Master's Thesis

Bounded Clustering Approach to Global Minimum Variance Portfolio (GMVP)

Jinwoo Park

HCC Lab

2019.12.30

HUMAN
CENTERED
COMPUTING
LABORATORY

Table of Contents

HUMAN
CENTERED
COMPUTING
LABORATORY

1. Introduction

HUMAN
CENTERED
COMPUTING
LABORATORY

Safe Investment with a GMVP

- Return & risk is a trade-off relationship
  : a portfolio with diversified assets can lower the risk without damaging return
- Recent stock market crashes arouse interests in **'safe investment'**
  : **Global Minimum Variance Portfolio** (GMVP) serves the need
- GMVP : targets to take as little volatility as possible without considering return[1)]
  → Needs only covariance matrix of return of assets as the input

Estimating the population covariance matrix

- The population covariance matrix : unknown in practice
  → needs to be estimated from historical data
- Sample covariance matrix commonly used : an unbiased estimator

Ex)

Daily returns are expressed in percent (%)

	AAPL	GOOGL	MSFT	FB
2017-01-31	-0.230206	-0.441839	-0.736988	-0.503894
2017-02-01	6.098063	-0.603519	-1.655066	2.232965
2017-02-02	-0.170874	0.370443	-0.644857	-1.793890
2017-02-03	0.427916	0.228534	0.807345	0.107001
2017-02-06	0.937403	0.181679	-0.062814	0.824553
2017-02-07	0.951723	0.926219	-0.329981	-0.166591
2017-02-08	0.387744	0.078386	-0.141889	1.790049
2017-02-09	0.287792	0.021690	1.136722	-0.044709
2017-02-10	-0.226552	0.577067	-0.093662	0.037274

→

	AAPL	GOOGL	MSFT	FB
AAPL	0.013850	0.005070	0.004824	0.006598
GOOGL	0.005070	0.009058	0.005742	0.007006
MSFT	0.004824	0.005742	0.008957	0.005368
FB	0.006598	0.007006	0.005368	0.012694

Sample covariance matrix computed from daily return of stocks

Problem: High Estimation Error

- Estimation error can be high in the sample covariance matrix & its inverse
  : Especially when number of assets is comparable to number of observation
- Estimation error cause problems
  : 1) Out-of-sample risk much higher than in-sample counterpart
  2) Worse out-of-sample performance than not optimized portfolio result

Baseline Model: Clustering Stocks

- **Divide and conquer** (Two-stage portfolio optimization)
  : - Clustering stocks for less number of features in a covariance matrix

- Stock clustering methods that have been proposed:
  1. Non-price information : accounting figures, Industry sectors
  2. Price : daily returns of stocks

Research Goal

Motivation

- Clustering based on non-price information (industry sector, etc)
  : Does not need to be related with stock price
  → Clustering quality deteriorates
- Clustering based on price
  : Too many stocks might be grouped in one cluster
  → Estimation error can still remain high after clustering

- Research Goal
  - **Find a price-based clustering algorithm to improve the portfolio performance more than methods already proposed**
  - **Should take care of both estimation error and clustering quality**

2. Related Works

1. Global Minimum Variance Portfolio (GMVP)

- **Introduced as a portfolio for taking least amount risk** *(H. Markowitz 1952)*
  - Only covariance matrix used, so less problematic to estimate (Merton, 1980)
  - Volatility of financial data shows a similar patter as it has historically (Engle 1982, Bollerslev 1986)
- **Better out-of-sample performance than optimized for return** (Jorion 1991, Chopra and Ziemba 1993)
- **Due to the estimation error of covariance matrix, GMVP might fail**
  - GMVP might not outperform randomly selected portfolio (Frankfurter et al, 1971)
  - A naïve equally weighted portfolio might outperform GMVP (DeMiguel et al, 2009)

2. Attempts to Decrease Estimation Error

- **Single or multi-factor models** : Structured but can be biased heavily (W. Sharpe 1963, E.F. Fama and et al. 1993)
- **Shrinkage estimator** : 'Compromise' between the unbiased and structured (Ledoit and Wolf 2003, Bodnar et al 2014)
- **Clustering approach** : 'Divide and conquer'
  - **Non-price information**
    - **Accounting figures** *(K. Marvin, 2015)*
    - **Industry sectors** *(M. Claeson, 2017)*
  - **Price**
    - **Same cluster if Pearson correlation coefficient of returns > 0.2** *(Z. Ren, 2005)*
    - **K-means clustering on daily return of stocks** *(S.R. Nanda et al. , 2010)*

