ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Econometrics and Operational Research

Improving the Extended Ultrametric Covariance Structure (EUCovS) for Gaussian Mixture Model Clustering: Application on Financial Ratios

Author:
Sjamiel Bagirov
Student ID number:
536308

Supervisor:
Dr. Carlo Cavicchia
Second assessor:
Dr. Nick Koning

3 July 2022

Abstract

In this paper, we extend the model-based partitioning algorithm (GMMEUCovS) of Cavicchia et al. (2022) and examine hierarchies between financial ratios and 1-year-ahead returns. Investors can use these findings to differentiate between the degree of importance of financial ratios. Utilizing data sets consisting of 65 different financial ratios and 1-year-ahead returns of companies in the S&P 500-index, we showed that numerous ratios are hierarchically linked with these returns. Our alternative procedure that transforms the parameterized covariance structure to be positive definite and ultrametric operates particularly competently with the high-dimensional financial data set. Although this procedure caused enlarged computational times, the canonical representation of Archakov & Hansen (2020) is able to partly offset it.

Keywords: *GMM, Hierarchical Clustering, Spectral Decomposition, Financial Ratios*

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

Financial ratios are a widely used tool by investors to determine whether a company is attractive to invest in. Popular examples of such ratios are the Price-to-Equity, Return-on-Invested-Capital and Debt-to-Equity ratios. One can expect that we can find clusters of stock companies based on these ratios as we have already identified sectors such as Energy, Services, Consumer Staples and Technology that differ substantially in ratio characteristics. Additionally, it would be interesting to cluster stocks based on these ratios where we also use annual returns as variable. That way, when clusters of stocks are formed, we can examine which financial ratios are the most hierarchically linked with the realized returns. Investors can use these findings to determine which changes in financial ratios they should pay more attention to. Generally, when we try to regress a company's returns on its financial ratios, we can end up with some significant ratios, but financial studies are not always consistent in determining which financial ratios have significant power. Mainly because it is dependent on the data; the used sample range and stock selection play a huge role. That is why it is worth studying whether groups of stocks can give a more reliable answer to the question "Which financial ratios are the most hierarchically linked with realized annualized returns?". We use 65 financial ratios tracked by Compustat-Capital IQ and WRDS for companies in the S&P 500 in 2021. Additionally, we compute 1-year-ahead returns for each corresponding publicly listed company using realized capital gains and dividend yields. When regarding these returns as variables, we can apply the partitioning algorithm to determine the importance of each variable. Not only that, but the covariances within and between groups of variables may display interesting relationships for investors.

As a result, we need a clustering model that is also able to display hierarchical structures among the variables within each cluster. The Gaussian Mixture Model with an extended ultrametric covariance structure (GMMEUCovS), given by Cavicchia et al. (2022) is therefore used. We alter the model in a way that we expect to be more precise and efficient (in terms of reduced running-time). Which is achieved by elegantly forcing the extended ultrametric covariance structure (EUCovS) to be positive definite and ultrametric and by making use of the canonical representation proposed by Archakov & Hansen (2020). Our alternative method of making the EUCovS positive definite and ultrametric consists of finding the nearest positive definite matrix (as described in Higham (1988)) and by making use of the adapted average linkage UPGMA algorithm for covariance matrices.

In sum, we propose the following formal research question: *"How does the improved Gaussian Mixture Model with an extended ultrametric covariance structure (GMMEUCovS) fit annual returns together with financial ratios?"* We answer this question by the following two sub-questions.

"How much more efficient does the GMMEUCovS fit annual returns together with financial ratios, when improving it with our alternative procedure of making the EUCovS positive definite an ultrametric and by making use of the canonical representation proposed by Archakov & Hansen (2020)?" and *"Where does the improved GMMEUCovS fit annual returns together with financial ratios?"* To our knowledge, there has been no research on using parameterized Gaussian Mixture Models on financial ratios and 1-year-ahead returns combined.

We found that there are a lot of financial ratios that are hierarchically linked with 1-year-ahead returns. Although thus no few ratios could be identified that have predictive power, we deduced that some ratios have explaining power with regards to the Price-to-Operating Earnings ratio. As prices are a substantial part of the calculation of realized returns (via the capital gain), the positive relationship between these ratios can provide insights into explanations for stock price changes. Furthermore, our proposed alternative procedure - which makes the parameterized covariance matrix positive definite and satisfies the ultrametric inequality - works well with our high-dimensional data set. This procedure did enlarge computation time, which is partly compensated for by the canonical representation of Archakov & Hansen (2020).

The remainder of this thesis is structured as follows: In Section 2, we outline the literature studies that are already done in the field of clustering financial ratios together with stock returns. Next to that, we provide the methodological background to model-based clustering algorithms. In Section 3, we explain the abstracted data and provide relevant data characteristics. This is followed by our mathematical models, algorithms and procedures in Section 4 where we display the used techniques to solve this optimization problem. Then, in Section 5, we present the findings where all computational results are displayed. Finally, Section 6 concludes this paper in conjunction with a discussion thereof.

2 Literature

Financial ratios are a widely used tool for investors (Barnes, 1987). They provide insightful data on the financial attractiveness of a company. Warren Buffett, arguably the world's most famous investor, compounded an average annual return of 20.1 % from 1965 through 2021 with his company 'Berkshire Hathaway'. Compared to 10.5% for the S&P 500 index, this out-performance is staggering [1]. It calls for studying his methods. During his value-investing approach to value companies and spot undervalued securities, he likes to use financial ratios as a tool for his investment thesis. For example, he likes to own companies that have a Current Ratio of greater than 1.5 and a Debt-To-Equity ratio of less than 0.5. And with him, a lot of other famous

[1] https://www.barrons.com/articles/warren-buffett-berkshire-hathaway-performance-51651885365

investors also make use of ratios such as Carl Icahn (with Price-To-Earning and Book-To-Market ratios) and Peter Lynch (with the PEG ratio). Nowadays, any person can find out a company's financial ratios in a few seconds with some clicks on the internet. As this type of data continues to become more and more accessible to the general public, it calls for researching their predictive performance regarding returns.
Öztürk & Karabulut (2018) showed that Earnings-to-Price and Net-Profit-Margin are good explanatory variables to regress on stock returns (measured with publicly listed companies from the İstanbul Stock Exchange). However, they found that another financial ratio is insignificant in explaining stock returns, the Current-Ratio. Lewellen (2004) also found some key/significant financial ratios to predict stock returns and discovered that some of them are time-dependent financial ratios such as the Book-To-Market ratio. On the contrary, Musallam (2018) found that their researched financial ratios were of insignificant power to explain stock returns (measured with 26 publicly listed companies in Qatar). This leads to uncertainty about which financial ratios work best for predicting stock returns. Until now, clustering stocks based on their financial ratios is usually done to generate diversified portfolios to reduce risk. Bin (2020) found that the combination of Sharpe ratios and the k-means clustering algorithm based on financial ratios can provide portfolios that generate greater returns than their market benchmark. Babu et al. (2012) found that clustering algorithms on stock price movements after financial news could generate abnormal returns compared to Support Vector Machines. However, we can also use hierarchical clustering models to include both realized returns (end of the year) and financial ratios (at the beginning of each year) to implicitly find financial ratios that are the most hierarchically linked with these returns. This way, we can indirectly pick financial ratios that have some predictive power concerning returns.

Clustering stocks, based on their financial ratios and returns, while partitioning the relationships between variables in each cluster can be regarded as an NP-hard problem (Křivánek & Morávek, 1986). When doing so with a Gaussian Mixture Model (GMM), we need an estimation of the covariance matrix to identify a cluster's orientation, shape and volume. When considering multidimensional data, this results in having to estimate a relatively large number of parameters, generally causing computational problems. For that, parameterizations of the covariance matrix in the GMM are researched to deal with the large data-parameters inefficiency. One of those parameterizations is done by a so-called eigendecomposition (Celeux & Govaert, 1995) which allows for different Gaussian Parsimonious Clustering Models (GPCMs) to be made. When we are dealing with high-dimensionality of parameters, the High Dimensional Data Clustering (HDDC) method (Bouveyron et al., 2007) is a robust clustering method that avoids overfitting of param-

eters and also uses the eigendecomposition of a covariance matrix. Another way to parameterize the covariance matrix is by using a mixture of factor analyzers model, which allows for Expanded Parsimonious GMMs (EPGMMs) to be made (McNicholas & Murphy, 2008). Cavicchia et al. (2022) introduced a new parameterization (which we implement in our research) that can flexibly describe the hierarchical structure among variables while displaying the clusters between observations. This method allows for a new parsimonious GMM to be made, the Gaussian Mixture Model with an extended ultrametric covariance structure (GMMEUCovS). The general concept of this extended ultrametric covariance structure is expressed in its ability to display and utilize the latent hierarchical structure among multidimensional variables from different clusters.
In Cavicchia et al. (2022), the GMMEUCovS was applied to the OECD Well-Being Data set (among another rather low-dimension real data set) to study its performance in terms of providing insights into the relationships among the indicators. Their success confirmed that the parsimonious model is not only applicable to synthetic (i.e. simulated) data but also on real data. Their research demonstrated that the GMMEUCovS can potentially also be used for multidimensional heterogeneous data.

3 Data

In this section, we discuss the data that we have used in this paper. In Section 3.1 we explain how the data on all the financial ratios and annual returns are gathered and what they consist of. There, we also work out how potential missing data are handled. In Section 3.2, we summarize relevant statistics on the researched stock companies and give insightful graphs on the data.

