AALTO UNIVERSITY

SEMINAR ON CASE STUDIES IN OPERATIONS RESEARCH

MS-E2177

Data-driven classification of stocks in asset portfolio management: Final report

---

Team Elo:

Tommi Kantala (Project manager)

Hannu Pantsar

Ilkka Särkiö

May 31, 2018

Contents

1 Introduction

Optimal portfolio management is one of the most studied problems in finance. It is also an important topic for the Finnish Mutual Pension Insurance Company Elo, which manages the pension security of approximately 700,000 Finnish citizens. Currently the market value of Elo's investment assets is approximately 23 billion e, which makes the stakes in portfolio management very high.

The goal of portfolio managers is to maximize returns and minimize risks. Minimization of risks is usually performed by diversifying the asset portfolio, which means including assets in the portfolio that presumably do not correlate with each other positively. Assumptions of positive mutual correlation are usually made for assets which belong in the same type of industry. Different industry classification methods for assets exist, such as the Global Industry Classification Standard (GICS). Use of GICS and other classification methods as a tool in portfolio optimization is questionable. Thus, it is beneficial to investigate whether better grouping methods for assets can be formulated.

Our task in this project is to group assets by using clustering methods and compare their correlation to grouping by using GICS. What Elo requires from us is a correlation matrix of the assets in a given dataset of asset performance indicators, which can be used to cluster the assets to any number of clusters. The performance indicators of assets can be for instance the returns or some descriptive equity data. The first approach to form the correlation matrix is to forecast the covariance of assets in the future with time series models and inspect how the forecasts correlate. If this approach is to be abandoned, the correlation matrix can be formed based on only the historical data of the assets.

After the correlation matrix is formed and a proper clustering method has been found, the resulting clusters and their correlations are to be compared

with their GICS classification. An important question is whether the resulting clusters outperform the GICS classifications in terms of correlation.

At the end of the project, we aim to have a script of some sort for the clustering that can easily be used by Elo. This means that Elo could relatively easily just plug in a dataset to the script that performs the clustering of assets for them. If a satisfying result is achieved, Elo could use the clustering based method as a supporting tool for portfolio diversification.

2 Literature review

2.1 Time series modelling

The time series modeling of stocks is of great academic and industrial interest. For stocks, there are three approaches, either to model directly returns, the volatility of the asset or both. In terms of forecasting, volatility is deemed to be slightly easier process than directly forecasting returns. For the returns, we could use autoregressive moving average (ARMA) models and their many derivatives, which try to model the mean process. The other approach is to use autoregressive conditional heteroskedasticity (ARCH) and their generalized versions called GARCH, which model the volatility process of an asset. These ARCH and GARCH models take into account the heteroskedasticity of the error term by modeling the variance of the time series. With these models, the forecast is that of volatility and not the original return. [1]

In ARCH models, the future volatility is forecasted based on the previous squared residuals, with the model estimating suitable weights for the selected period backwards. We denote returns as $r_t = \mu_t + a_t$, where r_t is the return of the asset at time t and it is assumed to consist of mean μ_t and a_t that is called shock or innovation. Thus a_t represents mean corrected asset return.

Formally, the ARCH model is defined as

$$a_t = \sigma_t \epsilon_t \tag{1}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + ... + \alpha_m a_{t-m}^2, \tag{2}$$

where $E(\epsilon_t) = 0, Var(\epsilon_t) = 1$, $\alpha_0 > 0$, $\alpha_i > 0$, $i > 0$. The parameters of the model are obtained via maximum likelihood estimation. [1, 2, 3]

The ARCH model is simple but also restricted. Extension of this, GARCH, allows declining weights to never go to zero, thus allowing whole data set to affect forecasts. [1]. We consider $a_t = r_t - \mu_t$ to be mean-corrected log return. For a_t then applies that

$$a_t = \sigma_t \epsilon_t \tag{3}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{4}$$

where ϵ_t is standard normal or Student-t distribution, $\alpha_0 > 0, \alpha_i, \beta_j \geq 0$ and $\sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) < 0$. Simple GARCH(1,1) model can be written as

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{5}$$

where $\alpha \geq 0, \beta_1 \leq 1, (\alpha_1 + \beta_1) < 1$. This can be interpreted, that large shocks contribute to more volatility, that is, a large shock is expected to be followed by another large shock. [3]

There are several derivatives of the standard GARCH model, such as exponential GARCH (eGARCH) and Glosten-Jagannathan-Runkle-GARCH (gjrGARCH). The models further relax the assumptions of GARCH models, often giving better models. For example, eGARCH allows the model to have asymmetric responses to positive and negatives shocks. It also uses logarithm of conditional variance which relaxes the positiveness constraint of the model coefficients. The model uses weighted innovation

$$g(\epsilon_t) = \begin{cases} (\theta + \gamma)\epsilon_t - \gamma E(|\epsilon_t|), & \text{if } \epsilon_t \geq 0 \\ (\theta + \gamma)\epsilon_t - \gamma E(|\epsilon_t|), & \text{if } \epsilon_t < 0, \end{cases} \tag{6}$$

which allows the responses to be asymmetrical. The eGARCH(p,q) model can be written also as

$$a_t = \sigma_t \epsilon_t \tag{7}$$

$$\ln(\sigma_t^2) = \alpha_0 + \frac{1 + \beta_1 B + ... + \beta_q B^q}{1 - \alpha_1 B - ... - \alpha_p B^p} g(\epsilon_{t-1}), \tag{8}$$

where B is lag operator $Bg(\epsilon_t) = g(\epsilon_{t-1})$. [3] Another way to consider the asymmetric impacts is to use gjrGARCH, which is based on the empirical observation, that negative shocks have stronger impact on the later observations. The model can be written as

$$a_t = \epsilon_t \sigma_t \tag{9}$$

$$\sigma_t^2 = \alpha_0 + (\alpha + \gamma I_{t-1})\epsilon_{t-1} + \beta \sigma_{t-1}^2, \tag{10}$$

where $I_{t-1} = \begin{cases} 0, & \text{if } a_{t-1} > \mu \\ 1, & \text{if } a_{t-1} < \mu \end{cases}$. If we set $\gamma = 0$, the model becomes standard GARCH. [4] The previous ARCH/GARCH models are all univariate. Notably, univariate models often work best with low values of p and q. Often $(1,1)$ models are a reasonable guess for an adequate model. [3]

As noted before, the univariate can be extended to the multivariate case, in which several GARCH processes have a relationship. Typically, these models are used to study volatility between markets, amplitude of impacts, indirect and direct correlations between asset volatilities, etc. The simplest model is a multivariate extension of GARCH, in which multivariate return series is$\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{a}_t$. The mean is conditional expectation of return series based on the past information $\boldsymbol{F}_{t-1}$. The volatility now becomes multivariate $\boldsymbol{\Sigma}_t = Cov(\boldsymbol{a}_t | \ \boldsymbol{F}_{t-1})$. The covariance matrix $\boldsymbol{\Sigma_t}$ is often re-parameterized as a Cholesky decomposition by using the positive definite form $\boldsymbol{\Sigma_t} = \boldsymbol{L}_t \boldsymbol{G}_t \boldsymbol{L}_t^{'}$ where $\boldsymbol{L_t}$ are lower triangle matrices with unit diagonal elements and $\boldsymbol{G_t}$ and diagonal matrix. The Cholesky decomposition performs a orthogonal transformation from $\boldsymbol{a}_t$ to $\boldsymbol{b}_t$ such that $\boldsymbol{b}_t = \boldsymbol{L}_t^{-1} \boldsymbol{a}_t$. The orthogonal transformation simplifies the likelihood function and makes it easier to estimate parameters. [3]

There are several different implementations of multivariate GARCH models. We previously saw a direct generalization, but the category also includes VEC, BEKK, Riskmetrics, (full) factor GARCH and flexible MGARCH. The problem with these models is that the number of parameters to be estimated is very high and thus models are not suitable except to cases where number of assets is very low. [3]

Another group of models is linear combinations of univariate GARCH models. These models include generalized orthogonal (goGARCH) models and latent factor models. In factor models, the co-movements of stocks are believed to be driven by a small number of common underlying variables. The linear combinations of univariate models assume that the time series is generated by a combination of one or many different univariate GARCH models, such as standard GARCH, eGARCH, APARCH or contemporaneous asymmetric GARCH. [3, 5]

The third category consists of nonlinear combinations of univariate GARCH models. These models cover models with constant and dynamic conditional correlation. Some popular models include copula-GARCH, constant conditional correlation GARCH (cccGARCH) and dynamic conditional correlation GARCH (dccGARCH). These models are less greedy in terms of parameters so that they can be fitted for a larger number of assets. There are short-comings in theoretical sense: results on stationarity, ergodicity and moments are not obtained as easily as in the other models we have covered, but the reduced number of parameters is more important in real life applications where the number of assets is large.. [5]

The dccGARCH model is an extension of cccGARCH, where the constant correlation between assets need not be constant and may vary in time. The dccGARCH model can be estimated consistently in two steps, by first building all individual univariate models for each time series, and then multivariate model is estimating. This allows larger number of assets to be used in feasible time. The DCC(1,1) model by Engle in 2002 is defined such that instead of

univariate variance we have variance matrix

$$H_t = D_t R_t D_t, \tag{11}$$

where $D_t = diag(h_{iit}^{1/2} ... h_{NNt}^{1/2})$, where h_{iit} is any univariate GARCH model and $R_t = diag(q_{11,t}^{-1/2} ... q_{NN,t}^{-1/2}) Q_t diag(q_{11,t}^{-1/2} ... q_{NN,t}^{-1/2})$, where the $N \times N$ symmetric positive definite matrix $Q_t = (q_{ij,t})$ is defined as $Q_t = (1 - \alpha - \beta)\bar{Q} + \alpha u_{t-1} u'_{t-1} + \beta Q_{t-1}$, where u_t, α, β are non-negative scalar parameters and $\alpha + \beta < 1$. The elements of $\bar{Q}$ (can be interpreted as pseudo-correlation matrix) can be estimated or set to an empirical counterpart. The model is limited in that conditional correlations must follow same dynamics, unless a relaxed flexible dccGARCH model is used (fdccGARCH). The dccGARCH and its extension and variations are relatively useful, because they work with larger number of assets. The dccGARCH model is also implemented in R in **rmgarch** package. [5, 6] Our observations is that the model is still not feasible with very large N, such as $N > 500$. [5, 6, 7]

