Delft University of Technology
Faculty Electrical Engineering, Mathematics & Computer Science
Delft Institute of Applied Mathematics

Nonparametric dependence modeling for financial markets using conditional Kendall's tau

(Dutch title: Nonparametrisch modelleren van de afhankelijkheid van financiële market met conditionele Kendall's tau)

Report on behalf of the
Delft Institute of Applied Mathematics
submitted in connection with

the degree of

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

JOB VLAK

Delft, The Netherlands
July 2021

Non-parametric dependence modeling for financial markets using conditional Kendall's tau

by

Job Vlak

Student Name	Student Number
Job Vlak	4859758

Supervisor:	Dr. A. Derumigny
Other committee member(s):	Dr. D. Kurowicka
Project Duration:	April, 2022 - July, 2022
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science, Delft

TU Delft

Preface

I would like to take this opportunity to thank Alexis for the more than excellent guidance during my Bachelor thesis. Every question I asked, you took the time to explain it completely from start to ending. At the end of the project I can say that I learned a lot from you, got intrigued by the world of financial mathematics and, above all, I got excited for programming in R. The latter I would not have expected to write down twelve weeks ago.

Job Vlak
Delft, July 2022

Abstract

In this thesis, we have examined conditional dependence in a financial context using conditional Kendall's tau (CKT). The conditional Kendall's tau is a measure of concordance between two random variables given some covariates. This thesis covers topics related to conditional Kendall's tau such as (conditional) copulas. We study non-parametric estimators of the conditional Kendall's tau using kernel density estimation and kernel regression. An application of the non-parametric estimator to the returns of thirteen different financial assets is finally provided. The assets consist of stock indices, bonds, futures and exchange rates. Further, we apply Principal Component Analysis (PCA) on the conditional Kendall's tau data matrix to increase the interpretability. In general, it seems that conditional dependence is slightly larger in the tails for all assets. Moreover, the conditional dependence for each group of assets is discussed. It seems that the degree of the conditional dependence relates to characteristics of an asset such as geographical properties and type of asset.

Contents

1

Introduction

In the field of dependence modeling, it is common to work with rank correlation coefficients. In this thesis, we use Kendall's tau. Contrary to usual linear correlations, it has advantages such as being scale-invariant [14]. Moreover, it can be explicitly written using underlying copulas. Kendall's tau can also be extended for the conditional setup. Surprisingly, their non-parametric estimates have been introduced in the literature only recently and their properties have not yet been fully studied in depth [8].

In view of applications to finance, conditional dependence is related to three-way interaction of financial assets. It tells us how the dependence evolves between the two assets when a conditioning variable is changing. Furthermore, it was shown that stock returns actually exhibit higher correlations during market declines than during market upturns [3]. This illustrates that analysing a model in which correlations depend on some conditioning variable is certainly relevant [22].

This thesis is structured as follows. Chapter 2 provides an overview of mathematical background. In order to understand the basis of this thesis, we touch upon the concept of (conditional) copulas. Furthermore, we discuss the idea of rank correlation. In particular, we examine Kendall's tau and its conditional setup. Next, we discuss the concept of kernel regression for the construction of non-parametric conditional estimators.

Chapter 3 is devoted to discussing the methodology of our research. First of all, we discuss several non-parametric estimators of the conditional Kendall's tau. We will define the conditional version of the averaging estimator that we use throughout this thesis. Next, we consider the theory behind Principal Component Analysis (PCA). This is a use-

ful statistical technique that analyses a multidimensional dataset with several dependent variables. It reduces dimensionality of such datasets and may increase interpretability.

Further, in Chapter 4 we perform an application of our methods to real data. The real data consists of thirteen different financial assets (variables). This includes eight stock indices from all over the world (AEX, DJI, EURO STOXX 50, CAC40, DAX, NASDAQ, Nikkei 225, S&P500). Moreover, we consider one US bond (FVX), one exchange rate (EURUSD), two futures related to oil prices (WTI and Brent) and the Bitcoin Price Index (BTC). First we estimate the conditional Kendall's tau for all assets using the package 'CondCopulas' in R. Then, we apply PCA and asses it by discussing principal components, scores, explained variance and the representation of the variables. Moreover, we cluster the PCA results on which we perform further analysis.

In Chapter 5 the conclusions of this research are summarized.

2

Preliminaries

In this chapter, we provide the necessary mathematical background for this report. First, some basic concepts of probability and statistics are explained. In section 2.2, copulas are introduced and defined, and their properties are discussed. Then, in section 2.4, rank-based correlation is explained, where Kendall's tau, in particular, is formally defined. Lastly, in Sections 2.5 and 2.6, Kernel density estimation and Kernel regression are introduced which are fundamental to understanding non-parametric estimation.

The proofs of the theorems in this chapter can be found in [5], [13], [14] and [17]. Besides, we use the formatting for all definitions and theorems from [22].

2.1. Basic notions

All concepts discussed in this subsection are from [5]. Let us first introduce the concept of conditional probability. This is essential to understand our main topic in this research: conditional Kendall's tau.

Definition 1 (Conditional PDF)**.** If X_1 and X_2 are continuous random variables in $\mathbb{R}$, with joint PDF $f(x_1, x_2)$, marginal density function f_1 of X_1, then the conditional probability density function of X_2 given $X_1 = x_1$ is defined to be

$$f(x_2|X_1 = x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \tag{2.1}$$

for all $x_1, x_2 \in \mathbb{R}$ such that $f_1(x_1) \geq 0$ and zero otherwise.

Note that this definition is for real-valued random variables. It can be easily extended for random variables in $\mathbb{R}^p$ for some integer $p \geq 2$. For this, we refer to [5]. Further,

there is a relation between conditional probability and independence: random variables X and Y are independent if and only if $f(x_2|X_1 = x_1) = f(x_2)$.

If X_1 and X_2 are discrete random variables, then the conditional PDF $f(x_2|x_1)$ is the conditional probability of the event $[X_2 = x_2]$ given the event $[X_1 = x_1]$. However, in the continuous case, $\mathbb{P}(X_1 = x_1)$ is defined, but equal to zero. The conditional PDF can therefore not be interpreted as a conditional probability in this case. Here, the conditional PDF can be thought of as assigning conditional "probability density" to arbitrarily small intervals $[x_2, x_2 + \Delta x_2]$. Thus, the conditional probability is the following

$$\mathbb{P}(a \leq X_2 \leq b|X_1 = x_1) = \int_a^b f(x_2|X_1 = x_1)\mathrm{d}x. \tag{2.2}$$

Similarly, we define the conditional CDF.

Definition 2 (Conditional CDF)**.** If X_1 and X_2 are random variables in $\mathbb{R}$, with joint PDF $f(x_1, x_2)$, then the conditional probability distribution function of X_2 given $X_1 = x_1$ is defined to be

$$F(x_2|X_1 = x_1) = \int_{-\infty}^{x_2} f(x_2|X_1 = x_1)\mathrm{d}x. \tag{2.3}$$

Next, we will explain conditional expectation which will be necessary to understand Kernel regression in section 2.4.

Definition 3 (Conditional expectation)**.** If X_1 and X_2 are continuous random variables with joint distribution function $f_{X,Y}(x, y)$, then the conditional expectation of X_2 given $X_1 = x_1$ is given by

$$m(x_1) = \mathbb{E}[X_2|X_1 = x_1] = \int_{-\infty}^{\infty} x_2 f(x_2|X_1 = x_1)dx_2 = \frac{\int x_2 f(x_2|X_1 = x_1)\mathrm{d}x_1}{f(x_1)}. \tag{2.4}$$

We will finish this subsection by explaining the probability integral transform which is linked to copulas. This topic describes that random variables from any given continuous distribution can be converted to random variables having a standard uniform distribution.

Theorem 1 (Probability Integral Transformation)**.** If a random variable X is continuous with CDF $F(x)$, then the random variable $U := F(x)$ has a standard uniform distribution, denoted as $U \sim \mathrm{UNIF}(0, 1)$

2.2. Copulas

The study of copulas and their applications in statistics is a rather modern phenomenon. From one point of view, copulas are functions that join multivariate distribution func-

tions to their one-dimensional marginal distribution functions. This is not a formal definition yet. Therefore, we will give two different definitions of copulas and state some of their fundamental properties. First, we provide a more analytical definition.

Definition 4 (P-dimensional copula)**.** Let $p \geq 2$ be an integer. A copula is a function $C : [0,1]^p \to [0,1]$ with the following properties:

1. For any $j = 1, \ldots, p$ and all $u_j \in [0,1]$, $\mathrm{C}(u_1, \ldots, u_{j-1}, 0, u_{j+1}, \ldots, u_p) = 0$.
2. For any $j = 1, \ldots, p$ and and all $u_j \in [0,1]$, $\mathrm{C}(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$.
3. C is a p-increasing, i.e. for each hyperrectangle $A = \Pi_{j=1}^{p}[a_j, b_j] \subseteq [0,1]^p$ the C-volume of A is non-negative:

$$\int_A \mathrm{d}C(\mathbf{u}) \geq 0.$$

Another definition is that copulas are the joint cumulative distribution functions of a multivariate random vector with uniform margins. This becomes clear when looking at Sklar's theorem, which is central to the theory of copulas and its applications in statistics.

Theorem 2 (Sklar's Theorem)**.** Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a random vector with joint cumulative distribution function F and univariate marginal CDFs $F_1, \ldots, F_p$. Then, there exists a copula C such that for all $x \in \mathbb{R}$,

$$F(x) = C(u_1, \ldots, u_p). \tag{2.5}$$

In addition, C is given for all $u \in [0,1]^p$ by

$$C(u_1, \ldots, u_p) = F(F_1^-(u_1), \ldots, F_p^-(u_p)), \tag{2.6}$$

where $F_1^-(u_j)$ is the inverse of F_j for $j = 1, \ldots, p$. Therefore, if X is continuous, then C is unique.

Theorem 2 shows that any joint CDF can be rewritten as a copula and marginal CDFs. Moreover, if the marginal CDFs are continuous on $\mathbb{R}$, and C is a copula (i.e. a distribution on $[0,1]$ with p uniform margins), then a joint CDF on $\mathbb{R}^p$ can be defined as [6]

$$F(\mathbf{x}) = C(F_1(x), \ldots, F_p(x)). \tag{2.7}$$

We refer to the class of all p-dimensional copulas with the notation $\mathcal{C}_p$. This set has the following important properties [6]:

1. Pointwise and uniform convergence are equivalent.
2. $\mathcal{C}_p$ is a convex and compact set.
3. Every copula is 1-Lipschitz
4. $(\mathcal{C}_p, <)$ is an ordered set, i.e. $C_1 < C_2$ if $C_1(\mathbf{u}) \leq C_2(\mathbf{u})$ for all $\mathbf{u} \in [0,1]^p$ and $C_1, C_2 \in C_p$

Basic concepts in probability, such as densities, also apply to the theory of copulas. If a copula exists and it has a density, then the probability density function of a copula can be obtained in the usual manner as follows,

$$c(u_1, \ldots, u_p) = \frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1 \ldots \partial u_p}. \tag{2.8}$$

From this, if the marginal densities $f_1, \ldots, f_p$ exist, it follows that the PDF of a random variable $\mathbf{X} = (X_1, \ldots, X_p)$ can be rewritten as

$$f(\mathbf{u}) = f_1(x) f_2(x) \ldots f_p(x) \cdot c(u_1, \ldots, u_p). \tag{2.9}$$

Now, we introduce the conditional setup in the context of the theory of copulas. Conditional copulas are important in applications such as time series and econometric models [6]. Moreover, conditional copulas are necessary to understand, to compute and to use conditional Kendall's. Therefore, we state the previous definition and theorem adjusted to the conditional version, which are rather similar to the unconditional case.

First of all, we introduce a random variable, say $\mathbf{Z} = (z_1, \ldots, z_d)$, which is the conditioning variable on $\mathbb{R}^d$. Conditional copulas form the link between the conditional joint CDF of a multivariate random vector with the conditional uniform marginals. The formal definition is as follows

Definition 5 (p-dimensional conditional copula)**.** Let $p \geq 2$ be an integer and $\mathbf{Z}$ be a conditioning random vector taking values in $\mathcal{Z} \subseteq \mathbb{R}^d$. A conditional copula is a measurable function $C : [0,1]^p \times \mathcal{Z} \to [0,1]$ such that for $\mathbb{P}_{\mathbf{Z}}$-almost every $\mathbf{z} \in \mathcal{Z}$, the following properties are satisfied:

1. For any $j = 1, \ldots, p$ and all $u_j \in [0,1]$, $\mathrm{C}(u_1, \ldots, u_{j-1}, 0, u_{j+1}, \ldots, u_p | \mathbf{Z} = \mathbf{z}) = 0$.
2. For any $j = 1, \ldots, p$ and and all $u_j \in [0,1]$, $\mathrm{C}(1, \ldots, 1, u_j, 1, \ldots, 1 | \mathbf{Z} = \mathbf{z}) = u_j$.
3. C is a p-increasing, i.e. for each hyperrectangle $A = \Pi_{j=1}^p [a_j, b_j] \subseteq [0,1]^p$ the C-volume of A is non-negative:

$$\int_A \mathrm{d}C(\mathbf{u} | \mathbf{Z} = \mathbf{z}) \geq 0.$$

Sklar's theorem can also be translated to the conditional setup. This will be as follows

Theorem 3 (Sklar's theorem - conditional version)**.** Let $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{Z}$ be a random vectors taking values in $\mathbb{R}^{\prime}$ and $\mathcal{Z} \subseteq \mathbb{R}^d$, respectively. Let the joint cumulative distribution function of $\mathbf{X}$ given $\mathbf{Z}$=$\mathbf{z}$, denoted by $F_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$, have the conditional univariate marginal CDFs $F_{1|\mathbf{Z}=\mathbf{z}}, \ldots, F_{p|\mathbf{Z}=\mathbf{z}}$. Then, there exists a conditional copula $C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ such that for all $\mathbf{x} \in \mathbb{R}^p$ and all $\mathbf{z} \in \mathcal{Z}$,

$$F_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(F^{-}_{1|\mathbf{Z}=\mathbf{z}}(u_1), \ldots, F^{-}_{p|\mathbf{Z}=\mathbf{z}}(u_p)) = C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(u_1, \ldots, u_p), \tag{2.10}$$

where $F^{-}_{j|\mathbf{Z}=\mathbf{z}}$ is the inverse of $F_{j|\mathbf{Z}=\mathbf{z}}$ for $j = 1, \ldots, p$. Therefore, if the conditional random variables $\mathbf{X}|\mathbf{Z} = \mathbf{z}$ are continuous, then $C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ is unique.

From Theorem 3, it follows that a conditional copula connects the conditional CDF with its conditional marginals. This is similar to what we have seen in the unconditional version, Theorem 2.

As becomes clear now, copulas let us separate the marginal distribution functions and the dependency structure of a multivariate distribution. There are a lot of different parametric classes of copulas that describe this dependence. Frequently used parametric classes are the Gaussian copula, t-copula, Gumbel's copula and Clayton copula. For more examples, we refer to [12] and [14].

2.2.1. Copulas in the financial market

The concept of copula has attracted more and more attention in finance and economics in recent years. For example, copulas have been widely applied in risk management.

Generally, a copula allows us to separate the univariate probability distribution of the variables (for example, a financial asset) from the interdependencies between it and other variables (other financial assets) defined by the copula. By doing so, one can model each variable separately and, on top of that, have a measure of the relations between those financial assets. Technically, this means that the univariate probability distribution, telling us the probabilities of outcomes of one financial asset, in particular, can be modelled by type of distribution of choice. Whereas another variable can be modelled using another type of probability distribution. By doing so, one can choose for each and any asset the most appropriate type of distribution, not influencing the relation between those assets. These interdependencies between those variables are represented by a multivariate probability distribution function, which tells us the joint outcomes of the variables. Looking at Sklar's theorem 2, it becomes clear that an p-dimensional

multivariate distribution, representing the copula of p financial assets, can capture all possible interdependencies between the p-variables [18].

There are risks connected to working with copulas. During the late 1990s, the CDOs appeared on the financial market. Collateralized Debt Obligations were new financial instruments which made it possible to form securities out of different types of debts, e.g. mortgages, via these derivatives. The correlation between defaults needed to be modelled in order to price these securities. David X. Li's Gaussian copula approach was used to model just that. The Gaussian copula, which is a common choice of the copula, is a helpful tool and relatively easy to fit. However, the Gaussian copula does not capture tail dependencies; risk in the tail is underestimated. In 2008, the crisis hit Wall street and the CDO market collapsed. Li's Gaussian copula model has been accused of increasing the intensity of the financial crisis [4].

2.3. Dependence measures

There are numerous ways to assess the dependence between random variables. In this section, we explain linear correlation and rank correlation, in particular Kendall's tau.

2.3.1. Correlation

Let us start with linear correlation which is most well-known in practice, the so-called Pearson's correlation. Pearson's coefficient measures the strength of linear correlation between two random variables. An advantage of Pearson coefficient is that it is computed easily compared to rank correlation. Other advantages are that it is closely related to the normal distribution and that linear operations are done easily. On the other hand, the Pearson coefficient has some obvious disadvantages. Most important, it can not be used for nonlinear dependence [16]. Pearson correlation coefficient is defined as the covariance of the two variables normalized to the product of the roots of the variance of both variables.

Definition 6 (Pearson's correlation coefficient)**.** Let X_1 and X_2 be real-valued random variables. The Pearson correlation coefficient is defined by

$$\rho_{X_1,X_2} := \frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}, \tag{2.11}$$

where σ is the standard deviation of the variables, respectively of X_1 and X_2.

Now let $\{(X_{1,1}, X_{2,1}), \ldots, (X_{1,n}, X_{2,n})\}$ be a collection of paired observations, where $X_{i,j}$

represents the j-th observation of the i-th variable. Then, the sample Pearson correlation coefficient is defined by

$$r_{X_1,X_2} = \frac{\sum_{i=1}^{n}(X_{1,i} - \bar{X}_1)(X_{2,i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{1,i} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^{n}(X_{2,i} - \bar{X}_2)^2}}, \tag{2.12}$$

where $\bar{X}_1$ and $\bar{X}_2$ denote the sample mean of respectively $\{(X_{1,1}, \ldots, X_{2,1})\}$ and $\{(X_{1,n}, X_{2,n})\}$.

Note that for heavy-tailed data, this coefficient will lead to misleading results because outliers will have a different impact on the numerator and denominator [22]. Therefore, it is necessary to introduce a more robust correlation coefficient.