HUMAN
CENTERED
COMPUTING
LABORATORY

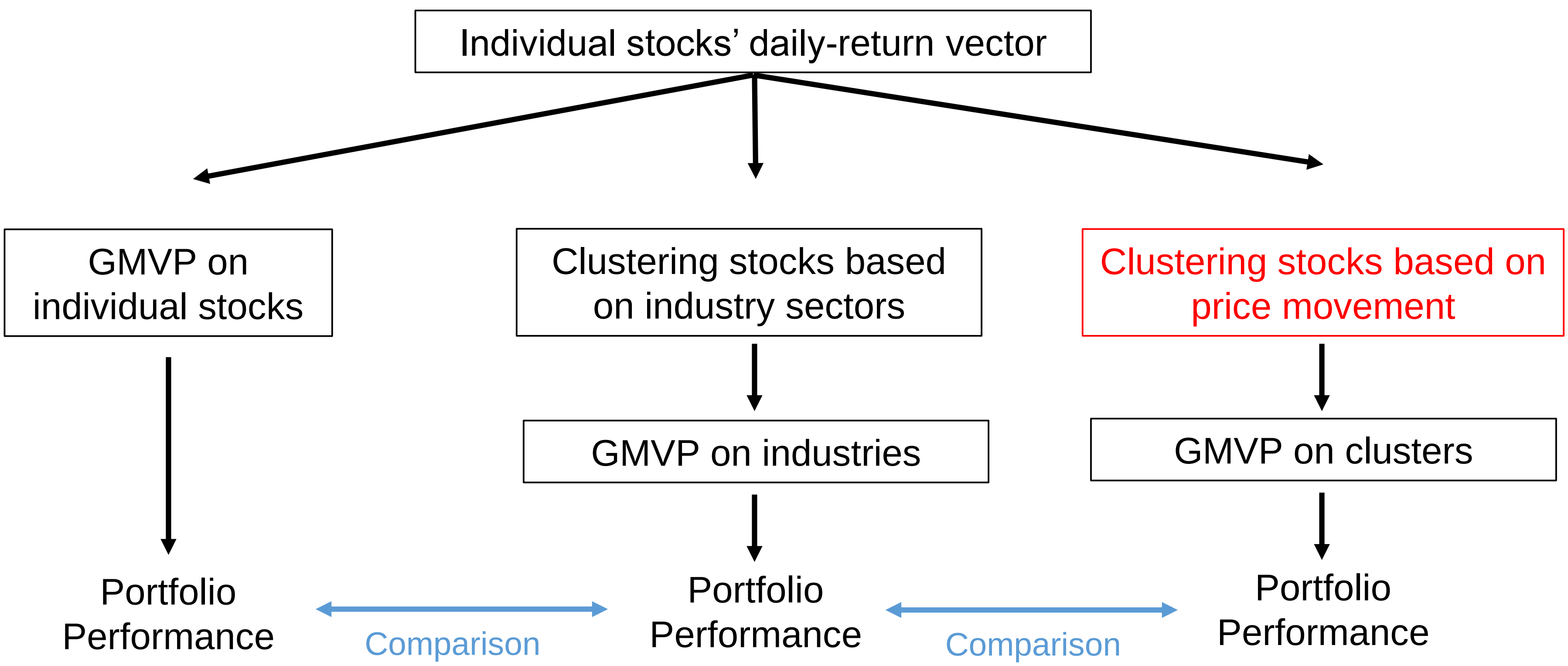
3. Clustering Algorithms

- Roughly divided into partitioning and hierarchical clustering
  1. Partitioning clustering: **K-means clustering** *(J.B. Macqueen 1966)*
  2. Hierarchical clustering : **Agglomerative hierarchical clustering** (Y. Rani 2013)
- **Constrained clustering on size**
  - K-means clustering with minimum cluster size (P.S. Bradley 2000)
  - K-means clustering with maximum cluster size (N. Ganganath et al. 2014)

