structure: Fixed weight or fixed shares
- Portfolio constraints: Positive weights, Sum to 100%
- Estimation objective: L1, L2, CVAR (Conditional value-at-risk) loss
- Regularization
- Rolling validation

To incorporate all these aspects in optimization, customized optimization algorithms are implemented and tested extensively

The final selected models are based on trade-off between model performance and model stability. While no information after Mar-2019 are used in model estimation (and selection), the final models show promising results in Apr-2019



Tracking Error (Measurement definitions below):

	Market Cap - Fixed weight	Market Cap - Fixed shares	Model (Only sum to 1)	Model (Positive + Sum to 1)
ETQ	0.006342	0.006610	0.006359	0.004161
MAD	0.002237	0.002281	0.003247	0.002564
RMSE	0.002901	0.002991	0.004283	0.003245
TESTD	0.002883	0.002973	0.004193	0.003241

Reference

Most of the analysis are based on the Denis Karlow dissertation "Comparison and Development of Methods for Index Tracking" on 2012. This dissertation will hereby refer as Karlow, 2012. The materials are available here (https://d-nb.info/1054242275/34) or on my github (https://github.com/taiqm/Upload/blob/master/Karlow.pdf)

Data

Prices and shares data are pulled from Yahoo Finance via yfinance/pandas_datareader packages

Russell 1000 constituents information are pulled from Norgate Data (https://norgatedata.com/ (https://norgatedata.com/))

Adjusted closed prices will be used for return calculation while closed prices will be used for market cap calculation (Yahoo closed prices already include split adjustment corporate actions but no dividend adjustment)

Data on daily frequency will be used in this research due to:

- The case study objective is to replicate daily closing
- Portfolio only has 25 assets so daily data provides enough samples without going to far back
- Higher frequency may contains more noise (Karlow, 2012 - Page 54)

Data period will be from July-2017 to Mar-2019 (since Russell rebalances on June annually and Norgate Data only available from 2018)

Tracking Error measurements

Multiple different tracking error measurements are calculated as followed (Karlow, 2012 - Page 58, 59, 65):

- Tracking error (TE): $TE_t = R_{I,t} - R_{P,t}$
- Mean absolute difference (MAD): $MAD = \frac{1}{T} * \sum_{t=1}^{T} |TE_t|$
- Root mean square error (RMSE): $RMSE = \sqrt{\frac{1}{T} * \sum_{t=1}^{T} TE_t^2}$
- Expost tracking quality (ETQ): $ETQ = \frac{1}{T} * \sum_{t=1}^{T} |R_{I,t}^c - R_{P,t}^c|$. With R_t^c is compound return up to time t

Portfolio structure

Both fixed weight and fixed shares portfolio structure are incoporated in the model selection framework

However due to fixed shares optimization cannot be solved by the solver directly, an adjustment method is applied on return to transform variable weight (i.e fixed shares) to fixed weight (Karlow, 2012 - Page 72)

The adjustment method (Karlow, 2012 - Page 75):

Correction at the end of the period	$R_{P,t} = \sum_{i=1}^{N} w_{i,T} \frac{V_{I,T}}{V_{I,t}} \frac{S_{i,t}}{S_{i,T}} r_{i,t}$

Portfolio constraints

There are two constraints considered:

- All weights sum to 100%
- All weights positive

The weight sum constraint is always enforced while positive weights constraint can be relaxed (Allow short sell)

Estimation objective

There are 3 different loss objectives are considered for model estimation:

- L1 loss: Minimize MAD
- L2 loss: Minimize RMSE
- CVAR loss: Minimize Conditional value-at-risk

$$CVAR_\alpha = \varepsilon + \frac{1}{T(1-\alpha)} \sum_{t=1}^{T} \max(0, |TE_t^*| - \varepsilon),$$

Regularization

To avoid overfitting and increase model generalization ability, regularization term is added in estimation

Due to weight sum constraint (and positive constraint), LASSO and Elastic-net are not suitable for regularization here

Therefore, Ridge regularization term is added. Regularized loss = Normal loss + $\lambda * \sum_{i=1}^{n} weight_i^2$

Rolling validation

To check model stability, rolling window k-fold validation is always conducted on all models

In-sample period will be 250 trading days

Validation period will be 25 trading days

With data from July-2017 to Mar-2019, we have 17-fold validation

Model selection

Take into account all the aspects above, 4386 models are generated based on various hyperparameters

Baseline models such as Market-capitalization weight are also included

Due to the models instability, model performance is converted to relative ranking within each validation iteration

The ratio between relative ranking average and relative ranking standard deviation are calculated

That ratio is ranked again and average across TE measurement to get the final score to select final model

Final results

Due to the constraints, there are two sub-groups so there are two final models:

1. With both sum and positive constraint:

    - Loss: L2
    - Regularization (Lambda): 0
    - Portfolio structure: Weight
    - Standardize: True

1. With only sum constraint:

    - Loss: L2
    - Regularization (Lambda): 0.004115
    - Portfolio structure: Shares (Adjusted_weight)
    - Standardize: True

Weight estimated (For Apr/2019 portfolio):

	Sum and Positive	Sum only
MSFT	2.666852e-10	-0.115331
AMZN	3.013651e-10	-0.009294
AAPL	3.775657e-02	0.095453
GOOG	5.062646e-10	0.015907
FB	1.222368e-01	0.142682
JNJ	3.470668e-02	0.096019
GOOGL	5.313660e-10	-0.001180
XOM	1.289621e-09	-0.036075
JPM	5.647675e-10	-0.168090
WMT	1.257573e-01	0.136257
V	8.003931e-10	0.119526
PG	9.885232e-02	0.121495
VZ	1.484465e-01	0.168916
BAC	9.879453e-02	0.230389
PFE	1.633320e-03	0.051716
MA	4.074717e-10	-0.061044
UNH	2.346146e-09	0.023735
INTC	3.538861e-02	0.070056
CVX	3.039191e-02	0.075665
CSCO	2.469115e-10	-0.133383
T	1.004719e-01	0.121642
BA	2.896280e-02	0.052884
MRK	7.637859e-10	-0.094021
HD	7.771377e-10	-0.034178
KO	1.366007e-01	0.130256

Code

Out[57]: Click here to toggle on/off the raw code.

```
[2020-03-10 01:19:22.219936] WARNING: Norgate Data: Unable to obtain valid status from Norgate Dat
a - perhaps NDU is not running?
[2020-03-10 01:19:22.224892] WARNING: Norgate Data: **PACKAGE VERSION WARNING*** You have version
(1.0.34) of the norgatedata package installed.  A newer version 1.0.35 is available and is a recom
mended upgraded.
[2020-03-10 01:19:22.226884] INFO: Norgate Data: NorgateData package v1.0.34: Init complete



MAE (Rescaling impact): 8.160305069134528e-06
MAE (Ridge impact) : 2.6382554356742425e-06
No standardized - Sum to 1 : 0.9999999999999984, All positive: True
Standardized - Sum to 1 : 1.0000000000000002, All positive: True
MAE (L1 loss): 0.00016639042197412988
Diff insample   : 8.392601094266237e-06
Diff validation : 3.687989620357931e-06
```