2.3.2. Rank Correlation

Contrary to Pearson's coefficient, rank correlation has the advantage of being invariant to (monotonic) changes in the underlying marginal distributions. Rank correlation coefficient is a non-parametric measure of the strength of the relationship between two rankings. It returns a value inside the interval $[-1, 1]$. Rank correlation is popular in the field of dependence modelling because of its advantages over linear correlation. Moreover, the correlation coefficient can be explicitly written using copulas. Examples of rank correlation coefficients are Kendall's tau, Spearman's rho and Blomqvist's coefficient.

2.3.3. Kendall's tau

In this thesis, only Kendall's tau will be used, in particular the conditional version. First, we define Kendall's tau and then we elaborate on this for the conditional setup.

For more than a century, Kendall's tau has become a popular dependence measure [8]. It quantifies the positive or negative dependence between two random variables in the interval $[-1, 1]$. Here, the value -1 corresponds to the perfect negative correlation and the value 1 corresponds to a perfect positive correlation. Further, note that the value of Kendall's tau is equal to zero if the variables are independent. However, the converse does not hold. So Kendall's tau being equal to zero does not necessarily mean that the variables are equal to zero [22]. Kendall's tau is based on concordance and discordance of the data. Informally speaking, a pair of random variables are concordant if large values of one variable tend to be associated with large values of the other and, contrarily, small values of one with small values of the other [14]. Formally speaking, it is defined as follows

Definition 7 (Concordance and discordance)**.** Let $X_{1,1:2}$ and $X_{2,1:2}$ be two independent copies of a random vector $\mathbf{X} \in \mathbb{R}^2$, then a pair is

- **condordant** when $(X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0$
- **discondordant** when $(X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0$

A pair of bivariate observations is concordant when both elements of one observation are either greater than or less than the corresponding elements of the other observation. A pair is discordant pair when only one of the elements of one observation is greater than the corresponding element of the other observation [22].

Let us define Kendall's tau as the difference between the probability of concordance and the probability of discordance of two independent versions of (X_1, X_2). Denoting by $C_{1,2}$ the unique underlying copula of (X_1, X_2) that is assumed to be continuous, their Kendall's tau can be directly defined as [6]

Definition 8 (Kendall's tau)**.** Let X_1 and X_2 be real-valued random variables. The population Kendall's tau of X_1 and X_2 is defined as

$$\tau_{X_1,X_2} := \mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0) - \mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0), \quad (2.13)$$

where $(X_{1,i}, X_{2,i})_{i=1,2}$ are to independent copies of (X_1, X_2). Further, we define the sample Kendall's tau for paired observations $\{(X_{1,1}, X_{2,1}), \ldots, (X_{1,n}, X_{2,n})\}$ by

$$\widehat{\tau_{X_1,X_2}} := \frac{2}{n(n-1)} \sum_{i_1 \leq i_2} sign((X_{1,i_1} - X_{1,i_2})(X_{2,i_1} - X_{2,i_2})). \quad (2.14)$$

It becomes clear that Kendall's tau will have a value inside the interval $[-1, 1]$. Moreover, note that the inequalities are strict because no such terms are corresponding to the cases $i = j$.

There are several alternative expressions for Kendall's tau equivalent to Definition 8, whenever the random variables are continuous.

$$\begin{aligned}
\tau_{X_1,X_2} :&= \mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0) - \mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0), \\
&= 2\mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2})) > 0) - 1, \\
&= 4\mathbb{P}(X_{1,1} < X_{2,1}, X_{1,2} < X_{2,2}) - 1, \\
&= 4\mathbb{P}(\mathbf{X}_1 < \mathbf{X}_2) - 1, \\
&= 1 - 4\mathbb{P}(X_{1,1} < X_{2,1}, X_{1,2} > X_{2,2}), \\
&= 4 \int C(u_1, u_2) \mathrm{d}C(u_1, u_2) - 1,
\end{aligned} \quad (2.15)$$

where $C(u_1, u_2)$ denotes the unique copula. Kendall's tau, therefore, provides much information about the underlying dependence structure. For convenience, the notation $\tau_{1,2}$ will be used instead of τ_{X_1,X_2}, whenever it is clear which variables are meant.

Finally, Kendall's tau has some advantages over Pearson's rho or other concordance measures worth mentioning [6].

- Kendall's tau is invariant for monotonic transformations.
- Kendall's tau always exists for any couple of marginal distributions.
- For one-dimensional families of copulas, the estimation of the parameter is equivalent to estimating Kendall's tau.

2.3.4. Conditional Kendall's tau

Now let us turn to the conditional Kendall's tau (CKT), having a multivariate covariate, say $\mathbf{Z} = (z_1, \ldots, z_p)$. In this research, we will consider only pointwise conditioning events. Using conditional Kendall's tau, the dependence between the variables X_1 and X_2 is measured given the p-dimensional vector of covariates $\mathbf{Z}$. Conditional Kendall's tau, and more generally conditional dependence measures, are of interest because they tell us the behaviour of the dependence between X_1 and X_2 when the covariate $\mathbf{Z}$ is changing. Conditional Kendall's tau is a conditional dependence measure used to predict whether a pair of observed random variables is concordant or discordant conditioned on $\mathbf{Z}$ [8] [22]. We define the conditional Kendall's tau similarly as in the previous subsection.

Definition 9 (Conditional Kendall's tau)**.** Let X_1 and X_2 be real-valued random variables and let $\mathbf{Z}$ be a random vector taking values in $\mathbb{R}^p$. For any point $\mathbf{z} \in \mathbf{Z}$, we define Kendall's tau of X_1 and X_2 conditional on $\mathbf{Z} = \mathbf{z}$ by

$$\begin{aligned}\tau_{X_1,X_2|\mathbf{Z}=\mathbf{z}} := \mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}) \\ - \mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}),\end{aligned}$$

where $(X_{1,i}, X_{2,i}, \mathbf{Z}_i)_{i=1,2}$ are independent copies of $(X_1, X_2, \mathbf{Z})$.

The conditional Kendall's tau has similar properties as we have seen in the previous subsection. For every point, $\mathbf{z} \in \mathbf{Z}$, the conditional Kendall's tau takes values inside the interval $[-1, 1]$.

Furthermore, when the conditional marginal distributions of X_1 and X_2 given $\mathbf{Z} = \mathbf{z}$ are continuous, the following alternative expressions of conditional Kendall's tau are equivalent to the one in Definition 9.

$$\tau_{X_1,X_2|\mathbf{Z}=\mathbf{z}} := 2\mathbb{P}((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2})) > 0|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}) - 1, \tag{2.16}$$

$$= 4\mathbb{P}(X_{1,1} < X_{2,1}, X_{1,2} < X_{2,2}|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}) - 1, \tag{2.17}$$

$$= 4\mathbb{P}(\mathbf{X}_1 < \mathbf{X}_2|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}) - 1, \tag{2.18}$$

$$= 1 - 4\mathbb{P}(X_{1,1} < X_{2,1}, X_{1,2} > X_{2,2}|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}), \tag{2.19}$$

$$= 4\int_{[0,1]^2} C_{1,2|\mathbf{Z}=\mathbf{z}}(u_1, u_2)\mathrm{d}C_{1,2|\mathbf{Z}=\mathbf{z}}(u_1, u_2) - 1, \tag{2.20}$$

where $C_{1,2|\mathbf{Z}=\mathbf{z}}(u_1, u_2)$ denotes the unique conditional copula. Finally, the advantages stated in the previous subsection also hold in the conditional setup.

Further, note that we are conditioning on a multivariate covariate $\mathbf{Z}$. Although in most textbooks conditioning is only defined with respect to one variable, the covariate could be multivariate as well. Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be random vectors both taking values in $\mathbb{R}^p$. We define $\mathbf{U} = (\mathbf{Z}_1, \mathbf{Z}_2)$, and $\mathbf{u} = (\mathbf{z}, \mathbf{z}) \in \mathbb{R}^{2p}$, then conditioning by $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z} \in \mathbb{R}^{2p}$ can be defined as conditioning with respect to the event $\mathbf{U} = \mathbf{u}$.

To understand where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ come from, note that, in the unconditional case of Kendall's tau, the observations $(X_{1,1}, X_{1,2})$ and $(X_{2,1}, X_{2,2})$ follow the same distribution F_{X_1,X_2}. In the conditional case in (2.20), we see it is coherent that the observations follow the same distributions $F_{X_1,X_2|\mathbf{Z}=\mathbf{z})}$.

Another way of looking at this, take for example three random vectors representing three different assets, each taking values in $\mathbb{R}^p$. We call them first asset $\mathbf{X}_1$, second asset $\mathbf{X}_2$ and the conditioning variable $\mathbf{Z}$, respectively. The conditioning variable of the random vector's observations $(X_{1,1}, X_{1,2})$ and $(X_{2,1}, X_{2,2})$ would not need to be independent in general. To be sure we compare the conditioned observations in the right way, we must ensure that both $\mathbf{Z}_1 = \mathbf{z}$ and $\mathbf{Z}_2 = \mathbf{z}$.

Defining estimators for a conditional Kendall's tau is less straightforward as we need to deal with the dependency on $\mathbf{Z}$. Therefore, we discuss kernel density estimation and kernel regression in the upcoming sections.

2.4. Kernel Density Estimation

We are using non-parametric estimation in this thesis. Non-parametric modelling favours generality. Given a set of minimal and weak assumptions it provides methods that are consistent for broad situations. This means it loses efficiency to some extent. Broadly

speaking, non-parametric modelling does not rely on parametric assumptions. Parametric modelling, on the other hand, favours efficiency. Given a model (a strong assumption on the data generating process), parametric inference delivers a set of methods (e.g. point estimation, confidence intervals, hypothesis testing) tailored for such a model. If the data generating process truly satisfies the assumptions, then it works. Otherwise, the methods may be inconsistent.

In a lot of situations, for instance, in the world of financial markets, knowledge of the data generation process is rarely known. That is the appeal of a non-parametric method: it will perform adequately no matter what the data generation process is. For that reason, non-parametric methods are useful [20].

Kernel density estimation is based on estimating the density function from the observed data. A density function f is simplest estimated from an independent and identical sample of observations $X_1, \ldots, X_n$ by a histogram. However, histograms depend on the bandwidth h and the origin t_0. The latter, we would like to avoid by letting the bins to be dependent on x (the point we want to estimate on f), rather than fixing this point beforehand. This is called a moving histogram and forms the basis for kernel density estimation. The presented formulas follow from [10]. This can be written as .... (problem with my overleaf code) ....

$$f(x;h) = \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}_{\{x-h\leq X_i\leq x+h\}}, \tag{2.21}$$

$$= \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2}\mathbb{1}_{\{-1\leq \frac{x-X_i}{h}\leq 1\}}, \tag{2.22}$$

$$= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{2.23}$$

where $K(z) = \frac{1}{2}\mathbb{1}_{\{-1<z<1\}}$, which is uniform on $(-1,1)$. Generally, $K(z)$ can be any density. Then $K(z)$ is known as a kernel. A kernel K is (most of the time) non-negative, symmetric, unimodal at zero (density has a single peak at zero) and satisfying $\int K = 1$. The histogram estimator can be considered as a sum of 'boxes' centred at the observations, the kernel estimator is a sum of 'bumps' placed at the observations. The kernel function K determines the shape of the bumps while the smoothing parameter h determines their width [20], hence it is called bandwith. More precise, the bandwidth h controls the sensitivity of the density estimates towards observations further away from z. The Gaussian and Epanechnikov kernels are common choices. However, the choice of kernel is not that important, since all estimates seem to have roughly the same shape

for different kernels having the same bandwidth. In fact, the bandwidth h is the crucial factor [10], [22].

This can be extended for a multivariate case. Then kernel density estimation is used to estimate multivariate densities f in $\mathbb{R}^p$ for a sample $X_1, \dots, X_n$, then the KDE at $\mathbf{X}$ is defined as

$$f(x; \mathbf{H}) := \frac{1}{n|\mathbf{H}^{\frac{1}{2}}|} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{H}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{X}_i)), \tag{2.24}$$

where $K_{\mathbf{H}}$ is the multivariate kernel and $\mathbf{H}$ is a $p \times p$ matrix representing the bandwidth in the multivariate case. This matrix is symmetric and positive definite. Again, the same properties as discussed before hold for a p-variate density function ([10], [20]).

2.5. Kernel Regression

Kernel regression is introduced to construct non-parametric conditional estimators which are necessary for estimating conditional Kendall's tau. The presented formulas follow from [10].

Recall Definition 3 for the definition of the conditional expectation. We are conditioning on some covariate $\mathbf{Z} \in \mathbb{Z}$. In kernel regression, the estimates of the densities f are computed by kernel density estimation from section 2.4. This comes down to the following

$$\widehat{f_{\mathbf{Z},Y}}(\mathbf{z}, y; h) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1\left(\frac{\mathbf{z} - \mathbf{Z}_i}{h}\right) \frac{1}{h} K_2\left(\frac{y - Y_i}{h}\right), \tag{2.25}$$

$$\widehat{f_{\mathbf{Z}}}(\mathbf{z}; h) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1\left(\frac{\mathbf{z} - \mathbf{Z}_i}{h}\right) \frac{1}{h} K_2\left(\frac{y - Y_i}{h}\right), \tag{2.26}$$

where K_1 and K_2 are kernel functions. Now the estimate of the conditional expectation is obtained by replacing the densities in Definition 3 with their estimates (2.25) and (2.26). This gives the following expression for estimating $m(\mathbf{z}) = \mathbb{E}[X|\mathbf{Z} = \mathbf{z}]$

$$\begin{aligned} \widehat{m(\mathbf{z}; h)} &:= \frac{\int x \widehat{f_{Y|\mathbf{Z}=\mathbf{z}}}(\mathbf{z}, x; h) dx}{\widehat{f_{\mathbf{Z}}}(\mathbf{z}; h))}, \\ &= \frac{\int x \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h}) \frac{1}{h} K_2(\frac{x - X_i}{h})}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h}}, \\ &= \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h}) \int x \frac{1}{h} K_2(\frac{x - X_i}{h})}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h})}. \end{aligned} \tag{2.27}$$

Then change of variables with $u_i = \frac{x - X_i}{h}$ can be used to compute the integral. Then the following will be obtained

$$\begin{aligned} \int x \frac{1}{h} K_2\left(\frac{x - X_i}{h}\right) \mathrm{d}x &= \int (hu + X_i) K(u) \mathrm{d}u, \\ &= h \int u K_2(u) \mathrm{d}u + X_i \int K_2(u) \mathrm{d}u, \\ &= X_i, \end{aligned} \tag{2.28}$$

where we used the properties of symmetry and $\int K = 1$ of a kernel. By combining (2.27) and (2.28) together, we obtain

$$\begin{aligned} \widehat{m(\mathbf{z}; h)} &:= \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h}) X_i}{\frac{1}{n} \sum_{i=1}^{n} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h})}, \\ &= \sum_{i=1}^{n} \frac{\frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_i}{h})}{\sum_{j=1}^{n} \frac{1}{h^d} K_1(\frac{\mathbf{z} - \mathbf{Z}_j}{h})} X_i, \end{aligned} \tag{2.29}$$

where the expression in front of the random variable X_i can be seen as a weighted average of observations $X_1, \ldots, X_n$. These are called the Nadaraya-Watson weights. This means that the Nadaraya-Watson weights are the local means of $X_1, \ldots, X_n$ about $Z = z$. There are different types of weights that can be used to construct an estimate. In this thesis, however, we will only consider the Nadaraya-Watson weights. The resulting estimator for the conditional expectation can be rewritten as

$$\widehat{m(\mathbf{z}; h)} := \sum_{i=1}^{n} w_{i,n}(\mathbf{z}) X_i, \tag{2.30}$$

where

$$w_{i,n}(\mathbf{z}) := \frac{K_h(\mathbf{z} - \mathbf{Z}_i)}{\sum_{i=1}^{n} K_h(\mathbf{z} - \mathbf{Z}_i)}, \tag{2.31}$$

where $K_h := \frac{1}{h^d} K(\,\cdot\,/h)$.

We will finish this section with a notion of the bias-variance trade-off and the curse of dimensionality.

First of all, bias is defined as the error between an estimator's expected value and the true value of the parameter being estimated. A high bias may cause missing relevant relations which results in a too simple model (underfitting). Underfitting happens when a model is unable to capture the underlying pattern of the data. These models usually

have a high bias [21]. Whereas the variance is defined as to what extent estimates are spread out from their average value. The variance explains the sensitivity in small fluctuations in a set. For Kernel estimations, in particular, high variance can be seen as a result of overfitting. The bias-variance trade-off explains, that when the bias is high, the variance is small and vice versa. As we have seen in the previous section, the bandwidth h controls the estimate's sensitivity towards observations Z_i that are further away from point z. Thus, reducing the bandwidth will decrease the estimator's bias and will increase its variance, which is the bias-variance trade-off [22] [10]. Hereby, larger sample sizes will allow for a smaller choice of bandwidth, since we have more data.

Lastly, note that the volume of the space $\mathbf{Z}$ grows exponentially fast when increasing the dimensionality of Z. This creates a decreasing density of observations within that space. This is called the curse of dimensionality. In other words, when the dimensionality increases, then the volume of the space increases so fast that the available data becomes sparse. To still obtain a reliable result, the amount of data often needs to grow exponentially with dimensionality. The immediate consequence is that we can only consider covariates of a few dimensions at most [22].

3

Methods

In this chapter, we set out the methods that we use to research how the conditional dependence is for the different assets by using conditional Kendall's tau (CKT). First, we will explain what data we use and what properties it has. Then, we introduce the estimators for a conditional Kendall's tau that we use throughout this report. Lastly, we will discuss Principal Component Analysis (PCA) which is a useful method to interpret high-dimensional data.

3.1. Financial Data

The original financial data used in this thesis is imported from Yahoo Finance. We will use the monthly returns of thirteen different assets from March 1986 until August 2020. The different assets, their types and abbreviations are listed below.

Type	Name	Abbreviation
European Stock Indices	France CAC 40 Stock Index	FCHI
	German DAX Index	GDAXI
	Amsterdam Exchange Index	AEX
	EURO STOXX 50 (European companies)	Eurostoxx
US Stock Indices	Dow Jones Industrial Average (30)	DJI
	Nasdaq Composite (Tech companies)	IXIC
	SP500 Index	SP500

Asian Stock Indices	Nikkei Index (Japan)	N225
Oil prices	West Texas Intermediate (oil, price per barrel)	WTI
	Brent Crude Oil (North sea, price per Barrel)	Brent
Debt and currency	5 year US Treasury Yield	FVX
	Price in Euros of 1 bitcoin	BTC.EUR
	1EUR in USD	EURUSD.X

We are interested in the conditional dependence between two different assets on each other, say X_1 and X_2, given some conditioning variable $\mathbf{Z} = \mathbf{z}$. The degree of conditional dependence is measured by conditional Kendall's tau. Given the thirteen assets, we want the estimates $\hat{\tau}_{X_i,X_j|\mathbf{Z}_k=\mathbf{z}_l^{(k)}}$ for all $i \neq j$ and $k \in \{1,\dots,13\}\backslash\{i.j\}$ and $l = 1,\dots,n$. Here, k corresponds to the 11 different conditioning variables and l represents the number of observations. Note that $\hat{\tau}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} = \hat{\tau}_{X_2,X_1|\mathbf{Z}=\mathbf{z}}$.