3. Method & Experiment

Experiment Overview

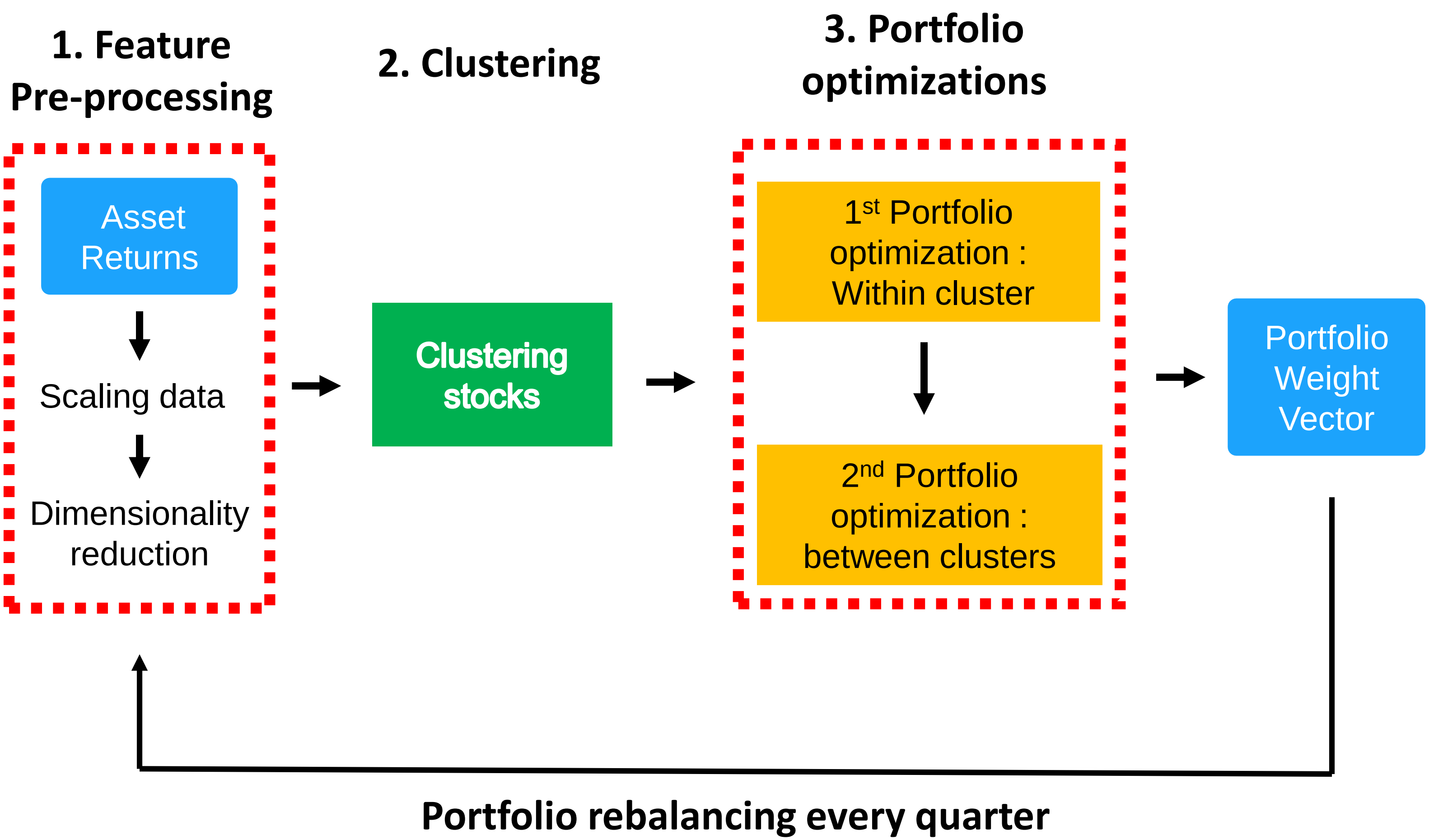
Compare the portfolio performance of each portfolio optimization method



Portfolio Performance Measures

- **Adjusted return**
  1. **Sharpe ratio** : Reward to risk (standard deviation) ratio.
  2. **Sortino ratio** : Reward to risk (downside standard deviation) ratio
- **Risk**
  3. **Standard deviation** : volatility of return
  4. **Downside Standard deviation** : standard deviation of return below threshold
  5. **Maximum Drawdown** : Maximum loss from a peak to a trough of a portfolio
  6. **Conditional Value at Risk** : Weighted average of the extreme losses in the tail of the distribution of possible or historical returns