3.1 Financial ratios and performance data

We abstract financial data from Compustat-Capital IQ and WRDS on annual realized returns and the financial ratios of all of the listed companies in the Standard and Poor's (S&P) 500 index for the period 2021. We measure the financial ratios at the beginning of the year (i.e. January 1st) such that relationships between financial ratios and returns can be made. This way, we can determine for each cluster of stocks which financial ratios (beginning of the year) are more hierarchically linked with realized annual returns (end of the year). The list of the 65 financial ratios can be found in Appendix Table 4 in Section 7.1. The list includes ratios that can be characterized by their economic intuition as follows [2]. Starting with the largest group of ratios, 22 different ratios are based on a company's financial soundness/solvency which indicates a company's resources to handle the long-term obligations, this includes ratios such as

[2] Obtained from WRDS Research Team, "WRDS Industry Financial Ratio"(2016)

the Total Debt to Equity Ratio and the Short-Term Debt/Long-Term Debt ratio. We consider 11 different valuation ratios that can suggest a company's economical attractiveness. Examples are the popular Price-To-Earnings and Price-To-Book ratios. Next to that, 15 ratios are based on a company's profitability ratios, indicating a company's ability to make profits. Popular examples of those ratios are the Gross Profit Margin and the After-tax Return on Invested Capital. 6 ratios are based on a company's efficiency, the degree of effectiveness in its utilization of assets and liabilities. Examples are Sales/Invested Capital and Asset Turnover. Lastly, a couple of other ratios are considered such as capitalization and liquidity ratios. We calculate the annual realized returns as follows.

$$R_{it} = \frac{P_{it} - P_{i(t-1)} + D_{it}}{P_{i(t-1)}}, \tag{3.1}$$

where R_{it} is the annual realized return for company i at time t, where i=1,...,$\sim$ 500 and where t corresponds to a certain year. Furthermore, P_{it} denotes the stock price of company i at year t (measured on 31 December) and D_{it} denotes the received dividends for shareholders at company i during year t. We do this for all the companies that are listed in the index during 2021 to avoid survivorship bias. Namely, at the time of writing, 4 companies have been added and 5 companies have been removed from the list since 31 December 2021. The removals were either due to being acquired by another company or due to market capitalization changes. Also, in the index constituents, there may be more than 500 companies listed in the S&P 500 as there can be companies with two share classes that are both big enough to enter the list but tend to be highly correlated in practice (financial ratios-wise and return-wise). For that reason, we omit the ones that are added later than the other. Popular examples are $GOOG & $GOOGL and $BRK-A & $BRK-B. Furthermore, when observing the data, we noticed that some companies lack a lot of financial ratios data. As we are dealing with annual ratios, we do not believe that the interpolation of these ratios is reasonable as they are extremely time-dependent. To still identify clusters, we filter the very least amount of companies out such that our data set only contains companies of which all financial data are available.

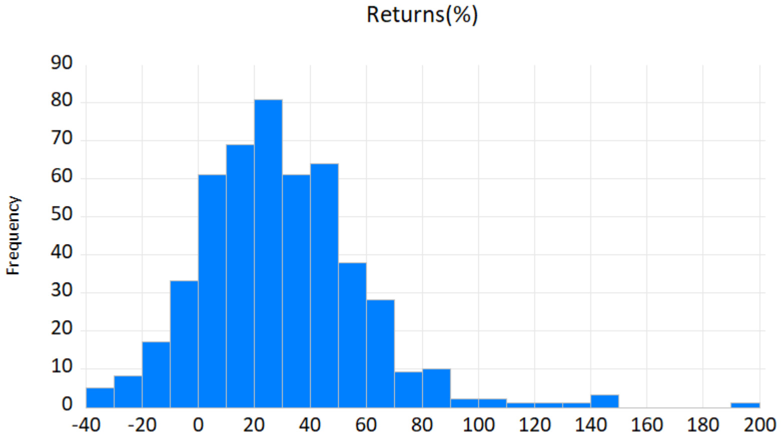
3.2 Characteristics and figures

In Table 1, we present the return data characteristics followed by a graph of the annualized realized returns for all the S&P 500 constituents in 2021 in Figure 1.

Table 1: Return statistics

Mean	**Median**	**St.Dev**	**Max.**	**Min.**	**Obs.**
27.95	25.93	28.72	191.78	-39.97	495

Figure 1: Returns visualization of S&P 500 companies in 2021



With an average mean return of roughly 28% in 2021, stockholders of companies in the S&P 500 generally have had a great year as a result of countries conquering COVID-19 and the corresponding trend-reversion in stock prices due to recovering investor confidence. However, with a standard deviation of almost 30%, the spread is wide. The biggest return came from the firm 'Devon Energy' which almost had a 200% growth in 2021. On the other hand, 'Penn National Gaming, Inc' was the worst performing stock which experienced a roughly 40% decline in shareholder value in 2021.

For the year 2021, the correlations between financial ratios and returns are listed in Appendix Table 5. The financial ratio with the highest (absolute) correlation is found the be the Book-to-Market ratio, with 0.29. This means that companies are expected to have relatively higher realized returns when they have had a relatively higher Book-to-Market ratio at the beginning of the year, certeris paribus. This finding is similar to the findings of Lewellen (2004). Still, we find that most correlations are generally low. This already indicates that a lot of these variables would have insignificant power in traditional OLS regressions on returns, which is similar with the findings of Musallam (2018). A possible explanation for this is that the correlation table is based on all stocks (i.e. non-clustered). This can cause hidden relationships between ratios and returns of certain clusters of stocks to fade away. This leads us to research whether hierarchically clustering methods can single out a few ratios that are the most hierarchically linked with returns.

4 Methodology

In this section, we explain the models and methods we use in our research. We start by describing the Gaussian Mixture Model in Section 4.1. In Section 4.2, we discuss the new parameterization of the covariance matrix given by Cavicchia et al. (2022) by explaining its procedure for finding the optimal clustering. There, we also propose a method that elegantly forces our covariance matrix to be positive definite without losing its ultrametric property. We show how we can make this procedure more efficient in Section 4.3 by making use of a canonical representation described in Archakov & Hansen (2020). Finally, in Section 4.4 we discuss some evaluating indices we use to justify optimal solutions.

4.1 Gaussian Mixture Models

We can cluster our data of returns and financial ratios of stock companies using a finite mixture model that, with a certain distribution, can model the density of this set of observations (denoted as $\mathbf{x}$) that get composed of G homogeneous clusters. The Gaussian Mixture Model (GMM) for model-based clustering assumes the following density.

$$f(\boldsymbol{x} \mid \boldsymbol{\Psi})=\sum_{g=1}^{G} \pi_g \phi\left(\boldsymbol{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), \tag{4.1}$$

where π_g is a non-negative mixing proportion (the a piori probability of an observation belonging to each corresponding cluster) that, together with the other mixing proportions, sum to one. That is, $0 \leq \pi_g \leq 1$ and $\sum_{g=1}^{G} \pi_i = 1$. Furthermore, the term $\phi\left(\boldsymbol{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$ denotes the density of a multivariate Gaussian distribution and $\boldsymbol{\Psi}$ is the general parameter vector. As we are dealing with P variables (with P=66 different variables (including returns)), the mean vector and covariance matrix ($\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ respectively) are both P-dimensional. In this GMM, the mean vector $\boldsymbol{\mu}_g$ determines the center of the ellipsoid while the covariance matrix $\boldsymbol{\Sigma}_g$ determines the orientation, shape and volume. In total, GMMs need to estimate $(G-1)+GP+GP(P+1)/2$ parameters: the mixing proportions, mean vectors and covariance matrix respectively where G is the number of clusters and P equals the number of variables.

4.2 New parameterization

In the research of Cavicchia et al. (2022), a parameterization of the covariance matrix is introduced to obtain a parsimonious model that is able to cluster data while also being able to display hierarchical relationships between variables. Their Extended Ultrametric Covariance Structure (EUCovS i.e. $\boldsymbol{\Sigma}_u$) is the proposed parameterization of the covariance matrix which is showed to be implementable into a GMM. The parameterization is defined as

$$\boldsymbol{\Sigma}_{\mathrm{u}}=\boldsymbol{V}\left(\boldsymbol{\Sigma}_W+\boldsymbol{\Sigma}_{\mathrm{B}}\right) \boldsymbol{V}^{\prime}-\operatorname{diag}\left(\boldsymbol{V} \boldsymbol{\Sigma}_{\mathrm{W}} \boldsymbol{V}^{\prime}\right)+\operatorname{diag}\left(\boldsymbol{V} \boldsymbol{\Sigma}_{\mathrm{V}} \boldsymbol{V}^{\prime}\right), \tag{4.2}$$

where $\boldsymbol{V}$ denotes the $(P \times Q)$ membership matrix that is constrained to be binary, row-stochastic and with non-empty groups, indicating whether a variable p corresponds to group q, with $p = 1, \ldots, P$ variables and $q = 1, \ldots, Q$ groups of variables. $\boldsymbol{\Sigma}_V$ symbolizes a diagonal $(Q \times Q)$ matrix where the elements on the diagonal correspond to the variances of the groups of variables. $\boldsymbol{\Sigma}_W$ is another diagonal $(Q \times Q)$ matrix but where the elements on the diagonal correspond to the covariances within the group of variables. Finally, $\boldsymbol{\Sigma}_B$ is a $(Q \times Q)$ matrix that is constrained to be symmetrical and where the diagonal elements are constrained to be 0 and where the off-diagonal elements correspond to the covariances between groups of variables.