Recently, there have been efforts to overcome the problem of estimating covariance matrices and GARCH models for very large N. A working paper by Engle in 2017 is one proposal to allow the use of dccGARCH with even $N > 1000$. In this paper, dccGARCH model is combined with a nonlinear covariance matrix shrinkage method derived from Random Matrix Theory. This makes it possible to efficiently estimate large covariance matrices of time series by correcting overfitting, which normally leads to small eigenvalues to be too small and large eigenvalues too large. In order to be computationally feasible, the likelihood function of dccGARCH process must be changed to composite likelihood method. The nonlinear shrinkage ensure that dccGARCH performs well by improving the correlation targeting matrix. The process now has three steps: 1) a univariate GARCH model is estimated for each asset 2) the unconditional correlation matrix is estimated and used for correlation targeting 3) The composite likelihood is maximized to estimate correlation dynamics. This approach has been showed to be feasible for $N = 1000$ assets. The implementation is based on an existing MATLAB library, which is

both modified and translated to C for increased performance. Best overall performance was achieved with the nonlinear shrinkage dccGARCH model when compared to other shrinkage methods.[8, 9]

There are also other methods than GARCH for modeling financial time series. Firstly, it is possible to combine ARMA models with GARCH model, such that also returns are predicted. Recently machine learning and deep learning have been utilized for the prediction of prices and returns. One possible method is to use semantic data and natural language processing (NLP) to forecast stock performance. Other methods include neural network approaches and deep learning applications. The forecasting with neural networks and deep learning can be done in many different ways. Some viable approaches are trend or price prediction based on fundamentals. Time series forecasting is not as popular, but it should be possible at least with deep learning methods. [10, 11, 12, 13]

2.2 Asset classification and clustering based on historical data

Portfolio managers have always sought to construct portfolios where the underlying risk is minimized by diversification of portfolio assets. This means that the portfolio should not be weighted too much with assets of which performance presumably correlate with each other. A common belief in finance is that assets which belong to same industry have the tendency to correlate with each other more and thus should be avoided having together in the same portfolio.

Different kinds of industry classifications for assets can be used for portfolio diversification, one popular being the Global Industry Classification Standard (GICS), which is maintained by MSCI and Standard & Poor's. GICS is a four-tiered, hierarchical industry classification system and its structure can

be seen in Figure 1 and the GICS sectors can be seen in Figure 2.

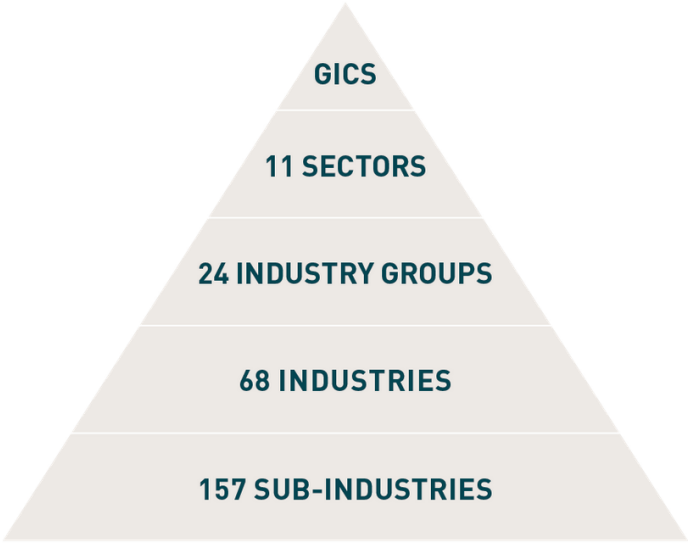

Figure 1: Structure of Global Industry Classification Standard (GICS). Retrieved from https://www.msci.com/gics on 24.4.2018. [14].

Sector	Industry Group
Energy	Energy
Materials	Materials
Industrials	Capital Goods
	Commercial & Professional Services
	Transportation
Consumer Discretionary	Automobiles & Components
	Consumer Durables & Apparel
	Consumer Services
	Media
	Retailing
Consumer Staples	Food & Staples Retailing
	Food, Beverage & Tobacco
	Household & Personal Products
Healthcare	Health Care Equipment & Services
	Pharmaceuticals, Biotechnology & Life Sciences
Financials	Banks
	Diversified Financials
	Insurance
Information Technology	Software & Services
	Technology Hardware & Equipment
	Semiconductors & Semiconductor Equipment
Telecommunication Services	Telecommunication Services
Utilities	Utilities
Real Estate	Real Estate

Figure 2: Main Sectors and Industry groups of GICS. Retrieved from https://www.marketindex.com.au/asx-sectors on 24.4.2018 and verified with the MSCI GICS data. [14, 15]

GICS classifications are annually reviewed and companies are classified quantitatively and qualitatively. Each company is assigned a single GICS classification at the sub-industry level according to its principal business activity for which revenue is used as a key factor. The GICS classification is more distinct than the other alternatives and is better supported by validity tests (the importance of industry classification in estimating concentration ratios). GICS classification data distributed by MSCI is used by over 200 global institutions, including nine of the top ten buy-side and sell-side firms. [14]

There is less literature than expected about the actual performance of GICS compared to different grouping methods of assets in terms of portfolio risk management. Assets can be grouped together by using methods which recognize the amount of comovement they have. Known factors for comovement of assets include Categories, Fundamentals, Trader habitat and Institutional investing. Categories are such classifications as GICS. Fundamentals are the actual value of the asset and measures of it, such as revenue, debt and market cap. Trader habitat can be traders only trading a subset of securities and lack of information and restrictions of traders. Institutional investing is investors trading assets owned by them in a similar fashion. [16]

It is notable to say that short and long run correlations of assets have increased after the 2008 financial crisis, which basically means that the comovements of assets have been less random from that point on. The most important remark however is that from a portfolio diversification point of view, different *clustering* methods have been more effective for grouping assets than industry classifications [17, 18].

Clustering is a mathematical method which groups a set of objects to different groups based on the similarity of the objects. Clustering methods can be divided to hard and soft clustering methods. In hard clustering each object is a part of one cluster and in soft clustering each object can be a part of several clusters with some weight.

In clustering of assets the similarity of assets is usually defined as the correlation of the asset performances over some time period. The usual metric for asset performance are the returns of the assets possibly adjusted with some logarithmic function. The correlation matrix for the matrix $\boldsymbol{X}$ of time series of returns is defined by

$$corr(\boldsymbol{X}) = (diag(\boldsymbol{\Sigma}))^{-\frac{1}{2}}\boldsymbol{\Sigma}(diag(\boldsymbol{\Sigma})^{-\frac{1}{2}} \tag{12}$$

where

$$\boldsymbol{\Sigma} = E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T] \tag{13}$$

is the covariance matrix. The elements of $corr(\boldsymbol{X})$ are the Pearson correlation coefficients between each of the variables in $\boldsymbol{X}$. In case of asset clustering these variables are different assets. Most asset clustering cases in literature use the Pearson correlation, but it has also been shown that using partial correlations instead can distinguish better firms in vastly different businesses. [17].

Correlation based clustering of assets can be performed using only the correlation matrix or alternatively by using different correlation based clustering methods as chaotic map algorithms and ultrametric correlation matrices [19].

In the literature assets have been usually clustered by using either agglomerative hierarchical clustering, k-means clustering or fuzzy C-set clustering. Most publications we found on asset clustering use agglomerative hierarchical clustering. This is understandable, because the literature suggests there is a hierarchical structure in the underlying stock return data [20]. Many of the industry classification methods like GICS are hierarchical which makes the comparison of hierarchical clustering to them a lot easier. Agglomerative hierarchical clustering is a hard clustering method performed using a

Phase	Action
1.	Set every variable to be its own cluster.
2.	Merge the clusters with the smallest distance inbetween them.
3.	Repeat step number 2 with recalculated distances until wanted amount of clusters.

Table 1: Agglomerative hierarchical clustering algorithm.

defined distance function to find the smallest distance between clusters. The algorithm can for this can be found from Table 1.

In case of correlation based clustering the distance is calculated between the rows of the correlation matrix. The distance metric $d(a,b)$ between two data points is usually defined to be the Euclidean distance

$$||a - b||_2 = \sqrt{\sum_i (a_i - b_i)^2}. \tag{14}$$

Other distance metrics exist such as squared euclidean distance, Manhattan distance, maximum distance and Mahalanobis distance. In literature using these for asset clustering is a rarity.

For the algorithm the distance between two clusters has to be also defined. This is called the Linkage criteria. The hierarchical clustering can be further named according to the linkage criteria used. Some linkage criteria for two clusters A, B are introduced in Table 2.

All of these linkage criteria have been used in the literature regarding asset clustering but the choice has been rarely explained. Most likely the choice of method has been done based on the results ability to distinguish assets. The complete-linkage clustering can be interpreted as the most conservative choice since the distance separating two assets within same cluster is always

Linkage method	Formula
Complete-linkage clustering	$max\{d(a,b) : a \in A, b \in B\}$.
Single-linkage clustering	$min\{d(a,b) : a \in A, b \in B\}$.
Average linkage clustering	$\frac{1}{\|A\|\|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b)$.

Table 2: Linkage methods for hierarchical clustering

shorter than distance between that cluster any other cluster [21]. Using single linkage method can lead to so-called "chaining effect", where distances between data points inside a cluster can be massive compared to distances between clusters [22] [19].

Performance of GICS against asset clustering has been studied a bit but not as much as was expected. According to “Clustering stocks using partial correlation coefficients” (Sean S. Jung, Woojin Chang): “in most cases, the firms in the same sector are not grouped together into a single cluster”. Partial correlations were used instead of Pearson correlations in the publication and the result "clearly indicates that the GICS sector classification is not the best one to divide the firms if one seeks to minimize the correlation in stock returns" [17].

According to correlation based comparison in "Industry classifications and return comovement" by Chan et al, assets with the same GICS classification tend to correlate the larger they are. With fewer assets the effect was less clear [23]. A multi-factor model with agglomerative hierarchical clustering has led to a result that cluster based strategies bring more profit than the strategies based on ICB industry classification, and that diversification by clustering is more beneficial than diversifying by ICB industry classification. The result is clearer in favor of clustering based approach when the market is more volatile [18].

3 Approach

3.1 Data

We received the data from ELO in batches. First we received a list of the companies with some meta data, such as names, tickers, countries et cetera. After this, we received the historical daily returns for approximately 3000 European stocks. The data only includes stocks that were listed on the last day included in the data set. Thus the further into the history we go, the more inputs are missing. This is called survivorship bias and the effect can be seen in Figure 3. We can also view this from other perspective to analyze how many companies are listed in the stock exchange. For example, we can see that the proportion of companies available grows very slowly on and some time after financial crisis of 2008. This means that the data is overly optimistic. This also limits the number of usable stocks, since we want to have as complete data as possible when analyzing the time series of returns and fundamentals. The fundamentals were received later and they were combined in R with the historical and meta data.