Experiment Flowchart



Experiment procedure

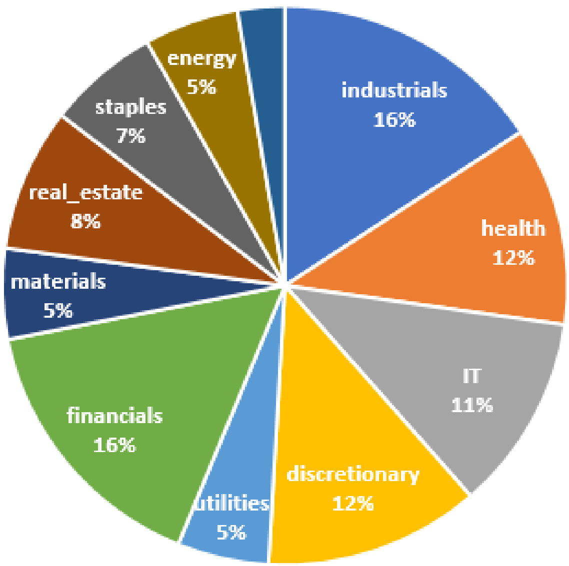
- **Scaling** : - Euclidean distance used in clustering algorithm
  - As such, clustering algorithm might perform when all features contribute equally (Standard scaling)
- **Dimensionality reduction**:
  - PCA : To reduce the noise of data
  - t-sne : To add non-linearity while reducing the dimensions
- **Clustering algorithm**:
  - K-means clustering / Hierarchical clustering
  - Bounded K-means clustering

Scaling		**Dim. reduction**		**Clustering**
- Standard Scaling (Normalization) - None	→	- PCA - t-sne - None	→	- K-means Clustering - Hierarchical Clustering - Bounded K-means Clustering

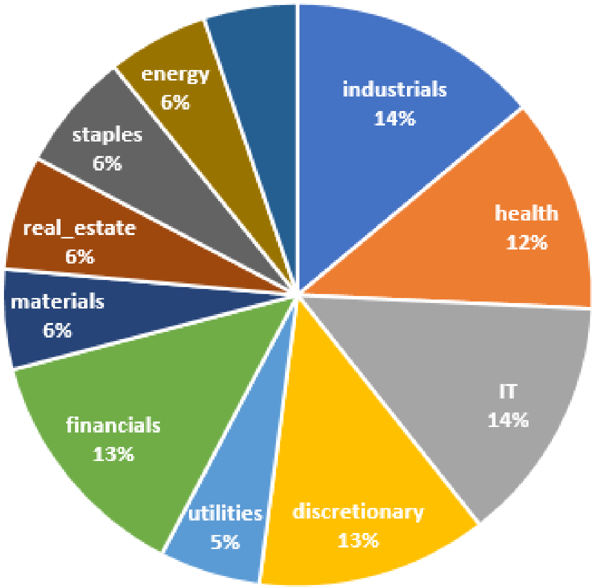
Dataset

- Data : **daily returns** of 590 companies in Russel 1000 stocks
  (which do not have missing values)
- Data period : 1999.11.02 ~ 2019.11.29
- Daily returns are split and dividend adjusted.
- Industry composition of 590 stocks is similar to that of S&P 500.