When generating the EUCovS parameterization as defined in formula 4.2, it is subjected to the following constraints

$$\begin{aligned}
&{}_B\sigma_{qh} \geq \min\left\{{}_B\sigma_{qs}, {}_B\sigma_{hs}\right\} q, h, s = 1, \ldots, Q, s \neq h \neq q; \\
&\min\left\{{}_W\sigma_{qq} : q = 1, \ldots, Q\right\} \geq \max\left\{{}_B\sigma_{qh} : q, h = 1, \ldots, Q, h \neq q\right\}; \\
&{}_V\sigma_{qq} > |{}_W\sigma_{qq}|(\sum_{l=1}^{p} v_{lq} - 1) + \sum_{\substack{h=1 \\ h \neq q}}^{Q} |{}_B\sigma_{qh}| \sum_{l=1}^{p} v_{lh} \quad q = 1 \ldots, Q,
\end{aligned} \tag{4.3}$$

Here, the first constraint displays the ultrametric inequality of matrix $\boldsymbol{\Sigma}_B$ which corresponds to the condition that for every $q, h, s = 1, \ldots, Q$, we are able to reorder triplet {q,h,s} such that ${}_B\sigma_{qh} \geq {}_B\sigma_{qs} = {}_B\sigma_{hs}$, i.e. such that the smallest two elements are equal (Cavicchia et al., 2022). The second constraint ensures that the smallest diagonal element of $\boldsymbol{\Sigma}_w$ is still larger than the largest off-diagonal element of $\boldsymbol{\Sigma}_B$ causing the column pointwise diagonal dominance in the summation of these two matrices to hold which helps $\boldsymbol{\Sigma}_u$ to satisfy the ultrametric constraint. The inequality in the last constraint, leads to the matrix $\boldsymbol{\Sigma}_u$ to have positive values on the diagonal, helping to satisfy the ultrametric inequality and ensures strict diagonal dominance such that our matrix is a (strict) extended ultrametric covariance matrix that is positive definite. This strict inequality may lead to the parameter $\boldsymbol{\Sigma}_V$ to be overestimated and display exploding behavior. For that, Cavicchia et al. (2022) provided two constraints that are used to replace this potentially troubling constraint.

$$\begin{gathered}
{}_V\sigma_{qq} \geq \max\left\{ |{}_W \sigma_{qq}|, |{}_B\sigma_{qh}|, h = 1, \ldots, Q, h \neq q\right\} \quad q = 1, \ldots, Q, \\
\boldsymbol{\Sigma}_{\mathrm{u}} = \boldsymbol{\Sigma}_{\mathrm{u}} + a\boldsymbol{I}_P, \text{ with } a > 0, \text{ and such that } \boldsymbol{\Sigma}_{\mathrm{u}} \text{ is positive definite.}
\end{gathered} \tag{4.4}$$

Here, the first constraint ensures that the largest element of $\boldsymbol{\Sigma}_u$ is always at least bigger than the absolute values of its off-diagonal elements, causing the column pointwise diagonal dominance of $\boldsymbol{\Sigma}_u$ to be satisfied, but not necessarily strictly. In the second constraint, α is equal to a small arbitrary positive ϵ (for example equal to 0.1^6) plus the absolute value of the smallest eigenvalue of $\boldsymbol{\Sigma}_u$ (as suggested by Cailliez (1983)), and $\boldsymbol{I}_P$ the $(P \times P)$ identity matrix. With the addition of the absolute value of the smallest eigenvalue, the eigenvalues of the original $\mathbf{\Sigma}_u$ become non-negative (following from the fact that the eigenvalues of $\mathbf{\Sigma}_u + \alpha \boldsymbol{I}_P$ are equal to the eigenvalues of $\mathbf{\Sigma}_u + \alpha$). Next to that, the addition of ϵ in α, ensures strict positive eigenvalues, i.e. a positive definite covariance matrix. Under these two constraints, this replacement does not alternate the identity of $\mathbf{\Sigma}_u$ as it is still an extended ultrametric covariance matrix and changes only a relatively few elements of $\mathbf{\Sigma}_u$, namely only the diagonal elements.

4.2.1 Alternative procedure

However, after the research done by Cailliez (1983), there has been more research done in the field of finding the nearest possible matrix that is positive definite. We can make use of a spectral decomposition of the original matrix and recompose it back, but with negative eigenvalues forced to be bigger than 0 (e.g. an ϵ bigger than 0), which can be applied on general symmetric matrices but also on correlation matrices (inspired by Higham (1988) and Higham (2002) that replaces the negative eigenvalues by 0 to get a positive semi-definite covariance matrix). This approach looks as follows[3]:

Step 1: Let $\boldsymbol{B} = (\mathbf{\Sigma}_u + \mathbf{\Sigma}_u^T)/2$, which is just $\mathbf{\Sigma}_u$ because it is symmetrical since we have imposed the first restriction in 4.3.

Step 2: Determine the spectral decomposition of $\boldsymbol{B}$, that is, find a projection matrix $\boldsymbol{R}$ such that we can write $\boldsymbol{B} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^T$ where $\boldsymbol{D}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{B}$. The orthonormalized columns of $\boldsymbol{R}$ span the eigenspace of $\boldsymbol{B}$ and are formed by the eigenvectors of $\boldsymbol{B}$. The inverse of this matrix, by definition of the orthonormality, is its transpose.

Step 3: Whenever an element (i.e. eigenvalue) of $\boldsymbol{D}$ is non-positive, replace this element with an arbitrary positive number ϵ (e.g. 0.1^6 to be consistent with the approach of Cavicchia et al. (2022)).

Step 4: Recompose $\mathbf{\Sigma}_u$ back by $\mathbf{\Sigma}_u = \boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}^T$ where $\tilde{\boldsymbol{D}}$ is the updated diagonal matrix containing only positive elements. The updated $\mathbf{\Sigma}_{\boldsymbol{u}}$ now only contains positive eigenvalues by force, making it, by definition, positive definite.

This results, in Frobenius norm, a unique approximate of $\mathbf{\Sigma}_u$ but it does not ensure that the ultrametric condition keeps holding as sometimes the off-diagonal elements of $\mathbf{\Sigma}_u$ may change. To deal with that, we apply the adapted average linkage UPGMA algorithm for covariance matrices. This hierarchically clustering algorithm uses average covariances from clusters such that the nearest positive definite matrix satisfies the ultrametric inequality. This approach may cause the eigenvalues, which we just updated to be positive, to be non-positive again as we have

[3] Summarized algorithm code available in Appendix 7.2.2, Algorithm 2

re-updated the matrix. Hence, we do another spectral decomposition and change the eigenvalues to be positive and recompose it back to find the nearest positive definite matrix. As we keep re-evaluating with the updated matrix, one can argue that this approach resembles a greedy heuristic of finding a positive definite matrix that is ultrametric. This way, however, we may potentially end up in a never-ending loop. So for this, we set a limit to the number of re-evaluations (e.g. 5 re-evaluations). If this limit is met, and we did not find a matrix that satisfies the ultrametric inequality and is the nearest positive-definite matrix to our original $\mathbf{\Sigma}_u$, then we still use the original solution of Cailliez (1983). This approach is elegant in the 3 following scenarios. 1. When there is no need for changing eigenvalues, i.e. our symmetric $\mathbf{\Sigma}_u$ already was positive definite. Then our approach does not update anything and we keep our original $\mathbf{\Sigma}_u$ as in Cavicchia et al. (2022). 2. When we have to re-evaluate a couple number of times, we damage the structure of the original covariance the least. This follows from the fact that our procedure keeps searching for the nearest positive definite matrix from the last (ultrametric) matrix. Therefore, our procedure can also be understood as a greedy heuristic to find a positive definite ultrametric matrix. And finally, 3. when we can not find a nearest positive definite matrix when imposing the ultrametric constraints, we use the solution of Cailliez (1983) so we end up damaging the structure of $\mathbf{\Sigma}_u$ equally (compared to the solution of Cailliez (1983)).

4.2.2 GMMEUCovS

This parameterization allows for a new Gaussian Mixture Model (GMM) with the assumption of an extended ultrametric covariance structure (GMMEUCovS). The density of the GMMEUCovS is as follows

$$
\begin{aligned}
f\left(\boldsymbol{x}_i \mid \boldsymbol{\Psi}\right) = & \sum_{g=1}^{G} \frac{\pi_g}{(2\pi)^{p/2}\left|\boldsymbol{V}_g\left(\mathbf{\Sigma}_{\mathrm{W}_g}+\mathbf{\Sigma}_{\mathrm{B}_g}\right)\boldsymbol{V}_g'-\operatorname{diag}\left(\boldsymbol{V}_g\mathbf{\Sigma}_{\mathrm{W}_g}\boldsymbol{V}_g'\right)+\operatorname{diag}\left(\boldsymbol{V}_g\mathbf{\Sigma}_{V_g}\boldsymbol{V}_g'\right)\right|^{1/2}} \\
& \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_g\right)'\left[\boldsymbol{V}_g\left(\mathbf{\Sigma}_{\mathrm{W}_g}+\mathbf{\Sigma}_{\mathrm{B}_g}\right)\boldsymbol{V}_g'-\operatorname{diag}\left(\boldsymbol{V}_g\mathbf{\Sigma}_{\mathrm{W}_g}\boldsymbol{V}_g'\right)+\operatorname{diag}\left(\boldsymbol{V}_g\mathbf{\Sigma}_{V_g}\boldsymbol{V}_g'\right)\right]^{-1}\right. \\
& \left.\left(\boldsymbol{x}_i-\boldsymbol{\mu}_g\right)\right\}.
\end{aligned}
\tag{4.5}
$$

Here, $\boldsymbol{x}_i$ is a random vector of length P, with $i = 1, \ldots, n$ ($\approx$ 500 S&P 500 firms) from a population consisting of G clusters. This parsimonious model still needs to estimate $(G-1)$ number of parameters for the mixing proportions and GP number of parameters for the mean vectors but it needs to estimate just $G(2Q+P-1)$ number of parameters for the covariance structure (Cavicchia et al., 2022). This is a reduction in comparison with the $GP(P+1)/2$ for a general GMM whenever the number of groups, Q, is smaller than $(P(P-1)+2)/2$ which is easily doable when we consider a lot of variables. To maximize the log-likelihood of this density, we maximize (as demonstrated according to Hathaway (1986)) $\boldsymbol{W}$ and $\boldsymbol{\Psi}$ in

$$\begin{aligned} \ell_{\mathrm{H}}(\boldsymbol{W}, \boldsymbol{\Psi}) = & \sum_{i=1}^{n} \sum_{g=1}^{G} w_{ig} \log \left(\pi_g \phi \left(\boldsymbol{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\mathrm{V}_g}, \boldsymbol{\Sigma}_{\mathrm{W}_g}, \boldsymbol{\Sigma}_{\mathrm{B}_g}, \boldsymbol{V}_g \right) \right) \\ & - \sum_{i=1}^{n} \sum_{g=1}^{G} w_{ig} \log \left(w_{ig} \right), \end{aligned} \tag{4.6}$$

where $\boldsymbol{W}$ is row-stochastic and has elements w_{ig} representing the probability of observation i belonging to cluster g. So it has elements between 0 and 1, with $i = 1, \dots n$ and $g = 1, \dots G$. We compute this maximization by the GMMEUCovS algorithm of Cavicchia et al. (2022) (shown in the Appendix Section 7.2.1) which is similar to the EM algorithm commonly utilized to estimate the parameters of a GMM. When the algorithm convergences, we can determine which observations correspond to which clusters by making use of the maximum a posteriori (MAP) approach on the matrix $\boldsymbol{W}$. This approach can determine the cluster membership matrix based on the computed probabilities w_{ig}. Furthermore, Cavicchia et al. (2022) argued that we generally should run the algorithm 20 times with different initialization points to augment the chance of reaching the global optimum solution at convergence.

