Temporal clustering of commodity market trading days: Clarifying volatility regimes with the Gaussian mixture model with extended ultrametric covariance structure

Erasmus University Rotterdam

Erasmus School of Economics
Bachelor Thesis: BSc Econometrics
H.J.F. van den Berg
502407
Supervisor: dr. C. Cavicchia
Second assessor: N.W. Koning

July 3, 2022

Abstract

The Gaussian Mixture Model with Extended Ultrametric Covariance Structure (GMMEUCovS, Cavicchia et al., 2022) can detect clusters of trading days in conditional volatility data of commodities. The introduced parsimonious parameterization of GMM is used to investigate the temporal clustering of financial time series data for the first time in literature. Daily price data from September 18, 2000, through July 31, 2020, is used to estimate conditional volatility of the log-returns for 23 commodities. The GMMEUCovS shows that it is able to detect unstable market periods, such as the crisis of 2008 and the covid-19 pandemic. It identifies broader concepts, which result in larger clusters relative to k-means clustering (Chen et al., 2021). Besides, the method pinpoints hierarchical covariance structures within the determined clusters. This paper is insightful for financial agents interested in commodity market behaviour.

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

The commodity markets represent a large amount of the total trade in goods. Gold alone already has a market cap of 11.6 trillion, according to CompaniesMarketCap.com (2022). Commodities are being used as hedging instruments for longer periods (Rehman et al., 2019). Furthermore, commodities are essential for development and growth. In the last years, financialization increased wildly in commodity markets (Tang and Xiong, 2012), which resulted in more comovement in the whole asset market (Ohashi and Okimoto, 2016). Comovement weakens the hedging benefits of commodities for other assets, for example, stocks or currencies. This shows the importance of investigating the behaviour of the commodity markets. In terms of hedging, the risk of commodities is an important part to investigate. Risk clustering, in terms of volatility, can be of great added value for financial agents. Clustering methods can determine clusters of periods with common market regimes.

Clustering is a form of unsupervised machine learning. There are several clustering techniques proposed in the literature. K-means clustering (MacQueen, 1967) is one of the most simple and popular clustering techniques. However, Gaussian Mixture Model (GMM) clustering is a substantial more theoretical and mathematical elegant technique.

Gaussian mixture models (Duda et al., 1973) are well-acknowledged models for modelling heterogeneous populations. The models assume that a heterogeneous population consists of a finite set of G homogeneous subpopulations that follow the Gaussian distribution. The GMM density assumes the following form

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

where each group of the mixture has the multivariate Gaussian density expressed by $\phi\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$ with a mean vector $\boldsymbol{\mu}_g$ consisting of p dimensions and a covariance matrix $\boldsymbol{\Sigma}_g$. The quantities $\pi_1, ...\pi_G$ are the mixing proportions (prior probabilities) such that $\pi_g >= 0$ and $\sum_{g=1}^{G} \pi_g = 1$ and $\boldsymbol{\Psi} = \{\pi_1, ..., \pi_G, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_G\}$ is the overall parameter vector, as stated by Cavicchia et al. (2022).

Even though GMM is a mathematically elegant method for solving clustering problems, it has one considerable disadvantage relative to other clustering techniques, such as hierarchical and k-means clustering. The vast number of parameters that need to be estimated results in computationally demanding problems when using sizable or high-dimensional datasets.

The number of parameters to be calculated can be reduced to tackle the computational demanding problem. The number of parameters that need to be determined can be split into three parts. The mixing proportions $G - 1$, the cluster means Gp and the covariance matrix of the clusters $Gp(p + 1)/2$, with a total amount of $G - 1 + Gp + Gp(p + 1)/2$ parameters. The calculation of the covariance matrix, existing of $Gp(p+1)/2$ parameters, is the most computational demanding part. To reduce this number of parameters, several parsimonious parameterizations of the covariance matrix are introduced in the literature.

Cavicchia et al. (2022) introduced a new GMM method with parameterization of the covariance matrix by assuming an *extended ultrametric covariance matrix* (GMMEUCovS) for each cluster. The advantage of this method relative to the factor analyzers (FA) methods is that the introduced method is more convenient when latent concepts are hierarchical linked. Cavicchia et al. (2022) use the same model as the constructed simultaneous model and its LS estimation of Cavicchia et al. (2020) to reconstruct a generic covariance matrix via an extended ultrametric covariance matrix associated with a hierarchy of concepts. The method proposed by Cavicchia et al. (2022) is an extended version of the method of Cavicchia et al. (2020), which only introduced the *ultrametric correlation matrix* that is usable for a non-negative correlation matrix.

This parsimonious parameterization introduced by Cavicchia et al. (2022) has not been used for time series data in earlier literature. The first valuable element of this study is the expanded usability of GMMEUCovS, namely the application of time-series data. The differences between the interpretation of time series data and the usual data used for this method, are the observations and the variables. Cavicchia et al. (2022) used different aspects of one country or one coffee bean as variables at one point in time, where the country and the coffee bean are observed in the same time period. When investigating time series, the different aspects, conditional volatility of the 23 commodities of the observed time periods will be used as variables and the time periods themselves as observations. The observed time periods will be clustered together based on the variables in that specific time period, the forecasted volatilities of the 23 commodities, and represent the commodity market regime.

Besides the fact that the usability of this method is extended to financial time series data, this clustering method is able to create valuable insights into different market regimes. The model clusters the time observations in different periods of common market regimes based on volatility. In addition to the clustering of periods, the method reveals the hierarchical covariance structures within the clustered market regimes.

For this commodity market study, daily prices of 23 commodities from September 18, 2000, to July 31, 2020, are used (5184 observations). The daily prices are converted into log-returns (5183 observations). Because volatility of assets can not be observed directly, a specific model is used to estimate this volatility. Conditional volatility forecasts for all 23 different commodities are estimated with the GJR-GARCH(1,1,1) model from the log-returns. The data is equivalent to the dataset used by Chen et al. (2021) and is extracted from Thomson Reuters DataStream.

The focus of this paper is investigating the meaningfulness of the application of the method proposed by Cavicchia et al. (2022). What is the interpretation of the obtained clusters? Do the formed clusters have a meaningful interpretation? How are the parsimonious GMM results related to the k-means clustering results from literature (Chen et al., 2021). Furthermore, what are the differences between this method and the k-means application? Finally, how do we interpret the obtained groups and hierarchical structures of the variables?

The GMMEUCovS method successfully identifies different market regimes in terms of low and high volatility. Compared to the k-means method used by Chen et al. (2021), it captures the broader concepts more clear. Within the formed clusters, it can pinpoint the covariance hierarchy structures. High volatile times have comparable structures concerning the amount of discordant and concordant groups. Some commodities are always in the same variable group regardless of the cluster.

The paper is structured as follows. Section 2 summarizes the insights into the commodity markets and unsupervised machine learning applications obtained from the literature. Section 3 introduces the used data. Next, Section 4 covers the used methods. Section 5 describes the obtained results. Finally, section 6 concludes and discusses the findings.

2 Literature review

This section provides a brief summary of findings from published literature. It explains the studied behaviour of commodity markets, and it gives a short overview of unsupervised machine learning applications on finance data.

2.1 Commodity markets

Commodities are essential for development and growth worldwide. Besides the simple use, commodities are being traded in commodity markets. Commodities are used as hedging instruments for longer periods (Rehman et al., 2019). From 2000 the financialization of the commodity markets increased wildly, which resulted in more comovement in the asset markets. This financialization caused changing reactions to market shocks relative to earlier stages of the market (Ohashi and Okimoto, 2016).

It is important to understand the changes in the commodity markets for investors that often use hedging techniques with commodities involved. However, following Fakhfekh and Hachicha (2021), gold is still one of the best hedging instruments. The correlation of gold with the stock markets is negative and diminishes when the stock prices decline. For risk purposes, this is important evidence for the safe-haven factor of gold (Creti et al., 2013).

The effects of oil price shocks on agricultural commodities (corn, soybeans, wheat, sugar and coffee) differ from the effects on metals (gold, silver and copper) (Ahmadi et al., 2016). Besides the difference in effect on the commodity groups, they show differences in effect among periods.

For example, Adams and Glück (2015) show that crude oil and copper show significant responses to changes in stock market risk, while aluminum and wheat do not show significant responses. To continue, they show the temporal differences in volatility spillover in the commodity markets. Before the bankruptcy of Lehman-Brothers in September 2008, there were, for both low and high volatility regimes, almost no significant risk spillovers. After this event, the start of the financial crisis of 2008, the volatility spillovers increased and became significant for both low and high volatility regimes.