When the data was used for correlation, some stocks were removed from the analysis if more than 20% of values were missing. One explanation for missing values is that the stock was not traded enough. Since all stocks are missing some values, for example due to bank holidays, a logarithmic 5-day moving average was calculated to smooth over the missing values. The fundamentals were also assumed to be lagged by approximately two months, since the figures are released in reports later than they have actually realized. The fundamentals data includes some outliers, such as companies that have much larger market cap than rest of the companies. To avoid problems in clustering caused by outliers, these are handled in the clustering phase by first standardizing them by subtracting the mean and dividing them with standard deviation and and capping the extreme values closer to rest of the

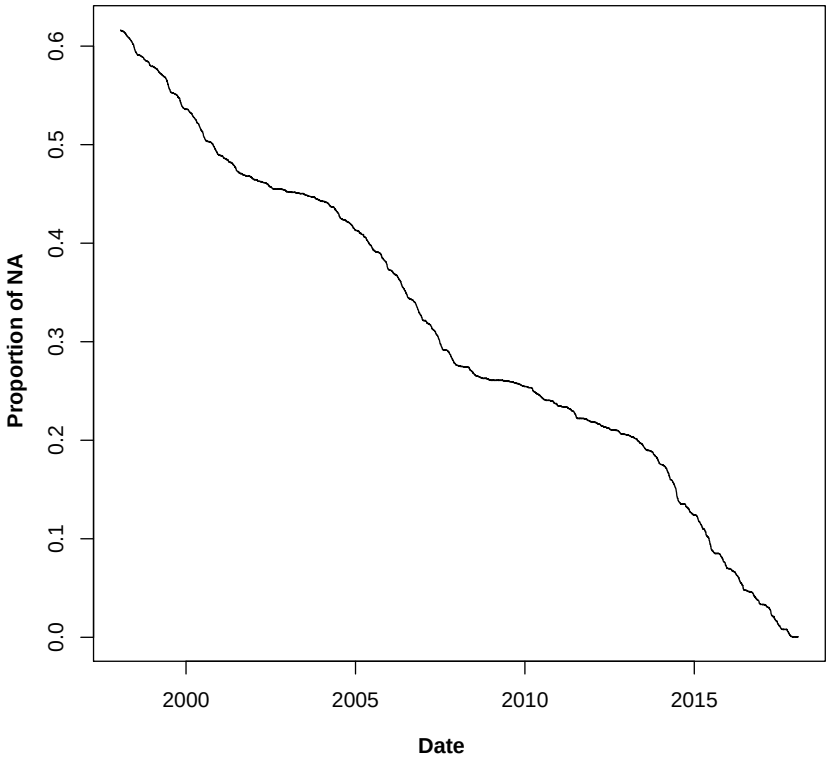

Figure 3: The proportion of NA in the historical returns. The portrays the proportion of companies not available compared to original n=3050.

data. Otherwise those outliers have huge distances which would dominate clustering and the outlier companies would create their own clusters that include only one stock. The code allows the user to select the fundamentals that are used to supplement the correlation matrix when clustering. The distances caused by differences in fundamentals data are larger than the distances measured from correlation matrix, but the correlation matrix is quite large (often in the range of 2000 by 2000 matrix) compared to the number of fundamentals. Thus they should have some sort of a balance.

3.2 GARCH modeling

The forecasting of covariance and correlation matrix was done in R using packages **rugarch** for univariate models and **rmgarch** for multivariate models. The dccGARCH models allowed us to first model all the univariate time series models. We tested standard GARCH, eGARCH and gjrGARCH. For the multivariate model we used dccGARCH and fdccGARCH. Most of the testing was done using 100 to 250 stocks. Several problems were encountered in modelling. First, we ran R on Ubuntu 16.04. We were not able to solve the root cause, but the performance on Linux was subpar compared to Windows. We suspect that the solver used for multivariate estimation was somehow incompatible or broken on Linux implementation. We also often found that univariate models did not converge, multivariate models reached singularity or the estimation process broke somehow with various error messages especially with larger number of stocks used. The libraries are quite high-level (on abstraction) and debugging was quite difficult. We were not able to solve what caused error messages such as "Non-numeric argument passed to binary operator". The modeling seemed to work up to 250 stocks reasonably well, especially on Windows.

3.3 Time series and fundamentals based clustering

The clustering process based on correlation matrices estimated from historical data and fundamentals was very data driven. We did do it parallel to the literature review but were not able to find clear consensus from the literature what would be the best approach, since different clustering methods, time windows, fundamentals and geolocations have been used. Since we were not able to form a forecasting process for the correlation matrix, we settled to do the clustering analysis based on the historical data and fundamentals. The clustering analysis can be performed for varying periods of time or windows (for example, three months of data selected from the time period of interest

is used for the correlation matrix) and the overlap of windows and total coverage (time period of interest) can be controlled. In each window, the latest fundamentals is used to supplement the correlation matrix. For comparative analysis, we calculated average within and inter cluster correlation. These figures can be then compared between different methods using box plots. We also calculated the overall correlation of the stocks and see how different clustering methods compare against it. We also extracted the mean and standard deviation of fundamentals for one clustering output to see if we can find any patterns explaining the output of clustering.

A critical information about stock clustering is how far into the future the clusters are valid and provide a means for portfolio management. If the clusters would remain similar for an extended period of time, clusterings made from data of consecutive periods of time should have similar clusters. To investigate this, cross tabulating consecutive clusterings and maximizing the trace of the resulting matrix can be calculated. If the clusterings would be exactly the same, only the diagonal of the resulting matrix should be non-zero. This, however, did not happen. In fact even with clusterings made from overlapping periods of time had very many off-diagonal elements.

We developed an alternative way to measure the stability and how long the clusters provide means for portfolio management: correlation tracking. A clustering was built from a 6 month period of time using both correlation of weekly averaged log returns as well as a set of fundamental variables as was done in the previous examples. Then, the within and inter cluster correlation was calculated every 30 or 5 days for 2 years or 3 months. The correlations were corrected with baseline correlation of the entire stockset. Identifying a point where the within cluster correlation declines to stable levels gives an estimate how long the clusterings provide added value on portfolio diversification.

4 Results

4.1 GARCH forecasting

Since we mostly did modeling on Linux and the process was computationally very heavy, we tried to forecast approximately three months separated from a four year long time series of daily running five day average log returns. In the example below, gjrGARCH models were fitted to each of the 250 univariate time series. After this, a fdccGARCH model, which allows the dynamics of between the stocks to change over time, was estimated on the fitted univariate time series. Both models used standard normal distribution for noise term. No external regressors were used.

To estimate the goodness of predicted correlation matrix we can compare the correlation heat maps between predicted and real three month correlation matrices. The predicted heat map can be seen in Figure 4 and the real correlation matrix can be seen in Figure 5. We can clearly see that the model dramatically underestimates the correlations between stocks. We can faintly see similar pattern emerging in the predicted correlation matrix heat map, but the effect is much weaker. We can also see the lacking predictive performance of the forecast in the histogram in Figure 6, which shows the difference between real correlation matrix and forecast. We can observe that the differences are quite normally distributed and we have quite a lot of large differences, especially when we consider that our correlation is between -1 and 1. Thus, even difference of 0.1 is quite large a difference. Based on our observations, it seems that forecasting large covariance matrices using basic models implemented in R packages are not accurate enough and fine tuning thousands of univariate time series would be quite tedious.

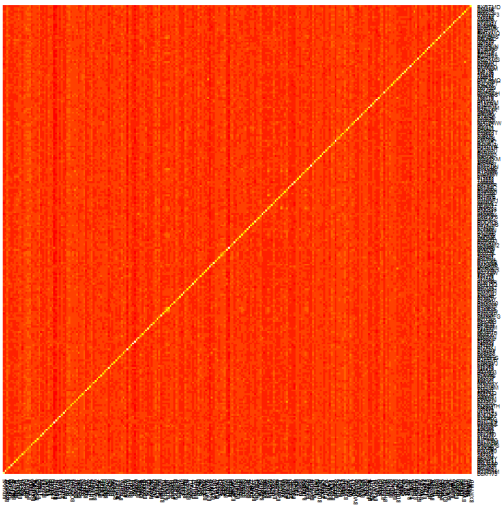

Figure 4: Predicted three month correlation matrix using gjrGARCH + fdccGARCH models for 250 stocks.

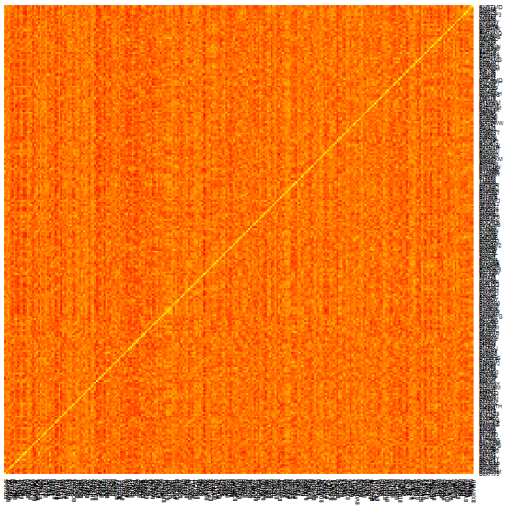

Figure 5: Real three month correlation matrix using gjrGARCH + fdccGARCH models for 250 stocks.

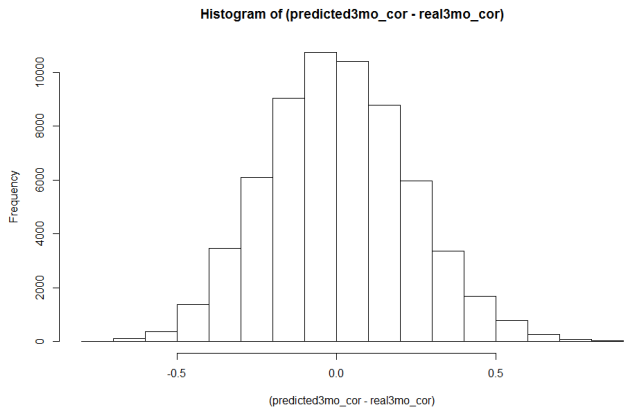

Figure 6: Difference between real correlation matrix and forecast.