To distinguish which combinations of numbers of clusters and groups of variables is optimal (in terms of best-fit w.r.t the number of parameters), we select by means of the Bayesian Information Criterion (BIC, Schwarz, 1978) given by

$$\mathrm{BIC} = 2\ell(\hat{\theta}) - \nu \log(n), \tag{4.7}$$

where ν is equal to the number of parameters, $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate of $\boldsymbol{\theta}$ generated by the algorithm and $\ell(\hat{\boldsymbol{\theta}})$ is the corresponding maximized log-likelihood. This way, the optimal model maximizes the BIC to pick the best combination of the number of clusters G and the number of groups of variables Q. The BIC is a widely used criteria for optimal data clustering since it has the feature of generating consistent solutions for the number of clusters (Keribin, 2000). When executing the algorithm, we need to select an upper-bound for the number of clusters G and for the number of groups of variables Q. Let us denote them by $maxG$ and $maxQ$ respectively. We determine these values by considering the eigenvalues for the $n \times n$ correlation matrix of observations and for the $P \times P$ correlation matrix of variables. We set $maxG$ ($maxQ$) equal to the number of eigenvalues that explain more than 1% of the total number of observations (variables). This way, we a execute a preliminary analysis to get a general idea about the dimensionality of the data set.

4.3 Canonical representation of EUCovS

The GMMEUCovS density function requires the computation of the determinant and inverse of the EUCovS matrix, which has nested blocks or submatrices. This computation can take a long time to finish as this P-dimensional covariance matrix can be quite large when we consider lots of variables. The canonical representation obtained by Archakov & Hansen (2020) facilitates these computations. This way, the log-likelihood in formula 4.6 can be easier generated as well, which is convenient as we evaluate this formula in each iteration of the GMMEUCovS algorithm. The canonical representation is a semi-spectral decomposition that can be applied to the covariance matrix with nested blocks. The matrix gets diagonalized apart from a single diagonal sub-matrix which is of dimension equal to the number of blocks in the original covariance matrix. In this sub-section, we explain this procedure of Archakov & Hansen (2020) while showing how we implement the canonical representation in the GMMEUCovS.

The EUCovS can be expressed as a block matrix that takes the following form

$$\boldsymbol{\Sigma}_u = \begin{bmatrix} \boldsymbol{B}_{[1,1]} & \boldsymbol{B}_{[1,2]} & \cdots & \boldsymbol{B}_{[1,Q]} \\ \boldsymbol{B}_{[2,1]} & \boldsymbol{B}_{[2,2]} & & \\ \vdots & & \ddots & \\ \boldsymbol{B}_{[Q,1]} & & & \boldsymbol{B}_{[Q,Q]} \end{bmatrix}, \tag{4.8}$$

where each $\boldsymbol{B}_{[j,k]}$'th block is of dimension $n_j \times n_k$ for $j, k = 1, \ldots, Q$. Each block is composed of the following elements

$$\boldsymbol{B}_{[j,j]} = \begin{bmatrix} \sigma_j^2 & \sigma_{jj} & \cdots & \sigma_{jj} \\ \sigma_{jj} & \sigma_j^2 & \ddots & \\ \vdots & \ddots & \ddots & \\ \sigma_{jj} & & & \sigma_j^2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}_{[j,k]} = \begin{bmatrix} \sigma_{jk} & \cdots & \sigma_{jk} \\ \vdots & \ddots & \\ \sigma_{kl} & & \sigma_{jk} \end{bmatrix} \quad \text{if } j \neq k. \tag{4.9}$$

$\boldsymbol{B}_{[j,k]}$ denotes the covariance between two groups of variables (and thus has identical elements) and σ_j^2 and σ_{jj} are the variance and covariance within the jth group of variables respectively.

Let the rotation matrix $\boldsymbol{R}$ be defined as

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{v}_{n_1} & 0 & \cdots & & \boldsymbol{v}_{n_1 \perp} & 0 & \cdots & 0 \\ 0 & \boldsymbol{v}_{n_2} & & & 0 & \boldsymbol{v}_{n_2 \perp} & & \vdots \\ \vdots & & \ddots & & & & \ddots & \\ 0 & \cdots & & \boldsymbol{v}_{n_Q} & 0 & \cdots & & \boldsymbol{v}_{n_Q \perp} \end{bmatrix}, \tag{4.10}$$

where $\boldsymbol{v}_{n_j}$ is a vector of dimension n_j that spans the eigenspace corresponding to the jth eigenvalue and $\boldsymbol{v}_{n_j \perp}$ is a matrix of dimension $(n_j \times (n_j - 1))$ that is orthogonal to $\boldsymbol{v}_{n_j}$ and orthonormal.

Per definition of orthonormality, $\boldsymbol{R}\boldsymbol{R}' = \boldsymbol{I}_n$, such that we can write $\boldsymbol{\Sigma}_u = \boldsymbol{R}\boldsymbol{R}'\boldsymbol{\Sigma}_u\boldsymbol{R}\boldsymbol{R}'$. Or, equivalently, $\boldsymbol{\Sigma}_u = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}'$ where we substitute $\boldsymbol{R}'\boldsymbol{\Sigma}_u\boldsymbol{R}$ for $\boldsymbol{D}$, the canonical form of $\boldsymbol{\Sigma}_u$ when rotated by $\boldsymbol{R}$. This diagonal matrix $\boldsymbol{D}$ takes the following form

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A} & 0 & \cdots & 0 \\ 0 & \lambda_1 \boldsymbol{I}_{n_1-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_Q \boldsymbol{I}_{n_Q-1} \end{bmatrix}, \tag{4.11}$$

where $a_{jk} = \begin{cases} \sigma_j^2 + (n_j - 1)\,\sigma_{j,j} & \text{for } j = k \\ \sigma_{j,k}\sqrt{n_j n_k} & \text{for } j \neq k \end{cases}$ and $\lambda_j = \sigma_j^2 - \sigma_{jj}$. These findings follow from the spectral decomposition of $\boldsymbol{\Sigma_u}$, using properties of the projection matrix (Archakov & Hansen, 2020, Theorem 1). The matrix $\boldsymbol{A}$ is not composed of blocks but does have dimension Q which is exactly equal to the block dimension of the original covariance matrix $\boldsymbol{\Sigma}_u$. The diagonal matrix $\boldsymbol{D}$ is related to the original $\boldsymbol{\Sigma}_u$ via sub matrix $\boldsymbol{A}$ and rotation matrix $\boldsymbol{R}$. Namely, the eigenvalues of $\boldsymbol{A}$ are also in $\boldsymbol{\Sigma}_u$ such that the rotation matrix $\boldsymbol{R}$, in its first Q columns, spans the eigenspace of $\boldsymbol{\Sigma}_u$ that is related with these eigenvalues that $\boldsymbol{A}$ and $\boldsymbol{\Sigma}_u$ have in common. The remaining columns of $\boldsymbol{R}$ are the residual eigenvectors of $\boldsymbol{\Sigma}_u$.

Archakov & Hansen (2020) also showed that, with this canonical representation of $\boldsymbol{\Sigma}_u = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}'$, we can transform the log-likelihood function corresponding to the density function for the multivariate Gaussian distribution. We implement this transformation into the density function for the GMMEUCovS. Now, when computing the log-likelihood at each iteration (Step 5), we turn to the following log-likelihood

$$\ell = \sum_{i=1}^{N} \log[\sum_{g=1}^{G} \pi_g 2\pi^{-p/2} \det \boldsymbol{D}^{-1/2} \times \exp\{\frac{-1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{R}\boldsymbol{D}^{-1}\boldsymbol{R}'(\boldsymbol{x}_i - \boldsymbol{\mu}_g)\}], \tag{4.12}$$

where Archakov & Hansen (2020) used that $\boldsymbol{\Sigma}_u^{-1} = (\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}')^{-1} = \boldsymbol{R}'^{-1}\boldsymbol{D}^{-1}\boldsymbol{R}^{-1} = \boldsymbol{R}\boldsymbol{D}^{-1}\boldsymbol{R}'$ (following from the orthonormality property of $\boldsymbol{R}$) and $\det \boldsymbol{R}\boldsymbol{D}'\boldsymbol{R} = \det \boldsymbol{\Sigma}_u = \det \boldsymbol{A} * \prod_{k=1}^{Q} \lambda_k^{n_{k-1}} = \det \boldsymbol{D}$. In the latter, Archakov & Hansen (2020) proved the second equality. In Appendix Section 7.2.3, we show that the third/last equality also holds.

Notice, with the canonical representation of $\boldsymbol{\Sigma}_u$ we are left to calculate the determinant of the $(Q \times Q)$ matrix $\boldsymbol{A}$ which is considerably easier than calculating the determinant of the $(P \times P)$ matrix $\boldsymbol{\Sigma}_u$ as in the original setting. When $Q \ll P$, we can determine the log-likelihood potentially much faster. Cavicchia et al. (2022) found with 15 variables that $G = 2$ and $Q = 4$ is the optimal combination of clusters of countries and groups of indicators - indicating that we may expect to find a solution such that $Q<P$.