In the unconditional case, we have $12 + 11 + \cdots + 2 + 1 = 78$ possible estimates of Kendall's tau. However, each combination of two assets X_1 and X_2 can be conditioned on the other 11 assets. Hence, in the conditional case, the total dataset is described by $78 \cdot 11 = 858$ variables having each 100 observations. This can be summarized in the following data matrix

$$A = \begin{pmatrix} \hat{\tau}_{X_1,X_2|\mathbf{Z}_3=\mathbf{z}_1^{(3)}} & \hat{\tau}_{X_1,X_3|\mathbf{Z}_2=\mathbf{z}_1^{(2)}} & \hat{\tau}_{X_1,X_4|\mathbf{Z}_2=\mathbf{z}_1^{(2)}} & \cdots & \hat{\tau}_{X_{12},X_{13}|\mathbf{Z}_{11}=\mathbf{z}_1^{(11)}} \\ \hat{\tau}_{X_1,X_2|\mathbf{Z}_3=\mathbf{z}_2^{(3)}} & \hat{\tau}_{X_1,X_3|\mathbf{Z}_2=\mathbf{z}_2^{(2)}} & \hat{\tau}_{X_1,X_4|\mathbf{Z}_2=\mathbf{z}_2^{(2)}} & \cdots & \hat{\tau}_{X_{12},X_{13}|\mathbf{Z}_{11}=\mathbf{z}_2^{(11)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\tau}_{X_1,X_2|\mathbf{Z}_3=\mathbf{z}_{100}^{(3)}} & \hat{\tau}_{X_1,X_3|\mathbf{Z}_2=\mathbf{z}_{100}^{(2)}} & \hat{\tau}_{X_1,X_4|\mathbf{Z}_2=\mathbf{z}_{100}^{(2)}} & \cdots & \hat{\tau}_{X_{12},X_{13}|\mathbf{Z}_{11}=\mathbf{z}_{100}^{(11)}} \end{pmatrix} \tag{3.1}$$

In fact, applying Kendall's tau assumes stationary data. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. A stationary process can be recognized by a flat-looking series, without trend and constant variance over time. Moreover, there are no periodic fluctuations (seasonality) [23].

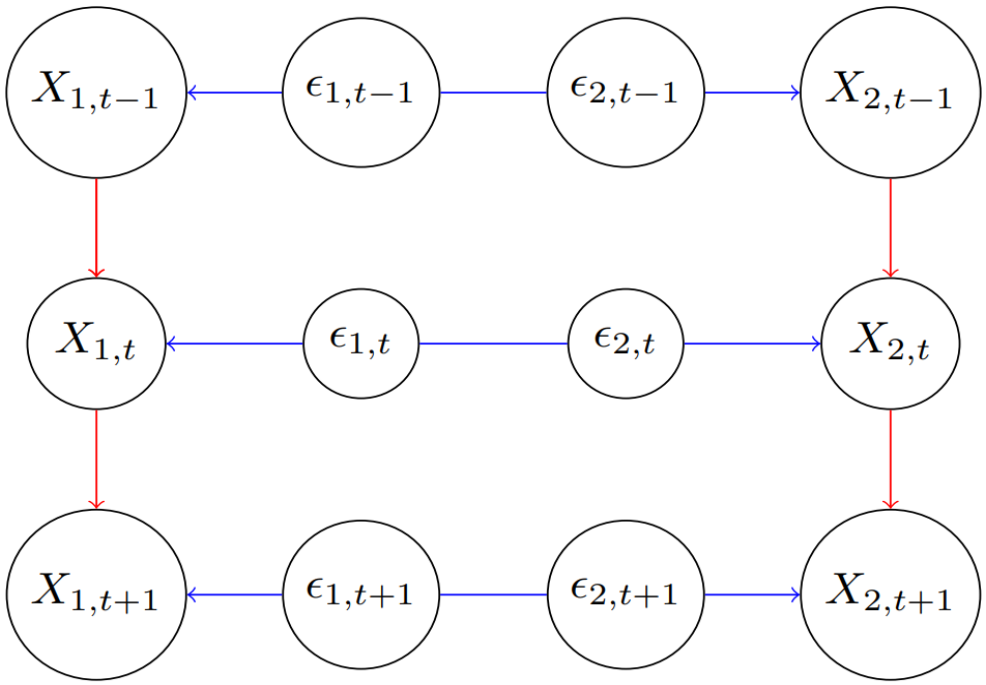
However, since the data of the financial assets is a time series, there is an influence of past observations on the next observations. This may result in, for instance, observations having unequal variance. This is called heteroskedasticity. Heteroskedasticity indicates

the data is non-stationary. Therefore, it could be useful to apply ARMA-GARCH filtering to remove the time dependencies between the observations. This could be done using the following model

$$\begin{cases} \sigma_{1,t} = \gamma_0 + \gamma_1 \epsilon_{1,t-1}^2 + \gamma_2 \sigma_{1,t-1}^2 \\ X_{1,t} = \alpha + \beta X_{1,t-1} + \sigma_{1,t} \epsilon_{1,t}, \end{cases}$$

where α, β are real-valued constants and σ and ϵ are random variables depending on time. The random variable σ represents the volatility of an asset at time t. The random variable ϵ represents the noise and is called innovation. In particular, the innovation is interesting since it is responsible for the unpredictable part of the behaviour of the assets. This process of filtering such that only the, in our case relevant dependence remains, is visualized in Figure 3.1.

Figure 3.1: Interdependence structure of two time series. The red arrows show the time-varying dependence between successive observations in time. The blue arrows show the dependence between two assets. The difference in the size of the circles is negligible.



In our research, we have not used ARMA-GARCH filtering. It is easier to use just the returns of the assets. For example, if the return on a given day of a given asset is 2%, then it seems that on that day the asset price is increased by 2%. However, it is important to understand how the dependence within our data works. We explicitly state that our data is not independent with respect to time.

3.2. Estimation of conditional Kendall's tau

In this section, we will explain the estimators for the conditional Kendall's tau. We formally define the estimators and state some related properties.

Defining estimators for a conditional Kendall's tau is less straightforward as we need to deal with the dependency on $\mathbf{Z}$. Non-parametric estimates of a conditional Kendall's tau have been introduced in the literature only a few years ago ([7], [8], [11], [22]). Therefore, some properties of estimates of a conditional Kendall's tau have been stated under too demanding assumptions. In particular, some assumptions were related to the estimation of conditional margins. However, this is not required because Kendall's tau are based on ranks. Therefore, we directly study non-parametric estimates $\hat{\tau}_{X_1,X_2|\mathbf{Z}=\mathbf{z}}$ without relying on the information about copulas. Therefore, none of the estimators that we will construct depend on estimation of conditional marginal distributions. In other words, we only have to conveniently choose the weights $w_{i,n}$ to obtain an estimator of the conditional Kendall's tau. This is coherent with the fact that conditional Kendall's taus are invariant with respect to conditional marginal distributions [8].

For the construction of non-parametric estimators of the conditional Kendall's tau, we first recall the equivalent expressions of the conditional Kendall's tau in the previous chapter, see (2.16).

$$\begin{aligned}\tau_{X_1,X_2|\mathbf{Z}=\mathbf{z}} :&= \mathbb{P}((X_{1,1}-X_{2,1})(X_{1,2}-X_{2,2}))>0|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z})\\ &\quad-\mathbb{P}((X_{1,1}-X_{2,1})(X_{1,2}-X_{2,2}))<0|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}),\\ &= 4\mathbb{P}(X_{1,1}<X_{2,1},X_{1,2}<X_{2,2}|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z})-1,\\ &= 1-4\mathbb{P}(X_{1,1}<X_{2,1},X_{1,2}>X_{2,2}|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}).\end{aligned}$$

Using the approach from [8], we introduce the following three kernel-based estimators of $\tau_{X_1,X_2|\mathbf{Z}=\mathbf{z}}$ where each line corresponds to the equivalent expressions for the conditional Kendall's tau above

$$\begin{aligned}\hat{\tau}^{(1)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} :&= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\Big(\mathbb{1}\{(X_{i,1}-X_{j,1})(X_{i,2}-X_{j,2})>0\}\\ &\quad-\mathbb{1}\{(X_{i,1}-X_{j,1})(X_{i,2}-X_{j,2})<0\}\Big),\end{aligned} \tag{3.2}$$

$$\hat{\tau}^{(2)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} := 4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\{X_{i,1}<X_{j,1},X_{i,2}<X_{j,2}\}-1, \tag{3.3}$$

$$\hat{\tau}^{(3)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} := 1-4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\{X_{i,1}<X_{j,1},X_{i,2}>X_{j,2}\}, \tag{3.4}$$

where $\mathbb{1}$ denotes the indicator function and $w_{i,n}$ is defined as the Nadaraya-Watson weights as in (2.31) given by

$$w_{i,n}(\mathbf{z}) := \frac{K_h(\mathbf{z} - \mathbf{Z}_i)}{\sum_{i=1}^{n} K_h(\mathbf{z} - \mathbf{Z}_i)},$$

where $K_h(\cdot) := \frac{1}{h^p} K_h(\cdot/h)$, for some kernel K on $\mathbb{R}^p$. The bandwidth sequence $h = h(n)$ is a sequence that converges to zero as $n \to \infty$. Throughout this thesis, we use the Epanechnikov kernel. As mentioned Section 2.4, recall that the choice of kernel does not really matter, only the choice of bandwidth. Of course, there are alternative weights such as local linear and Priestley-Chao, that would lead to different results [8].

The estimators $\hat{\tau}^{(1)}_{1,X_2|\mathbf{Z}=\mathbf{z}}, \hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}}$ and $\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}$ look similar, but they are nevertheless different. In case of $i = j$, the the estimator $\hat{\tau}^{(s)}_{1,X_2|\mathbf{Z}=\mathbf{z}}$ will return values in different subsets of the interval $[-1, 1]$ for $i = 1, 2, 3$. In fact, the estimators will have values in the following intervals

$$\begin{aligned}
\hat{\tau}^{(1)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} &\in [-1 + s_n, 1 + s_n], \\
\hat{\tau}^{(2)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} &\in [-1, 1 + 2s_n], \\
\hat{\tau}^{(3)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} &\in [-1 + 2s_n, 1],
\end{aligned}$$

where s_n denotes the sum of squared weights, $s_n := \sum_{i=1}^{n} w^2_{1,n}(\mathbf{z})$. According to Derumigny and Fermanian [8], there exists almost surely a direct relationship between these estimators given by

$$\hat{\tau}^{(1)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} = \hat{\tau}^{(2)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} + s_n = \hat{\tau}^{(3)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} - s_n. \tag{3.5}$$

An event is said to happen almost surely if it happens with probability one. In other words, the set of possible exceptions may be non-empty, but it has probability zero.

We prefer a rescaled estimator, such that it takes values in the entire interval $[-1, 1]$, over the estimators we have seen [22]. Consequently, we define a rescaled estimator taking values in $[-1, 1]$ by

$$\begin{aligned}
\tilde{\tau}^{(1)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}} &:= \frac{\hat{\tau}^{(1)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}}}{1 - s_n}, && (3.6) \\
&= \frac{\hat{\tau}^{(2)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}}}{1 - s_n} + \frac{s_n}{1 - s_n}, && (3.7) \\
&= \frac{\hat{\tau}^{(3)}_{X_1,X_2|\mathbf{Z}=\mathbf{z}}}{1 - s_n} - \frac{s_n}{1 - s_n}. && (3.8)
\end{aligned}$$

This estimator has been implemented in the programming language R by the package CondCopulas, see [9]. In Appendix A, the reader can find the code used to compute the estimates for the conditional Kendall's tau.

3.2.1. Choice of the conditioning event

The interval on which we condition is constructed using quantiles. A quantile determines how many values in a distribution are above or below a certain limit. In this thesis, we define the conditioning event to be an equidistant sequence starting at q_{10} and ending at q_90 for 100 points. Here q_{10} represents the point after which 10% of the data is distributed. Similarly, q_{90} represents the point after which 90% of the data is distributed. This choice of the interval is robust to outliers. Indeed, one outlier could stretch out the entire interval too much. This procedure is used to construct the conditioning event for the subset of all conditioning variables.

For estimating CKT, on the other hand, for the complete dataset, we use a sequence of equidistant quantiles for 100 points, starting from q_{10} and ending at q_{90}. This construction of the interval takes even more into account the distribution of points in a dataset.

3.2.2. Choice of the bandwidth

As mentioned in Chapter 2, the choice of bandwidth h is crucial for the behaviour of the conditional Kendall's tau. It determines for the most part what the curve will look like. A too small bandwidth will return a bumpy curve which shows a lot of individual peeks. This is called undersmoothing. This is not realistic since it is not likely that the correlation shifts instantly from a large value to a small value. On the other hand, a too large bandwidth will return a flattening curve. In other words, the curve will look like a unimodal distribution and hide all non-unimodal distribution properties. This is called oversmoothing.

The choice of a proper bandwidth is hard. Ideally, it would be preferable to choose the bandwidth h for the conditional Kendall's tau for each combination of assets separately. However, this is tedious and time-consuming work. There are several clever ways, 'rules of thumb', that can be used to determine a sufficiently working bandwidth. For instance, Silverman's rule of thumb for one-dimensional data or the more generalized Scott's rule of thumb for d-dimensional data [2].

$$\hat{h}_{Silverman} = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}}, \tag{3.9}$$

where σ is the standard deviation of the distribution and n is the amount of data points.

$$\hat{\mathbf{h}}_{Scott} = \frac{1}{h^{d+4}}\Sigma^{1/2}, \tag{3.10}$$

where Σ is the covariance matrix and d the number of dimensions.

Although they are simple, they have limitations. Scott's rule requires the data from the normal distribution. The Silverman's rule is more robust but only works well for normal distributed data as well and distributions close to normal.

Therefore, a non-parametric bandwidth selector would be more likely. According to [8], a common way for kernel methods is to choose h as the minimizer of the cross-validation criterion.

$$CV(h) := \frac{2}{n(n-1)} \sum_{i,j=1}^{n} \left(g_k(\mathbf{X}_i, \mathbf{X}_j) - \hat{\tau}^{(h,k)}_{-(i,j),1,2|\mathbf{Z}=(\mathbf{Z}_i+\mathbf{Z}_j)/2} \right)^2 K_h(\mathbf{Z}_i - \mathbf{Z}_j), \tag{3.11}$$

for $\tilde{\tau}_{1,2|\mathbf{Z}=\cdot}$. Here, g_k is defined as the indicator function in (3.2) for $k = 1, 2, 3$.

The easiest way to choose a bandwidth h is based on visual inspection. This is often a well-functioning method. In this research, we have determined the bandwidth manually for each variable. The bandwidth is adjusted after inspecting the plots of the estimates of the conditional Kendall's tau such that all curves look well-behaved. Our choice of bandwidths for each conditioning variable is listed below

Financial Asset (Variable)	Abbreviation	Bandwidth h
France CAC 40 Stock Index	FCHI	$h = 0.0009$
Amsterdam Exchange Index	AEX	$h = 0.001$
German DAX Index	GDAXI	$h = 0.001$
EURO STOXX 50	Eurostoxx	$h = 0.001$
Dow Jones Industrial Average	DJI	$h = 0.0009$
Nasdaq Composite	IXIC	$h = 0.0007$
SP500 Index	SP500	$h = 0.001$
Nikkei Index	N225	$h = 0.001$
Oil prices West Texas Intermediate	WTI	$h = 0.003$
Brent Crude Oil	Brent	$h = 0.003$
Debt and currency 5 year US Treasury Yield	FVX	$h = 0.0025$
Price in Euros of 1 bitcoin	BTC.EUR	$h = 0.005$
1EUR in USD	EURUSD.X	$h = 0.0005$

In Appendix A, we have included our visual inspection analysis for each asset. Here, our choice for bandwidths is supported by several plots.

3.3. Principal Component Analysis (PCA)

As mentioned in Section 3.1, we have a dataset consisting of 858 variables and 100 observations each. It is hard to interpret and visualise this high-dimensional dataset. Therefore, we introduce Principal Component Analysis.

Principal component analysis (PCA) is a multivariate statistical technique that analyses multidimensional datasets with several dependent variables. The goal is to reduce the dimensionality of such datasets, to increase interpretability and, at the same time, to preserve as much as possible of the information contained in the original dataset. To achieve these goals, this technique computes new uncorrelated variables called principal components which are obtained as linear combinations of the original variables multiplied by weights, so-called factor scores. Thus, the first principal component is intuitively defined as the linear combination of observed variables, which has the maximum variance. Then, the second component is orthogonal to the first principal component and captures the second largest variance. This process iterates for all principal components.

Note that, the PCA method is particularly useful when the variables within the dataset are highly correlated. Correlation indicates that there is redundancy in the data. Due to this redundancy, PCA can be used to reduce the original variables into a smaller number of new variables explaining most of the variance in the original variables [19].

3.3.1. Mathematical background

In the following subsection, we provide a more rigorous explanation of PCA which is based on the article of Abdi and Williams (2010) [1] and the report of Nguyen [15]. In short, we recognize four steps in Principal Component Analysis. Now each step is discussed and applied to the dataset containing all estimates of the conditional Kendall's tau of all different assets.

Step 1 - Constructing a data matrix

The data $p \times n$ matrix, denoted by A, consists of p variables representing the columns and n observations representing the rows. In our dataset we have $p = 858$ variables and $n = 100$ observations. Hence, our data matrix is constructed as in (3.1)

Step 2 - Preprocessing the data matrix

Almost always, the data is preprocessed before analysing. First of all, the columns of A will be centered so that the mean of each column is equal to zero. To obtain this we subtract the mean μ_j of each variable per column.

If, in addition, each element of A is divided by the square root of the identity matrix $\sqrt{\mathbf{I}}$, the method is referred to as the covariance PCA. Recall that the covariance is defined as $Cov(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$. Then the matrix $\mathbf{X}^\top\mathbf{X}$ is called the covariance matrix. A covariance matrix is a square matrix giving the covariance between each pair of variables X_i and X_j. Any covariance matrix, denoted as Σ, is symmetric and its main diagonal contains variances, i.e. the covariance of each element with itself.