4.2 Clustering

4.2.1 Comparison of hierarchical clustering, GICS and FactSet classifications.

Figure 8 shows boxplots for average within and inter correlation of classification based on historical returns in selected window, fundamental data combined with historical returns, GICS and FactSet. Both hierarchical clustering methods use 12 cluster to match the number of GICS clusters. The FactSet uses 20 sectors. The selected time period starts from 31-12-2017 and runs two years backwards with window length of 4 and overlap of 1 months. Total of 8 windows are used for each clustering method.

As shown in from the Figure 8, the hierarchical clustering seems to create slightly clearer difference between average within and inter cluster correlation. Even though the FactSet has more clusters than GICS and hierarchical methods, it still cannot separate cluster very well and hierarchical methods provide more separation for clusters. With hierarchical clustering in later windows we can see that we can create clusters that have very low average correlation with rest of the clusters. However, the performance greatly depends on the selected window. We will address this behavior in Section 4.2.3.

If we increase the number of cluster and granularity of the industry classification, we see some positive reactions in terms of correlation. Figure 9 show how number of outliers sharply increases when the number of clusters is increased. Especially the hierarchical clustering benefits from more clusters. We see much stronger separation in the within and inter correlation during in windows where the correlations are weaker. When GICS and FactSet are used, the number of outliers increases sharply.

Hierarchical clustering introduces a lot more variation in the within and inter cluster correlation as GICS and FactSet clusters. Exact reason is difficult to

determine but the reason probably lies in the more deviating cluster sizes of hierarchical clustering. This leads to larger differentiation in the clusters. An example and comparison of clustering sizes in a typical hierarchical clustering of 12 clusters and GICS sectors is in figure 7.

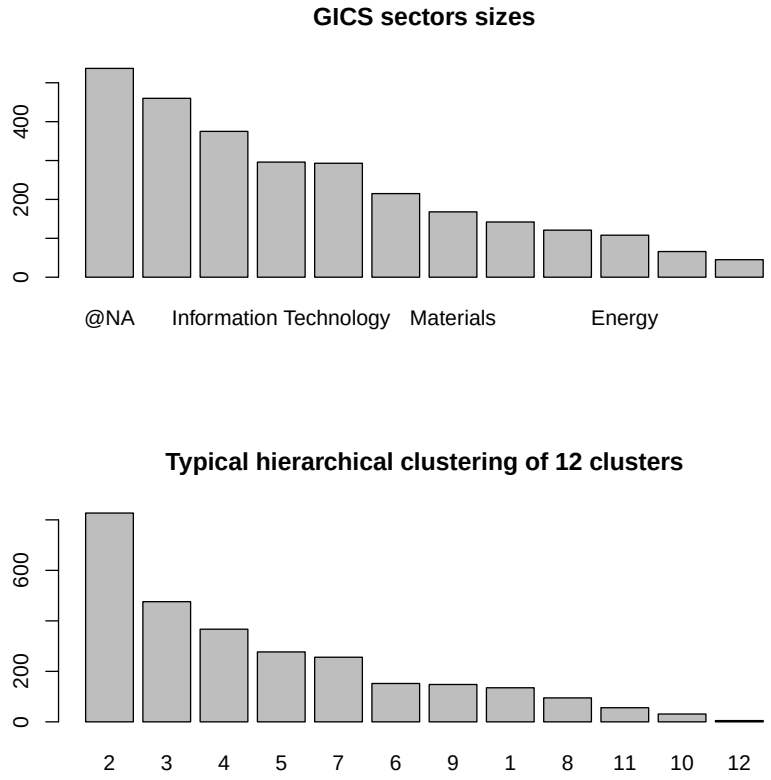

Figure 7: Cluster sizes in typical GICS sector clustering and 12 cluster hierarchical clustering. Hierarchical clusters have more deviation in cluster sizes.

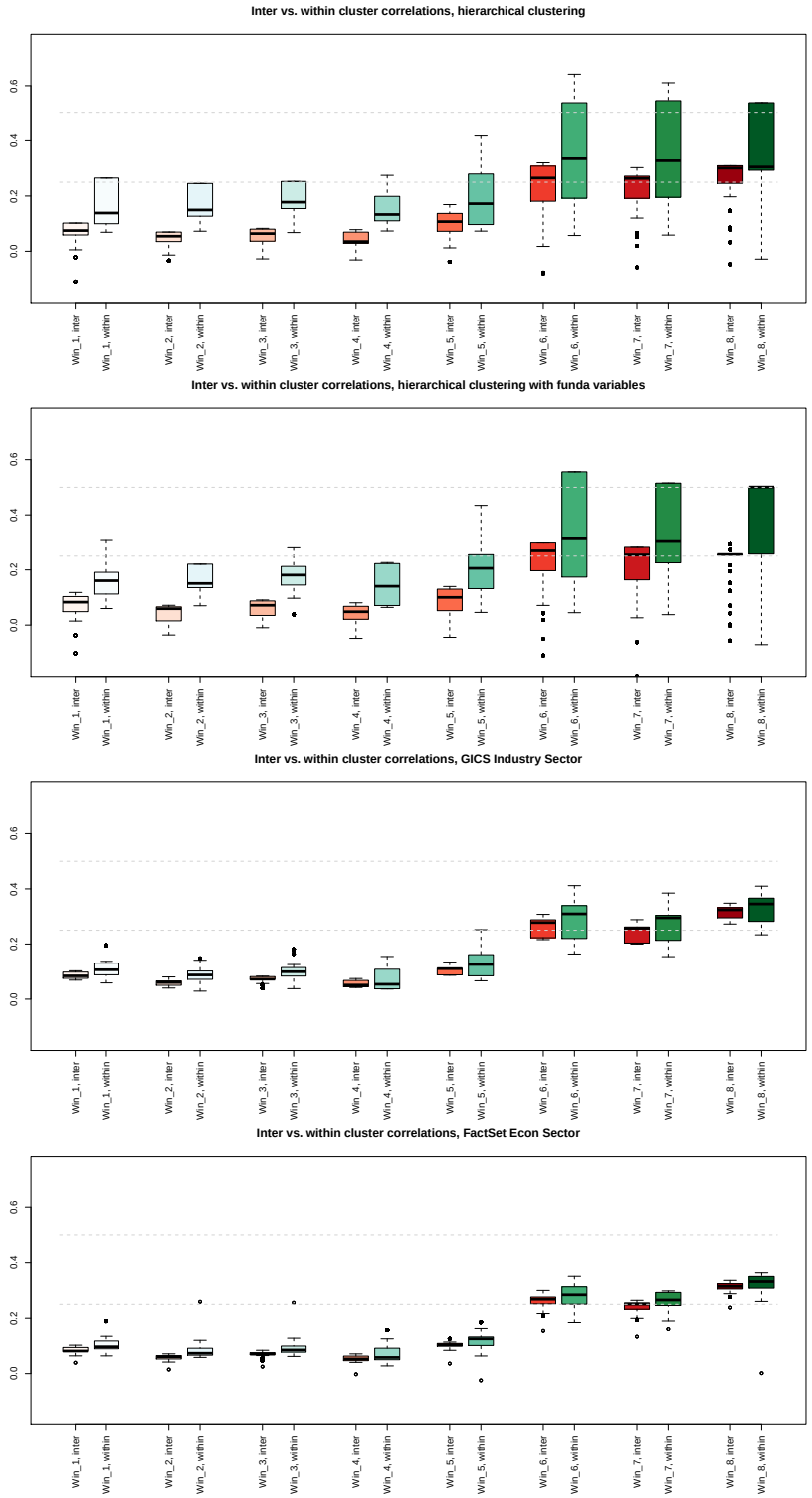

Figure 8: Boxplot of within and inter cluster correlation for different methods using two years of data and four month windows with one month of overlap resulting total of eight clusters. 12 clusters and GICS sectors, 20 FactSet Econ sectors. Gradient color scale has no significant purpose.

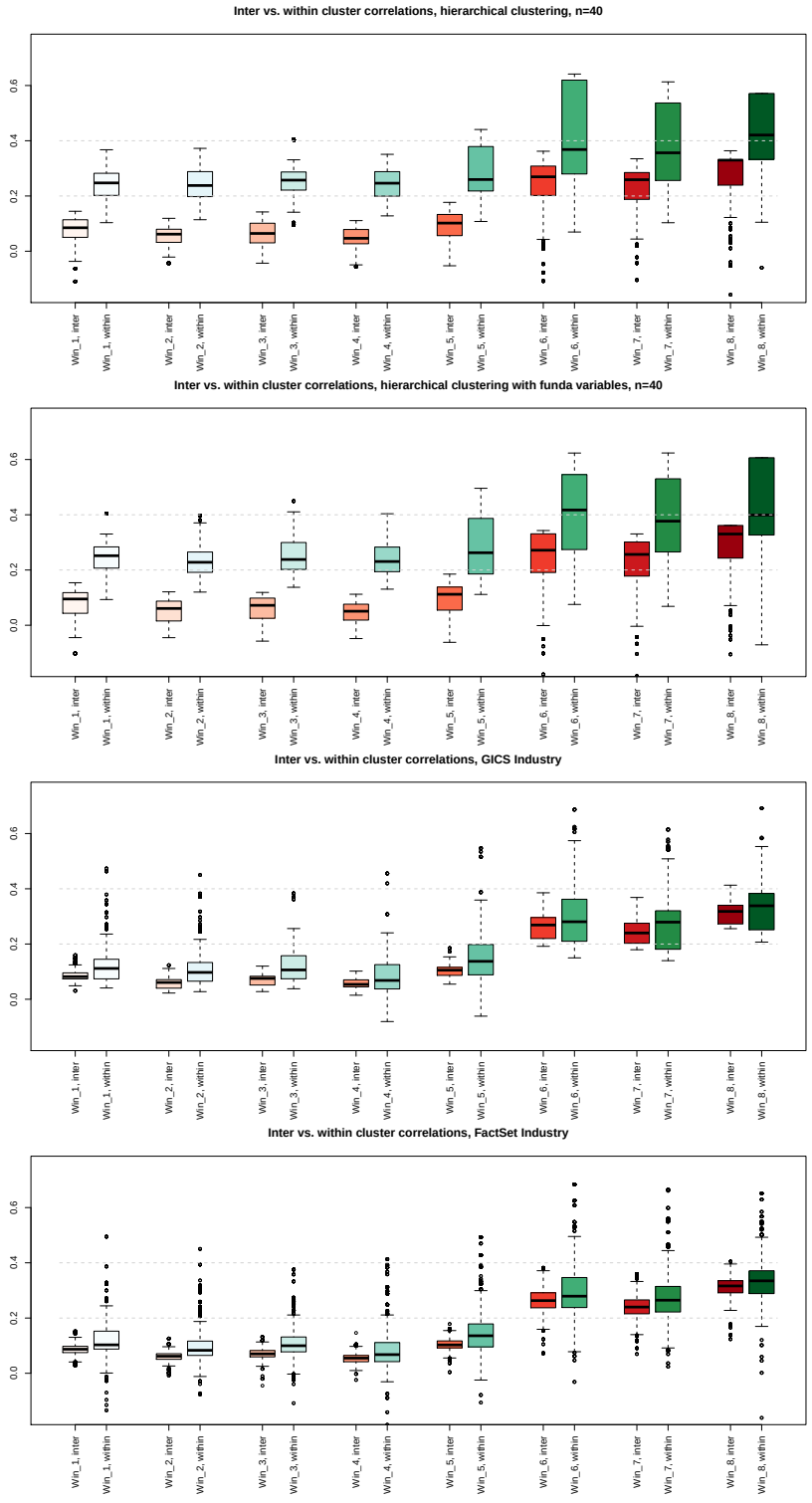

Figure 9: Boxplot of within and inter cluster correlation for different methods using two years of data and four month windows with one month of overlap resulting total of eight clusters. 40 clusters, 69 GICS industries and 168 FactSet industries. Gradient color scale has no significant purpose.