Next, the connectedness of the commodities' directional volatilities increased substantial during the global crisis due to the spread of the coronavirus (Umar et al., 2021). To be more specific, in the directional volatilities connectedness, tin, gold, nickel, lead, and aluminum markets function as net transmitters during specific returns and volatility periods. Contrarily, several industrial and precious metal markets, such as copper, zinc and platinum, function as a net receivers.

The above-described aspects of behaviour of commodity markets are broad and mixed. The aspects imply that there exist different periods of commodity market regimes through time, and within these regimes, there can be hierarchical covariance structures discovered in the variables that are valuable to investigate.

2.2 Unsupervised machine learning for finance

Several applications of unsupervised machine learning on economic or financial topics are studied. The subjects vary from natural language processing to unsupervised machine learning on quantitative financial data.

Different economic policy uncertainty indexes are introduced using unsupervised machine learning on news articles called natural language processing (Azqueta-Gavaldón, 2017, Azqueta-Gavaldon et al., 2020). In other studies, socio-economic factors are used in combination with unsupervised machine learning to analyze the influence of the factors on the choice of the water source (Tiyasha et al., 2021)

In addition to the economic applications, unsupervised machine learning is a technique that can also analyze patterns in financial data. Shokry et al. (2020) have already proved that these techniques can recognize hidden money laundering networks. Moreover, the application for tax fraud detection is shown by De Roux et al. (2018). Furthermore, time series clustering is included in the research of Chen et al. (2021) with the k-means clustering method.

However, in addition to Chen et al. (2021), scarcely anything is published about unsupervised machine learning techniques applied to financial time series data. They are one of the pioneers on this subject. Their analysis of k-means clustering directly inspired this work. The unexplored subject of financial time series data is expanded by this research by applying GMMEUCovS (Cavicchia et al., 2022) to a financial time series. Besides the capability of clustering, the method is also capable of detecting hierarchical structures within the clusters, another topic that is scarcely investigated in the literature.

3 Data

This research uses the dataset that Chen et al. (2021) used in order to compare the obtained results from GMMEUCovS with the already published results of the k-means clustering method. The dataset contains a wide variety of commodities from the group of precious metals, base metals, energy and agricultural-based commodities. The data contains daily prices from September 18, 2000, to July 31, 2020, for the 23 commodities shown in table 1. This forms 5185 observations of daily returns for each of the 23 commodities. There are no missing values. However, one interesting negative value on April 20, 2020, for WTI is -37.63. For all commodities, the observations from this date are deleted, because the negative value causes undefined outcomes when calculating the log-return. This results in a dataset of 23 commodities with 5184 observations. For this research, the daily prices are transformed into log-returns, which results in a series of 5183 volatilities on trading days from September 19, 2000, until July 31, 2020, for the 23 commodities of interest. This data is obtained by Chen et al. (2021) from the Thomson Reuters DataStream.

Table 1: In dataset included commodities sorted by commodity group

group name	name
precious metals	gold, silver, platinum, palladium
base metals	copper, zinc, tin, lead, nickel, aluminum
energy	brent, West Texas Intermediate crude (WTI), Gasoil, Gasoline
agricultural	soya oil, palm oil, wheat, corn, soybeans, coffee, cocoa, cotton, lumber

4 Methodology

This section elaborates on the methods used and the outcomes' evaluations.

This research investigates the temporal relationships in the financial time series of commodities with unsupervised machine learning techniques. The parsimonious GMM method GMMEUCovS (Cavicchia et al., 2022) is used to create insights into the financial time series with a less computational demanding process relative to the classic GMM method. Besides, this specific GMM method can reveal hierarchical covariance structures within the clusters, which can be insightful.

Unsupervised machine learning techniques are highly suitable in case of handling quantitative numeric data. Hence, applying these techniques to numeric financial data should be appropriate. The two methods of interest in this study, namely GMM and k-means clustering, belong both to the class of unsupervised machine learning techniques.

First, the data is converted into estimated conditional volatility with a variant of the GARCH model. The trading dates are clustered based on the volatility forecasts with the GMMEUCovS model. The clusters are evaluated when the best fit model is found.

4.1 K-means clustering

K-means clustering is a simple and popular clustering method in which observations are allocated to the nearest mean (cluster center). It is a popular method because of the simplicity of its use. However, this technique has its disadvantages. For example, it is highly sensitive to outliers, the outcome depends heavily on the starting conditions(Likas et al., 2003), and different size and density clusters are not handled (Arora et al., 2016). Hence, we are investigating more mathematically elegant clustering methods like GMM clustering.

4.2 Gaussian Mixture Model clustering

The application of Gaussian mixture model clustering is often used for a wide variety of topics. For example, the GMM clustering methods are used for bike-sharing clustering and classification by Jia et al. (2019), and customer segmentation by Jang et al. (2021), to image clustering and classification by Permuter et al. (2006).

Since the GMM estimations are computational demanding, several parsimonious regressions in literature are proposed. The so-called Gaussian Parsimonious Clustering Models (GPCMs) impose geometric features on the cluster covariance structure or constrain covariance components across clusters to be equal or unequal (Celeux and Govaert, 1995). There is a substantial amount of Parsimonious GMMs (PGMMs) proposed in the literature. Several are based on eigen-decompositions of the covariance matrix. Besides, there are PGMMs that use Factor Analysis (FA) (McLachlan et al., 2003) or probabilistic principal component analyzers (Tipping and Bishop, 1999)to create PGMMs. For high-dimensional datasets, Bouveyron et al. (2007) proposed a model for High Dimensional Data Clustering (HDDC), where GMM is based on eigen-decomposition. This model specifically reduces the number of distinct eigenvalues, which results in less computational demanding GMM solutions.

In addition to these PGMMs, Cavicchia et al. (2022) introduced a new method that also detects the relationships among variables called the Gaussian Mixture Model with Extended Ultrametric Covariance Structure (GMMEUMCovS).

4.3 GMMEUMCovS

This GMMEUMCovS is a parameterization of the covariance matrix by using an extended ultrametric covariance matrix (EUCM) for each cluster. This EUCM is an extension of the UCM that was proposed by Cavicchia et al. (2020) in order to include generic covariance structures next to non-negative structures. This method detects hierarchical covariance structures in the clusters. Because of the grouped variables, less parameters must be estimated, which leads to less computational demanding processes that also show hierarchical covariance structures of the variables. How the algorithm is established is briefly explained in the following four subsections.

4.3.1 Notation

With the aim of comfortable reading, the used notation in this paper is adopted from Cavicchia et al. (2022) and defined as follows:

n, p, G, Q Number of observations, variables, clusters, and groups of variables, respectively.

$\Sigma = \left[\sigma_{jl}\right]$ Covariance matrix of order p.

$\mathbf{V} = \left[\sigma_{jl}\right]$ $(p \times Q)$ membership matrix, where $v_{jq} = 1$ if the j th variable belongs to the q th group; $v_{jq} = 0$ otherwise. It is binary and row-stochastic, i.e., with one non-zero element per row, identifying a partition of variables in Q groups.

$\mathbf{\Sigma}_{\mathrm{V}} = \left[{}_{V}\sigma_{qq}\right]$ Diagonal matrix of order Q with diagonal entries representing variances of the groups of variables.

$\mathbf{\Sigma}_{\mathrm{W}} = \left[{}_{W}\sigma_{qq}\right]$ Diagonal matrix of order Q with diagonal entries representing covariances within groups of variables.

$\mathbf{\Sigma}_{\mathrm{B}} = \left[{}_{B}\sigma_{qh}\right]$ Matrix of order Q with off-diagonal entries representing covariances between groups of variables, and diagonal ones equal to zero.

Cavicchia et al. (2022) extended Cavicchia et al. (2020) from the use of a non-negative ultrametric correlation matrix to a generic covariance matrix. The non-negativity constraint is relaxed in this case. Let us summon the properties of a covariance matrix:

(i) symmetry: $\Sigma = \Sigma'$

(ii) non-negativity of the diagonal: $\sigma_{jj} \geq 0$ for all $j = 1, ..., p$

(iii) positive semi-definiteness: $\boldsymbol{x}'\mathbf{\Sigma}\boldsymbol{x} \geq \boldsymbol{x} \in \mathbb{R}^{p}$

A matrix that only satisfies the following properties does not directly meet the conditions of an (extended) ultrametric covariance matrix. Cavicchia et al. (2022) added the following restrictions to satisfy the Extended Ultrametric Covariance Matrix (EUCM) restrictions:

(iv) ultrametric inequality: $\sigma_{jl} \geq \min\{\sigma_{jh}, \sigma_{lh}\}$,for all $j, l, h = 1, ..., p$

(v) diagonal dominance: $\sigma_{jj} \geq \sum_{\substack{l=1 \\ l \neq j}}^{p} |\sigma_{jl}|$ for $j = 1, \ldots, p$

When properties (i), (ii), (iv), and (v) hold, the matrix satisfies all the conditions to be named a Weak Extended Ultrametric Covariance Matrix. When properties (ii) and (v) are strictly satisfied, the matrix meets the Strict Extended Ultrametric Covariance Matrix conditions.