The variables could have been measured in different units. This may cause issues when analysing and comparing them. Therefore, we standardize the data by dividing each variable by its norm. To obtain this, each column is divided by $\sqrt{\frac{\sum_{i=1}^{p}(\hat{\tau}_{X_i,X_j|\mathbf{Z}=\mathbf{z}_k})^2-\mu_j}{\mathbf{I}}}$. This preserves the correlations but ensures that the total variance equals one. In this case, the analysis is referred to as a correlation PCA. Then, the matrix $\mathbf{X}^\top\mathbf{X}$ is a correlation matrix, having all diagonal entries equal to one and the other entries are given by $\frac{\sum_{i=1}^{p}(\hat{\tau}_{X_i,X_j|\mathbf{Z}=\mathbf{z}_l})(\hat{\tau}_{X_i,X_j|\mathbf{Z}=\mathbf{z}_l})}{\sqrt{\sum_{i=l}^{I}(\hat{\tau}_{X_i,X_j|\mathbf{Z}=\mathbf{z}_l})^2-\mu_j}\sqrt{\sum_{i=1}^{I}(\hat{\tau}_{X_i,X_k|\mathbf{Z}=\mathbf{z}_l})^2-\mu_k}}$. Indeed, the covariance matrix is the correlation matrix of the standardized random variables for $i = 1, \ldots, 100$, $j = 1, \ldots, p$ and $l = 1, \ldots, 100$. Further, note that the correlation matrix is symmetric because the correlation between X_i and X_j is the same as the correlation between X_j and X_i [19].

The data we have imported are returns that are not centered, standardized and stationary. However, this is not necessary because CKTs are directly comparable since they are all between -1 and 1. So there is no need to normalize them to make them comparable since they are already comparable without any need for any transformation. Therefore, we use covariance PCA.

Step 3 - Computing eigenvalues and corresponding eigenvectors of the correlation matrix

In PCA, the components are obtained from the Singular Value Decomposition (SVD) of the $n \times p$ matrix A. Singular Value Decomposition is a factorization of a rectangular matrix into three simpler matrices, say $\mathbf{P}, \Delta, \mathbf{Q}^\top$. In other words, the singular value decomposition is given by $\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^\top$, where

1. $\mathbf{P}$ is an $p \times p$ matrix of the orthonormal eigenvectors of $\mathbf{X}^\top\mathbf{X}$.
2. $\mathbf{Q}$ is the transposed $n \times n$ matrix of the orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^\top$.

3. Δ is a rectangular diagonal $p \times n$ matrix of the singular values. The singular values can be found by taking the square root of the eigenvalues of both $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}\mathbf{X}^\top$.

Every principal component, the new variables, can be written as a linear combination of the original variables. The values of these new variables are called factor scores. Geometrically, the factor scores are the *projections* of the observations onto the principal components. The factor scores, denoted as $\mathbf{F}$, are following from

$$\mathbf{F} = \mathbf{P}\Delta.$$

Here, the matrix $\mathbf{Q}$ gives the coefficients or weights of the linear combinations. In other words, the matrix $\mathbf{Q}$ is the projection matrix that *projects* the original values onto the principal components.

Step 4 - Further analysing of the eigenvalues and corresponding eigenvectors
Eigenvalues indicate the amount of variance explained by each factor. Eigenvectors are the weights that could be used to calculate factor scores. The eigenvector of the correlation matrix corresponding to the largest eigenvalue corresponds to the first principal component. The second principal component corresponds to the eigenvector of the second largest eigenvalue of the correlation matrix. We iterate this process for each principal component.

We need to figure out how many components are necessarily needed. According to Abdi and Williams (2010), this problem is still open, but there are some useful rules. First, we investigate this by plotting the principal components against their amount of variance they explain. In general, the first few principal components account for the vast majority of the variance of the variables.

Another way to find the ideal number of principal components is to choose only the components whose eigenvalues are larger than the average. In general, it seems that these principal components have an eigenvalue larger than one. However, it should be noted that this rule may lead to ignoring important information.

Since the $n \times n$ matrix $\mathbf{Q}$ can be seen as the projection matrix, the matrix with the transformed values is found by multiplying matrix A with $\mathbf{Q}$. By knowing that $\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^\top$ and $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ holds, we obtain

$$\mathbf{X}\mathbf{Q} = \mathbf{P}\Delta\mathbf{Q}\mathbf{Q}^\top\mathbf{Q} = \mathbf{P}\Delta.$$

3.3.2. (Geometric) Interpretation of PCA

Besides the linear algebraic approach we have seen in the previous section, it is useful to provide a geometric approach to interpret the PCA results in Chapter 4.

It is not possible to visualize the data $p \times n$ data matrix consisting of all combinations of conditional Kendall's tau. However, imagine that we are considering three different assets. We can make a 3D scatter plot of these three variables. Now, PCA first computes the best fitting line for all data points. When the direction of the best-fit line is found we can mark the location of each observation along the line. We find the 90-degree projection of each observation onto the line. The distance from the origin to this projected point on the best fitting line is called the score. Each observation gets its own score value which can be both negative or positive. This line is in the direction of maximum variance, meaning that the variance of these scores will be maximal. This is called the first principal component.

Then, we could find the second principal component. By rotating the second principal component's direction vector we ultimately find a direction that gives the greatest variance in the score values when projected on the second principal component. Note that this vector also starts at the origin. Rotating is allowed in all direction if and only if it keeps perpendicular to the first principal component.

Once we have computed the principal components, we plot them against each other. In the figure below an example is given for the entire dataset of 858 data points. The data points are constructed as a linear combination of the first two principal components (PC1 and PC2). In other words, $\hat{\tau}_{X_i,X_j|\mathbf{Z}=\mathbf{z}} = v_i\phi_1(\mathbf{z}) + v_j\phi_2(\mathbf{z})$, where (v_i, v_j) are the score values and $\phi_1(\mathbf{z}), \phi_2(\mathbf{z})$ correspond to the first two principal components. The principal components could be seen as a curve for all $\mathbf{Z} = \mathbf{z}$. PC1 and PC2 are the most relevant patterns of the conditional Kendall's tau for all different combinations of assets. A large value of v means that a principal component contributes a lot to the shape of the conditional Kendall's tau for specific assets, either in a positive or negative manner. A small value of v, on the other hand, means that a principal component contributes only to a small extent.

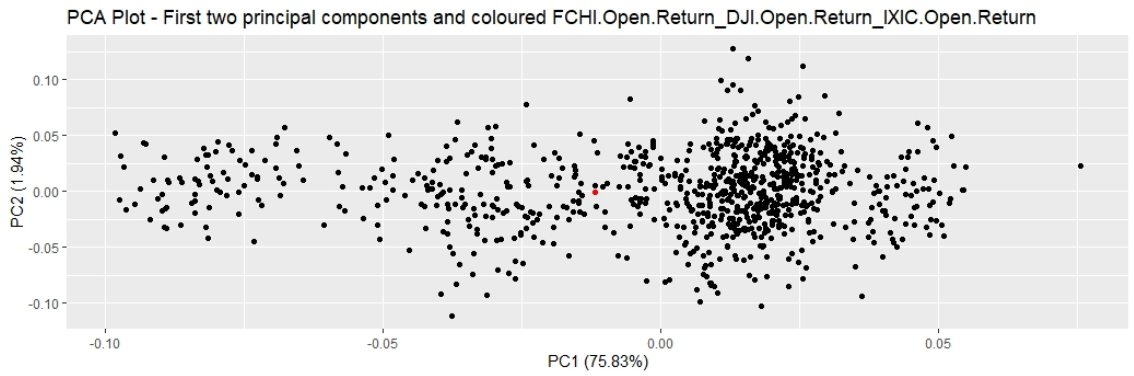


Figure 3.2: Score plot of PC1 and PC2 for the enitre dataset. The percentage next to the axis titles is the amount of variance explained by the corresponding principal component. The red dot corresponds to the conditional Kendall's tau of FCHI and DJI given IXIC.

From Figure 3.2 it becomes clear that $v_1 = v_{1,FCHI,IXIC|\mathbf{Z}=\mathbf{z}} = -0.012$ and $v_2 = v_{2,FCHI,IXIC|\mathbf{Z}=\mathbf{z}} = 0$. This results in a CKT curve that corresponds to $\hat{\tau}_{FCHI,IXIC|\mathbf{Z}=\mathbf{z}} \simeq -0.12\phi_1(\mathbf{z})$. Note that estimated CKT is not exactly equal to $-0.12\phi_1(\mathbf{z})$. The CKT is a linear combination of all PCs, not just PC1 and PC2.

Lastly, it is important to have a look at three concepts that help us interpret the results of our Principal Component Analysis. These three concepts are explained below and will be used in the next chapter.

- Explained variance of each principal component
- Quality of representation of all variables per component
- Clustering

Explained Variance of the principal components

The explained variance is the ratio between variance that is attributed by each component and the total variance. The percentage of variance explained by a component, tells the percentage of information in the dataset that this component preserves. If the number is 60%, it means that they preserve 60% of the information in the dataset. Hence, the lower the total variance is, the higher is the information loss. Interestingly, the fact is that PCA never tells if the information lost was relevant or irrelevant.

We will use explained variance of each component to evaluate the usefulness of each component and to choose how many components to use. An example is provided below

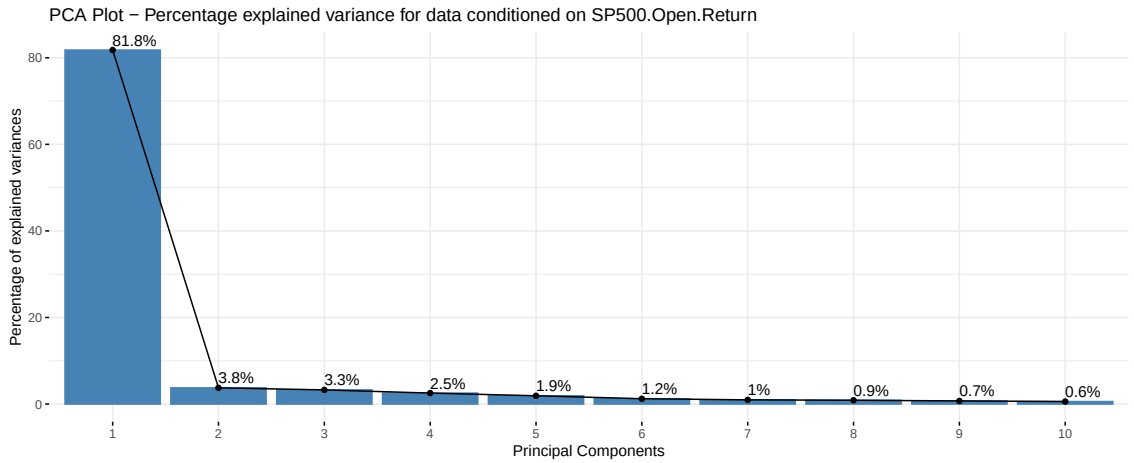


Figure 3.3: The explained variance for the subset of conditioning variable SP500. PC1 captures most of the variance, 81.8%. PC2, PC3, PC4 are around 3%-4% after which the explained variance in percentage decreases fast.

Quality of representation using cos2

The quality of representation of the variables is estimated using cos2. It's calculated as the squared factor scores, which are the coordinates of the PCA plot. A high cos2 indicates a good representation of the variable on the principal component. A low cos2, on the other hand, indicates that the variable is not perfectly represented by the PCs. If a variable is perfectly represented by, for instance, only two principal components. Then, the sum of the cos2 on these two PCs is equal to one. Note that for some of the variables, more than 2 components might be required to perfectly represent the data [19].

Clustering

We will use clustering to find meaningful patterns in the results obtained by PCA. The clusters are computed using the package *Factoextra* in R using the command *fviz_cluster* in combination with the function *pam*. In Appendix C, the code is provided that we use to compute and to analyse the clustered PCA results.

4

Results Simulation Study

In this chapter, we present an application of the methods to the financial data discussed in Chapter 3. In Section 4.1, we show the results of both the estimation of CKT as the PCA analysis applied to the entire dataset. We will perform a deep-dive on the data of the subset of one specific conditioning variable in Section 4.2. We have chosen the asset Dow Jones Index (DJI). Then, we elaborate on the clustering analysis. In Section 4.3 we cluster with respect to the variables X_1 and X_2. Whereas in Section 4.4, we cluster with respect to the conditioning variable $\mathbf{Z}$.

4.1. Results Complete Dataset

In Figure 4.1, all different curves of conditional Kendall's tau are plotted together. The more overlap the curves have, the darker the colour is. It allows us to see patterns across the set of CKTs. For instance, we see heavier fluctuating values for values of the conditional Kendall's tau in the tails compared to in the middle. This indicates a stronger correlation, either positive or negative, in the tails. In the middle, around q_{50}, it seems that there is almost no correlation for most combinations of assets. However, for a smaller group of combinations of assets, there is certainly a positive correlation in the middle. These are the curves fluctuating between values 0.5 and 0.75. Moreover, the curves below zero seem to have a parabolic trend. For positive valued curves, the parabolic shape seems to be mirrored on the x-axis. Besides, quite a few positive valued curves do not seem to have a parabolic shape. These curves appear to have a larger positive correlation.

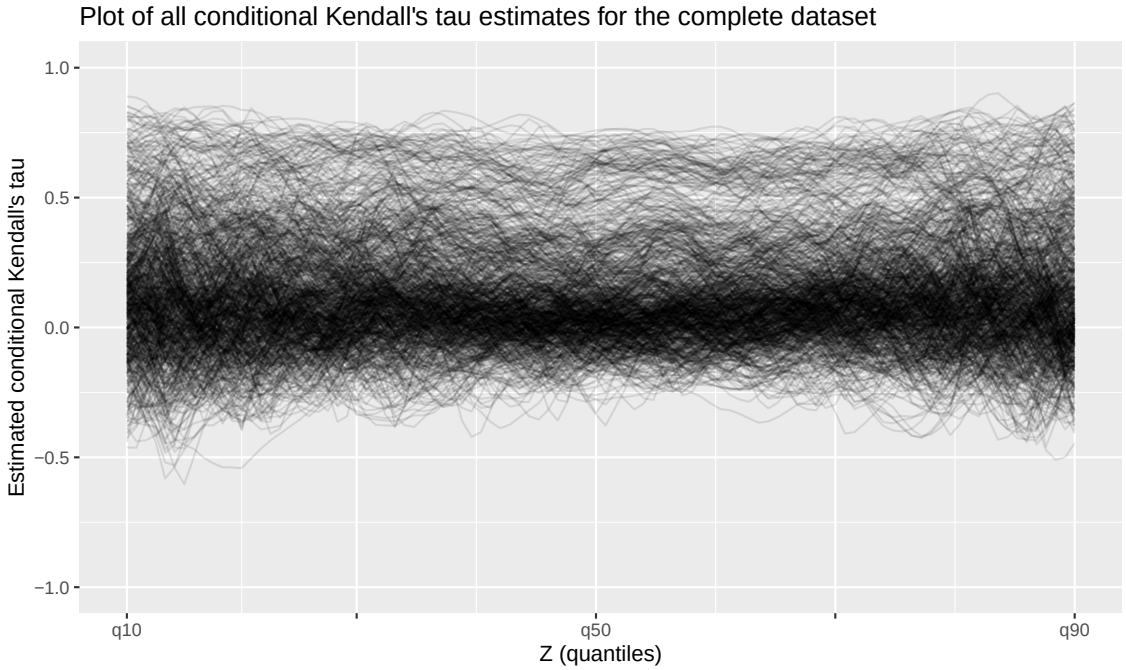


Figure 4.1: All estimates of conditional Kendall's tau are plotted together for all financial variables. On the y-axis, we have the values of the coefficient which are between $[-1, 1]$. On the x-axis the conditioning event in quantiles.

Minimum and maximum values

The minimum and maximum value of the curve of CKT are shown in Figure 4.2. There seems to be a clear relation: the larger the minimum value of the CKT, the larger its maximum value.

The black lines are defined as $y = x$ and $y = -x$. We have included these lines to illustrate the differences in size of minimum and maximum values. This could tell us more about whether curves are flat or fluctuating to some degree. If the minimum and maximum value are equal, it would have been on the line $y = x$. If a value has opposite sign, it will be on the line $y = -x$. The latter is the case for some combinations of variables. We see that the closer the points are to the line $y = x$, the more flattening the curve is. On the other hand, the further away the points are, the more heavily fluctuating the curve could be.

According to Figure 4.2, the points are properly distributed diagonally. There are no points distributed differently. This indicates the choice bandwidth is good. Indeed, when a bandwidth is chosen too small, it would have been spread across the top left of the graph. When a bandwidth is chosen too large, on the other hand, the points would have been closer to or on the line $y = x$.

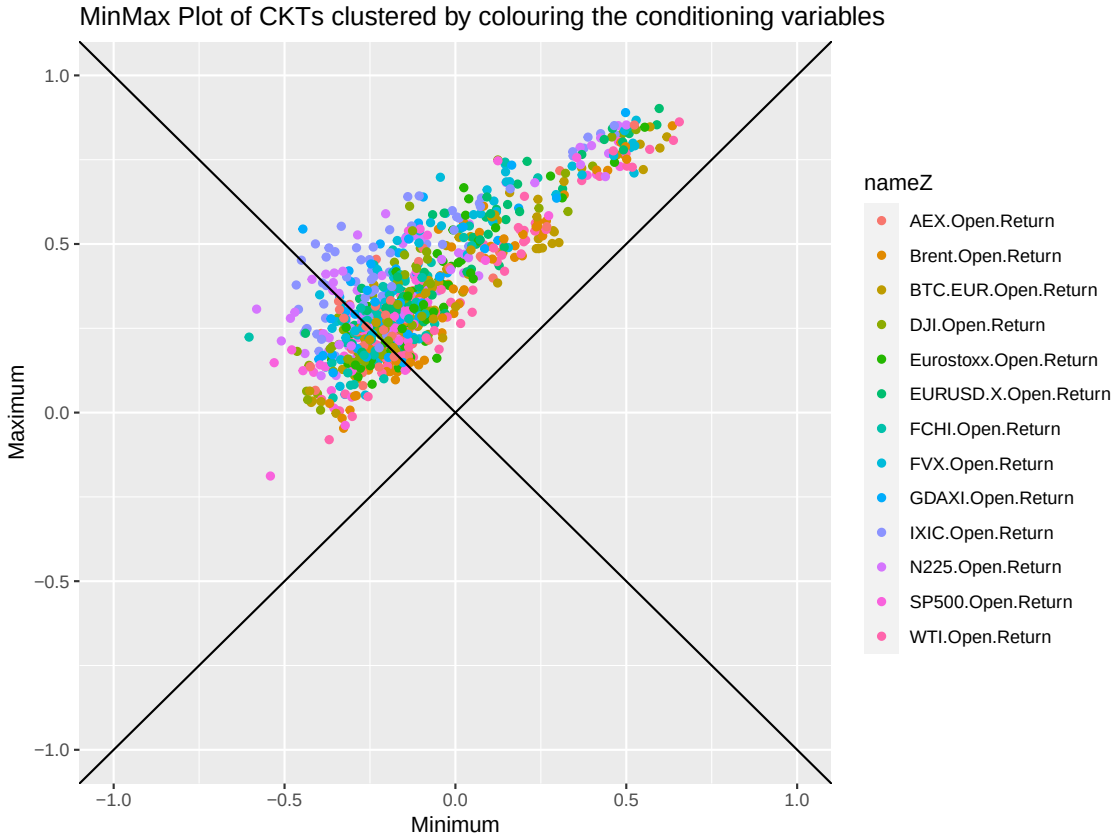


Figure 4.2: Plot of the minimum and maximum value of each possible combination of variables in the complete dataset.