Industries of stocks used for experiment

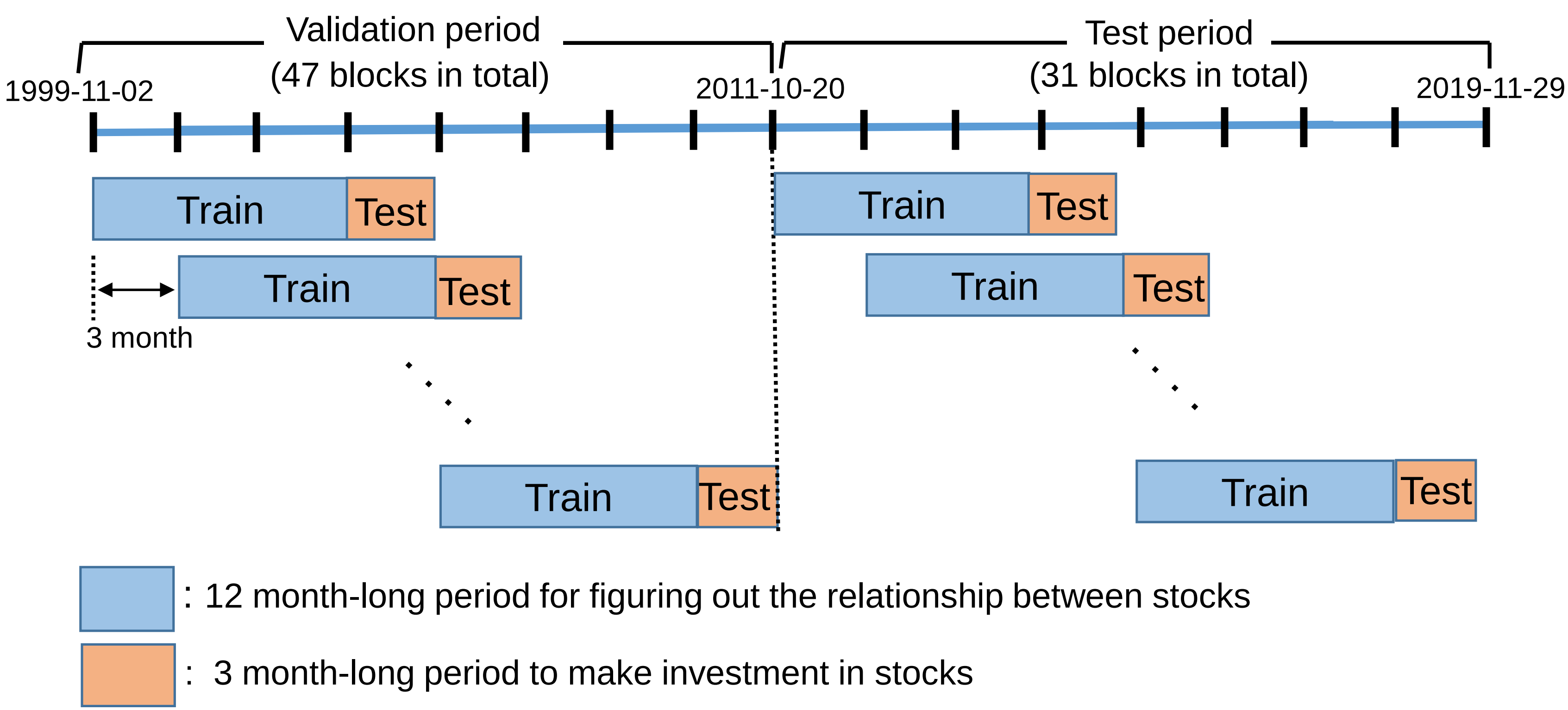


Industries of S&P 500



How to Feed Data for Portfolio Rebalancing

- How to handle data while optimizing a portfolio optimization



4. Results

HUMAN
CENTERED
COMPUTING
LABORATORY

Result Summary (1) : Comparison of Models

- **Standard deviation of portfolio daily returns (annualized)**

Clustering	pre-processing	Scaling method	Validation std	Test std
GMVP on individual stocks			0.1075	0.0946
GMVP on industry sectors			0.0913	0.0845
K-means Clustering	Not used	Standard Scaled	0.1009	0.0954
		Raw data	0.1902	0.2053
	PCA	Standard Scaled	0.0989	0.0911
		Raw data	0.2205	0.1916
	t-sne	Standard Scaled	0.0935	0.0825
		Raw data	0.0967	0.0829
Hierarchical Clustering	Not used	Standard Scaled	0.1197	0.1019
		Raw data	0.1084	0.0973
	PCA	Standard Scaled	0.1235	0.1067
		Raw data	0.1313	0.1032
	t-sne	Standard Scaled	0.0948	0.0838
		Raw data	0.0954	0.0864
Bounded K-means Clustering	Not used	Standard Scaled	0.0906	0.0825
		Raw data	**0.0886**	**0.0798**
	PCA	Standard Scaled	0.0906	0.0822
		Raw data	0.0900	0.0805
	t-sne	Standard Scaled	0.0925	0.0872
		Raw data	0.0905	0.0862

Models with the best performance
: Bounded K-means clustering with raw-data without using dimensionality reduction

Result Summary (2) : Estimation Error

- Comparison of in-sample performance and out-of-sample performance

- Portfolio estimation error (annualized)

	In-sample Std	Out-of-sample Std	Difference
stock-based GMVP	0.0489	0.0946	93.51%
Industry-based GMVP	0.046	0.0846	83.79%
Cluster-based GMVP	**0.0462**	**0.0798**	**72.73%**

  - **In-sample Std** : the mean of standard deviations of Train data
  - **Out-of-sample Std** : the mean of standard deviations of Test data