4.3.2 Extended ultrametric covariance structure

The parameterization of the extended ultrametric covariance matrix of Cavicchia et al. (2022) is defined as follows:

$$\mathbf{\Sigma}_{\mathrm{u}} = \mathbf{V}\left(\mathbf{\Sigma}_{\mathrm{W}} + \mathbf{\Sigma}_{\mathrm{B}}\right)\mathbf{V}' - \operatorname{diag}\left(\mathbf{V}\mathbf{\Sigma}_{\mathrm{W}}\mathbf{V}'\right) + \operatorname{diag}\left(\mathbf{V}\mathbf{\Sigma}_{\mathrm{V}}\mathbf{V}'\right) \tag{2}$$

subject to constraints

$$\mathbf{V} = \left[v_{jq} \in \{0, 1\} : j = 1, \ldots, p, q = 1, \ldots, Q\right]; \tag{3}$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \text{ i.e. } \sum_{q=1}^{Q} v_{jq} = 1 \quad j = 1, \ldots, p; \tag{4}$$

$$\mathbf{\Sigma}_{\mathrm{B}} = \mathbf{\Sigma}_{\mathrm{B}}', \operatorname{diag}\left(\mathbf{\Sigma}_{\mathrm{B}}\right) = \mathbf{0}, {}_{B}\sigma_{qh} \geq \min\left\{{}_{B}\sigma_{qs}, {}_{B}\sigma_{hs}\right\} q, h, s = 1, \ldots, Q, s \neq h \neq q; \tag{5}$$

$$\min\left\{{}_{W}\sigma_{qq} : q = 1, \ldots, Q\right\} \geq \max\left\{{}_{B}\sigma_{qh} : q, h = 1, \ldots, Q, h \neq q\right\}; \tag{6}$$

$$V\sigma_{qq} > |_{W}\,\sigma_{qq}\left|\left(\sum_{l=1}^{p} v_{lq} - 1\right) + \sum_{h=1}^{Q}\right|_{B} \sigma_{qh} \mid \sum_{l=1}^{p} v_{lh} q = 1, \ldots, Q, \tag{7}$$

$$\mathbf{\Sigma}_{\mathrm{u}} = \mathbf{\Sigma}_{\mathrm{u}} + a\mathbf{I}_p, \text{with } a > 0, \text{and such that } \mathbf{\Sigma}_{\mathrm{u}} \text{ is positive definite,} \tag{8}$$

where diag($\boldsymbol{X}$) is a matrix with only the diagonal entries of matrix $\boldsymbol{X}$ and zeros for the off-diagonal entries. $\mathbf{1}_k$ is a unitary vector of order k. $\boldsymbol{I}_k$ is an identity matrix of order k. Besides, the factor a is equal to the absolute value of the smallest eigenvalue of $\mathbf{\Sigma}_u$ plus an arbitrarily small constant to prevent computational singularity. When all the components of equation 2 satisfy constraints (3)-(8), $\mathbf{\Sigma}_u$ belongs to the Extended Ultrametric Covariance Matrices family.

4.3.3 GMMEUCovS algorithm

Cavicchia et al. (2022) found an approach to connect the Extended Ultrametric Covariance Structure with the Gaussian Mixture Model. They created an algorithm that estimates the GMM method and the

EUCovS simultaneous. This algorithm clusters the observations with a parsimonious parameterization while detecting the hierarchical covariance structures of the derived clusters.

Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ be a random sample of p-dimensional random vectors, in our case the 23 volatility estimates of the commodities for the 5183 observations, which are drawn from a population of G subpopulations. Each density of $\boldsymbol{x}_i$, conditional on the membership of the population, is drawn from a multivariate Gaussian with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_{u_g}$.

The algorithm to estimate the GMMEUCovS introduced by Cavicchia et al. (2022) consists of initialization and four repeating steps.

Initial step: Initialize the values for $\widehat{\boldsymbol{W}} = [\hat{w}_{ig}]$ and $\widehat{\boldsymbol{V}} = [\widehat{\boldsymbol{V}}_g]$. With these initial matrices and the observations, the values for $\widehat{\boldsymbol{\pi}} = \left[\widehat{\pi}_g\right], \widehat{\boldsymbol{\mu}} = \left[\widehat{\boldsymbol{\mu}}_g\right], \widehat{\boldsymbol{\Sigma}}_{\mathrm{V}} = \left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{V}_g}\right], \widehat{\boldsymbol{\Sigma}}_{\mathrm{W}} = \left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{W}_g}\right], \widehat{\boldsymbol{\Sigma}}_{\mathrm{B}} = \left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{B}_g}\right]$ are determined. They are computed in the order of the text above with the formulas that are included in Cavicchia et al. (2022). The constraints are applied in the following order: constraints 5, 6, 7. When the $\boldsymbol{\Sigma}_{u_g}$ is computed, restriction 2 is used. Finally, $\widehat{\boldsymbol{W}} = [\hat{w}_{ig}]$ is computed according to the formulas of Cavicchia et al. (2022). $\widehat{\boldsymbol{W}} = [\hat{w}_{ig}]$ and $\widehat{\boldsymbol{V}} = [\widehat{\boldsymbol{V}}_g]$ can be initialized randomly. However, to decrease computing time, a k-means clustering approach for initializing $\widehat{\boldsymbol{W}}$ is preferred. For the initialization of $\widehat{\boldsymbol{V}}$, an Ultrametric Covariance Matrix algorithm for the weighted covariance matrix $\boldsymbol{S}_g$ to find the partition of variables in $\widehat{\boldsymbol{V}}_g$ for $g = 1, ..., G$ can improve the computation time of the whole algorithm.

After the initialization, there are four steps that need to be repeated.

Step 1: for every $g = 1, ..., G$, $\widehat{\pi}_g$ is updated given $\hat{w}_{ig}$, $i = 1, ..., n$.

Step 2: for every $g = 1, ..., G$, $\widehat{\boldsymbol{\mu}}_g$ is updated given $\hat{w}_{ig}$, $i = 1, ..., n$.

Step 3: $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{V}_g}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{W}_g}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{B}_g}$ and constraints 5, 6, 7 are executed in the same order as the initialization. $\widehat{\boldsymbol{\Sigma}}^{(t)}_{u_g}$ is computed with the rest of the Sigma's with respect to constraint 8. Furthermore, $\widehat{\boldsymbol{V}}^{(t)}_g$ is determined with $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{V}_h}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{W}_h}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{B}_h}$ for h < g, and $\widehat{\boldsymbol{\Sigma}}^{(t-1)}_{\mathrm{V}_g}, \widehat{\boldsymbol{\Sigma}}^{(t-1)}_{\mathrm{W}_g}, \widehat{\boldsymbol{\Sigma}}^{(t-1)}_{\mathrm{B}_g}$, for $h > g, g, h = 1, ..., G, h \neq g$;

Step 4: for every $g = 1, \ldots, G$ and $i = 1, ..., n$ $\hat{w}_{ig}$ is updated given $\left\{\hat{\pi}^{(t)}_g, \hat{\boldsymbol{\mu}}^{(t)}_g, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{V}_g}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{W}_g}, \widehat{\boldsymbol{\Sigma}}^{(t)}_{\mathrm{B}_g}, \mathbf{V}^{(t)}_g\right\}$

Stopping rule: Compute $\ell_{\mathrm{H}}\left(\widehat{\boldsymbol{W}}^{(t+1)}, \widehat{\boldsymbol{\Psi}}^{(t)}\right)$ and repeat **Step 1-4** if

$$\frac{\ell_{\mathrm{H}}\left(\widehat{\boldsymbol{W}}^{(t+1)}, \widehat{\boldsymbol{\Psi}}^{(t)}\right) - \ell_{\mathrm{H}}\left(\widehat{\boldsymbol{W}}^{(t)}, \widehat{\boldsymbol{\Psi}}^{(t-1)}\right)}{\left|\ell_{\mathrm{H}}\left(\widehat{\boldsymbol{W}}^{(t)}, \widehat{\boldsymbol{\Psi}}^{(t-1)}\right)\right|} > \epsilon, \tag{9}$$

where ϵ is an arbitrary small positive constant, and t < T, the maximum number of iterations.

All the necessary formulas and exact initializations are elaborated by Cavicchia et al. (2022).