In Figure 4.2 there could be recognized two clusters. On the one hand, we see most of the points have a negative minimum value around the interval $[-0.5, 0.0]$ and a positive maximum value around the interval $[-0.5, 0.5]$. Intuitively, this tells us that the average CKT will be around zero. Figure 4.1 reinforces this intuition since most of curves of the CKTs seem to be constant around zero. On the other hand, some points have positive minimum value around 0.5 and a large positive maximum value around the interval $[0.5, 1]$. Now, the average of CKTs will be positive. Again, Figure 4.1 reinforces this intuition since there are quite some curves of CKTs that take on larger positive values.

In Figure 4.2, we grouped the points based on all conditioning variables. It appears that the CKTs behave according to the same pattern. This indicates there seems no clear pattern per conditioning variable visible.

4.1.1. Results PCA for the complete dataset

In the next section, we will analyse the PCA results on the complete financial dataset. We will first research its principal components (PC). Then, we will asses the quality of

representation of the variables by PC1 and PC2. Lastly, the results of clustering of the score plot will be discussed. This will be the basis for Sections 4.2 and 4.3.

Principal components

As mentioned before, the curve of a CKT is constructed as linear combination of the principal components. So the shape that the components have, may tell us the shape of the curve of a CKT.

In Figure 4.3 the first six principal components are plotted which seem to be smooth and stable. It becomes clear that especially the first five components are relevant since component number six (plot F) does not show a consistent trend. The influence of PC1 on a CKT is that it will contribute to less correlation in the tails compared to the middle. Note that we are looking at a negative interval on the y-axis. However, the scores (v_i, v_j) could also be negative, resulting in a curve that is mirrored on the x-axis. Further, PC2 has quite a large positive contribution for the left tail, after which the curve ultimately becomes constant around zero. This means that for larger values of $\mathbf{Z}$ there is much less influence of PC2 on the CKTs.

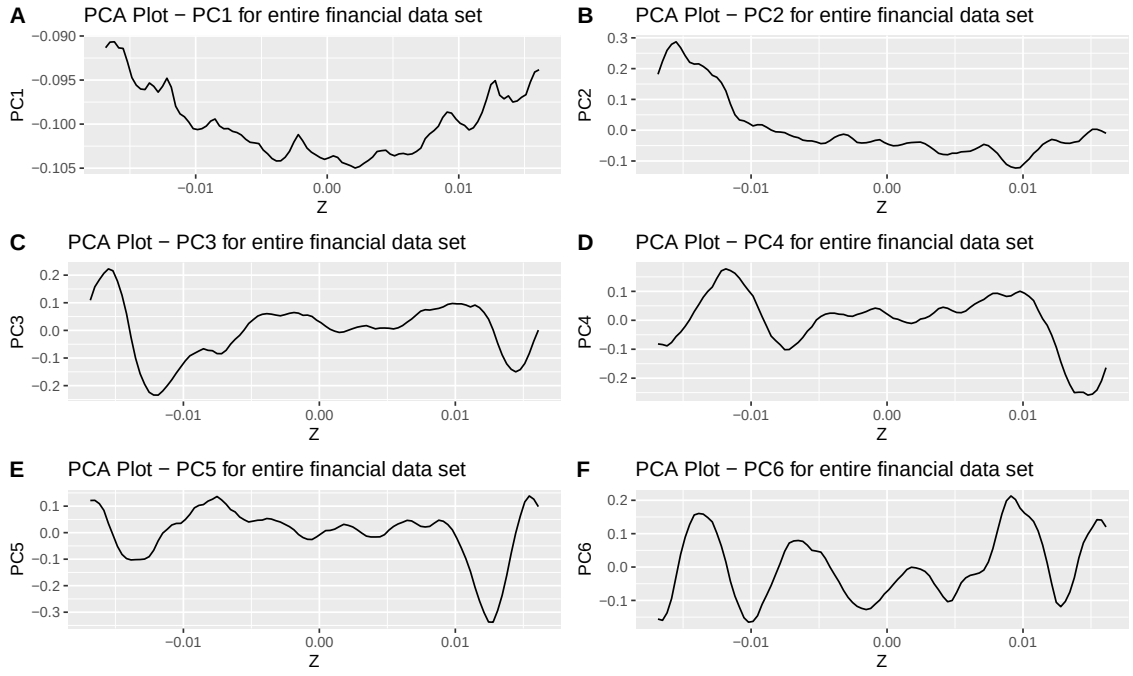


Figure 4.3: The curves of the first six principal components for the complete dataset.

Note that the curves of the CKTs in 4.1 are linear combinations of all principal components. Since we are looking at only the first two principal components, it gives an indication of the structure of the CKTs. Therefore, it does not allows us to explain the exact curves of CKTs in Figure 4.1.

Note that the average correlation for PC2, PC3 and PC4 are approximately around zero. Since principal components are uncorrelated, just PC1 determines the average of the correlation. PC2, PC3, PC4 (and the other components until PC100) still contribute to the shape of the curve of a CKT.

Furthermore, it is important to know how much variance of the original dataset is captured by the principal components. Figure 4.4 tells us the percentage of explained variance of each component. Clearly, PC1 captures most of the variance. This indicates that most of the observations are distributed across the first dimension. Often a percentage of 80% of the total variance works sufficiently, i.e. we will choose component one and two.

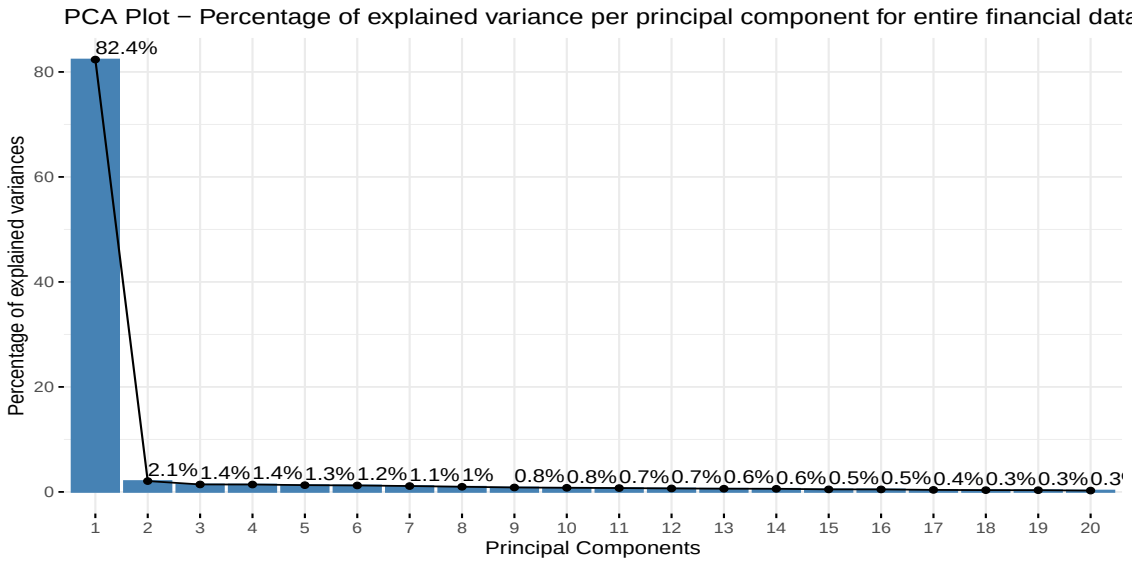


Figure 4.4: Percentage of explained variance of the first twenty principal components of the complete dataset.

Quality of the representation

In Figure 4.5 the quality of the representation of the variables by PC1 and PC2 are shown. In the tails we see high cos2 values which indicates a very good representation of the variables on the first two principal component. In the middle of the plot, around the origin, we see a low cos2. This indicates that these variables are not perfectly represented by the PCs. It is noticeable that these points seem to have a vertical pattern. This is coherent with the fact that PC1 explains most of the variance.

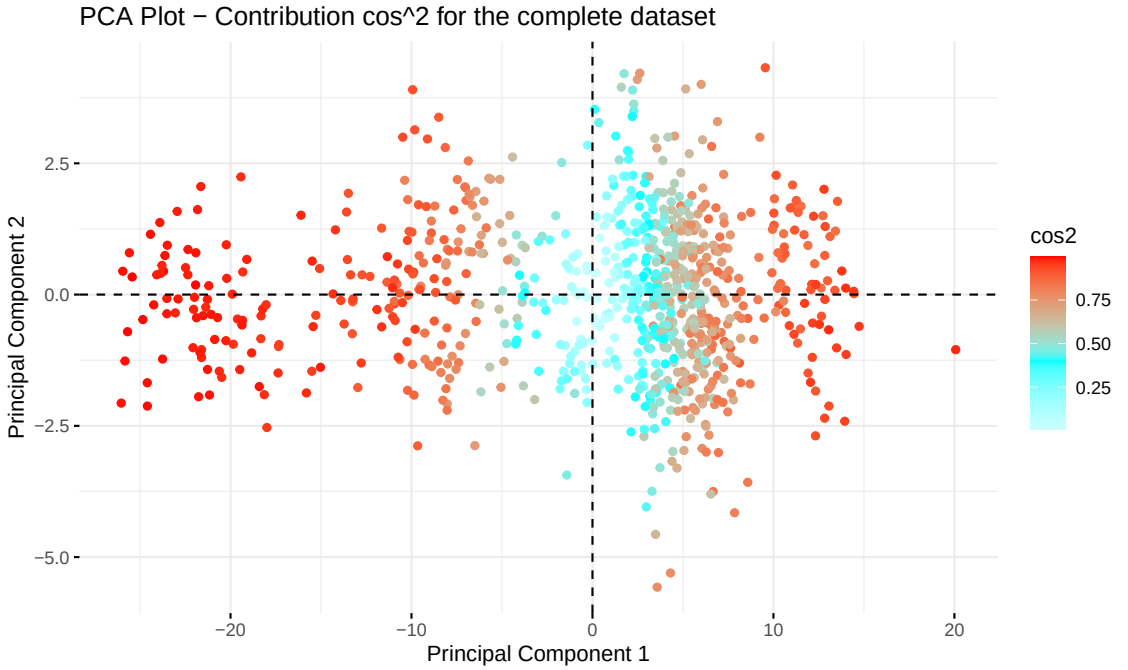


Figure 4.5: Representation of the variables by PC1 and PC2 for the complete dataset. The larger the value for cos2, the better the representation is.

4.1.2. Clustering

To what extent a curve of an estimated CKT will look as its principal components, is determined by the factor scores. In Figure 4.6 the plots of all factor scores for PC1 and PC2 are shown grouped by the conditioning variable.

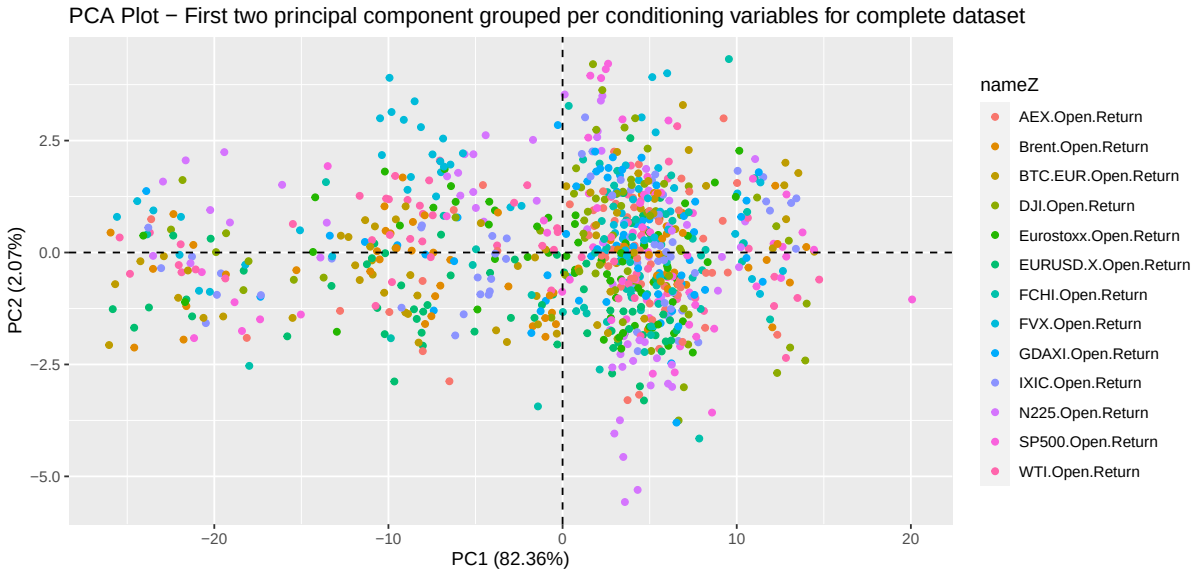


Figure 4.6: Score plot for the complete dataset coloured for its conditioning variable.

It appears there is no clear trend for the scores when clustering on the conditioning variable. Further, there seems to be a high density of factor scores in the rectangle $[0, 10] \times [-2.5, 2.5]$. This indicates that a lot of CKTs will probably look like a linear combination of PC1 and PC2 multiplied by these scores. Remember that the representation of these variables by PC1 and PC2 is not very good, see Figure 4.5.

Note that the y-axes of the plots in Figure 4.3 consist of intervals of small values. Whenever, scores take on small values, this results in small contributions of PC1 and PC2 to the curves of the CKT. This corresponds with our findings in Figure 4.1, where most curves fluctuate around zero for all $\mathbf{Z} = \mathbf{z}$. Moreover, the estimated CKTs have slightly larger positive values in the left tail. This corresponds to the presence of PC2 which has a single peak for small values of $\mathbf{Z}$.

Furthermore, there seems to be four clusters appearing in the complete dataset. In Figure 4.7 these four clusters are plotted. The clusters are vertically composed along the x-axis (PC1 axis). The differences in the clusters in Figure 4.7 are caused by the large influence of PC1 on the variables. Recall that PCs are ranked by how much variance of the original dataset they describe. PC1 reveals the most variation, while PC2 reveals the second most variation. Therefore, differences among clusters along PC1 axis are actually larger than along PC2 axis.

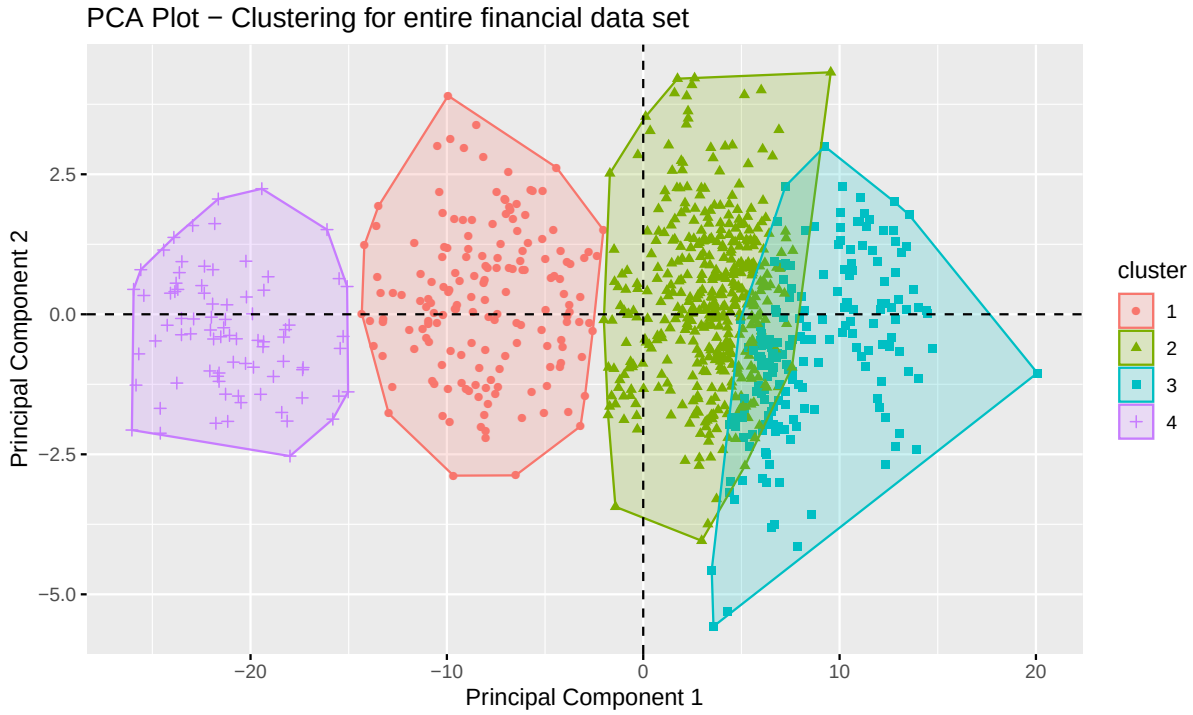


Figure 4.7: Score plot of PC1 and PC2 clustered in four groups applied to the entire financial dataset.

Cluster 1 includes the points that have the largest negative scores for PC1. For PC2 the points both have positive and negative scores. It seems that the scores for PC2 are quite evenly distributed in this cluster. That is, there are not only positive or negative scores corresponding to PC2. Cluster 4 has similar properties, but for smaller negative scores for PC1. According to Figure 4.3 PC1 is a curve with only negative values, indicating a negative correlation. Note that negative factor scores result in mirroring the PC1 on the x-axis, resulting in a positive correlation. This is coherent with the curves of the CKTs having positive correlation in Figure 4.1.