When setting $\xi(\boldsymbol{x}_i|\boldsymbol{\mu}_g, \boldsymbol{D}_g, \boldsymbol{R}_g) = 2\pi^{-p/2} \det \boldsymbol{D}^{-1/2} \times \exp\left\{\frac{-1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g\right)' \boldsymbol{R}\boldsymbol{D}^{-1}\boldsymbol{R}'\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g\right)\right\}$ then Equation 4.12 can be rewritten as $\ell = \sum_{i=1}^{N} \log[\sum_{g=1}^{G} \pi_g \times \xi(\boldsymbol{x}_i|\boldsymbol{\mu}_g, \boldsymbol{D}_g, \boldsymbol{R}_g)]$. This rewriting enables us to again use Hathaway (1986) to rewrite this even further into

$$\ell_{\mathrm{H}}(\boldsymbol{W}, \tilde{\boldsymbol{\Psi}}) = \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig} \log(\pi_g \xi(\boldsymbol{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{D}_g, \boldsymbol{R}_g) - \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig} \log\left(w_{ig}\right). \tag{4.13}$$

This way, in every computation of Step 5 of the GMMEUCovS algorithm of Cavicchia et al. (2022), we can use $\xi(\cdot)$ instead of $\psi(\cdot)$ for potentially faster computation.

4.4 Evaluating the optimal cluster

After we have identified a certain optimal combination of the number of clusters G, and the number of groups of variables Q (based on BIC values), we want to determine whether the formed clusters are "good". Regarding this clustering analysis, literature often uses (one of) the following evaluation metrics; the Davies–Bouldin index and the Dunn index. We compute these metrics to justify our optimal solution found by comparing BIC values. The purpose of these internal evaluation indexes is to identify the quality of the obtained clusters. Hence we use them for justification of our optimal solution. This is done by computing the values of these two metrics for all possible combinations of G and Q. Note, as these two metrics differ in merits and drawbacks, we should carefully analyze the outcomes.

The Davies-Bouldin index (Davies & Bouldin, 1979) can be formulated as:

$$\mathrm{DB} = \frac{1}{G}\sum_{g=1}^{G} \max_{h \neq g}\left(\frac{\sigma_g + \sigma_h}{\mathrm{d}(\mu_g, \mu_h)}\right), \tag{4.14}$$

where $g, h = 2, \ldots, G$ (number of clusters), σ_g is the average distance of the elements in cluster g to centroid μ_g and $\mathrm{d}(\mu_g, \mu_h)$ denotes the distance between cluster g and h. We thus neglect the solutions for when $G = 1$ to avoid dividing by 0. Furthermore, as distances are strictly non-negative, the lower-bound of this index is 0. This index evaluates the degree of intra-cluster distances (i.e. the numerator, which we want to be as small as possible) relative to distances between clusters (i.e. the denominator, which we want to be as large as possible). Hence, based on this metric, the best clusters should have the lowest DB value.

The Dunn-Index (Dunn, 1973), which is a commonly used internal evaluation metric for clustering analysis, can be formulated as:

$$\mathrm{D} = \frac{\min_{1 \leq g < h \leq G} \mathrm{d}(g, h)}{\max_{1 \leq k \leq G} \mathrm{d}'(k)}, \tag{4.15}$$

where $g, h, k = 2, \ldots, G$ (number of clusters),$\mathrm{d}(g, h)$ indicates the distance between cluster g

and h and $\mathrm{d}'(k)$ equals the intra-cluster distance of elements in cluster k. Again, as distances are strictly non-negative, the lower-bound of this index is 0. This metric, similar to the Davies-Bouldin, compares distances between clusters with distances inside clusters to identify well-separated clusters. The Dunn-Index metric, however, puts relatively more weight on outlier clusters which is noticeable in its denominator: the cluster for which the intra-distances are maximized. With this index, a partitioning is suggested to be the best one when it has the highest value (well-separated (highly-densed) clusters such that the numerator (denominator) is as big (small) as possible). This merit can also be seen as a drawback: when all the clusters generally partition the data set well but 1 cluster performs bad (i.e. has big intra-distances), then the metric may suggest that the found partitioning is generally bad as the denominator is substantially high.

5 Results

In Section 5.1, we first determine the best number of maximum re-evaluations in our alternative procedure of making $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ positive definite. Next, we measure the potentially improved efficiency (in terms of running time) with our high-dimensional financial data set using the canonical representation of the EUCovS. In Section 5.2, we apply the algorithm to our financial data set to study the hierarchy between financial ratios and returns.

5.1 Precision and Efficiency

We measure the number of re-evaluations needed in our alternative procedure to determine the best maximum number of re-evaluations. We try all combinations of $G = 1$ and $Q = 1, \ldots, 24$. Starting with 500 as the maximum number of re-evaluations, our alternative procedure of making $\boldsymbol{\Sigma}_u$ positive definite first checks whether the eigenvalues already are positive which would, per definition, imply that our symmetric square EUCovS matrix is positive definite. This happens roughly 28% of the time. In the other 72% of the the time, our procedure could find in 77% of those cases a nearest positive definite matrix that satisfies the ultrametric inequality in exactly 1 iteration (so without having to re-evaluate). Hence, our procedure can find a positive definite ultrametric matrix in $0.28 + 0.72 * 0.77 \approx 83\%$ of times in 1 iteration. Occasionally, our procedure has to re-evaluate because our ultrametric matrix is no longer positive definite. We found that the number of such re-evaluations is highly dispersed. In most cases ($\approx 90\%$)), a feasible matrix is found in less than 5 re-evaluations. However, in the other cases, no feasible matrix is found and after 500 re-evaluations, we still end up using the procedure of Cailliez (1983). Thus, to reduce computation time, we advise to take 5 (or $\sim$1% of observations) as the maximum number

of re-evaluations. In sum, our procedure can return a positive definite ultrametric matrix $\mathbf{\Sigma}_u$ in $0.83 + 0.17 * 0.9 \approx 98\%$ of the time using less than 5 re-evaluations, which points out that the original covariance structure can be altered with relatively little damage due to the low number of iterations. Although both the original method of Cailliez (1983) and our alternative procedure concluded the same optimal solution, the difference will be more and more noticeable when the data dimensionality increases. That is, the approach of Cailliez (1983) alters every diagonal element of the covariance matrix quite heavily by increasing them with the absolute value of the minimum eigenvalue whereas our approach uses the nearest positive definite matrix.
Next, we measure the improved efficiency using the canonical representation on $\mathbf{\Sigma}_u$.

To make such a comparison, we differentiate between 4 different scenarios, which are:

- Scenario 1 (base scenario): Without the mathematical extensions, i.e. purely replicating the GMMEUCovS of Cavicchia et al. (2022).
- Scenario 2: Using only the alternative procedure of making $\mathbf{\Sigma}_u$ positive definite as an extension.
- Scenario 3: Using only the canonical representation of the EUCovS as an extension.
- Scenario 4: Using both extensions (i.e. Scenario 2+3).

The computation times are listed in Table 2.

Table 2: Computation times for all 4 different scenarios

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**
Running Time (sec)	3361	8655	3112	7429

From Scenario 2, we notice that the alternative procedure of making $\mathbf{\Sigma}_u$ positive definite has a substantial enlarged effect on the computation time. This is mostly due to the time-consuming process of making the large $P \times P$ matrix $\mathbf{\Sigma}_u$ ultrametric using the adapted average linkage UPGMA algorithm. This is partly compensated by the canonical representation, as in Scenario 4, but it still causes a significant time increase as opposed to the base scenario. However, it is clear that the canonical representation indeed reduces computation times as the running time for Scenario 3 is a 7.4% decrease compared to the running time of Scenario 1.
In sum, there seems to be a trade-off between precision and efficiency. The alternative procedure of making $\mathbf{\Sigma}_u$ positive definite often uses a small number of re-evaluations, thereby damaging the original structure of $\mathbf{\Sigma}_u$ the least. This comes at the cost of enlarged computation time due to the time-consuming adapted average linkage UPGMA algorithm to let our positive definite matrix satisfy the ultrametric inequality. To partly compensate for this enlarged computation time, the

canonical representation helps reduce running-time, but it does not reduce it significantly more than the case where we do not incorporate the alternative procedure at all.

5.2 Financial Ratios

For our financial data set, the optimal combination of the number of clusters G and the number of groups of variables Q consists of $G = 1$ and $Q = 6$ (with $g = 1, \dots, 9$ and $q = 1, \dots, 24$ tested). The soft K-means initialization also found that one cluster has the highest probability for all the observations. As our data set consists of large-cap stock companies listed in the USA, this absence of clusters could have been predicted as their ratios are expected to be correlated with each other to some degree. We presume that this absence of clusters evades when also including for example low-cap European stock companies in the data set. The optimal solution is based on the BIC values in Table 3, which display the highest value for G=1 and Q=6.