4.3.4 Model selection

For the choice of G and Q, there are several selection criteria available such as the Bayesian Information Criterion (BIC, Schwarz, 1978), the Akaike information criterion (AIC, Akaike, 1973), and the Integrated

Completed Likelihood (ICL, Biernacki et al., 2000). The BIC and ICL perform comparable for composing the number of mixture components and groups of variables (Scrucca et al., 2016). Steele and Raftery (2010) show that the BIC criterion outperforms other criteria like AIC for Gaussian Mixture Models. Because of these findings and the fact that Cavicchia et al. (2022) use the BIC, the preferred model is the model that maximizes the BIC value for specified G and Q values. For comparablility with Chen et al. (2021), G is forced to the same value as k for k-means, namely $G = k = 9$. The BIC, for a model with parameter vector $\boldsymbol{\theta}$, is given by

$$\text{BIC } = 2\ell(\hat{\boldsymbol{\theta}}) - v \log n, \tag{10}$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and v is the number of free parameters in the Gaussian Mixture Model.

4.4 GJR-GARCH modelling

Volatilities cannot be observed directly, resulting in an estimation of the conditional volatilities. When investigating the volatility of the 23 commodities, different GARCH models can be used. Nugroho et al. (2019) show that the GJR-GARCH (equivalent to TGARCH) model outperforms alternative autoregression models such as GARCH, log-GARCH and GARCH-M on empirical data. Hence, the GJR-GARCH(1,1,1) process (GLOSTEN et al. (1993)) using the Student's t distribution is chosen to improve the comparability with Chen et al. (2021). The model for each commodity is defined as

$$\varepsilon_t = R_t \tag{11}$$

$$\sigma_t^2 = \omega + \lambda_1 \varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-1}^2 I_{[\varepsilon_{t-1}<0]} + \beta_1 \sigma_{t-1}^2, \tag{12}$$

where R_t represents the log-return, ω represents the long-run average of the volatility per commodity, λ_1 represents the effect of a positive shock, $\lambda_1 + \gamma_1$ represent the shock of a negative return, and β_1 represents the autoregressive component of the volatility.

4.5 Performance evaluation

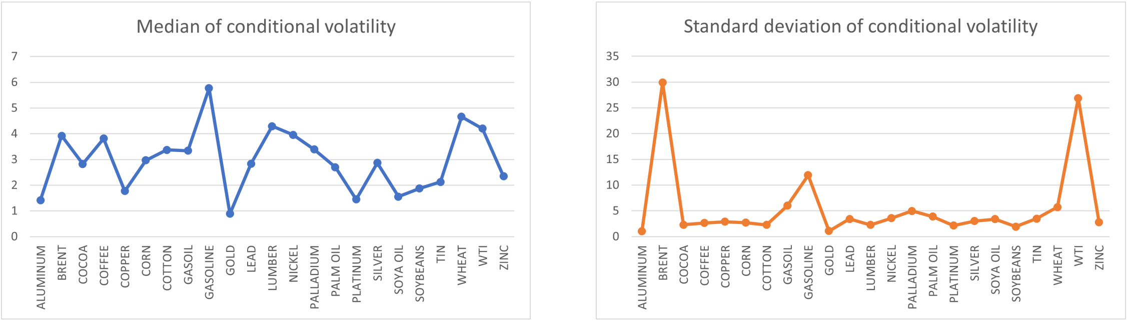
The individual clusters are evaluated, in the essence that periods of high volatility, like the covid-19 period and the financial crisis in 2008, will be highly distinguished from the other periods. In addition to this evaluation, the clusters are compared to the clustering results of Chen et al. (2021) in terms of comparable time clusters and identifying abnormal periods.

5 Results

This section reflects the results obtained with the GMMEUCovS method applied on conditional volatilities of commodities. A short preliminary analysis is done. After that, the used GJR-GARCH(1,1,1) models are discussed. Finally, the results of temporal clustering of trading days are explained and compared with Chen et al. (2021).

5.1 Preliminary analysis

Figure 1 visualizes the median and the standard deviation of the conditional volatility of the commodities. Figure 1a shows a low median for the metals aluminum, gold, copper, and platinum and agricultural-based commodities such as soya oil and soybeans. Brent, gasoline, lumber, nickel, wheat, and WTI show a high median. Regarding the standard deviation of the conditional volatility, energy-based commodities such as brent, gasoil, gasoline, and WTI have relatively higher standard deviations as shown in figure 1b. Precious metals, such as gold and platinum, show low standard deviation. Besides, cocoa, coffee, corn, cotton, lumber, soybeans and zinc show low standard deviation. The commodities with low standard deviation have a relatively stable conditional volatility over time.



(a) Median

(b) Standard deviation

Figure 1: Median and standard deviation or the conditional volatility forecasts

5.2 GJR-GARCH(1,1,1) results

All the GJR-GARCH(1,1,1) models are fitted and used to estimate the conditional volatility. Nickel has the highest amount of kurtosis of $v = 6.824$. However, this is a workable amount. The estimation output of the GJR-GARCH(1,1,1) model on the log-returns of nickel is displayed in table 2.

Table 2: GJR-GARCH(1,1,1) results of nickel

Variable	Coefficient	Standard error	Z-statistic	P-value
Omega	0.048	0.013	3.715	0.000
Lambda	0.043	0.007	6.275	0.000
Gamma	-0.002	0.008	-0.228	0.820
Beta	0.948	0.007	145.365	0.000
DOF-t-dist	6.824	0.648	10.525	0.000
Log-likelihood	-11086.15		BIC	4.286140

5.3 GMMEUMCovS results on conditional volatilities

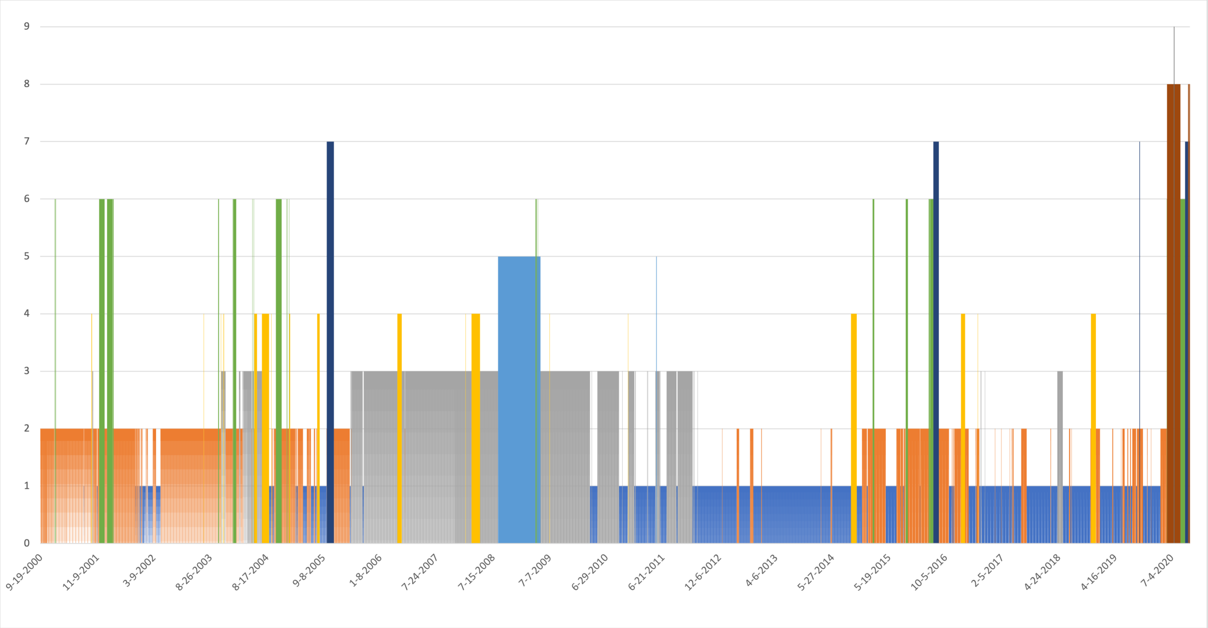
$G = 9$ and $Q = 7$ give the optimal combination concerning the BIC value of the number of clusters $G = 9$ (due to comparability with Chen et al. (2021)) and the options for the number of variable groups $Q = 1, ..., 23$. The results for the GMMEUCovS algorithm with $G = 9$ and $Q = 7$ applied to the commodity time-series data are shown in the next subsections.

5.3.1 Temporal clustering of trading days

The algorithm determined 9 clusters of trading days. Trading days with similar combinations of volatilities of the 23 commodities are grouped. The clustered periods are visualized in figure 2. The size of the clusters is shown in 3 through the number of observations per cluster.