Cluster 2 consists of the points having small values for PC1. Since PC1 determines the average correlation, this means the corresponding CKTs look like a constant line close to zero. This indicates that the CKTs in this cluster do not correlate. Lastly, cluster 3 consists of points having a negative correlation. Positive scores for PC1 indicates a negative correlation since PC1 a curve that takes only negative values. However, note that cluster 3 overlaps cluster 2 on the left side. This means cluster 3 consists also of points that correspond to curves that seem to fluctuate around zero rather than around negative values.

The clustered scores can be associated directly with the plot of all CKTs in Figure 4.1. This gives us Figure 4.8.

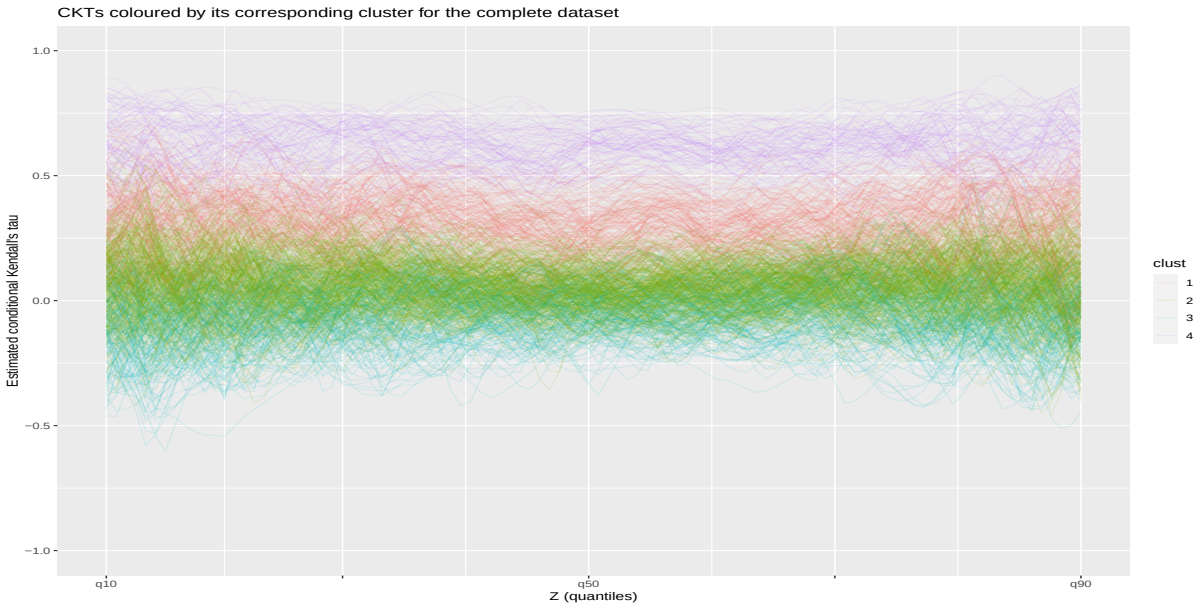


Figure 4.8: All estimates of conditional Kendall's tau grouped using four clusters.

Indeed, the previously described results per cluster correspond with Figure 4.8. Now, it is clear that all CKTs seem to be constant fluctuating around its average value. Recall

that this average value of the CKTs is determined just by PC1. Only in the tails, the correlation seems a little bit larger. For cluster 1,2 and 4 this means a relatively larger positive correlation. Whereas for cluster 3 this means a relatively larger negative correlation. This is coherent with the parabolic shape of PC1.

Next, it is interesting to look at what combinations of assets are in what clusters and how they interact with each other. In order to examine this, we have created a connected multigraph $G = (V, E)$ which explains the relation between the assets per cluster. Here, the set of vertices (V) consists of the financial assets in a specific cluster. The set of edges (E) consists of the connections between any two assets. A connection represents a score of an individual for one conditioning variable. For example, if GDAXI and SP500 are linked to each other, it means that this cluster contains the point of GDAXI and SP500 for some conditioning variable. The information of the conditioning variable is contained in the edge and it is not explicitly mentioned which conditioning variable is considered. However, the conditioning variable is examined using the code provided in Appendix B.6 Clustering. The number of edges is the the number of combinations between the assets for any conditioning variable. For example, when the vertices of GDAXI and SP500 are connected by five edges, this means there are five scores of GDAXI and SP500 in a cluster for five different conditioning variables. Note that for each combination of two assets there are 11 edges in total.

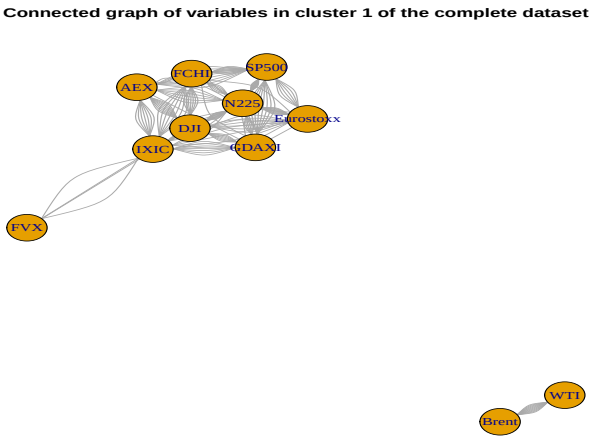


Figure 4.9: Multigraph $G_1 = (V, E)$ representing the scores and their underlying relation in cluster 1 applied to the complete dataset. The assets EURUSD and BTC are not present in this cluster.

The first cluster corresponds to small to moderate positive correlation. In Figure 4.9 we recognize two different multigraphs separated from each other. The large multigraph,

on the top side of Figure 4.9, consists of 9 assets connected by a lot of edges. Note that N225 and Eurstoxx are connected by 11 edges, i.e. for any conditioning variable the conditional dependence between N225 and Eurostoxx is moderately positive. The smaller second multigraph consists only of Brent and WTI having eleven edges with each other. This means all combinations of Brent and WTI given any conditioning variable are clustered together. This indicates that there is a positive conditional dependence between Brent and WTI which is reinforced by Figure 4.10

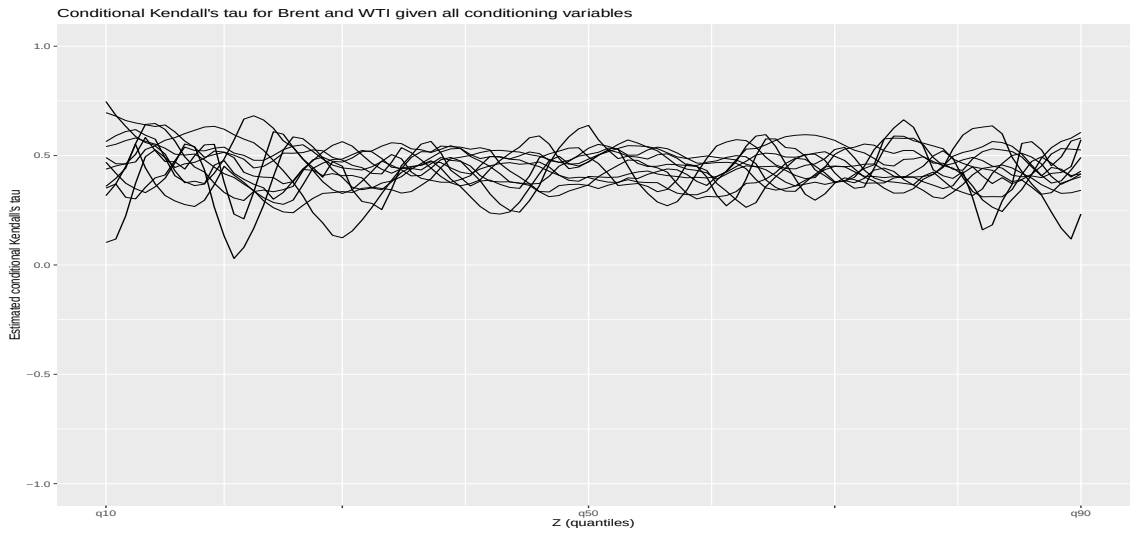


Figure 4.10: Conditional Kendall's tau for Brent and WTI for all conditioning variables. The conditional dependence between Brent and WTI is moderatly strong. The CKTs fluctuate between 0.2 and 0.7 and it is mostly around 0.45

Cluster 2 corresponds to the data that is not to almost not correlated. In Figure 4.11 the relation between the assets in cluster 2 can be seen. All assets occur at least once in this multigraph. We see in Figure 4.11 that FVX is connected to all other assets for almost all conditioning variables. This means for FVX that there is almost no correlation between the other assets and itself. The remaining connections of between FVX and other assets are in cluster 3. Similarly, all assets connected to BTC lie for almost all conditioning variables in cluster 2. The other connections are in cluster 3. However, the scores are still around the interval $[4.5, 8]$ for which there is almost full overlap with cluster 2. Only the connection with EURUSD is in cluster 3, without overlap of cluster 2 (around $x = 10$). This indicates a slightly negative correlation between BTC and EURUSD for most conditioning variables. This also applies to Brent with the exception of its connection with WTI. We have seen in Figure 4.9 that for all conditioning variables the CKTs of Brent and WTI are located entirely in cluster 1 with a moderate positive conditional dependence.

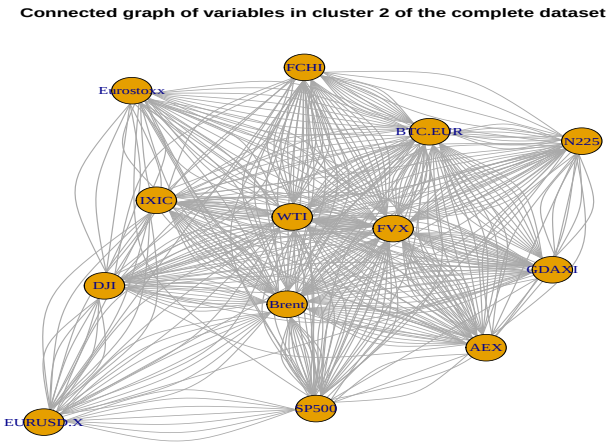


Figure 4.11: Multigraph $G_2 = (V, E)$ representing the scores and their underlying relation in cluster 2 applied to the complete dataset.

Next, in Figure 4.12 we see the scores in cluster 3 visualised as a multigraph. The scores in cluster 3 correspond to data that is not to slightly negatively correlated.

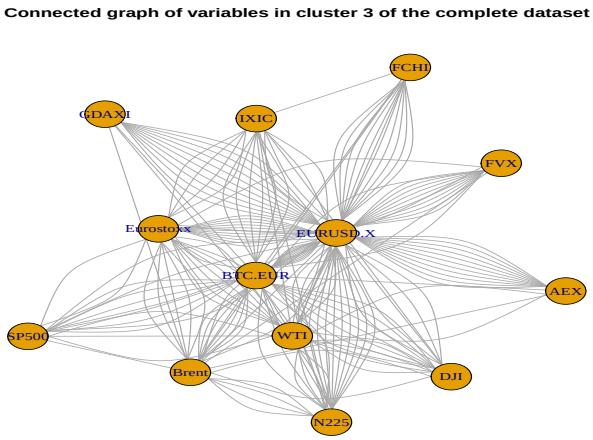


Figure 4.12: Multigraph $G_3 = (V, E)$ representing the scores and their underlying relation in cluster 3 applied to the complete dataset.

In fact, EURUSD is connected to the European assets (AEX, Eurostoxx, FCHI and GDAXI) for almost all conditioning variables. Also for N225, FVX and BTC almost all connections with EURUSD are in this cluster. The other connections are present to a lesser extent in both cluster 2 and 3. This indicates that EURUSD has no to small negative correlation with all assets. In particular, a more negative correlation when combined with European assets. This is because the scores of of the individuals of

EURUSD with a European asset are located more on the right of the x-axis in Figure 4.7. Recall that larger positive scores result in larger negative correlation.

Lastly, cluster 4 consists of the individuals that have the most positively correlated CKTs. In Figure 4.13 we recognize two groups. The multigraph in the upper left corner consists of all European assets which are connected to each other. The European assets are linked with each other for all conditional variables except for the European assets itself. For example, AEX misses two links with Eurostoxx, which are exactly the scores for AEX and Eurostoxx conditioned on GDAXI and FCHI. This indicates that the European have a strong positive correlation between each other given other assets outside of Europe.

The multigraph in the lower right corner, on the other hand, consists of only US assets (DJI, IXIC, SP500). SP500 with IXIC, and SP500 with DJI are in this cluster for all conditioning variables. The scores of DJI and IXIC are not for all conditioning variables this cluster. In fact, the scores of DJI and IXIC for all conditioning variables are spread out over both cluster 1 and 4. It is surprising that only the score for DJI and IXIC given SP500 lies in cluster 4. In fact, it has the largest positive value for the x-axis, around $x = 20$. This indicates the largest average negative correlation.

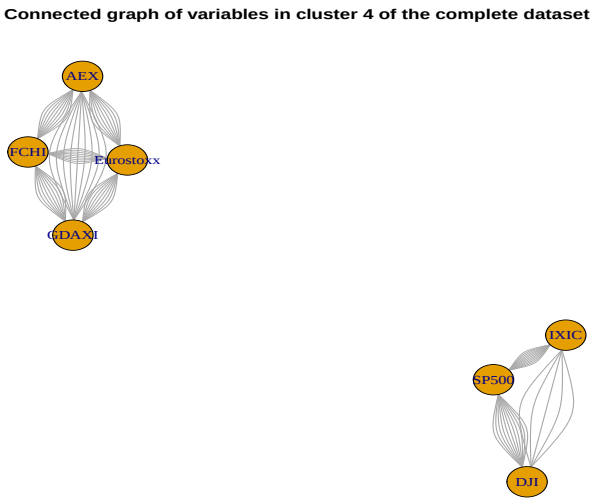


Figure 4.13: Multigraph $G_4 = (V, E)$ representing the scores and their underlying relation in cluster 4 applied to the complete dataset.

The analysis on the multigraphs is done by using the code in Appendix B. With the code we created the table *dataclust1* with all information on the scores for any combination of assets.

4.2. Results deep-dive on one single asset: Dow Jones Index (DJI)

In this section, we will perform a deep-dive on one single conditioning asset: Dow Jones Index (DJI). We examine the subset of the conditioning variable DJI using the same analysis done in Section 4.1.

In Figure 4.14 we see the estimated conditional Kendall's tau for all combinations of variables conditioned on DJI. Most of the curves of the CKTs seem to be quite constant around zero. Only in the tails, there are larger fluctuations which may be both positive or negative. A few of the CKTs have a stronger positive correlation. These are the curves around $\hat{\tau} = 0.5$ which seem to be quite constant, also in the tails. The pattern of the estimated CKTs conditioned on DJI do not differ much from to the pattern of the CKTs in the entire dataset from Figure 4.1.

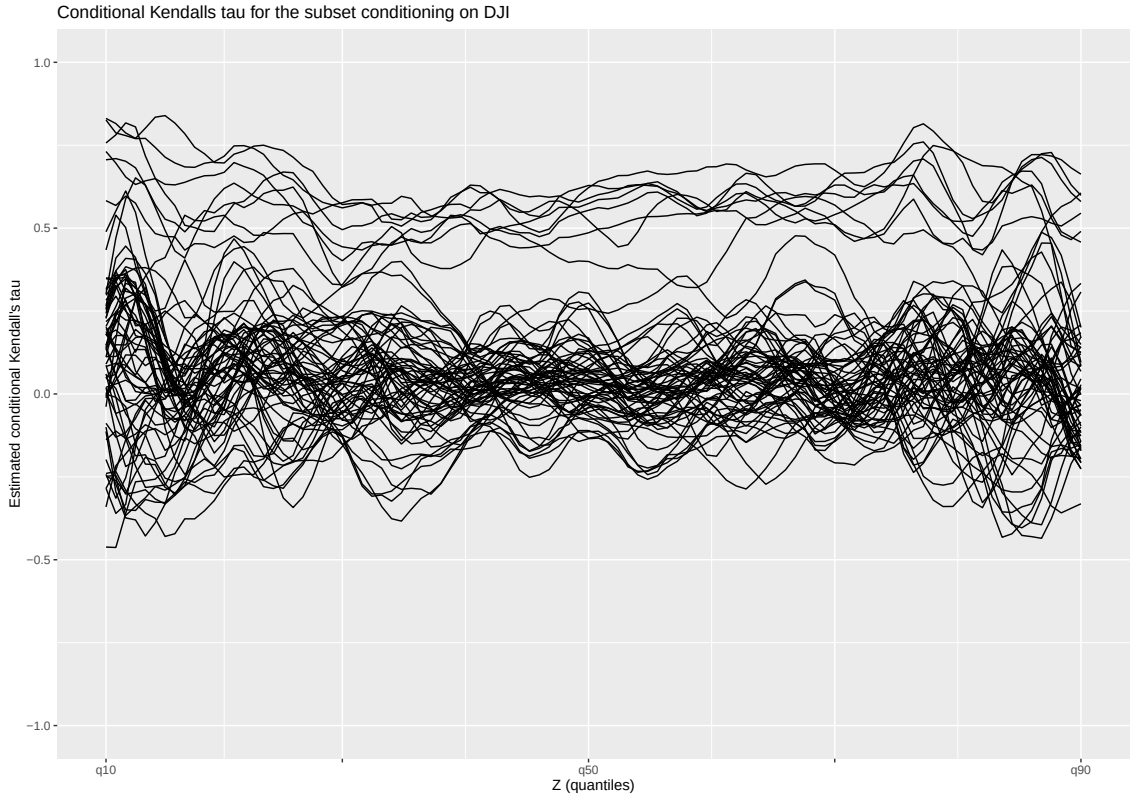


Figure 4.14: Estimated conditional Kendall's tau for all combinations of assets conditioned on DJI.

Note that all curves seem smooth and stable. They do not fluctuate too heavily and are not too flat. This indicates we have used a good choice of bandwidth for DJI.

Minimum and maximum values

There appear to be two clusters for the points in Figure 4.15. This is similar to the complete dataset in Figure 4.2. On the one hand, we have a cluster consisting of a negative minimum and a positive maximum in the interval $[0, 0.5]$. These points correspond to a curve of a CKT that fluctuates around zero. Intuitively, the average of this CKT will be around zero and implies that there is almost no correlation. The other cluster consists of points with both a positive minimum and maximum value. This time the maximum value is larger, namely it is in the interval $[0.5, 1]$. These points correspond to a curve of a CKT that fluctuates around a positive value. Intuitively, this curve has a positive valued average which indicates a positive correlation. Indeed, these two clusters of minimum and maximum values strongly correspond to the two groups of curves in Figure 4.14.