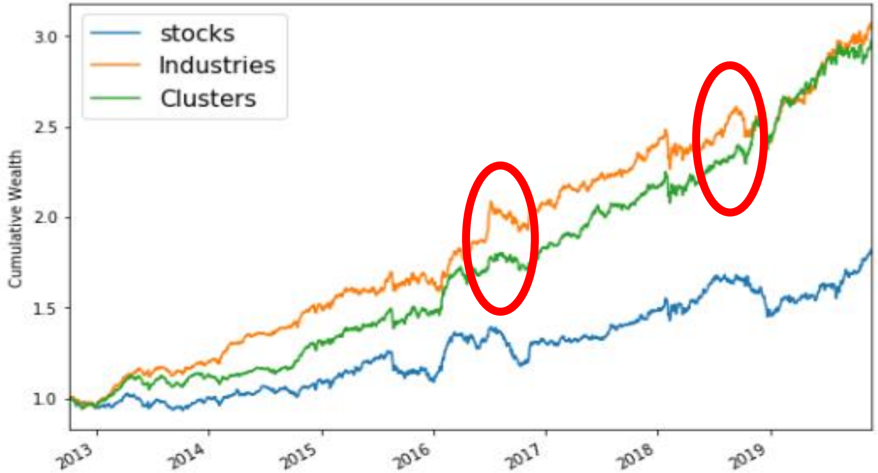
Result Summary (3) : Portfolio Performance

- Comparison of out-of-sample portfolio performance

- Portfolio performance (annualized)

	Sharpe Ratio	Sortino Ratio	Std	Downside Std	Maximum DrawDown	CVaR
stock-based GMVP	0.8963	1.2915	0.0946	0.0686	-15.69%	-1.12%
Industry-based GMVP	1.8232	2.5207	0.0848	0.0637	-9.36%	-0.97%
Cluster-based GMVP	**1.8316**	**2.5726**	**0.0803**	**0.0608**	**-8.21%**	**-0.93%**

- Cumulative wealth graph



5. Discussion

HUMAN
CENTERED
COMPUTING
LABORATORY

Three Points Related with the Experiment

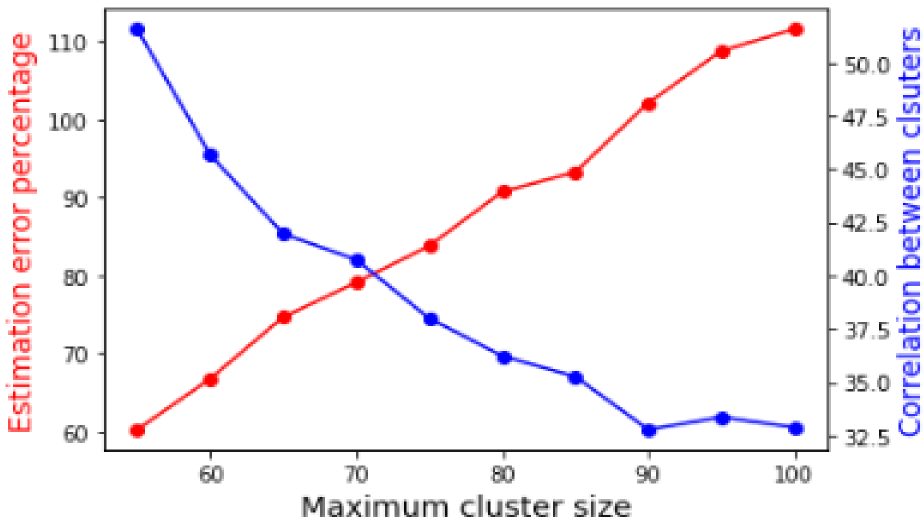
1. Trade-off between ‘**estimation error**’ and ‘**correlation between clusters**’
   - Maximum cluster size can control these two values

2. Both affects the portfolio optimization performance
   - Need to find where to compromise for the best portfolio performance
   - Implies that bounded clustering algorithm is needed

3. Dimensionality reduction and scaling improves portfolio performance
   - Improvement comes from decreased estimation error

1. Trade-off Relationship Found in Clustering

- Found a trade-off caused by maximum cluster size
  : **estimation error of covariance matrix** and **clustering quality**