Table 3: BIC values

	BIC				**G**				
Q	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	-249210.380	-149247.278	-145753.220	-2968.756	-140758.413	-139565.907	-138615.067	-138558.537	-137786.246
2	-126470.200	151139.800	154293.200	154161.000	149363.500	151829.300	149573.500	146908.600	111435.300
3	153743.300	153275.700	-21765.400	155320.200	150185.400	147681.100	152310.500	151826.700	147438.500
4	-21155.040	153458.440	155759.810	153959.910	153562.340	153135.780	150614.330	152512.040	152551.320
5	156052.700	154588.000	154821.600	155120.100	154371.400	154024.400	141478.300	145706.400	145429.400
6	**156242.300**	153655.500	144168.200	149914.300	155899.900	155100.900	154280.700	151869.700	150603.400
7	155210.800	153828.500	153578.000	152715.000	155687.500	152352.400	145129.000	137215.700	152017.200
8	148819.840	154000.480	147151.620	74942.750	146554.320	152326.740	58455.500	114925.970	147761.880
9	142390.890	140792.900	125286.370	45962.460	138516.680	143272.700	98226.350	142891.580	142634.740
10	135996.410	135512.720	-35214.830	36568.800	138720.600	140105.090	90890.330	106634.240	135473.280
11	129497.710	132558.590	109290.750	43344.580	132660.500	125253.670	114016.930	85233.040	134055.760
12	123070.260	131004.210	93679.840	98598.660	130020.590	125779.150	108448.890	90674.780	124512.130
13	116130.031	119480.720	9686.839	103587.110	108772.227	113169.684	102841.315	114446.228	117570.126
14	110218.570	109673.260	53940.320	55748.700	116697.110	110432.670	56457.810	96900.120	106071.230
15	103946.141	112059.132	-5343.755	58600.246	102597.206	109199.857	97984.092	88180.755	110154.037
16	97264.680	78693.470	86840.040	48532.920	97434.890	102730.530	86060.440	75365.960	103356.980
17	93151.930	92429.490	61081.470	62361.700	99399.780	89573.590	71470.400	73646.670	99595.930
18	84513.430	97109.930	46514.160	65815.230	92235.740	94865.430	71716.890	71979.880	88616.000
19	77733.830	94147.510	70694.890	43985.370	87395.490	86765.800	77071.920	79356.650	89503.910
20	71226.880	89469.680	65557.590	34722.610	79773.260	81811.580	54312.380	81404.600	78295.110
21	64646.993	66650.163	-3911.085	16700.958	83619.841	76046.592	69332.074	54909.994	71011.331
22	58122.050	79674.740	31665.400	17559.510	63553.520	67576.780	48979.410	42455.030	70539.260
23	59886.580	-161653.780	42237.880	48544.140	72500.800	63345.000	47008.280	41450.550	67562.070
24	-164980.540	-174714.680	36704.360	41491.140	59324.680	61032.030	45116.470	40469.830	64710.530

We notice that the BIC values generally decrease for higher G and Q. Thus, generally, the penalty for taking on more clusters and groups of variables outweighs the increase in (log-)likelihood. To justify this solution, we turn to the Davies-Bouldin and Dunn indices values. These are listed in Appendix Table 6 and 7 respectively. Note, as we are dealing with $G = 1$ as the optimal solution, their values are infinite (as the distance between a group with itself is 0 such that we are dividing between 0). This can also be the case for higher-order clusters. Namely, when 1 soft cluster dominates the whole data-set, it reduces to, theoretically, 1 hard cluster (which happens quite often in our case as the optimal solution is, indeed, 1 cluster). To still get some justification for our optimal solution, we determine whether the solutions for $Q = 6$ are better than other solutions for Q. We do this by considering their average percentile rank, where we substitute the Inf/NaN values in these higher order clusters for the best solution found. Note, as these indices already should be interpreted with caution due to their drawbacks, we do not put too much weight on this approach. Still, based on Tables 6 and 7, the solution for $Q = 6$ is in the 33rd percentile (based on the Davies-Bouldin-Index values) and in the 70th percentile (based on the Dunn-Index values). Thus, based on the Dunn-Index, our optimal solution of $Q = 6$ can be partly justified. As this metric puts relatively more weight on outlier clusters, one can argue that this metric may be more applicable in this situation. That is, given that a substantial number of higher-order clusters deduce to (the same) 1 theoretical cluster, we are more interested in the appearances when this is not the case, i.e. the outliers. For Q=6, we have relatively many higher-order clusters deducing to 1 theoretical cluster, which justifies our solution of $G = 1$ and $Q = 6$.

Using 1 cluster and 6 groups of variables, we obtain the following groups.

- Group 1 (Interest Solvency): After-tax Interest Coverage and Interest Coverage Ratio.
- Group 2 (Singleton group): Sales/Stockholders Equity.
- Group 3: All the other 57 financial ratios + annual returns.
- Group 4 (Singleton group): Price/Book.
- Group 5 (Operating Earnings Valuation) : Price/Operating Earnings (Basic, Excl. EI), After-tax Return on Average Common Equity and Pre-tax return on Net Operating Assets.
- Group 6 (Singleton group): Shillers Cyclically Adjusted P/E Ratio.

Thus, the algorithm finds that many ratios are hierarchically linked with 1-year-ahead returns as the corresponding group (Group 3) consists of 57 ratios. While no few ratios thus have substantial predictive power, from Group 5 we can deduce which ratios have at least some explaining power. That is, Group 5 contains a valuation ratio, the Price/Operating Earnings, and two profitability ratios, the After-tax Return on Average Common Equity and the Pre-tax return on Net Operating Assets. Their positive covariances (see the heat map in Figure 2) displays positive relationships with each other. Therefore, when these two profitability ratios increase in value, the Price/Operating Earnings ratio is also expected to increase. This can happen when the Operating Earnings fall and/or the price increases. Hence, the two profitability ratios and Operating Earnings together can explain changes in stock prices, which is a significant part of the return calculation in formula 3.1. Next to that, this relationship can also identify whether a stock price is undervalued (overvalued). That is, when all components of the relationship have risen (fell) over time, except for the component 'Price', we can expect markets to correct for this imbalance to restore the relationship by the so-called 'Invisible Hand'. Furthermore, Group 2 has a negative covariance with Group 3. This indicates that Sales per dollar of Stockholders' Equity at the beginning of the year has a negative relation with annual realized returns (end of the year) which is in Group 3. Their negative correlation also confirms this relationship.

The covariances within and between groups of variables are displayed into a heat map in Figure 2 and the corresponding hierarchical tree is shown in Figure 3.

Figure 2: Heat map of variable covariances

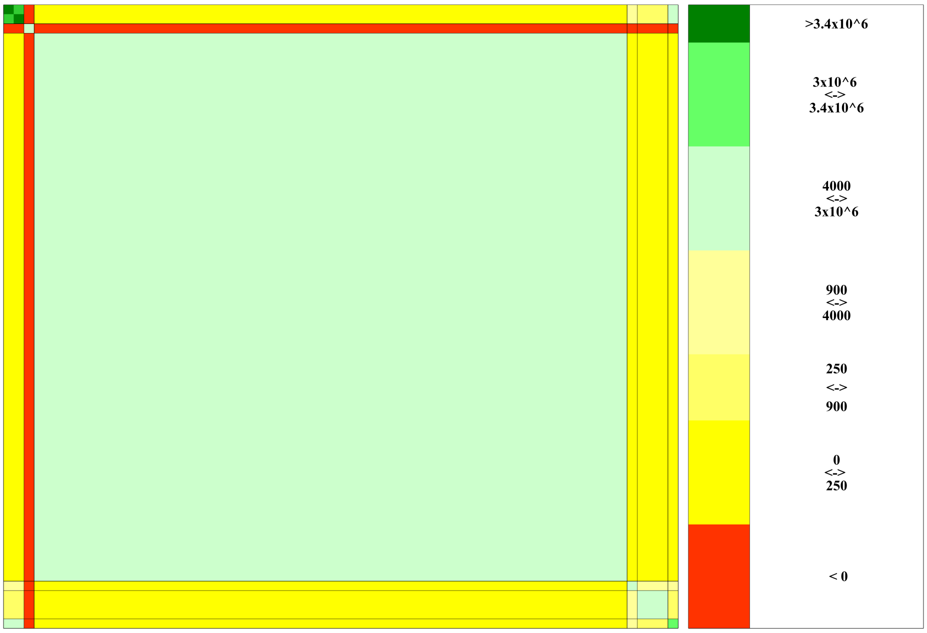
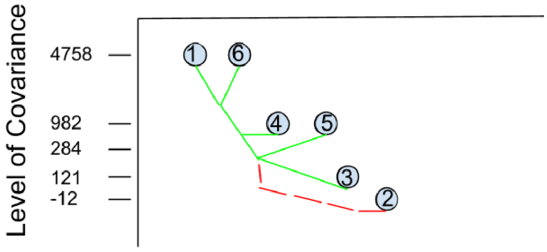


Figure 3: Hierarchical tree path diagram between groups of variables, unscaled



From this, we can also notice that Group 2 not only has a negative relation with Group 3 but also with the other groups of variables, making it discordant from the other groups. Furthermore, the covariances between Group 3 and the other groups (excluding Group 2) are rather small, which points out that the variables in Group 1, 4, 5 and 6 at the beginning of the year 2021, matter relatively little for annual realized returns compared to the variables in Group 3.

6 Conclusion and Discussion

In this paper, we researched *"How does the improved Gaussian Mixture Model with an extended ultrametric covariance structure (GMMEUCovS) fit annual returns together with financial ratios?"*. We found that returns are hierarchically linked with relatively many ratios. Thus, no few ratios stood out to possess extraordinary predictive power. However, we found that Price-to-Operating Earnings, After-tax Return on Average Common Equity, and Pre-tax return on Net Operating Assets are the most hierarchically linked with each other with positive relationships. As the stock price is a substantial part of the return calculation, these variables possess explanatory power in stock price changes. Furthermore, we improved the Gaussian Mixture Model with an extended ultrametric covariance structure (GMMEUCovS, proposed by Cavicchia et al. (2022)) by elegantly forcing the extended ultrametric covariance structure to be positive definite and ultrametric. The former is done by finding the nearest positive definite matrix, using its spectral decomposition and the latter is done by using the adapted average linkage UPGMA algorithm. While this approach results in enlarged computation time, the canonical representation (proposed by Archakov & Hansen (2020)) was able to partly compensate for it. Hence, the improved Gaussian Mixture Model with an extended ultrametric covariance structure fits annual returns together with financial ratios well and efficiently. With this research, investors can now establish which ratios are more/less important for predicting 1-year-ahead returns. Besides, given the positive relationship between Price-to-Operating Earnings, After-tax Return on Average Common Equity and Pre-tax return on Net Operating Assets, investors can spot undervalued securities when all parts of this relationship have risen, except the price. Furthermore, at the cost of enlarged computation time, our alternative procedure of making the extended ultrametric covariance structure positive definite paves the way for other Gaussian Mixture Models to also be computed more efficiently. That is, in cases when the parameterized covariance function has to be transformed such that it becomes positive definite, our greedy heuristic finds such matrix in relatively few iterations. For further research, besides utilizing more data years, educated guesses for alternative ratios with predictive power may be worthwhile to examine such that more accurate results may appear. Examples could be environmental scores or the percentage of women on boards and in the C-Suite. Methodological-wise, instead of estimating the membership matrix $\boldsymbol{V}$ row-by-row in the GMMEUCovS algorithm, one could improve this step more accurately when there is a priori knowledge of the number of variables per group. Namely, by regarding the estimation as a mixed integer linear programming (MILP) problem with binary, row-stochastic and fixed size constraints in the groups of variables. This may even reduce computation times as this could find optimal solutions faster instead of trying every row-composition of $\boldsymbol{V}$.