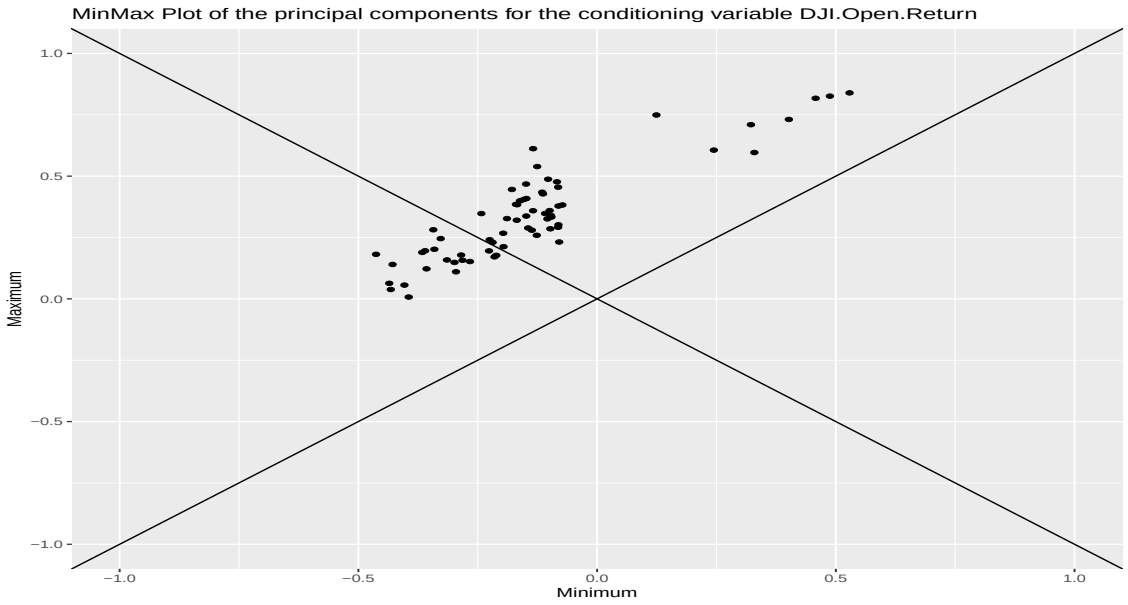


Figure 4.15: Minimum and maximum value of each curve of a CKT. On the x-axis the minimum value is given and on the y-axis the maximum is given.

4.2.1. Results PCA for conditioning on DJI

Principal components

In Figure 4.16 the first four principal components are plotted. First of all, note that these principal components are different from the ones we have seen in Section 4.1. The reason for this is that PCA is now applied to a subset of the variables conditioned on DJI instead of to the entire dataset. This results in different PCs and scores. In fact, a score still corresponds to the same curve of the CKTs but as a new linear combination between different principal components multiplied by different scores.

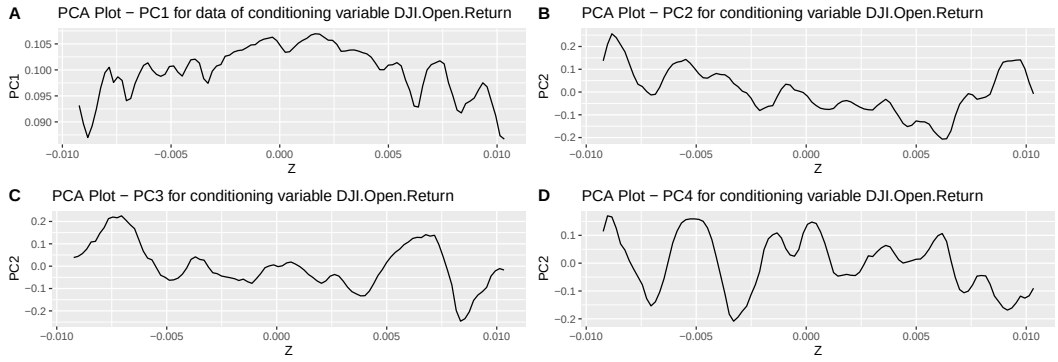


Figure 4.16: The curves of the first four principal components for the subset of conditioning variable DJI.

In Figure 4.14 we see the first three principal components describe a trend. PC4 shows too little consistent behaviour and will not be very useful for interpreting the data. Here, PC1 is very similar to the one of the entire dataset but mirrored on the x-axis. Because of the small interval on the y-axis, it seems that the PC1 from Figure 4.16 is a nearly constant line for a positive value. Further, PC2 starts for positive values and gradually decreases. In the right tail, PC2 rises rapidly from negative values to approximately zero. PC3 seems similar to PC2 but it is slightly different in the right tail.

Note that the average correlation for PC2, PC3 and PC4 are approximately around zero. Since principal components are uncorrelated, just PC1 determines the average correlation in the curve of a CKT. Of course, PC2, PC3, PC4 (and the other components until PC100) contribute to the shape of the curve of a CKT.

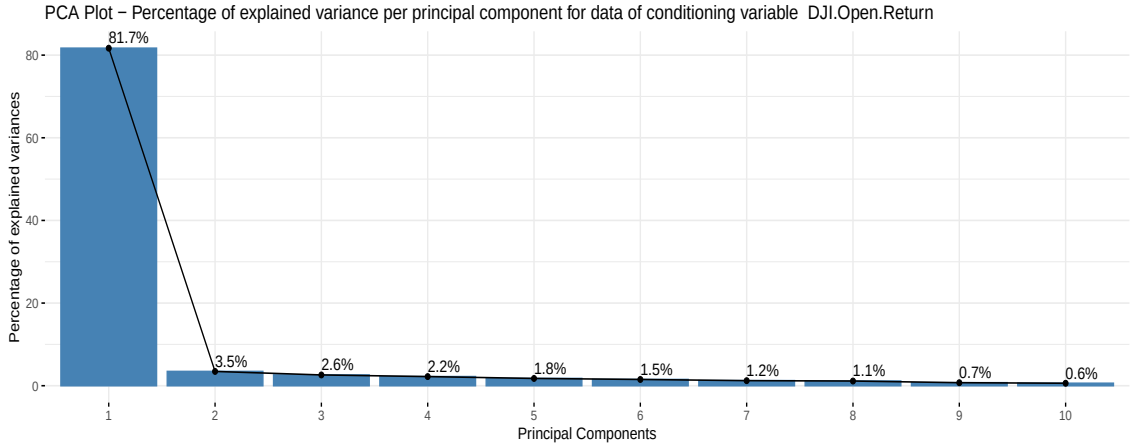


Figure 4.17: Percentage of explained variance of the first ten principal components of the subset of conditioning variable DJI.

It is important to know how much variance of the original dataset is explained by the

principal components. Figure 4.17 tells us the percentage of explained variance of each component. Clearly, PC1 maximizes the variance most. This indicates that most of the observations are distributed across the first dimension. Often a percentage of 80% of the total variance works sufficiently. Here, choosing the first two components explains 85.2% of the variance.

Quality of the representation

In Figure 4.18 the quality of the representation of the variables by PC1 and PC2 are shown. In the tails, there is a high value for cos2 which indicates a very good representation of the variables on the first two principal component. In the middle of the plot, around the origin, we see a low value for cos2. This indicates that these variables are not really well represented by the PCs. It is noticeable that these points seem to have a vertical pattern. This is because of the stronger influence of the variables on PC1.

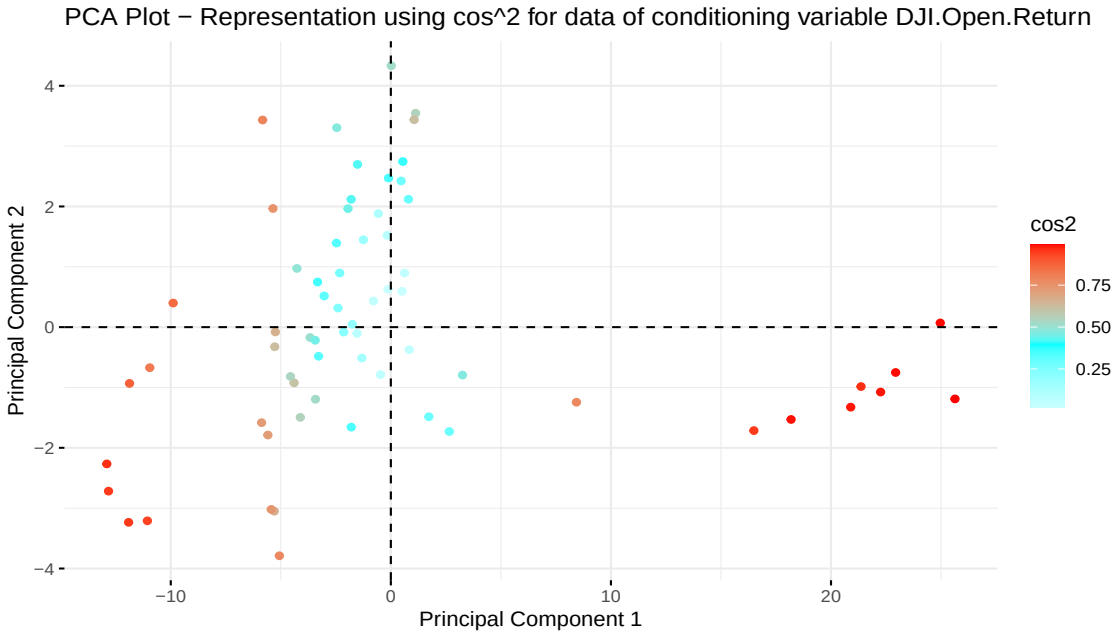


Figure 4.18: Representation of the variables by PC1 and PC2 for the subset of conditioning variable DJI. The larger the value for cos2, the better the representation is.

4.2.2. Clustering

In Figure 4.19, the scores of PC1 and PC2 of each combination conditioned on DJI are plotted. The names of the combination of assets are included. Since the principal components and scores are different for this subset of DJI, the distribution of points is differently compared to the score plot of the complete dataset in Figure 4.6.

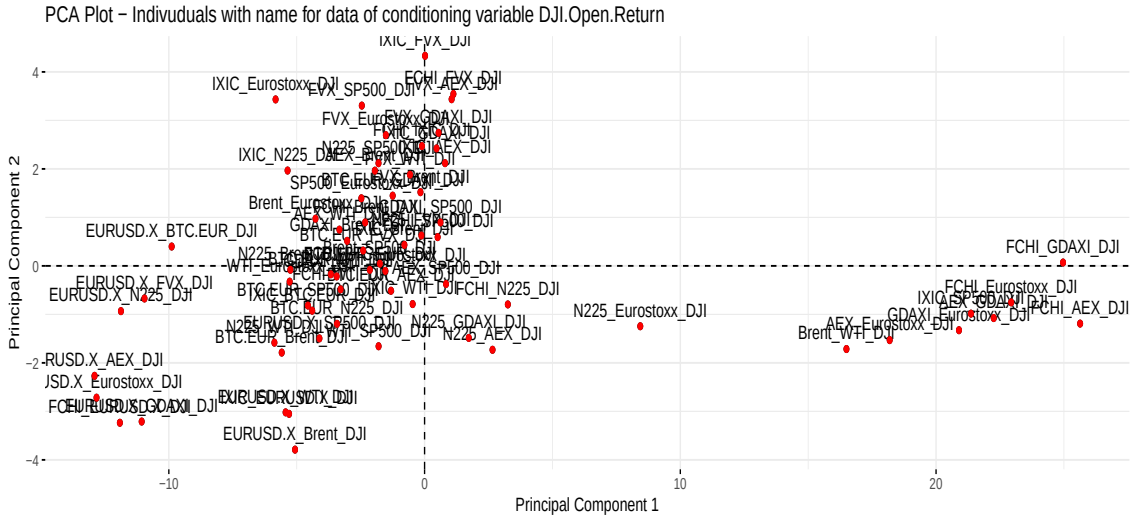


Figure 4.19: Score plot of the subset of conditioning variable DJI using name labels.

From Figure 4.19, it becomes clear we could recognize three different clusters in the subset of conditioning variable DJI. The clusters are constructed using the same approach as for the complete dataset.

The first cluster contains the largest number of scores. The scores are around the origin and often have small negative values on the x-axis. Similarly as in the case of the complete dataset, PC1 seems to be almost constant. In this case, PC1 is almost constant with a small positive average correlation. For small negative scores, this indicates that the average correlation will be around zero and slightly negative. This corresponds to the majority of the CKTs fluctuating around zero, as can be seen in Figure 4.14.

Cluster 2 consists of a few points that have a larger negative value for the PC1 score. This indicates that these points may correspond to negatively correlated combinations of assets. Furthermore, all points have a negative values on the y-axis as well except one single score. From Figure 4.19 becomes clear that this point belongs to the combination EURUSD BTC conditioned on DJI.

Lastly, the third cluster consists of a few points that have large values on the x-axis. In other words, PC1 contributes to great extent in the linear combination representing the original observation. Now, this corresponds to the few curves of CKTs that fluctuate around values larger than zero, as can be seen in Figure 4.14. Notice that the points are distributed along the x-axis (horizontal) instead of along the y-axis (vertical). This is different compared to the clustering for the complete dataset in Figure 4.7.

Note that Figure 4.18 tells us the representation of the points in cluster 1 are not as good as both cluster 2 and 3.

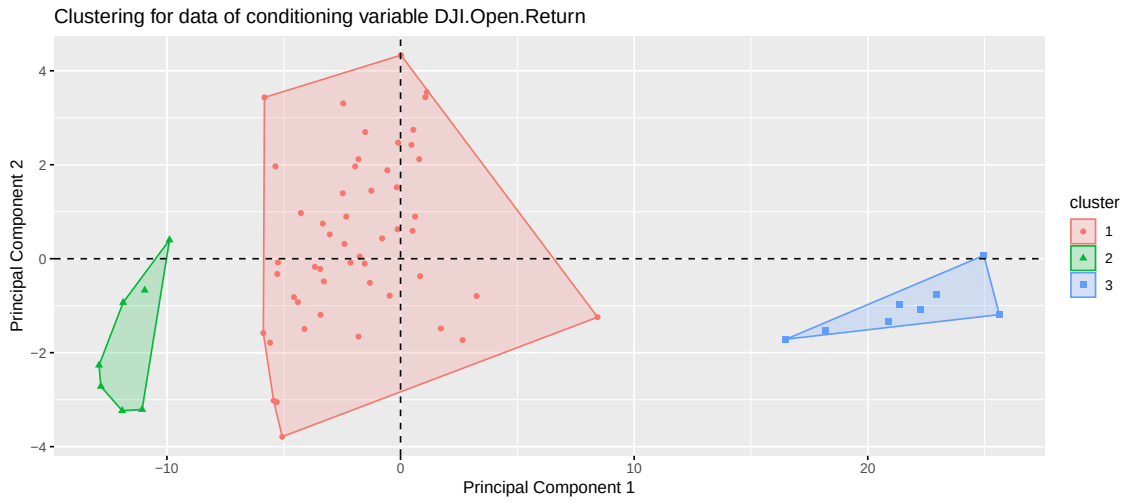


Figure 4.20: Score plot of PC1 and PC2 clustered in three groups applied to the subset of conditioning variable DJI.

Next, it is interesting to look at what combinations of assets are in what clusters and how they interact with each other. Again, we have created a connected multigraph $G_{DJI} = (V, E)$ which explains the relation between the assets per cluster. Here, the set of vertices (V) and the set of edges (E) are defined similarly as in Section 4.1. Note that for each combination of two assets there is now only one edge because the conditioning variable is DJI all the time.

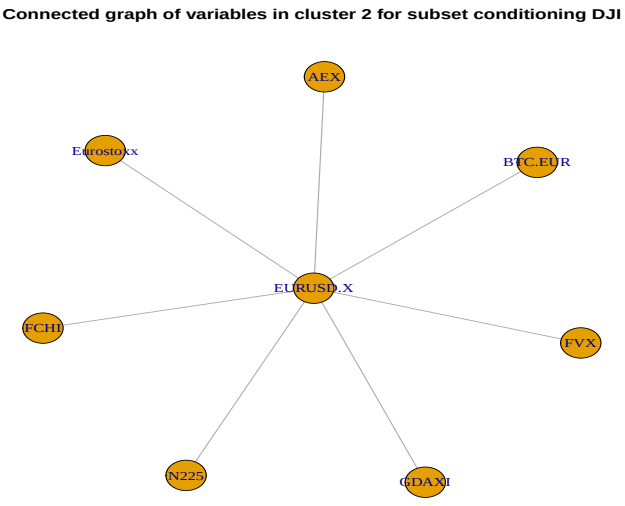


Figure 4.21: Multigraph $G_{DJI,2} = (V, E)$ representing the scores and their underlying relation in cluster 2 applied to the subset of the conditioning variable DJI. The US assets (SP500, IXIC) and futures related to oil prices (Brent, WTI) are not present in this cluster.

In Figure 4.21 the combinations are shown for cluster 2 which all have EURUSD. Interestingly, the the combinations of EURUSD with Brent, WTI and the US assets IXIC and SP500 are not in this cluster, but in cluster 1. These combinations are all around $x = -5$ and around $y = -3$. This means, when conditioning on DJI, EURUSD is a little negatively conditionally dependent with European assets and to a lesser extent for US assets.

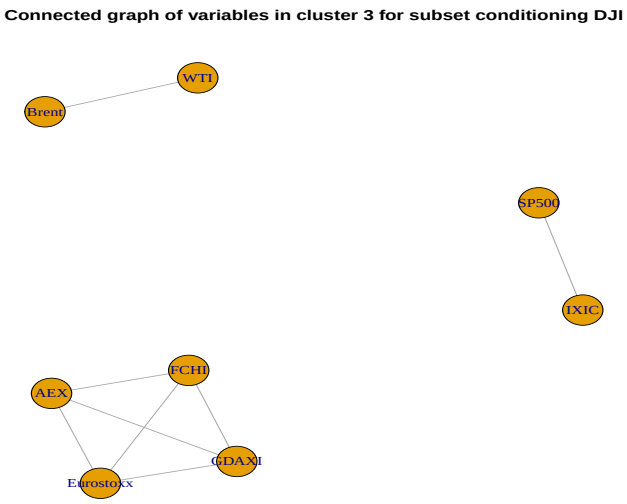


Figure 4.22: Multigraph $G_{DJI,3} = (V, E)$ representing the scores and their underlying relation in cluster 3 applied to the subset of the conditioning variable DJI. The assets N225, FVX, EURUSD and BTC are not present in this cluster.