4.2.2 Financial figures of hierarchical clusters

We also studied the relative financial figures of clusters created with hierarchical clustering using data from 2017 with window of one year and using last financial figures from 2017. Note that the clustering is based on limited fundamental values, that is, the outlier values are capped to 5th or 95th percentile in order to avoid them forming their own clusters of one stock, but in the Table 3 the non-capped values are used.

Table 3 shows the mean and standard deviation of EBITDA-%, FreeCashFlow (FCF)-%, NetMargin, MarketCap per Turnover, MarketCap per Earnings and NetDebt per Equity. First, we can see that the size of the clusters varies, the smallest consisting of only four stocks and the largest including 856 stocks. If we look this clusters figures in the Table, we notice that the standard deviations are very large compared to mean values. If we consider how hierarchical clustering works, we can assume that in the smallest cluster some outlier stocks are grouped, which would explain the small size of the cluster. Nevertheless, the standard deviations tend to be still quite large, which could indicate that most of the clusters include "abnormal" stocks. It it very difficult to find clear patterns in the figures that could explain the results of the clustering but we can try.

Cluster 1 seems to have quite high and positive MarketCap per Turnover but not as strong negative Market Cap per Turnover. This cluster might include large companies that have been unprofitable. The standard deviation is large which is possibly explained by outlier companies. Cluster 10 is quite close to the average, so maybe it includes very "average" companies. Cluster 11 has very negative EBITDA-% and FCF-%, indicating that very badly performing companies might be clustered into same cluster at least to some extent. In cluster 3 the negative MarketCap per Turnover might sound impossible but it is actually correct depending on the company structure and accounting rules. Clearly, there are clusters with different values for each

relative fundamental even though interpretation is difficult. It might be that for example looking at EBITDA-% we might have clusters where we have companies with positive EBITDA-%, clusters with close to zero EBITDA-% and clusters with very negative EBITDA-% but the effects are not clear enough to draw any definitive conclusions.

But as said before, it is very tedious to find strong root causes for the clustering output. If one were to analyze the root causes, it would be beneficial to use smaller subsample, where the effect might be clearer. The outliers and mixture of companies with positive and negative values make the interpretation from means and standard deviations difficult.

Table 3: Relative fundamentals of 12 clusters based on return and fundamental data of stocks in 2017

V1	N	EBITDA-% mean (sd)	FCF-% mean (sd)	NetMargin mean (sd)	MarketCap per Turnover mean (sd)	MarketCap per Earnings mean (sd)	NetDebt per Equity mean (sd)
1	100	-7.76 (75.08)	-5.88 (54.91)	0.00 (0.36)	168.39 (1592.28)	-24.61 (473.51)	0.51 (4.20)
2	856	-0.19 (5.35)	-0.32 (4.03)	0.06 (1.09)	7.53 (86.06)	24.50 (247.59)	-0.03 (6.04)
3	754	-1.13 (35.19)	0.12 (4.26)	0.16 (0.70)	-6.22 (251.23)	25.83 (84.23)	0.50 (1.13)
4	343	-0.13 (3.04)	-1.47 (22.69)	-0.02 (1.33)	54.37 (887.11)	29.76 (135.66)	0.16 (1.51)
5	61	-0.02 (1.07)	-0.26 (1.05)	0.05 (0.33)	6.78 (12.50)	34.59 (227.30)	-0.09 (0.95)
6	208	-2.08 (27.87)	-1.64 (19.21)	0.04 (0.39)	22.69 (195.57)	11.61 (448.42)	0.10 (1.05)
7	99	0.04 (4.08)	-0.38 (1.92)	-0.01 (0.43)	16.68 (76.67)	364.12 (2515.96)	0.81 (6.57)
8	77	-5.43 (35.32)	-6.03 (33.62)	-0.21 (1.70)	25.01 (97.03)	-2.87 (90.46)	0.24 (1.50)
9	76	-0.66 (3.05)	-1.51 (4.78)	-0.26 (1.95)	8.60 (21.55)	-6.97 (555.92)	0.76 (3.57)
10	123	1.46 (16.53)	-3.95 (26.34)	0.08 (0.39)	38.01 (336.54)	30.74 (108.46)	0.18 (1.35)
11	64	-13.70 (101.19)	-13.15 (100.35)	0.29 (2.15)	79.97 (561.52)	-165.52 (1347.04)	0.30 (4.21)
12	4	-9.75 (19.66)	-9.83 (19.70)	-0.21 (0.54)	95.01 (188.50)	-3.52 (21.84)	-0.15 (0.67)
Avg.	230	-3.27 (27.29)	-3.69 (24.41)	0.00 (0.95)	43.07 (358.88)	26.47 (521.37)	0.27 (2.73)

4.2.3 Baseline correlation

As addressed in the previous Section, we noticed that the correlation of the stocks seems to change over time. In Figure 10 we have the mean correlation development from 2016 to 2018. We can see the mean correlation of the stocks vary strongly over time. Notably, since the 2017, there has been only very weak correlation in our stock universe. We also wanted to study the effect of a financial crisis. In the same Figure we see that the correlation after the mid-2008 was quite strong due to the bear market, but all together the correlations from 2007 to 2009 were moderately strong.

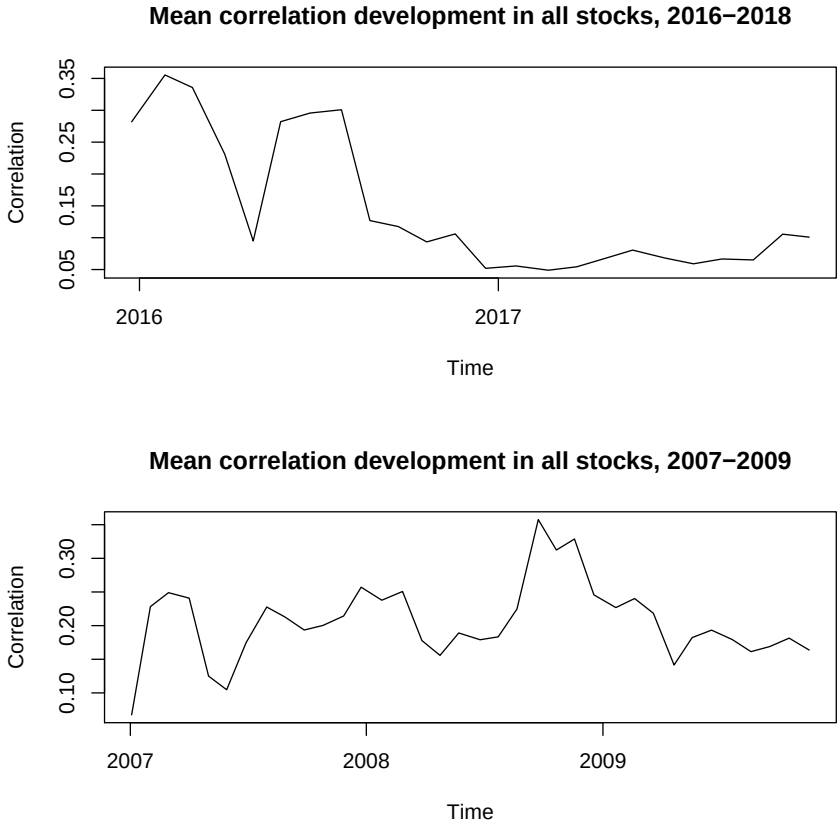

Figure 10: Total mean correlation from 2016 to 2018 and from 2007 to 2009

We also studied the yearly effects. In Figure 11 we can see mean correlation development for first quarter of 2016 and 2008. The year 2016 has been a rollercoaster ride, as the mean correlation has been bouncing between 0.1 and 0.5, before it has settled down to 0.1 range. In 2008 we can possibly notice the effect of financial crisis, because in mid-february 2008 the mean correlation soars from nearly zero (notice the small dip just before the jump) to 0.5. This is caused by most of the stocks crashing, but it also indicates that some stocks might have not correlated strongly with other stocks.