References

Archakov, I., & Hansen, P. R. (2020). A canonical representation of block matrices with applications to covariance and correlation matrices. *arXiv preprint arXiv:2012.02698*.

Babu, M. S., Geethanjali, N., & Satyanarayana, B. (2012). Clustering approach to stock market prediction. *International Journal of Advanced Networking and Applications*, *3*(4), 1281.

Barnes, P. (1987). The analysis and use of financial ratios. *Journal of Business Finance dan Accounting*, *14*(4), 449.

Bin, S. (2020). K-means stock clustering analysis based on historical price movements and financial ratios.

Bouveyron, C., Girard, S., & SCHMID, C. (2007, 09). High-dimensional data clustering. *Computational Statistics Data Analysis*, *52*, 502-. doi: 10.1016/j.csda.2007.02.009

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, *48*(2), 305–308.

Cavicchia, C., Vichi, M., & Zaccaria, G. (2022). Gaussian mixture model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification*, 1–29.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781-793. doi: https://doi.org/10.1016/0031-3203(94)00125-6

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224–227.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

Hathaway, R. J. (1986, March). Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters*, *4*(2), 53-56.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, *103*, 103–118.

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, *22*(3), 329–343.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 49–66.

Křivánek, M., & Morávek, J. (1986). Np-hard problems in hierarchical-tree clustering. *Acta informatica*, *23*(3), 311–323.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*, *74*(2), 209-235. doi: https://doi.org/10.1016/j.jfineco.2002.11.002

McNicholas, P. D., & Murphy, T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, *18*(3), 285–296.

Musallam, S. R. (2018). Exploring the relationship between financial ratios and market stock returns. *Eurasian Journal of Business and Economics*, *11*(21), 101–116.

Öztürk, H., & Karabulut, T. A. (2018). The relationship between earnings-to-price, current ratio, profit margin and return: an empirical analysis on istanbul stock exchange. *Accounting and Finance Research*, *7*(1), 109–115.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

7 Appendix

7.1 Data

Table 4: List of all the 65 different financial ratios considered

Financial ratios

- Shillers Cyclically Adjusted P/E Ratio
- Book/Market
- Enterprise Value Multiple
- Price/Operating Earnings (Basic, Excl. EI)
- Price/Operating Earnings (Diluted, Excl. EI)
- P/E (Diluted, Excl. EI)
- P/E (Diluted, Incl. EI)
- Price/Sales
- Price/Cash flow
- Dividend Payout Ratio
- Net Profit Margin
- Operating Profit Margin Before Depreciation
- Operating Profit Margin After Depreciation
- Gross Profit Margin
- Pre-tax Profit Margin
- Cash Flow Margin
- Return on Assets
- Return on Equity
- Return on Capital Employed
- Effective Tax Rate
- After-tax Return on Average Common Equity
- After-tax Return on Invested Capital
- Pre-tax return on Net Operating Assets
- Pre-tax Return on Total Earning Assets
- Gross Profit/Total Assets
- Common Equity/Invested Capital
- Long-term Debt/Invested Capital
- Total Debt/Invested Capital
- Capitalization Ratio
- Interest/Average Long-term Debt
- Interest/Average Total Debt
- Cash Balance/Total Liabilities
- Inventory/Current Assets
- Receivables/Current Assets
- Total Debt/Total Assets
- Total Debt/EBITDA
- Short-Term Debt/Total Debt
- Current Liabilities/Total Liabilities
- Long-term Debt/Total Liabilities
- Profit Before Depreciation/Current Liabilities
- Operating CF/Current Liabilities
- Cash Flow/Total Debt
- Free Cash Flow/Operating Cash Flow
- After-tax Return on Total Stockholders Equity
- Total Liabilities/Total Tangible Assets
- Long-term Debt/Book Equity
- Total Debt/Total Assets
- Total Debt/Capital
- Total Debt/Equity
- After-tax Interest Coverage
- Interest Coverage Ratio
- Cash Ratio
- Quick Ratio (Acid Test)
- Current Ratio
- Inventory Turnover
- Asset Turnover
- Receivables Turnover
- Payables Turnover
- Sales/Invested Capital
- Sales/Stockholders Equity
- Research and Development/Sales
- Avertising Expenses/Sales
- Labor Expenses/Sales
- Accruals/Average Assets
- Price/Book

Table 5: Correlations between financial ratios at the beginning of 2021 and realized returns in 2021.

	Returns
Shillers Cyclically Adjusted P/E Ratio	0.07
Book/Market	0.29
Enterprise Value Multiple	0.05
Price/Operating Earnings (Basic, Excl. EI)	0.10
Price/Operating Earnings (Diluted, Excl. EI)	0.09
P/E (Diluted, Excl. EI)	0.09
P/E (Diluted, Incl. EI)	0.10
Price/Sales	-0.07
Price/Cash flow	0.11
Dividend Payout Ratio	0.02
Net Profit Margin	-0.12
Operating Profit Margin Before Depreciation	-0.11
Operating Profit Margin After Depreciation	-0.20
Gross Profit Margin	-0.14
Pre-tax Profit Margin	-0.13
Cash Flow Margin	-0.04
Return on Assets	0.02
Return on Equity	0.03
Return on Capital Employed	-0.03
Effective Tax Rate	0.06
After-tax Return on Average Common Equity	0.06
After-tax Return on Invested Capital	0.01
After-tax Return on Total Stockholders Equity	0.06
Pre-tax return on Net Operating Assets	-0.10
Pre-tax Return on Total Earning Assets	-0.08
Gross Profit/Total Assets	-0.02
Common Equity/Invested Capital	0.14
Long-term Debt/Invested Capital	-0.16
Total Debt/Invested Capital	-0.17
Capitalization Ratio	-0.15
Interest/Average Long-term Debt	0.01
Interest/Average Total Debt	0.03
Cash Balance/Total Liabilities	-0.05
Inventory/Current Assets	0.07
Receivables/Current Assets	0.10
Total Debt/Total Assets	-0.16
Total Debt/EBITDA	0.06
Short-Term Debt/Total Debt	-0.10
Current Liabilities/Total Liabilities	0.02
Long-term Debt/Total Liabilities	0.00
Profit Before Depreciation/Current Liabilities	-0.01
Operating CF/Current Liabilities	0.17
Cash Flow/Total Debt	0.05
Free Cash Flow/Operating Cash Flow	-0.02
Total Liabilities/Total Tangible Assets	-0.06
Long-term Debt/Book Equity	-0.12
Total Debt/Total Assets2	-0.18
Total Debt/Capital	-0.13
Total Debt/Equity	-0.08
After-tax Interest Coverage	-0.05
Interest Coverage Ratio	-0.05
Cash Ratio	-0.03
Quick Ratio (Acid Test)	0.02
Current Ratio	0.07
Inventory Turnover	0.09
Asset Turnover	0.09
Receivables Turnover	0.10
Payables Turnover	-0.03
Sales/Invested Capital	0.06
Sales/Stockholders Equity	-0.04
Research and Development/Sales	-0.02
Avertising Expenses/Sales	-0.16
Labor Expenses/Sales	-0.11
Accruals/Average Assets	0.10
Price/Book	-0.07
Returns	1.00

7.2 Methodology

7.2.1 GMMCovS algorithm

When we test a certain combination of the number of clusters G and groups of variables Q, we perform the following algorithm of Cavicchia et al. (2022) that optimally divides variables into groups and helps with clustering observations.