Cluster 2 dominates in the first five years of the 21st century. At the beginning of 2006, cluster 3 arises. This cluster shows, on average more volatility than the years before, which matches with the upcoming uncertainty in that period. In September 2008, the month when Lehman Brothers (Britannica, 2022) was declared bankrupt, cluster 5 arose. The overall conditional volatility of this cluster is the largest among all the derived clusters. After the financial crisis of 2008, the markets were still unstable as shortly before the crisis, and these observations additionally fall in cluster 3. When the effects of the crisis begin to die out, the market regime is directed to the less volatile one from the beginning of the first decade of the 21st century. This results in cluster 1, comparable with cluster one except for a few instances. From 2015 through the beginning of 2020, clusters 1 and 2 are starting to alternate, which looks like the market is coming closer to its original position at the beginning of the 2000s. After this relatively stable period, cluster 8 is distinguished. Cluster 8 starts on March 20th, 2020, one day before the World Health Organization declares the sars-covid-19 pandemic. This cluster shows a substantial increase in volatility for approximately half of the commodities.

Figure 2: Clusters of trading days based on conditional volatilities



Clusters 1 and 2 are the most common clusters through the first two decades of the 21st century. The overall average conditional volatility is comparable between the clusters. However, the volatilities are distributed to different commodities. Precious metals, silver, platinum, and palladium, base metals, tin, lead, and nickel, energy-based commodities, WTI, gasoil, gasoline, agricultural-based commodities, soya oil, corn, and cotton, are less volatile after the crisis (cluster 1) than before (cluster 2). The following commodities have comparable volatility: gold, copper, aluminum, brent, palm oil, soybeans, coffee, and cocoa. Zinc, wheat, and lumber have a larger volatility in cluster 1.

The financial crisis of 2008 is captured by cluster 5. This cluster shows larger conditional volatilities for all the commodities except for palladium relative to clusters in non-crisis periods. Palladium has almost the lowest conditional variance of all possible clusters in cluster 5. When viewed from the cluster perspective, as shown in figure 7b, gold and palladium have the lowest conditional volatility. This could be proof of the hedging properties of precious metals gold and palladium. Furthermore, cotton, nickel, brent, and cocoa are the most volatile commodities of this period.

Cluster 8 covers the beginning of the covid-19 pandemic. This cluster shows higher overall conditional volatility relative to other clusters. Cotton, gasoline, lead, nickel, and silver have substantial volatility. Besides, gold, platinum, aluminum, lumber, and soybeans are as low as in market regimes without crises. However, the volatilities during this crisis are different from the volatilities of the 2008 crisis.

Two crises are distinguished in the observation period by different clusters. Hence, it implies that there are differences among the crises. The differences are visualized in figure 7(a). Cluster 8 (covid-19) shows relative lower volatilities for aluminum, brent, cocoa, coffee, copper, corn, gasoil, gold, lumber,

Table 3: Total number of observations per cluster

cluster	number of observations	prior probability
1	1833	0.349
2	1453	0.282
3	1207	0.234
4	194	0.039
5	186	0.036
6	170	0.033
7	73	0.015
8	64	0.012
9	3	0.001
Total	5183	1.000

platinum, soybeans, tin, wheat, and zinc to cluster 5 (2008 crisis). The other commodities show lower volatilities in cluster 8 relative to cluster 5.

5.3.2 Comparison with k-means

The applied GMMEUCovS algorithm to cluster the trading days in terms of conditional volatility distinguished comparable clusters with the k-means solution of Chen et al. (2021). Cluster 2 in this paper is comparable with cluster 1 in their paper. Cluster 3 in this study is comparable with their cluster 7. These display the market regime before and after the 2008 crisis. Besides, cluster 1 in this study is comparable with cluster 3 in their study, the stable period after the crisis. The cluster that captures the covid-19 crisis, cluster 8 in this paper is almost of the same length as cluster 8 of Chen's publication. Chen et al. (2021) distinguished cluster 2, the period right before and right after the financial crisis in 2008, from the other clusters around that period. GMMEUCovS does not capture this cluster. The GMMEUCovS method has formed three large clusters, where the observations of the k-means method are more equally distributed over the clusters. The GMMEUCovS method captures broader concepts of market regimes, which causes more clear groups with more observations. Clusters 1 and 2 capture the broad stable times, clusters 5 and 8 capture the broad volatile periods, all with clusters including more observations than the k-means results. The GMMEUCovS gives comparable results in distinguishing market regimes with the k-means approach. The GMMEUCovS captures broader and more extended periods of low and high volatile market regimes. The less obvious market regimes are less represented in the GMMEUCovS results (fewer observations included).

5.3.3 GMMEUCovS commodity covariance hierarchies for each cluster

The GMMEUCovS algorithm (Cavicchia et al., 2022) has another vital feature: the ability to detect hierarchical covariance structures within the clusters. In this subsection, the covariance hierarchies are explained in the following structure. The two ’normal’ clusters, 1 and 2, are evaluated in terms of variable groups, and the two clusters are compared in terms of similarities among the variable groups of the clusters. After that, the two ’crisis’ clusters 5 and 8 are evaluated in terms of variable groups, and the two ’crisis’ clusters are compared in terms of similarities among variable groups of the clusters. Finally, the four interesting clusters are compared altogether. All the graphs and tables are in the appendix.

The variable covariance hierarchy structure in calm times from 2000 until 2005, where cluster 2 is the most common, is shown in figure 4 and table 5: Group 2 and 3 are merged first. A broader group is formed together with group 6, including almost all base metals. Hereafter, in the following order, the broadest group in the hierarchy is formed when groups 7, 4, 5, and 1 are included. All the groups form concordant relationships.

During stable times from 2010 through 2019, cluster 1 is the most common. The variable group hierarchy structure is shown in figure 3 and table 4. Group 3 and 4 are merged, the same as group 1 and 6. Then broader groups are created in the following order, first, groups 3 and 4 with group 7. This group contains almost all agricultural-based commodities. The broadening of the groups is followed by group 5 and groups 1 and 6. The broadest group is the group that includes group 2, which has a negative covariance between the other variable groups, which shows that only the last two groups are discordant, and the other groups are concordant with each other.

The most specific groups, groups 1 and 6 merged for cluster 1, and groups 2 and 3 merged for cluster 2, have four similar commodities: silver, lead, cotton, and soybeans. When the following broader group is investigated, group 7 in cluster 1 and group 6 in cluster 2 are added, coffee is also a common commodity in both groups of the clusters. Besides, zinc, tin, and coffee are combined in one group for both stable clusters.

Cluster 5 is observed at the time of the financial crisis of 2008. This cluster shows a different EUCovS than the structure of more stable periods such as clusters 1 and 2 as shown in figure 5 and table 6. Groups 1 and 2 and groups 4 and 7 are merged. Group 6 is broadened by groups 4 and 7. Groups 1 and 2 broaden the latter. This broad group contains almost all base metals and all agricultural-based commodities except wheat. This broad group can be aggregated with group 3, with a discordant relationship, so negative covariance. Finally, group 5 creates the broadest group, with an additional negative covariance, which shows the discordant relationships between group 5 and the previous group.

During the beginning of the covid-19 crisis in 2020, cluster 8 appeared. This cluster is a cluster with high overall volatility. Although this is also a crisis, the covariance structure of the variables is different as shown in figure 6 and table 7. Clusters 6 and 7 are merged. Together with group 4 and thereafter group 5, they form a broader variable group. Groups 1 and 3 are merged and represent the largest part of the agricultural-based commodities, energy-based commodities, and base metals. These two groups form a broader variable group. The two groups have a negative covariance, which shows that the groups have a discordant relationship. Finally, group 2 can be added with another discordant relationship.

Both two crisis EUCovS show three general discordant groups. They also form a group with many standard variables. Groups 1, 2, 4, 6, and 7 of cluster 5 have fourteen common commodities together with groups 1 and 3 of cluster 8: silver, nickel, gasoil, gasoline, soya oil, corn, coffee, cotton, palladium, platinum, zinc, aluminum, palm oil, lead. The other broad concordant groups are different. Gasoil and gasoline are included in the same variable groups for volatile periods.

When inspecting the variable groups among the clusters, there are a few interesting results. Cotton and silver are always in the same variable group. This combination is supplemented by lead in stable times and supplemented by nickel in volatile times. Brent and lumber are also in the same group in every market regime, supplemented by copper for stable periods. Besides, gasoline, soya oil, and corn are always included in the same group, regardless of the market regime.

Next to the variables in each group, there are different relationships between blocks between stable and non-stable periods. The more stable periods have more concordant groups than the more unstable periods. Volatile clusters show more discordant relationships between the blocks.