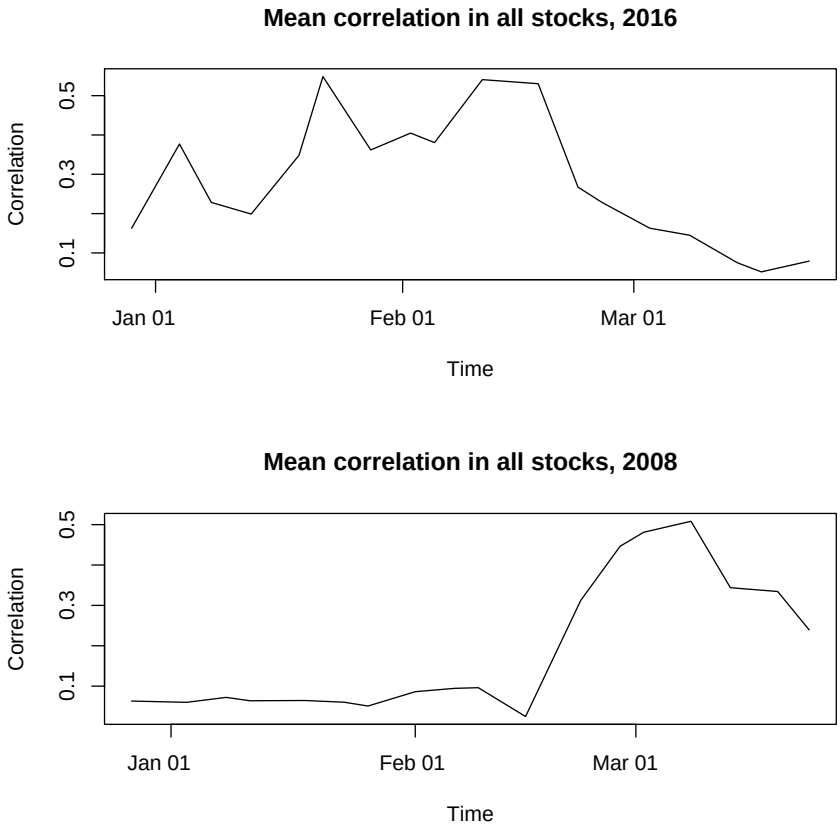

Figure 11: Total mean correlation in 2016 and 2008

4.2.4 Correlation tracking

We tracked correlations for hierarchical clustering in 2016-2018, monthly for two years and every 5 days for 3 months in 2016 to provide tracking with different resolutions. Another tracking was calculated monthly for 2007-2009 and every 5 days for 3 months in 2008. Clustering size was either 12 or 40. Correlation tracking for the same tracking periods and resolutions was performed also for GICS sectors and industries for comparison

Figure 12 shows how the within and inter correlation of the clusters evolves starting from 2016, using six months of data to form the clusters and rolling window of one month to track the correlation. The line represents a weighted average based on size. The baseline correlation has been dominant in both hierarchical data based clustering and in GICS clusters. The weak effect that was achieved with hierarchical clustering quickly diminishes, within correlation shrinks and inter correlation approaches zero, too. With GICS, the within correlation seems to be more stable and the inter cluster correlation remains through the time period at zero. There are large clusters in data based cluster and in GICS clusters (companies with no GICS classification) which most likely have poor performance due to large size and their weight also pulls average correlation to lower level than expected. If we increase the number of clusters for both GICS and hierarchical clustering, we obtain higher within correlation and lower inter correlation. We can see this effect in Figure 13. The number of hierarchical clusters is lower, since already now half of the clusters have less than 30 stocks. In long-term perspective, the GICS seems to perform better, though in short-term perspective the hierarchical clustering offers higher correlation. It feels natural that if we increase the number of clusters we get more granularity and thus more separation between the clusters. But we can also notice that our implementation with this data produces few large clusters and many small ones.

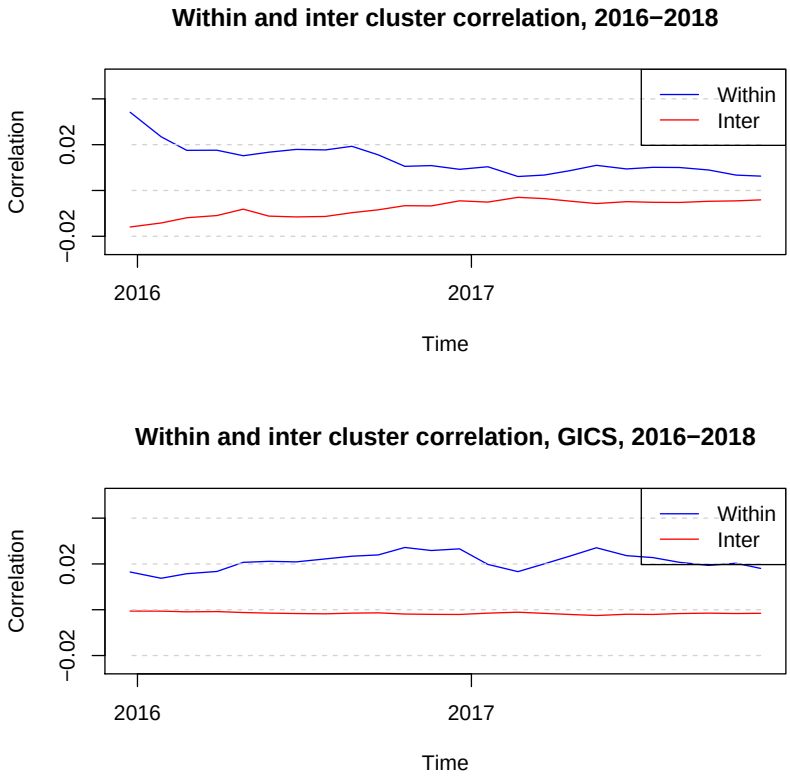

Figure 12: Correlation tracking from 2016 to 2018 using rolling one month window for tracking, with clusters formed using six months of data, 12 GICS clusters and 12 hierarchical clusters. Baseline correlation removed.

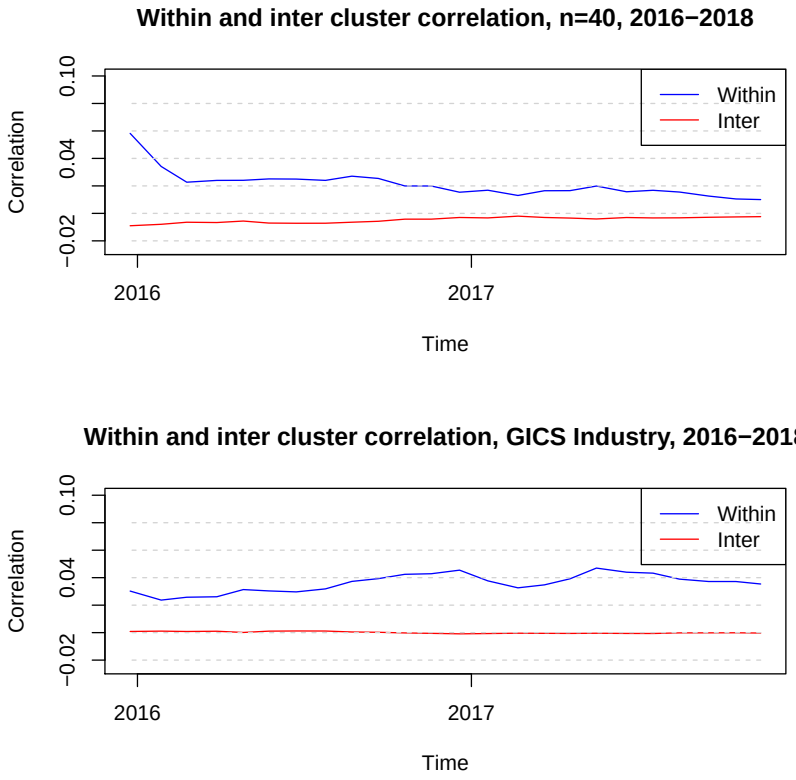

Figure 13: Correlation tracking from 2016 to 2018 using rolling one month window for tracking, with clusters formed using six months of data, 68 GICS clusters and 40 hierarchical clusters. Baseline correlation removed.

Figure 14 shows the tracking in 2016 with shorter rolling window and the clustering is based only on three months of data. In this one the hierarchical clustering seems to work slightly better compared to longer window and also when compared to GICS. However, the correlation is very weak. If we increase the number of cluster as in the previous case,there is quite good improvement on the short-term performance. Long-term performance is on par with GICS. This can be seen in Figure 15.

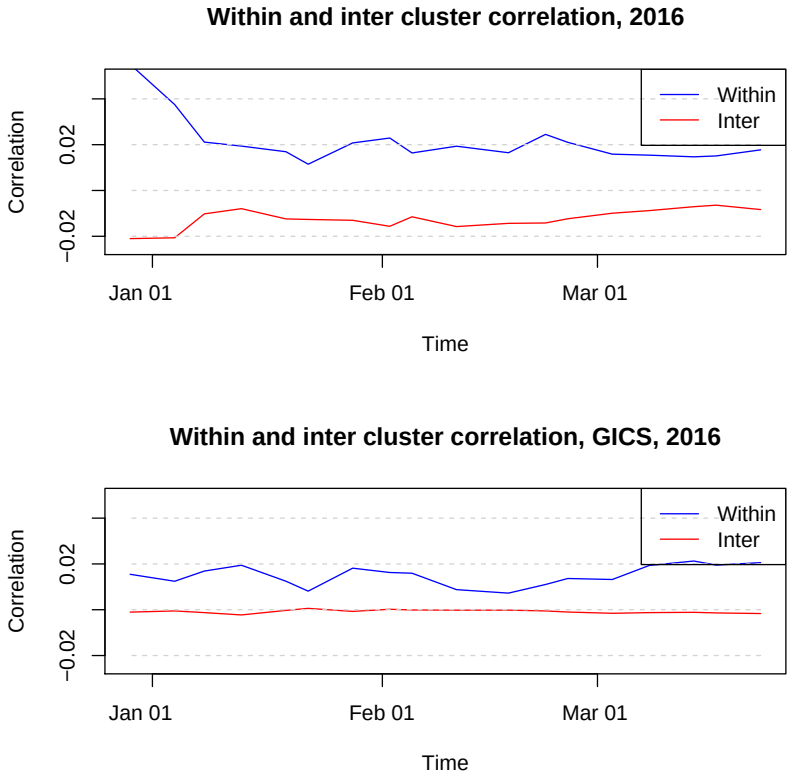

Figure 14: Correlation tracking from 2016 to 2018 using rolling five day window for tracking, with clusters formed from three months of data, 12 GICS clusters and 12 hierarchical clusters. Baseline correlation removed.

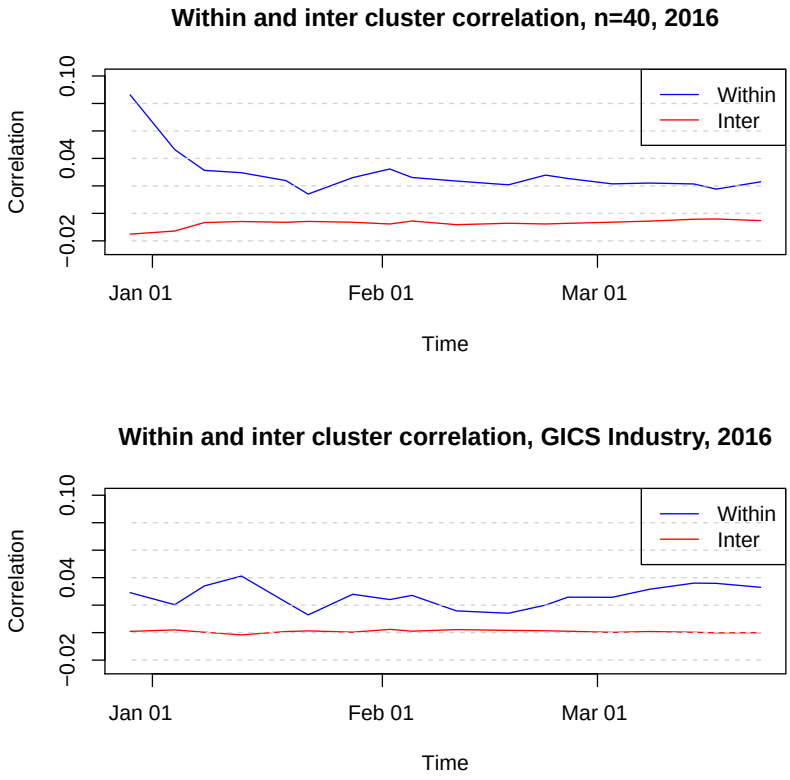

Figure 15: Correlation tracking from 2016 to 2018 using rolling five day window for tracking, with clusters formed from three months of data, 68 GICS clusters and 40 hierarchical clusters. Baseline correlation removed.