Algorithm 1: GMMEUCovS algorithm

```
Input: Number of clusters $G$ and number of groups of variables $Q$, $\epsilon$, max. number of
       iterations (MAX)
Step 0: values for $\hat{W}$ and $\hat{V}$ are selected for initialization. Use soft k-means with $k = G$
    for the initialization of the $\hat{\mathbf{W}}$ matrix. For the initialization of $\hat{\mathbf{V}}$, run the adapted
    UCM algorithm applied on S. The initial values for $\hat{\pi}$ are selected by maximizing 4.6
    with respect to $\pi_g$ which results in $\widehat{\pi_g} = \frac{n_g}{n}$ with $n_g$ being equal to the sum of all the
    $\widehat{w_{ig}}$ corresponding to $g = 1, \ldots, G$. Similarly, when maximizing 4.6 with respect to $\mu_g$,
    $\widehat{\Sigma_V}$, $\widehat{\Sigma_W}$ and $\widehat{\Sigma_B}$, it obtains their initial values. That is, $\widehat{\mu_g} = \frac{\sum_{i=1}^{n} \widehat{w_{ig}} x_i}{n_g}$,
    $\hat{\Sigma}_{V_g} = \left(\hat{V}_g' \hat{V}_g\right)^{-1} \hat{V}_g' \operatorname{diag}(S_g) \hat{V}_g$,
    $\hat{\Sigma}_{W_g} = \left[\left(\hat{V}_g' \hat{V}_g\right)^2 - \hat{V}_g' \hat{V}_g\right]^{-1} \operatorname{diag}\left[\hat{V}_g' \left(S_g - \operatorname{diag}\left(\hat{V}_g \hat{\Sigma}_{V_g} \hat{V}_g'\right)\right) \hat{V}_g\right]$ and
    $\hat{\Sigma}_{\mathrm{B}_g} = \hat{V}_g^{+} S_g \left(\hat{V}_g'\right)^{+}$ subject to the constraints corresponding to each parameter. For
    the estimation of $\Sigma_{B_g}$, it requires the other parameters in the log-likelihood (w.r.t $\Sigma_{B_g}$)
    to be fixed and the ultrametric condition to be satisfied which is solved by the adapted
    average linkage UPGMA algorithm for covariance matrices. Furthermore, $\hat{V}_g^{+}$ denotes
    the Moore-Penrose generalized inverse of $\hat{V}_g$.
Test $= 1,$ iteration $= 0$
while Test $\leq \epsilon$ and iteration $\leq$ MAX do
    iteration = iteration +1
    Step 1: with these initial values, the initial chosen composition of W is updated by
    $\hat{w}_{ig} = \frac{\hat{\pi}_g \phi\left(x_i | \hat{\mu}_g, \hat{\Sigma}_{V_g}, \hat{\Sigma}_{\mathrm{W}_g}, \hat{\Sigma}_{\mathrm{B}_g}, \hat{V}_g\right)}{\sum_{h=1}^{G} \hat{\pi}_h \phi\left(x_i | \hat{\mu}_h, \hat{\Sigma}_{V_h}, \hat{\Sigma}_{\mathrm{W}_h}, \hat{\Sigma}_{\mathrm{B}_h}, \hat{V}_h\right)}$, which is the updated probability of the i'th
    observation belonging to a cluster g.

    Step 2: the values for $\hat{\pi}$ are updated again using $\widehat{\pi_g} = \frac{n_g}{n}$ but now with the
    updated composition of $\hat{W}$.

    Step 3: the values for $\mu_g$ are updated again using $\widehat{\mu_g} = \frac{\sum_{i=1}^{n} \widehat{w_{ig}} x_i}{n_g}$,
    $\hat{\Sigma}_{V_g} = \left(\hat{V}_g' \hat{V}_g\right)^{-1} \hat{V}_g' \operatorname{diag}(S_g) \hat{V}_g$ also with the updated composition of $\hat{W}$.

    Step 4: compute the EUCovS $\hat{\Sigma u}$ by first updating $\widehat{\Sigma_V}$, $\widehat{\Sigma_W}$ and $\widehat{\Sigma_B}$ as in Step 1
    where $\widehat{\Sigma_B}$ is computed such that it fulfills the ultrametric condition as described in
    constraints 4.3 (by selecting the closest matrix (in Frobenius norm) that satisfies
    this condition). Then, estimate the binary and row-stochastic membership matrix
    $\hat{V}_g$ row-by-row by computing 4.6 with respect to $\Sigma_u$ given the estimated $\widehat{\Sigma_V}$, $\widehat{\Sigma_W}$
    and $\widehat{\Sigma_B}$ and decide whether a certain variable p should be in group Q by comparing
    which selection generates the highest log-likelihood.

    Step 5: Test $= \frac{\ell_{\mathrm{H}}\left(\hat{W}^{(t+1)}, \hat{\Psi}^{(t)}\right) - \ell_{\mathrm{H}}\left(\hat{W}^{(t)}, \hat{\Psi}^{(t-1)}\right)}{\left|\ell_{\mathrm{H}}\left(\hat{W}^{(t)}, \hat{\Psi}^{(t-1)}\right)\right|}$. Here, $\ell_{\mathrm{H}}\left(\hat{W}^{(t+1)}, \hat{\Psi}^{(t)}\right)$ denotes the
    log-likelihood to be estimated as in formula 4.6
```

7.2.2 Alternative procedure

Algorithm 2: Greedy Heuristic of making $\boldsymbol{\Sigma}_u$ positive definite and ultrametric

```
Input: $\boldsymbol{\Sigma}_u$, $\epsilon$, max. number of iterations ($MAX$)
Output: $\tilde{\boldsymbol{\Sigma}}_u$
Set $\tilde{\boldsymbol{\Sigma}}_u = \boldsymbol{\Sigma}_u$, $i = 1$
while $\tilde{\boldsymbol{\Sigma}}_u$ not positive definite and $i \leq MAX$ do
    $i$=$i+1$
    $\boldsymbol{B} = \tilde{\boldsymbol{\Sigma}}_u$
    Generate R s.t. $\boldsymbol{B} = \boldsymbol{RDR}^T$ where $\boldsymbol{R}$ is orthonormal and $\boldsymbol{D} \in \mathbb{R}^{P \times P}$ a diagonal
      matrix with eigenvalues of $\boldsymbol{B}$ on its diagonal
    if $\min(\boldsymbol{D}) < 0$ then
        for $k \leftarrow 1$ to $P$ do
            if $\boldsymbol{D}[k,k] < 0$ then
                $\boldsymbol{D}[k,k] = \epsilon$
    $\boldsymbol{C} = \boldsymbol{RDR}^T$
    Generate $\tilde{\boldsymbol{\Sigma}}_u$ using the the adapted average linkage UPGMA algorithm for $\boldsymbol{C}$
```

7.2.3 Expression for $\det D$ in the canonical representation of Σ_u

D is a block diagonal matrix composed as the direct sum of A, $\lambda_1 I_{n_{1-1}}, \lambda_2 I_{n_{2-1}}, \ldots, \lambda_Q I_{n_{Q-1}}$ (or formally, D is $A \oplus \lambda_1 I_{n_{1-1}} \oplus \lambda_2 I_{n_{2-1}} \oplus \ldots \oplus \lambda_Q I_{n_{Q-1}}$) we can compute the determinant of D as $\det D = \det A \times \det \lambda_1 I_{n_{1-1}} \times \det \lambda_2 I_{n_{2-1}} \times \ldots \times \det \lambda_Q I_{n_{Q-1}}$. Note that $\lambda_j I_{n_{j-1}}$ is λ_j with multiplicity n_{j-1} such that it's determinant is equal to $\lambda_j^{n_{j-1}}$. Therefore, for the computation of D we can equivalently write

$$\det D = \det A * \prod_{j=1}^{Q} \lambda_j^{n_{j-1}} \tag{7.1}$$

which is, due to the canonical representation, equivalent to the determinant of the block matrix Σ_u, as Archakov & Hansen (2020) showed for a general block matrix.

7.3 Results

Table 6: Davies-Bouldin Index values for 2021

	G							
Q	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	3.942	4.850	1.764	1.570	1.532	1.243	1.084	1.100
2	NaN	NaN	NaN	0.045	0.018	1.059	0.018	3.880
3	NaN	17.220	1.024	0.203	1.696	NaN	0.534	4.021
4	NaN	1.442	NaN	0.045	0.018	0.018	4.739	4.111
5	NaN	4.986	0.018	NaN	NaN	NaN	0.018	NaN
6	2.203	4.327	NaN	NaN	NaN	2.115	NaN	NaN
7	NaN	NaN	NaN	0.018	NaN	5.296	1.497	NaN
8	1.442	NaN	0.045	2.115	2.236	0.018	0.203	0.096
9	NaN	NaN	0.045	NaN	3.331	0.045	0.018	2.622
10	2.115	NaN	0.045	0.018	0.096	6.038	0.018	0.018
11	1.442	NaN	0.045	0.018	5.599	0.018	0.045	0.668
12	1.442	0.018	0.018	1.024	0.018	0.567	0.018	0.018
13	1.953	0.018	0.018	0.453	0.453	0.018	0.130	11.928
14	NaN	0.018	0.045	1.024	NaN	0.096	0.018	1.889
15	1.771	0.018	0.203	NaN	0.018	0.018	0.045	0.018
16	NaN	0.018	0.045	0.182	0.045	0.534	0.057	2.833
17	NaN	NaN	0.018	1.497	0.310	0.018	0.045	3.311
18	1.024	NaN	0.018	0.096	0.018	0.096	0.018	0.096
19	1.333	NaN	0.018	0.018	1.079	0.018	0.018	1.484
20	1.031	0.018	0.018	3.073	0.057	0.655	3.448	0.118
21	2.115	0.018	0.018	1.042	0.018	5.641	0.018	0.096
22	0.921	0.018	0.846	4.677	2.805	0.057	0.096	0.120
23	14.128	0.018	0.018	1.333	2.921	1.213	0.950	1.219
24	NaN	NaN	0.018	0.018	0.462	0.598	0.428	0.732

Table 7: Dunn-Index values for 2021

	G							
Q	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0.007	0.017	0.020	0.021	0.031	0.040	0.041	0.031
2	Inf	Inf	Inf	1.036	3.181	0.104	3.181	0.003
3	Inf	0.053	0.154	0.189	0.008	Inf	0.106	0.002
4	Inf	0.106	Inf	1.036	3.181	3.181	0.010	0.003
5	Inf	0.002	3.181	Inf	Inf	Inf	3.181	Inf
6	0.043	0.002	Inf	Inf	Inf	0.043	Inf	Inf
7	Inf	Inf	Inf	3.181	Inf	0.012	0.013	Inf
8	0.106	Inf	1.036	0.043	0.013	3.181	0.189	0.366
9	Inf	Inf	1.036	Inf	0.017	1.036	3.181	0.027
10	0.043	Inf	1.036	3.181	0.366	0.004	3.181	3.181
11	0.106	Inf	1.036	3.181	0.008	3.181	1.036	0.032
12	0.106	3.181	3.181	0.154	3.181	0.013	3.181	3.181
13	0.013	3.181	3.181	0.107	0.107	3.181	0.045	0.008
14	Inf	3.181	1.036	0.154	Inf	0.366	3.181	0.027
15	0.005	3.181	0.189	Inf	3.181	3.181	1.036	3.181
16	Inf	3.181	1.036	0.183	1.036	0.106	0.774	0.005
17	Inf	Inf	3.181	0.013	0.155	3.181	1.036	0.017
18	0.154	Inf	3.181	0.366	3.181	0.366	3.181	0.366
19	0.055	Inf	3.181	3.181	0.055	3.181	3.181	0.008
20	0.107	3.181	3.181	0.149	0.774	0.129	0.015	0.339
21	0.043	3.181	3.181	0.235	3.181	0.003	3.181	0.366
22	0.059	3.181	0.189	0.004	0.024	0.774	0.366	0.045
23	0.002	3.181	3.181	0.055	0.005	1.223	1.338	0.811
24	Inf	Inf	3.181	3.181	2.234	2.107	2.216	1.577