6 Conclusion and Discussion

This study expands the application of the GMMEUCovS algorithm to financial time-series data for the first time in the literature. This could be the first step for more broad applications of the method on time-series data. However, is this application meaningful?

First of all, does the model detect interpretable clusters? The model can distinguish more volatile times, such as the financial crisis of 2008 and the covid-19 crisis, from more stable times. The more stable periods are captured by clusters 1 and 2. Although their overall conditional volatility is comparable, it is distributed differently before and after the 2008 crisis. Precious metals excluding gold, base metals tin, lead, nickel, copper, and aluminum, and all agricultural-based commodities excluding wheat and lumber are equal or less volatile after the crisis of 2008 (cluster 1) than before (cluster 2). Remarkable is that all the energy-based volatilities are strictly less volatile after the 2008 crisis than before.

The model distinguishes crisis times from more stable times in cluster 5 (2008 crisis) and cluster 8 (covid-19 pandemic). On average, these clusters have more volatility relative to ’normal’ periods. However, the two volatile clusters differ from each other in the distribution of volatilities. Cluster 8 is shorter than cluster 5, implying that the impact of covid-19 is relatively shorter than that of the 2008 crisis. However, this can be investigated by adding more trading days to the data set after July 2020.

How are the results related to the k-means clustering results from Chen et al. (2021), and are there differences? Both clustering techniques have comparable performance in distinguishing volatile periods from less volatile periods. The GMMEUCovS method can detect broader and longer clusters in the first decades of the 21st century. The other clusters whithout a prominent market regime are smaller than the k-means clusters.

What is the interpretation of the obtained groups and hierarchical covariance structures of the variables? Cotton and silver commodities are always in the same variable group for the broad clusters. This combination is thickened by lead in stable periods and nickel in volatile periods. Brent and lumber form always a group in broad clusters, supplemented by copper in stable periods. Additionally, regardless of the market regime, gasoline, soya oil, and corn are always included in the same variable group. Finally, stable periods have more concordant variable groups, and unstable periods have more discordant groups.

Thus, the GMMEUCovS method distinguishes volatile crisis periods from broader, more stable ones. The method captures broader concepts better than the k-means application. In addition to other parsimonious GMM clustering techniques, it can also detect the covariance hierarchy structures of the variables. Thus, the extended application of the method proposed by Cavicchia et al. (2022) to financial time-series data is the first step in investigating time-series data with (parsimonious) GMM methods.

For further research, it can be interesting to apply other (parsimonious) GMM methods to the same dataset. It can be interesting to measure and compare the performance of the individual clustering techniques, such as k-means and GMM, in terms of Dunn index (Dunn, 1974) or silhouette coefficient (Rousseeuw, 1987).

To optimize the outcomes of this research, the estimation of the GJR-GARCH(1,1,1) models can be improved. The log-return equation can be determined for every commodity instead of assuming the log-return equals the shock.

References

Z. Adams and T. Glück. Financialization in commodity markets: A passing trend or the new normal? *Journal of Banking Finance*, 60:93–111, 2015. ISSN 0378-4266. doi: https://doi.org/10.1016/j.jbankfin.2015.07.008. URL https://www.sciencedirect.com/science/article/pii/S0378426615002022.

M. Ahmadi, N. Bashiri Behmiri, and M. Manera. How is volatility in commodity markets linked to oil price shocks? *Energy Economics*, 59:11–23, 2016. ISSN 0140-9883. doi: https://doi.org/10.1016/j.eneco.2016.07.006. URL https://www.sciencedirect.com/science/article/pii/S0140988316301785.

H. Akaike. Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest*, 1973.

P. Arora, Deepali, and S. Varshney. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78:507–512, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.02.095. URL https://www.sciencedirect.com/science/article/pii/S1877050916000971. 1st International Conference on Information Security Privacy 2015.

A. Azqueta-Gavaldon, D. Hirschbühl, L. Onorante, and L. Saiz. Economic policy uncertainty in the euro area: an unsupervised machine learning approach. *Available at SSRN 3516756*, 2020.

A. Azqueta-Gavaldón. Developing news-based economic policy uncertainty index with unsupervised machine learning. *Economics Letters*, 158:47–50, 2017. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2017.06.032. URL https://www.sciencedirect.com/science/article/pii/S0165176517302598.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000. doi: 10.1109/34.865189.

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional discriminant analysis. *Communications in Statistics - Theory and Methods*, 36(14):2607–2623, 2007. doi: 10.1080/03610920701271095. URL https://doi.org/10.1080/03610920701271095.

E. Britannica. 2022. URL https://www.britannica.com/event/bankruptcy-of-Lehman-Brothers.

C. Cavicchia, M. Vichi, and G. Zaccaria. The ultrametric correlation matrix for modelling hierarchical latent concepts. *Advances in Data Analysis and Classification*, 14(4):837–853, May 2020. ISSN 1862-5347. doi: 10.1007/s11634-020-00400-z. Not EUR.

C. Cavicchia, M. Vichi, and G. Zaccaria. Gaussian mixture model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification*, Feb. 2022. ISSN 1862-5347. doi: 10.1007/s11634-021-00488-x. Publisher Copyright: © 2022, Springer-Verlag GmbH Germany, part of Springer Nature.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5): 781–793, 1995. ISSN 0031-3203. doi: https://doi.org/10.1016/0031-3203(94)00125-6. URL https://www.sciencedirect.com/science/article/pii/0031320394001256.

J. M. Chen, M. U. Rehman, and X. V. Vo. Clustering commodity markets in space and time: Clarifying returns, volatility, and trading regimes through unsupervised machine learning. *Recources Policy*, 2021.

CompaniesMarketCap.com. Market cap of gold (precious metal). 2022. URL https://companiesmarketcap.com/gold/marketcap/.

A. Creti, M. Joëts, and V. Mignon. On the links between stock and commodity markets' volatility. *Energy Economics*, 37:16–28, 2013. ISSN 0140-9883. doi: https://doi.org/10.1016/j.eneco.2013.01.005. URL https://www.sciencedirect.com/science/article/pii/S0140988313000078.

D. De Roux, B. Perez, A. Moreno, M. d. P. Villamil, and C. Figueroa. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 215–222, 2018.

R. O. Duda, P. E. Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4 (1):95–104, 1974. doi: 10.1080/01969727408546059. URL https://doi.org/10.1080/01969727408546059.

J. A. G. A. Fakhfekh, M. and N. Hachicha. "hedging stock market prices with wti, gold, vix and cryptocurrencies: a comparison between dcc, adcc and go-garch models". *International Journal of Emerging Markets*, ahead-of-print No., 2021. URL https://doi.org/10.1108/IJOEM-03-2020-0264.

L. R. GLOSTEN, R. JAGANNATHAN, and D. E. RUNKLE. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801,

1993. doi: https://doi.org/10.1111/j.1540-6261.1993.tb05128.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1993.tb05128.x.

M. Jang, H.-C. Jeong, T. Kim, and S.-K. Joo. Load profile-based residential customer segmentation for analyzing customer preferred time-of-use (tou) tariffs. *Energies*, 14(19), 2021. ISSN 1996-1073. doi: 10.3390/en14196130. URL https://www.mdpi.com/1996-1073/14/19/6130.

W. Jia, Y. Tan, L. Liu, J. Li, H. Zhang, and K. Zhao. Hierarchical prediction based on two-level gaussian mixture model clustering for bike-sharing system. *Knowledge-Based Systems*, 178:84–97, 2019. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.04.020. URL https://www.sciencedirect.com/science/article/pii/S0950705119301935.

A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(02)00060-2. URL https://www.sciencedirect.com/science/article/pii/S0031320302000602. Biometrics.

J. MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics Data Analysis*, 41(3):379–388, 2003. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(02)00183-4. URL https://www.sciencedirect.com/science/article/pii/S0167947302001834. Recent Developments in Mixture Model.

D. B. Nugroho, D. Kurniawati, L. P. Panjaitan, Z. Kholil, B. Susanto, and L. R. Sasongko. Empirical performance of GARCH, GARCH-m, GJR-GARCH and log-GARCH models for returns volatility. *Journal of Physics: Conference Series*, 1307(1):012003, aug 2019. doi: 10.1088/1742-6596/1307/1/012003. URL https://doi.org/10.1088/1742-6596/1307/1/012003.

K. Ohashi and T. Okimoto. Increasing trends in the excess comovement of commodity prices. *Journal of Commodity Markets*, 1(1):48–64, 2016. ISSN 2405-8513. doi: https://doi.org/10.1016/j.jcomm.2016.02.001. URL https://www.sciencedirect.com/science/article/pii/S2405851315300489.

H. Permuter, J. Francos, and I. Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2005.10.028. URL https://www.sciencedirect.com/science/article/pii/S0031320305004334. Graph-based Representations.