We also studied the same effect from 2007 to 2009. This period includes the financial crisis of 2008, which was very strong bear market for stocks. In Figure 16 we can see the correlation tracking with six months of data used for clustering and one month rolling window for tracking, with 12 hierarchical and GICS clusters. The weak correlation diminishes quickly and after this the performance is very similar to GICS, though the GICS might have slightly stronger reaction to 2008 financial crisis and strong downwards trend even with the baseline correlation removed. If we increase the number

of hierarchical clusters to 40 and GICS clusters to 68, we notice that again the hierarchical clustering outperform GICS in within cluster correlation, but only for very short period. This can be seen in Figure 17.

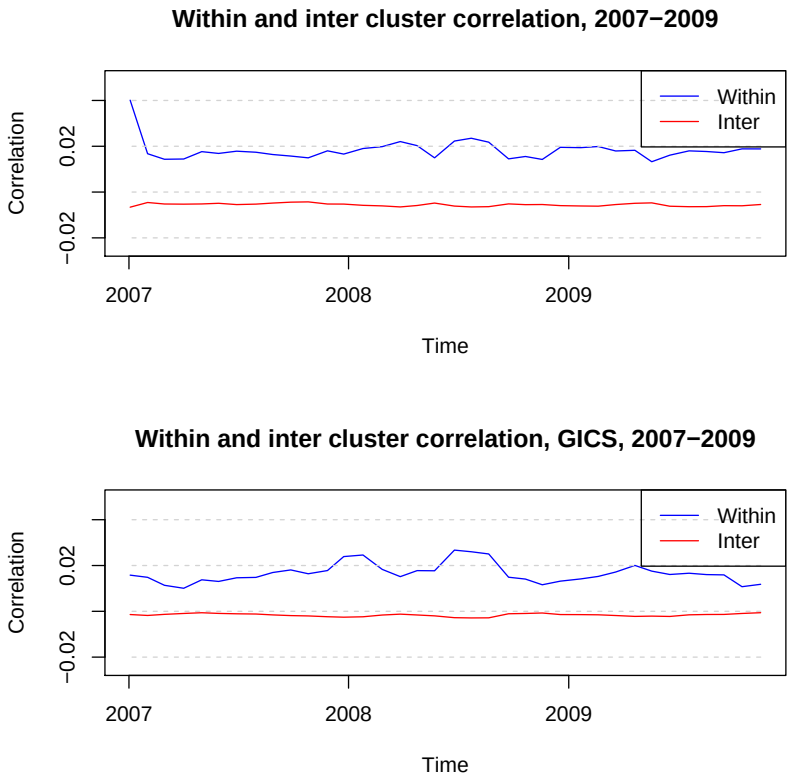

Figure 16: Correlation tracking in 2008 using rolling one month window for tracking, with clusters formed using six months of data, 12 GICS clusters and 12 hierarchical clusters. Baseline correlation removed.

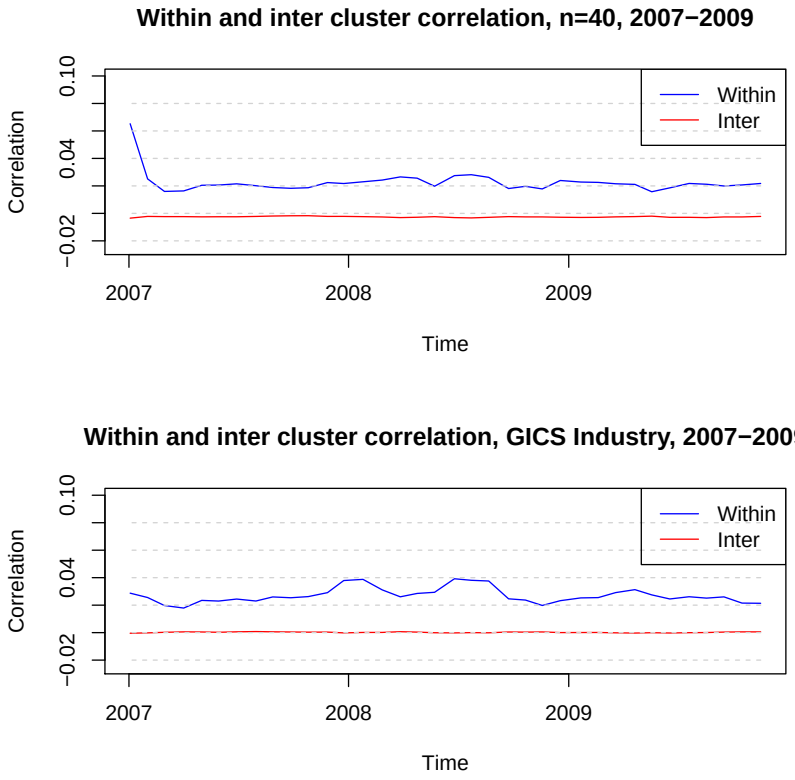

Figure 17: Correlation tracking in 2008 using rolling one month window for tracking, with clusters formed using six months of data, 68 GICS clusters and 40 hierarchical clusters. Baseline correlation removed.

Figure 18 shows the correlation tracking for year 2008 alone, using 12 hierarchical and GICS clusters with five day window and clusters formed from three months of data. From this we can see that in the first quarter of 2008 the GICS and hierarchical clustering have behaved differently. The hierarchical clustering has decreasing within correlation until mid-February where the correlation starts to rise again, whereas the GICS does not have similar concave pattern. In Figure 19 we can see the correlation tracking for 68 GICS clusters and 40 hierarchical clusters. Both hierarchical and GICS behave in same manner as with less clusters, but both have slightly higher within

cluster correlation. In this case the data driven clustering does not seem to outperform the GICS clustering as strongly.

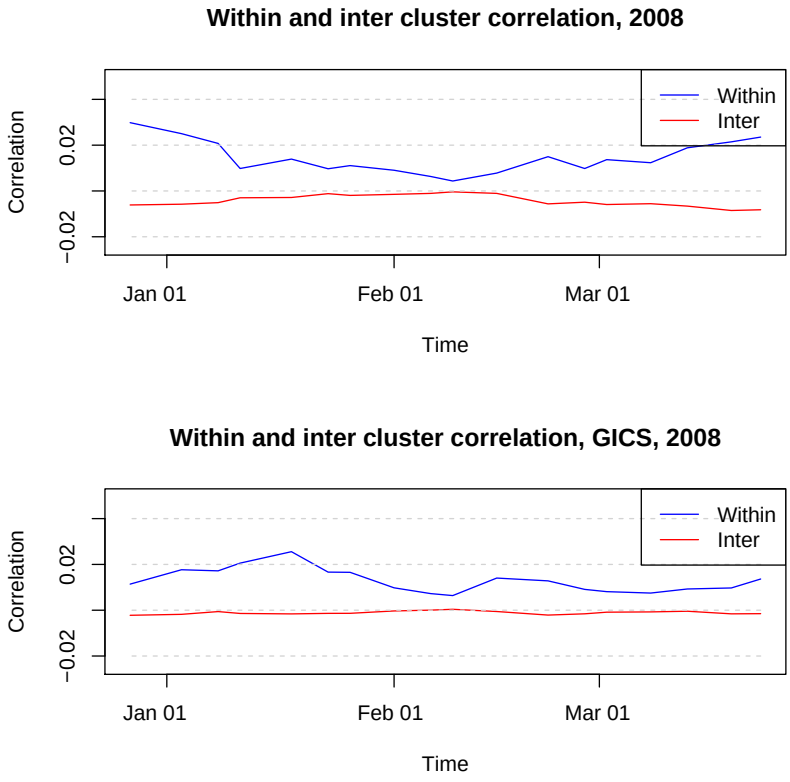

Figure 18: Correlation tracking in 2008 using rolling five day window for tracking, with clusters formed from three months of data, 12 GICS clusters and 12 hierarchical clusters. Baseline correlation removed.

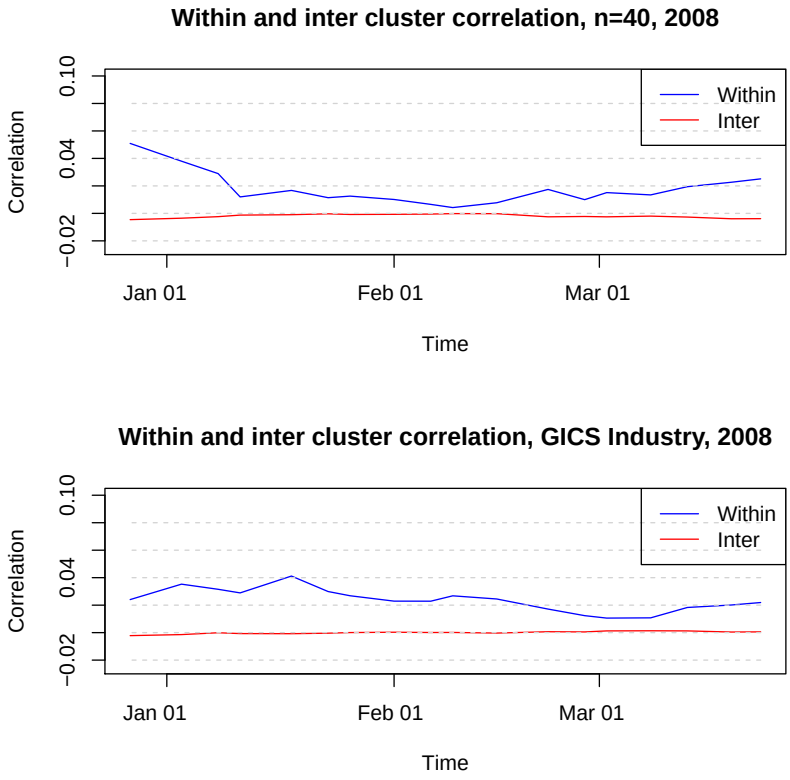

Figure 19: Correlation tracking in 2008 using rolling five day window for tracking, with clusters formed from three months of data, 68 GICS clusters and 40 hierarchical clusters. Baseline correlation removed.