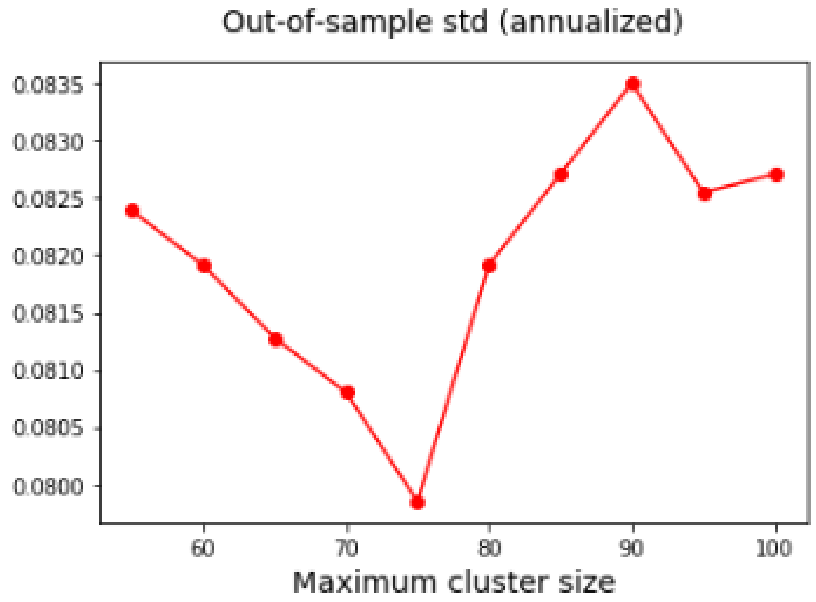


Experimented with raw data without dimensionality reduction

- As the maximum cluster size increases,
  - The dimensionality of covariance matrix increases → **Bigger estimation error**
  - The clustering quality improves → **Smaller correlation between clusters**

2. Where to Set the Maximum Cluster Size

- Out-of-sample performance is decided both by these two components



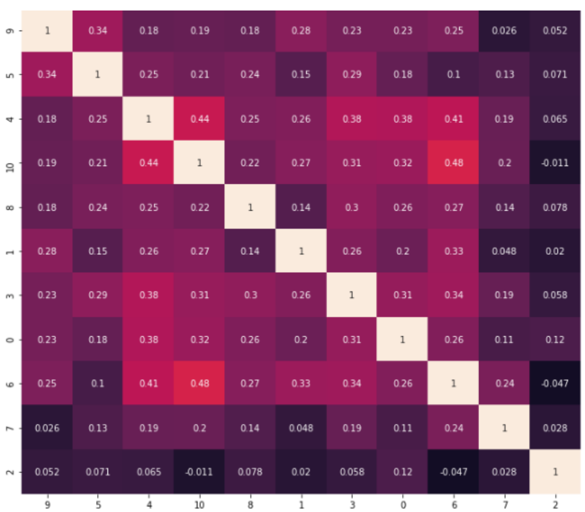
Experimented with raw data without dimensionality reduction

- Need to find where to set the maximum cluster size for the best performance
- Need to use clustering methods where maximum clustering size can be manually controlled to find the compromise

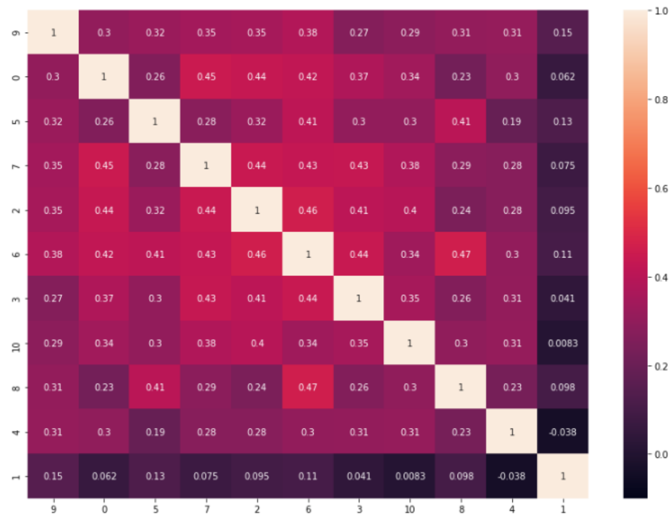
Clustering Without the Maximum Size Constraint

- Unbounded clustering algorithms focus only on clustering quality
- Better clustering quality, but poor portfolio performance due to estimation error

Ordinary K-means clustering



Bounded K-means clustering



Experimented with raw data without dimensionality reduction

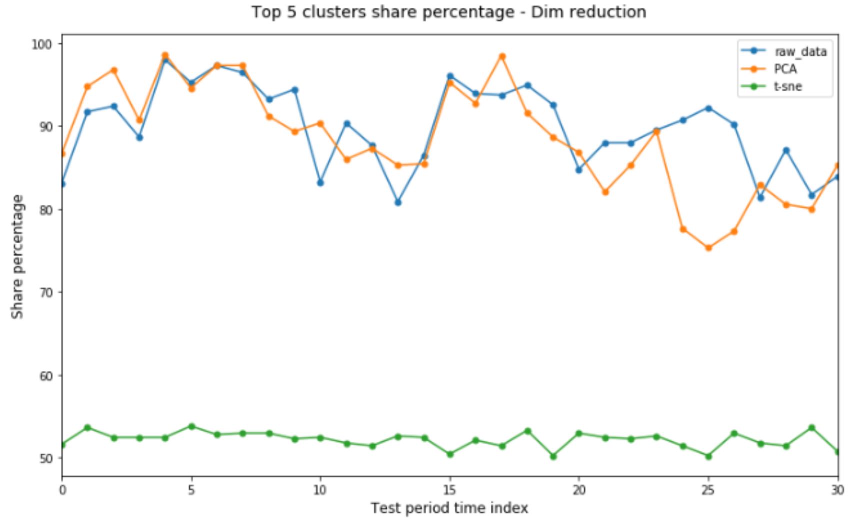
- The estimation error offsets benefits coming from better clustering quality