M. U. Rehman, E. Bouri, V. Eraslan, and S. Kumar. Energy and non-energy commodities: An asymmetric approach towards portfolio diversification in the commodity market. *Resources Policy*, 2019.

P. Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20: 53–65, 11 1987. doi: 10.1016/0377-0427(87)90125-7.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364. URL http://www.jstor.org/stable/2958889.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.

A. E. M. Shokry, M. A. Rizka, and N. M. Labib. Counter terrorism finance by detecting money laundering hidden networks using unsupervised machine learning algorithm. In *Proc. 13th IADIS Int. Conf.(ICT), Soc. Hum. Beings (ICT), 6th IADIS Int. Conf. Connected Smart Cities (CSC), 17th IADIS Int. Conf. Web Based Communities Social Media (WBC), 14th Multi Conf. Comput. Sci. Inf. Syst.(MCCSIS), 13th IADIS Int. Conf. ICT, Soc. Hum. Beings (ICT), 6th IADIS Int. Conf. Connected Smart Cities (CSC), 17th IADIS Int. Conf. Web Based Communities Social Media (WBC), 14th Multi Conf. Comput. Sci. Inf. Syst.(MCCSIS IADIS)*, pages 89–97, 2020.

R. J. Steele and A. E. Raftery. Performance of bayesian model selection criteria for gaussian mixture models. *Frontiers of statistical decision making and bayesian analysis*, 2:113–130, 2010.

K. Tang and W. Xiong. Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6):54–74, 2012. doi: 10.2469/faj.v68.n6.5. URL https://doi.org/10.2469/faj.v68.n6.5.

M. E. Tipping and C. M. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 02 1999. ISSN 0899-7667. doi: 10.1162/089976699300016728. URL https://doi.org/10.1162/089976699300016728.

T. Tiyasha, S. K. Bhagat, F. Fituma, T. M. Tung, S. Shahid, and Z. M. Yaseen. Dual water choices: The assessment of the influential factors on water sources choices using unsupervised machine learning market basket analysis. *IEEE Access*, 9:150532–150544, 2021. doi: 10.1109/ACCESS.2021.3124817.

Z. Umar, F. Jareño, and A. Escribano. Oil price shocks and the return and volatility spillover between industrial and precious metals. *Energy Economics*, 99:105291, 2021. ISSN 0140-9883. doi: https://doi.org/10.1016/j.eneco.2021.105291. URL https://www.sciencedirect.com/science/article/pii/S0140988321001961.

7 Appendix

7.1 R code

The code is structured as follow: Main.R includes the main steps of the algorithm. The main function runs starts the algorithm. This function calls step0, step1, step2, step3, step4, which are also included in in the main file. All the steps make use equations. Those equations are programmed in the file equations.R. The equations are named equation_xx, where xx equals the number of the equation in the paper of Cavicchia et al. (2022). Furthermore, some initializations are programmed in this file, the names of the functions speak for themselves.

If you would like to run my algorithm, you have to load the dataset first and open file "Main.R", after that, you can choose G and Q, whereafter you can run the main function in line 13. so only the first 13 lines are used to start the algorithm.

7.2 Graphs and tables

Figure 3: Hierarchy of 7 variable groups of cluster 1: path diagram representation

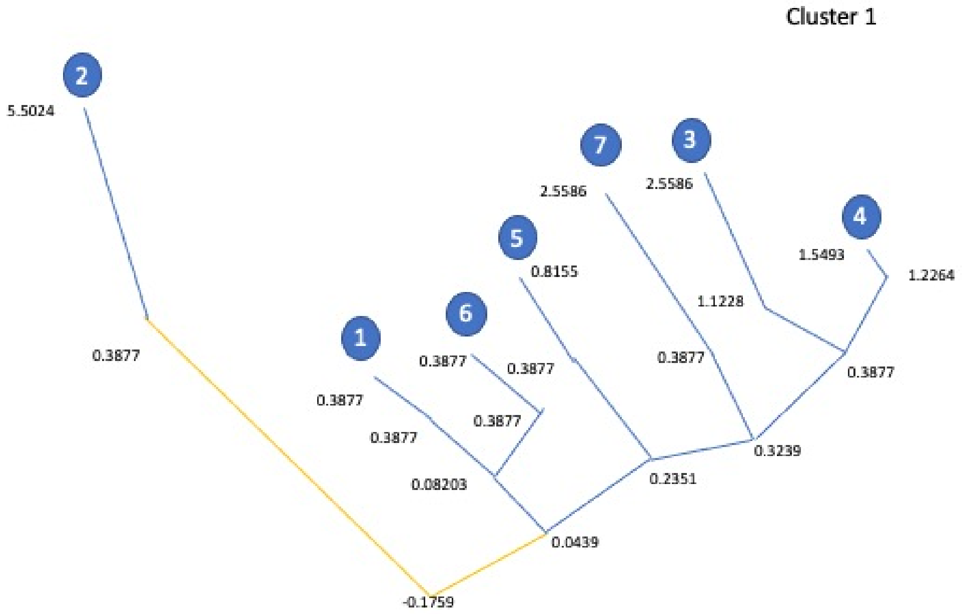


Table 4: Variable groups pinpointed by GMMEUCovS for cluster 1

cluster	variable group	included variables
1	1	palm_oil
	2	palladium
	3	zinc, tin, soybeans, coffee, cocoa
	4	silver, lead, cotton
	5	gold, platinum, copper, brent, gasoil, lumber
	6	aluminum
	7	nickel, wti, gasoline, soya_oil, wheat, corn

Figure 4: Hierarchy of 7 variable groups of cluster 2: path diagram representation

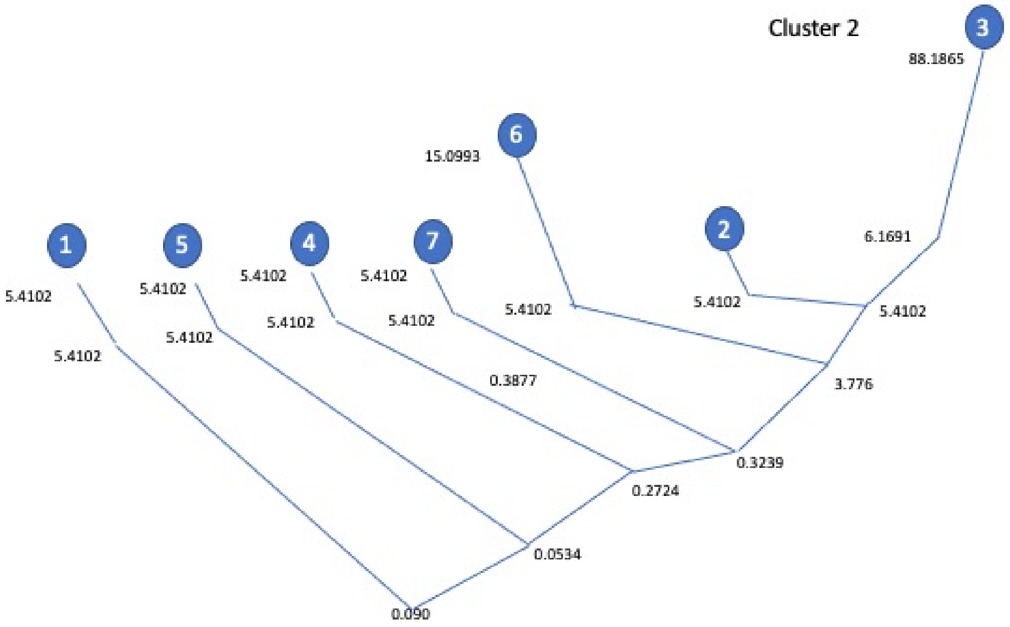


Table 5: Variable groups pinpointed by GMMEUCovS for cluster 2

cluster	variable group	included variables
	1	gasoline, soya_oil, corn
	2	copper, brent, palm_oil, wheat, soybeans, lumber
	3	silver, lead, cotton
2	4	gold
	5	platinum, palladium, nickel, cocoa
	6	zinc, tin, wti, gasoil, coffee
	7	aluminum

Figure 5: Hierarchy of 7 variable groups of cluster 5: path diagram representation