We also tested earlier another approach to correlation tracking based on confusion matrix. This representation gives some rough idea how well results of two clustering outputs from different time windows resemble each other. An example of this can be seen in Table 4. which shows that the clustering is not stable. The rows are ordered to maximize the diagonal. The elements represent number of common stocks in two consecutive outputs of clustering algorithm. This is a arbitrary example but the results are very similar in all cases. The clustering is not stable, the stocks move from one cluster to another the consecutive windows.

Table 4: Typical example of confusion matrix produced from two consecutive clusterings. The rows are ordered to maximize the common stocks on the diagonal. The elements present number of common items in clusters. The clustering is very unstable, since assets do not remain in same clusters between clustering runs, but scatter to completely new clusters.

	1	2	3	4	5	6	7	8	9	10	11	12
4	14	11	4	10	22	20	9	8	8	9	4	6
5	7	14	37	18	29	14	4	11	5	2	9	13
6	17	26	90	37	145	46	30	18	26	28	21	27
11	8	3	11	13	18	6	6	3	12	1	3	4
3	4	16	102	36	287	62	18	13	9	11	4	8
2	9	20	52	41	115	63	16	9	25	25	9	16
10	9	5	12	14	10	5	9	3	7	7	0	8
8	13	23	53	21	93	36	27	27	28	14	8	17
7	1	0	14	9	12	4	5	6	17	6	1	3
9	8	8	26	14	39	12	15	6	21	21	6	12
12	2	1	4	1	4	1	1	2	4	1	0	2
1	17	16	22	25	28	13	20	9	26	24	5	24

5 Conclusions

Based on the results obtained we can draw some conclusions. The correlation and covariance forecasting based on GARCH models is currently not feasible for large number of stocks. There are better methods being developed and implemented at least in the academia using MATLAB and C. These should be kept in mind and reviewed again in the future.

The clustering based on historical returns and fundamentals of the stocks seems to be slightly better option for diversification than using solely GICS. At least in the data set used, the GICS classification provided no class for

many stock. FactSet provides their own classification which is complete, however, we saw no dramatic improvement in the performance using more complete data. The added granularity did increase the within cluster correlation but the clustering is not very stable. Based on results in correlation tracking, the achieved clustering in the project provides superior correlation performance for a time period of about a month from the end of the clustering data time period when compared to GICS. This would mean that portfolio allocations based on hierarchical clustering should be updated monthly with new data to provide sound basis for portfolio diversification.

The work accomplished so far is not exhaustive and plenty of new questions emerged during the process. We recommend that further work should focus on experimenting with different clustering parameters to find a model which has best performance in retaining high within correlation over time. Suitable parameters to tweak are e.g. length of the time period used for clustering, number of clusters, hierarchical clustering distance and linkage selection, selection of fundamental variables, normalization of fundamental variables, possible weighting of fundamental variables and data outlier detection and removal both with respect to fundamental variables as well as returns data. This is something that can be done straight off with the functions and ready script code made for the project with also novice knowledge on the matter.

Reasons on why we had somewhat lower overall stock correlation than some of the papers in the literature presented could be further investigated. It would be also interesting to study how much historical data and certain fundamentals explain of the variance of the stock. The list could be continued to even further.

More advanced development could be about testing other clustering methods besides hierarchical clustering such as k-means. Follow up on scalable and efficient GARCH models for forecasting the correlation structure of the stocks can also be fruitful.

6 References

[1] R. Engle, "Garch 101: The use of ARCH/GARCH models in applied econometrics," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 157–168, 2001.

[2] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982. [Online]. Available: http://www.jstor.org/stable/1912773

[3] R. Tsay, *Analysis of Financial Time Series*, ser. Wiley Series in Probability and Statistics. Wiley, 2005. [Online]. Available: https://books.google.fi/books?id=ddL4tTLb_08C

[4] G. Ali, "EGARCH, GJR-GARCH, TGARCH, AVGARCH, NGARCH, IGARCH and APARCH models for pathogens at marine recreational sites," *Journal of Statistical and Econometric Methods*, vol. 2, no. 3, pp. 57–73, 2013.

[5] L. Bauwens, S. Laurent, and J. V. K. Rombouts, "Multivariate GARCH models: a survey," *Journal of Applied Econometrics*, vol. 21, no. 1, pp. 79–109, 2006.

[6] A. Ghalanos, *rmgarch: Multivariate GARCH models.*, 2015, r package version 1.3-0.

[7] R. Engle, "Dynamic conditional correlation," *Journal of Business & Economic Statistics*, vol. 20, no. 3, pp. 339–350, 2002. [Online]. Available: https://doi.org/10.1198/073500102288618487

[8] R. F. Engle, O. Ledoit, and M. Wolf, "Large dynamic covariance matrices," Working Paper Series, University of Zurich, Department of Economics, Tech. Rep., 2017. [Online]. Available: http://www.econ.uzh.ch/static/wp/econwp231.pdf

[9] O. Ledoit, M. Wolf *et al.*, "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *The Annals of Statistics*, vol. 40, no. 2, pp. 1024–1060, 2012.

[10] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," *Artificial Intelligence Review*, Oct 2017. [Online]. Available: https://doi.org/10.1007/s10462-017-9588-9

[11] M. Abe and H. Nakayama, "Deep learning for forecasting stock returns in the cross-section," *arXiv preprint arXiv:1801.01777*, 2018.

[12] S. Kobayashi and S. Shirayama, "Time series forecasting with multiple deep learners: Selection from a bayesian network," *Journal of Data Analysis and Information Processing*, vol. 5, no. 03, p. 115, 2017.

[13] A. Navon and Y. Keller, "Financial time series prediction using deep learning," *arXiv preprint arXiv:1711.04174*, 2017.

[14] Morgan Stanley Capital International (MSCI), *Global Industry Classification standard*, 2018. [Online]. Available: https://www.msci.com/documents/10199/4547797/MSCI_GICS_Overview.pdf/00036370-db84-4d04-8180-0f4686abe7b5

[15] Australian portfolio investment, *GICS Sectors and Industries*, 2018. [Online]. Available: https://www.marketindex.com.au/asx-sectors

[16] N. Barberis, A. Shleifer, and J. Wurgler, "Comovement," *Journal of Financial Economics*, vol. 75, no. 2, pp. 283–317, 2005.

[17] S. S. Jung and W. Chang, "Clustering stocks using partial correlation coefficients," *Physica A: Statistical Mechanics and its Applications*, vol. 462, pp. 410–420, 2016.

[18] K. Gautam, "Does a cluster based factor model perform better than an industry based factor model? An investigation of European equities," Master's thesis, EDHEC Business School, Nice, 2012. [Online]. Avail-

able: https://new.cfasociety.org/france/Documents/Quant%20Awards%202012/Kumar%20Gautam-Research%20Paper-Quant%20Awards.pdf

[19] N. Basalto, R. Bellotti, F. De Carlo, P. Facchi, and S. Pascazio, "Clustering stock market companies via chaotic map synchronization," *Physica A: Statistical Mechanics and its Applications*, vol. 345, no. 1-2, pp. 196–206, 2005.

[20] R. N. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 11, no. 1, pp. 193–197, 1999.

[21] Z. Sun, "Institutional clientele and comovement," Working paper. University of California, Irvine - Paul Merage School of Business, 2015. Available at: https://dx.doi.org/10.2139/ssrn.1332201 note=2015 note=https://dx.doi.org/10.2139/ssrn.1332201.

[22] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2785–2797, 2015.

[23] L. K. Chan, J. Lakonishok, and B. Swaminathan, "Industry classifications and return comovement," *Financial Analysts Journal*, vol. 63, no. 6, pp. 56–70, 2007.

A Self-assessment

The project work was mostly carried out by making several sprints of few days instead of constant working throughout the spring. There were multiple periods of over one week when nothing was done for the project. This was mostly due to tight calendar of our team members and bad scheduling where the approaching deadline of deliverables became the force driving us to do something for the project work.

In general, the project work turned out to involve a bigger workload than expected. Even the clustering of assets, which in the beginning of the project was supposed to be the one of the easiest parts, turned out to be quite troublesome. Our knowledge of these methods was from previous courses and was limited which led us to believe that implementing them to any type of data set would be fairly doable. As we dug in to these methods and the given data set it became clear to us that there are so many things that could and should be accounted for when using certain types of methods for our data set that we would not have the time to do even a nearly perfect analysis. At the end of the project there were lots of different things that we could have investigated but did not have the time to do so. The approach of using forecasting models to form the correlation matrix was abandoned too late. With proper literature review and model testing as stated in the project plan this approach could have been abandoned nearly a month earlier. Our team could have benefited on at least trying giving itself more strict deadlines for action points.

Some of the scheduling problems we had were caused by slow communication both from our and ELO side. We did not ask ELO as much questions as we could have and some data deliveries took longer than expected, in which case we possibly should have requested more actively the materials from client.

All team members were working part time in their own respective jobs and also had other courses for the spring. The project work was not prioritized by

us as much as these things and other aspects of life. Week before and after the 1st of May celebration in Finland had us basically abandon the project work for a while. The core of our project work, the R-code, was mostly written by only one team member. This was a bottleneck which led to other two team members not able to continue the project work sometimes even though they could have had the time for it. Also since the team members had scheduling issues, communication between team members regarding to who is on which phase of their work was not always successful.

The fact that the course is graded only in terms of passed or failed had a negative effect on our motivation of finishing the project work properly. Since we are not graded with the usual number scale we had the impression that our performance will not be thoroughly assessed, which led to decreased motivation. The course was also one of the last courses in university for all the team members and with all the team members being eager to fully start the work life there might have not been as much motivation for the course compared to earlier courses.

The project work was successful in teaching us different types of forecasting and clustering methods and how to implement them on a massive data set. We also learned not to make assumptions on workload caused by different methods before performing a thorough review on these methods. Most importantly the project work taught us project skills like how to scope a project properly and how important it is to stay in schedule so that possible surprises do not ruin the project flow. In future projects our team members will be able to internally and externally communicate and keep track of what should be done next much better.