3. Impact of Dimensionality Reduction

- Clustering with t-sne performs better than others

Clustering	pre-processing	Scaling method	Validation std	Test std
K-means Clustering	Not used	Standard Scaled	0.1009	0.0954
		Raw data	0.1902	0.2053
	PCA	Standard Scaled	0.0989	0.0911
		Raw data	0.2205	0.1916
	t-sne	Standard Scaled	0.0935	0.0825
		Raw data	0.0967	0.0829

- t-sne creates more balanced cluster size → less estimation error



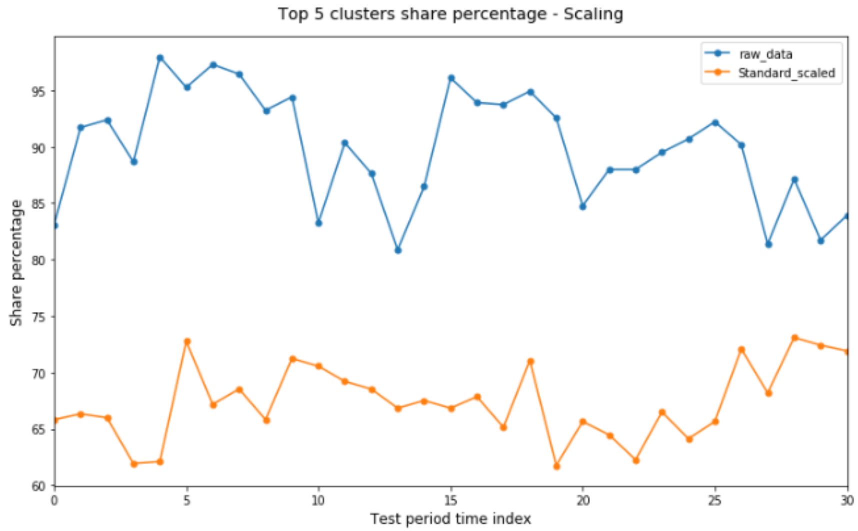
Experimented with K-means clustering with raw data

Impact of Scaling method

- Clustering with standard scaling performs better than raw-data

Clustering	pre-processing	Scaling method	Validation std	Test std
K-means Clustering	Not used	Standard Scaled	0.1009	0.0954
		Raw data	0.1902	0.2053
	PCA	Standard Scaled	0.0989	0.0911
		Raw data	0.2205	0.1916
	t-sne	Standard Scaled	0.0935	0.0825
		Raw data	0.0967	0.0829

- Standard scaling creates more balanced cluster size → less estimation error



Experimented with K-means clustering without dimensionality reduction

Conclusion

- To improve the performance of GMVP, estimation error needs to be reduced
- When applying clustering approach to GMVP,
  - Trade-off between **estimation error** and **correlation between clusters**
  - Both affects the portfolio performance, so needs to be controlled
- Bounded K-means clustering can find a compromise for the best performance
  - Improves the out-of-sample portfolio performance by controlling the trade off
  - Allows better prediction of out-of-sample volatility by decreasing the gap between the out-of-sample risk and in-sample counterpart
- Scaling and dimensionality reduction methods can improve the performance, but better if we can control the maximum clustering size more precisely

6. Appendix

Global Minimum Variance Portfolio

- Finding the asset weights that minimize the portfolio variance (risk), given the covariance matrix of assets.

$$W_{GMV} = \underset{w}{argmin} \{W^T \Sigma \mathrm{W} \,; W^T \cdot 1_N = 1\}$$
$$= \frac{\boldsymbol{\Sigma}^{-1} 1_N}{1_N^T \boldsymbol{\Sigma}^{-1} 1_N}$$

* W_{GMV} is an asset allocation vector that we try to find.

* ***W* =** $(w_1, \ldots, w_n)^T$ is a vector of portfolio weights

* **Σ** is a variance covariance matrix of assets (stocks)

* 1_N is a N dimensional vector of ones