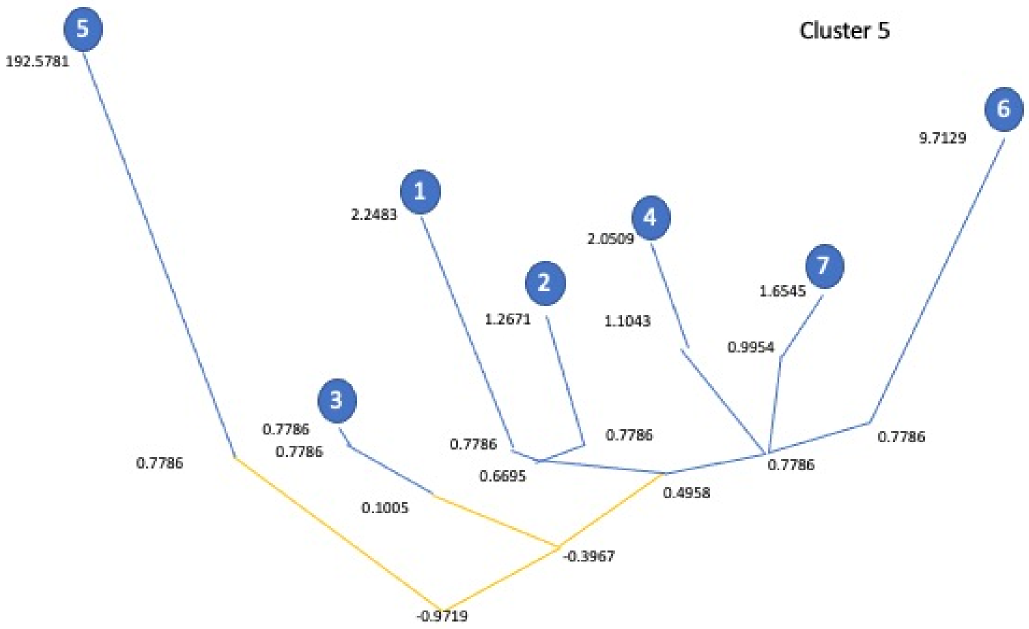


Table 6: Variable groups pinpointed by GMMEUCovS for cluster 5

cluster	variable group	included variables
5	1	silver, nickel, gasoil, gasoline, soya_oil, corn, coffee, cocoa, cotton
	2	palladium
	3	tin, wheat
	4	zinc, aluminum, palm_oil, soybeans
	5	gold, brent, lumber
	6	copper, lead, wti
	7	platinum

Figure 6: Hierarchy of 7 variable groups of cluster 8: path diagram representation

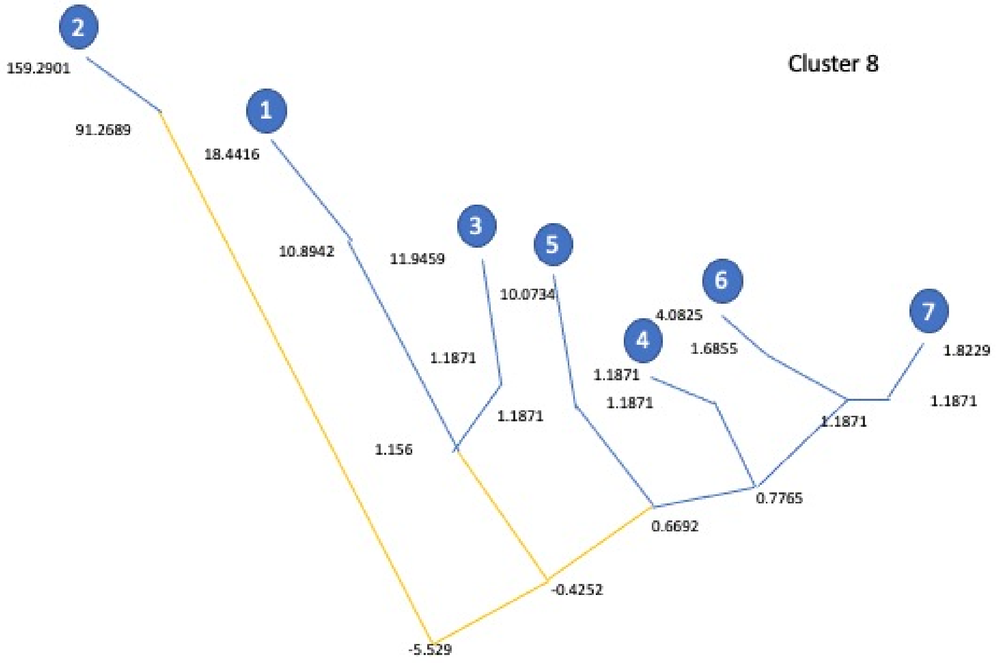
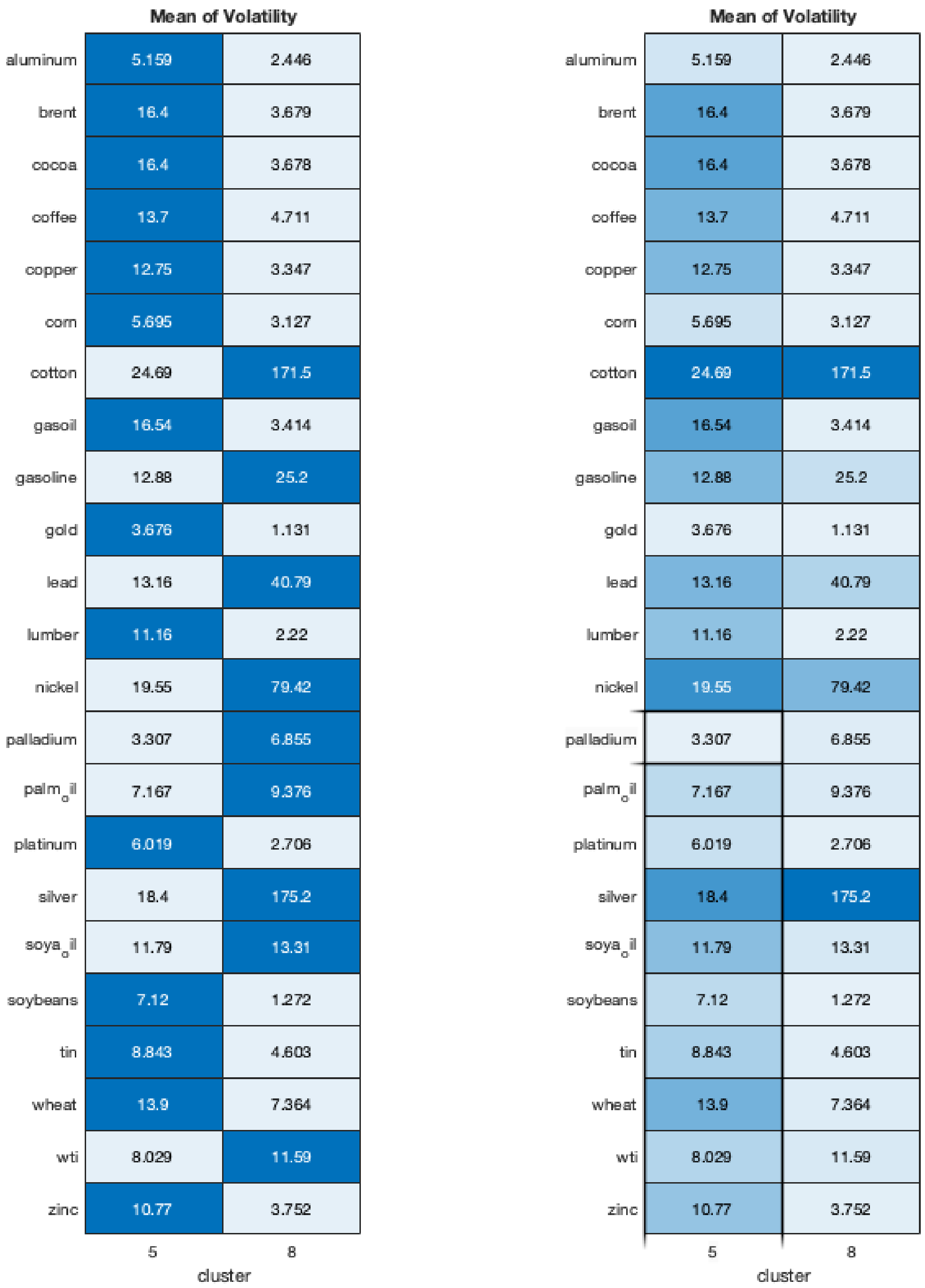


Table 7: Variable groups pinpointed by GMMEUCovS for cluster 8

cluster	variable group	included variables
8	1	platinum, palladium, zinc, tin, aluminum, brent, gasoil, gasoline, soya_oil, wheat, corn, coffee, lumber
	2	wti
	3	silver, lead, nickel, cotton, palm oil
	4	soybeans
	5	copper
	6	cocoa
	7	gold



(a) scaled rows

(b) scaled columns

Figure 7: Heatmaps of average conditional volatility of cluster 5 and 8