In Figure 4.22, we recognize three separated connected graphs for cluster 3. In this cluster the most positively correlated assets are given. The combinations Brent and WTI, SP500 and IXIC, and the European assets with each other all conditioned on DJI give a positive correlation. This corresponds with the relationship of the assets in cluster 1 and 4 of the complete dataset, see Figures 4.9 and 4.13.

Next, in Figure 4.23 the connections between the assets conditioned on DJI are shown. Recall that cluster 1 contains scores corresponding to slightly negatively and not correlated CKTs. The dependence between the European assets and assets outside of Europe conditioned on DJI are in this cluster. Furthermore, all combinations for Brent and WTI are in this cluster, except the combination Brent and WTI. The latter combination is positively correlated. Besides, we see that all possible combinations for N225, FVX and BTC are in cluster 1. Apparently, there is almost no dependence between these assets and all other assets conditioning on DJI.

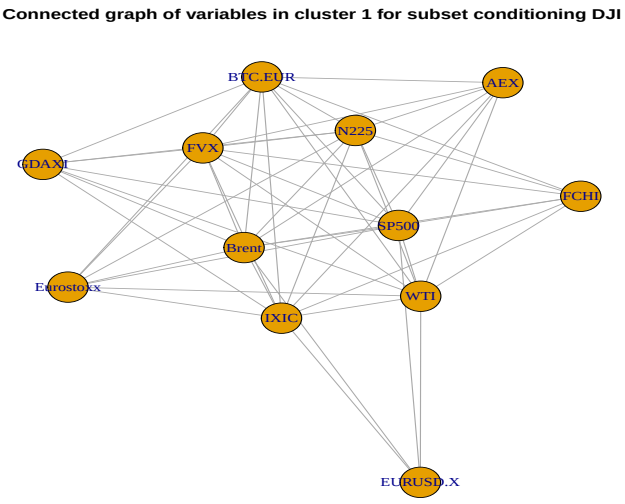


Figure 4.23: Multigraph $G_{DJI,1} = (V, E)$ representing the scores and their underlying relation in cluster 1 applied to the subset of the conditioning variable DJI.

Lastly, the scores for both IXIC and SP500 with the other assets are included in cluster 1, except the one with each other. Namely, the dependence of SP500 and IXIC given DJI is positively correlated. The curve of this CKT is given in Figure 4.24.

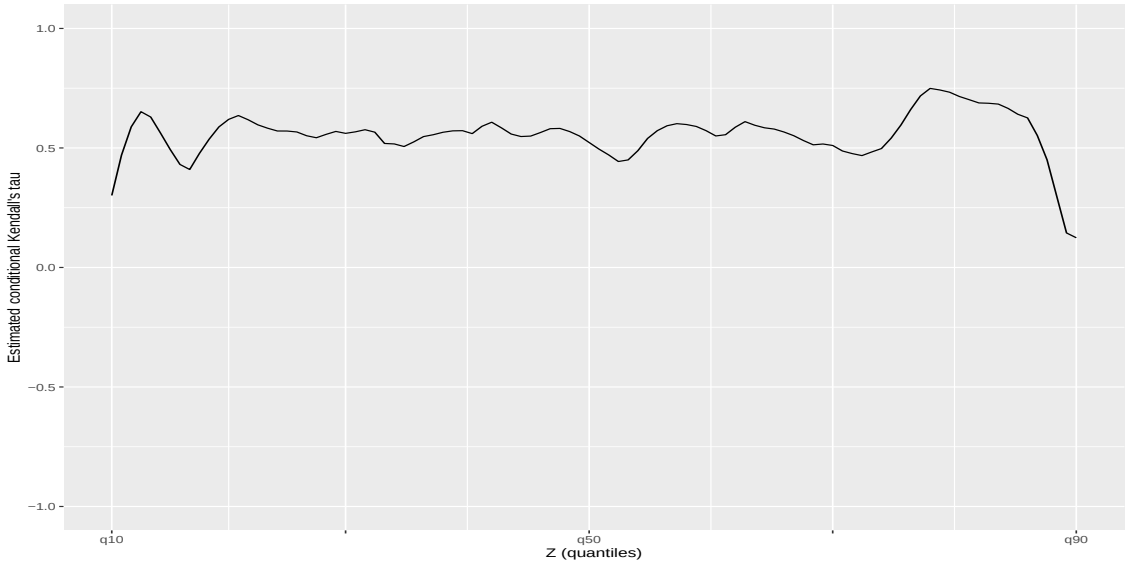


Figure 4.24: Curve of the conditional Kendall's tau for SP500 and IXIC given DJI.

The four clusters used in Section 4.1 can be applied on the plot of all estimates of CKT. This gives us Figure 4.25. This is coherent with the results in Section 4.1. Note that the combinations N225 and Eurostoxx given DJI and Brent WTI given DJI belong to

cluster 3 in Figure 4.20 whereas to cluster 1 in Figure 4.25 below.

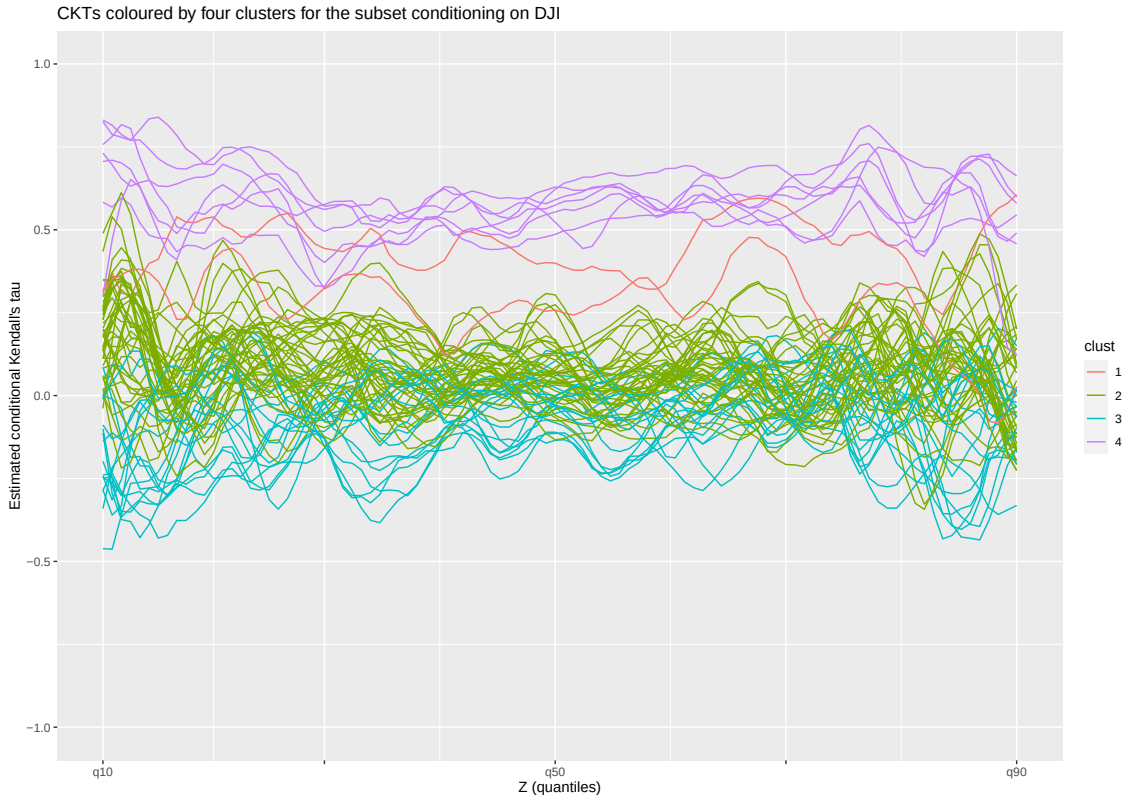


Figure 4.25: Curves of the CKTs conditioned on DJI using the clustering approach from Section 4.1.

4.3. Results clustering with respect to X_1 and X_2

It is interesting to see how the dependence is between any combination of two assets, say X_1 and X_2, for all conditioning variables. We can show this by plotting only the CKT curves of two variables conditioned on all variables instead of all CKT curves together as in Figure 4.8. This results in 78 subplots, since we have $12+11+\cdots+2+1 = 78$ possible combinations of X_1 and X_2. Note that every subplot contains 11 curves corresponding to all possibilities where we can condition on. We still apply the clustering of Section 4.1. This provides insight into the extent to which there is negative, no or positive conditional dependence.

In Figures 4.26 until 4.28, all 78 subplots are shown. We will discuss the subplots ordered from largest positive correlation (cluster 4) to smallest negative correlation (cluster 3).

To start, we see that cluster 4 consists of the combinations of European and US assets with themselves. In case of the European assets, for some conditioning variables the average correlation is smaller.

Then, cluster 1 contains mostly curves that are a combination of one European asset and one US asset given any conditioning variable. Also combinations between one European or US asset and N225 are contained in cluster 1 for almost any conditioning variable. Interestingly, for all conditioning variables the CKT of Brent and WTI is contained in cluster 1 as well.

Next, cluster 2 contains the points corresponding to curves that have no or slightly positive correlation, fluctuating around zero. For both FVX and BTC it holds that their CKTs with almost all other assets given any conditioning value are in this cluster. Note that there are indeed curves of BTC in cluster 3, but do overlap with cluster 2. In other words, even though the curves are in cluster 3, they indicate no correlation. Furthermore, note that FVX has in combination with IXIC and WTI some curves corresponding to other clusters. However, these curves still fluctuate around zero and indicate there is almost no correlation. Lastly, almost all CKTs for Brent and WTI in combination with another asset given all conditional variables are fluctuating around zero. However, only for the conditional dependence of Brent and WTI it takes large positive values.

Lastly, cluster 3 contains both curves with negative average correlation as curves fluctuating around zero with almost no correlation. The latter group overlaps with cluster 2. It becomes clear that EURUSD is strongly present in this cluster in combination with all other assets for most conditioning variables. EURUSD is negatively correlated in combination with AEX, BTC, Eurostoxx, FVX, GDAXI, N225 and FCHI given any conditioning variable. In combination with Brent, DJI, IXIC, SP500 and WTI the CKTs

are more fluctuating around zero rather than around negative values.

The results described above are coherent with our results in Section 4.1.

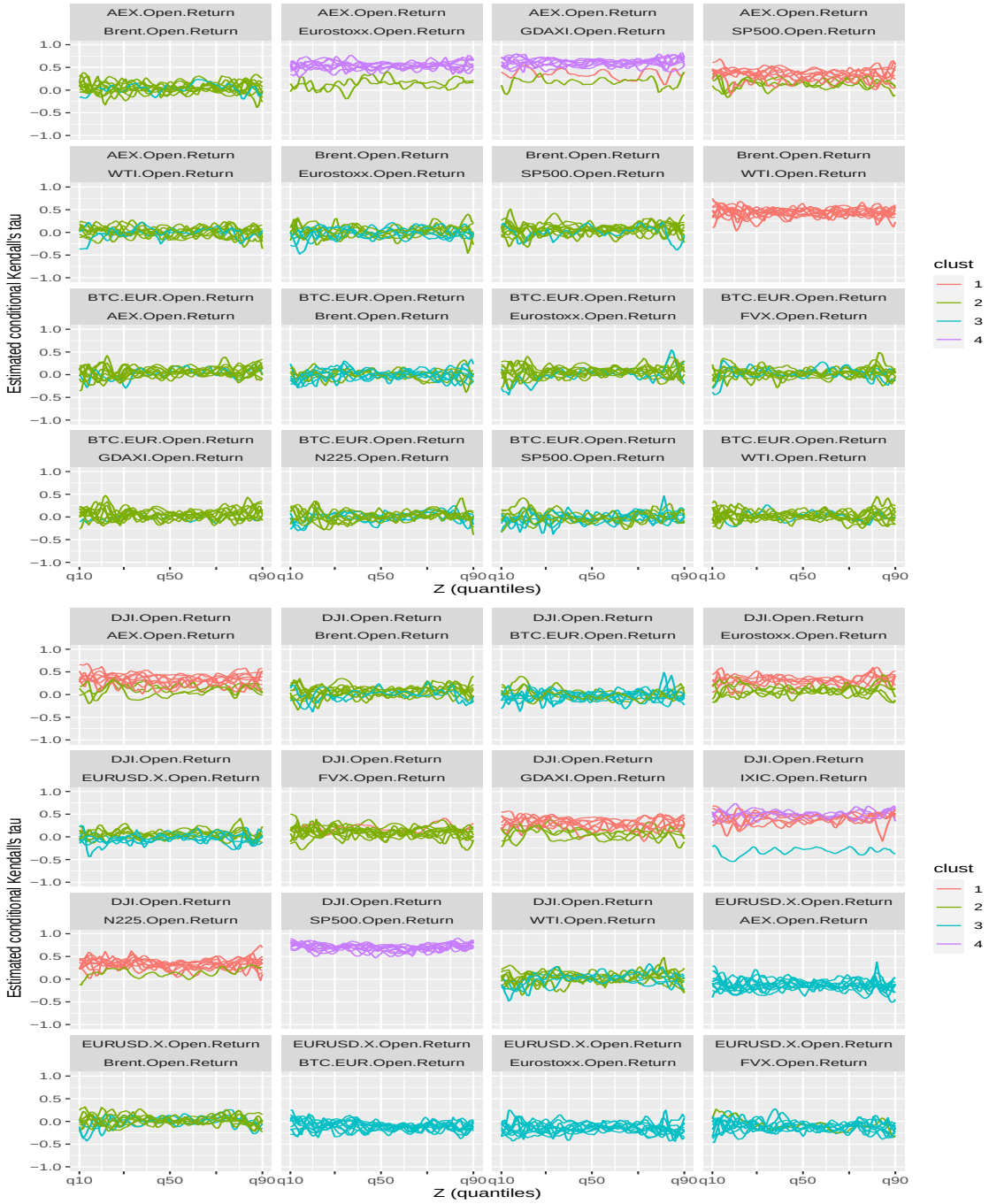


Figure 4.26: CKTs of a combination of two selected variables for all conditioning variables. The subplots are ordered in alphabetical order. (1-32)

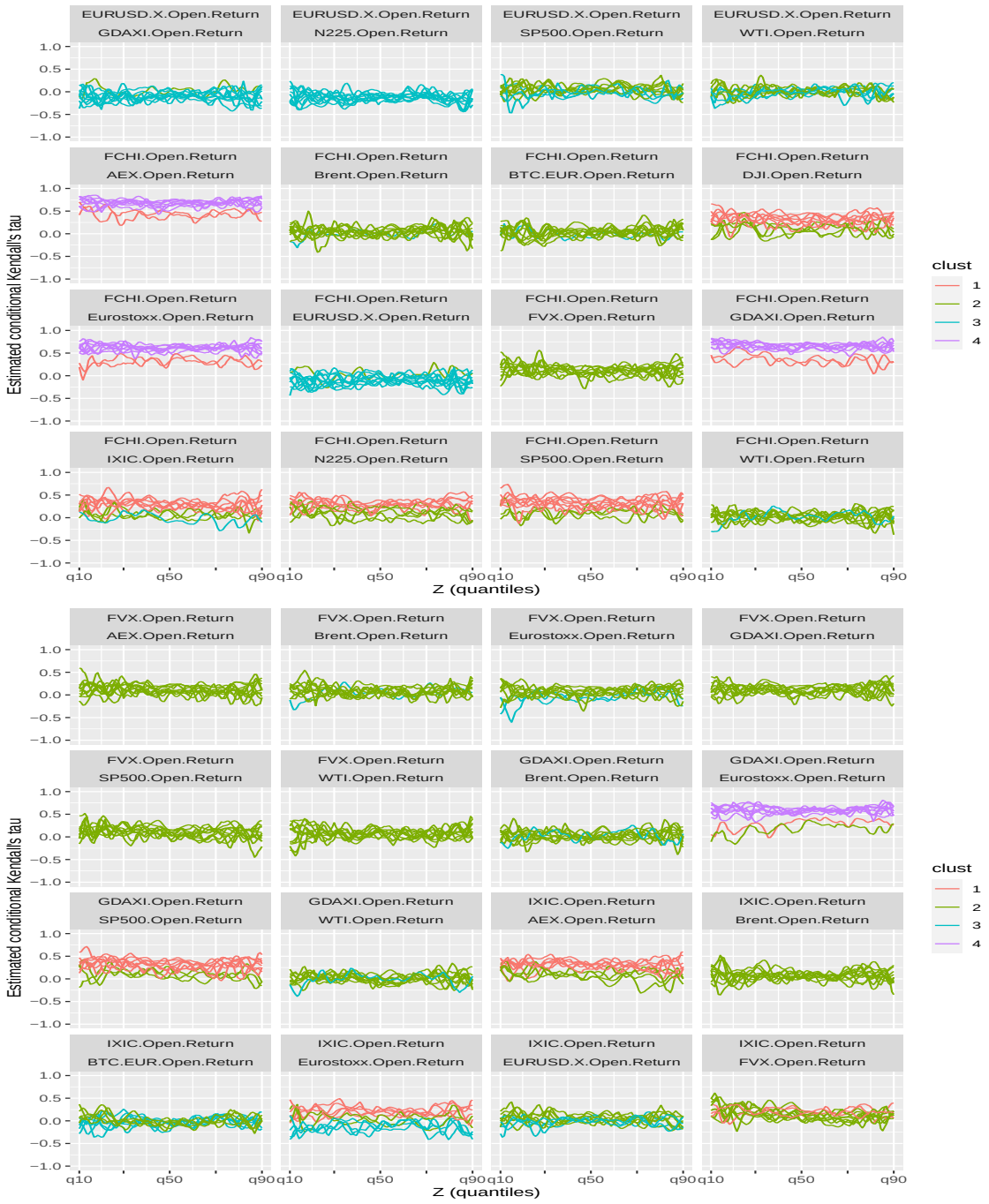


Figure 4.27: CKTs of a combination of two selected variables for all conditioning variables. The subplots are ordered in alphabetical order. (33-64)

Figure 4.28: CKTs of a combination of two selected variables for all conditioning variables. The subplots are ordered in alphabetical order. (65-78)

4.4. Results clustering with respect to conditioning variable Z

In Figure 4.29 the curves of the CKTs for all combinations of any two assets given a fixed conditioning variable are plotted. The same clustering approach is used as for the complete dataset in Section 4.1. It becomes clear from Figure 4.29 that for each subsets of a conditioning variable, there is a different number of CKTs in the four clusters.

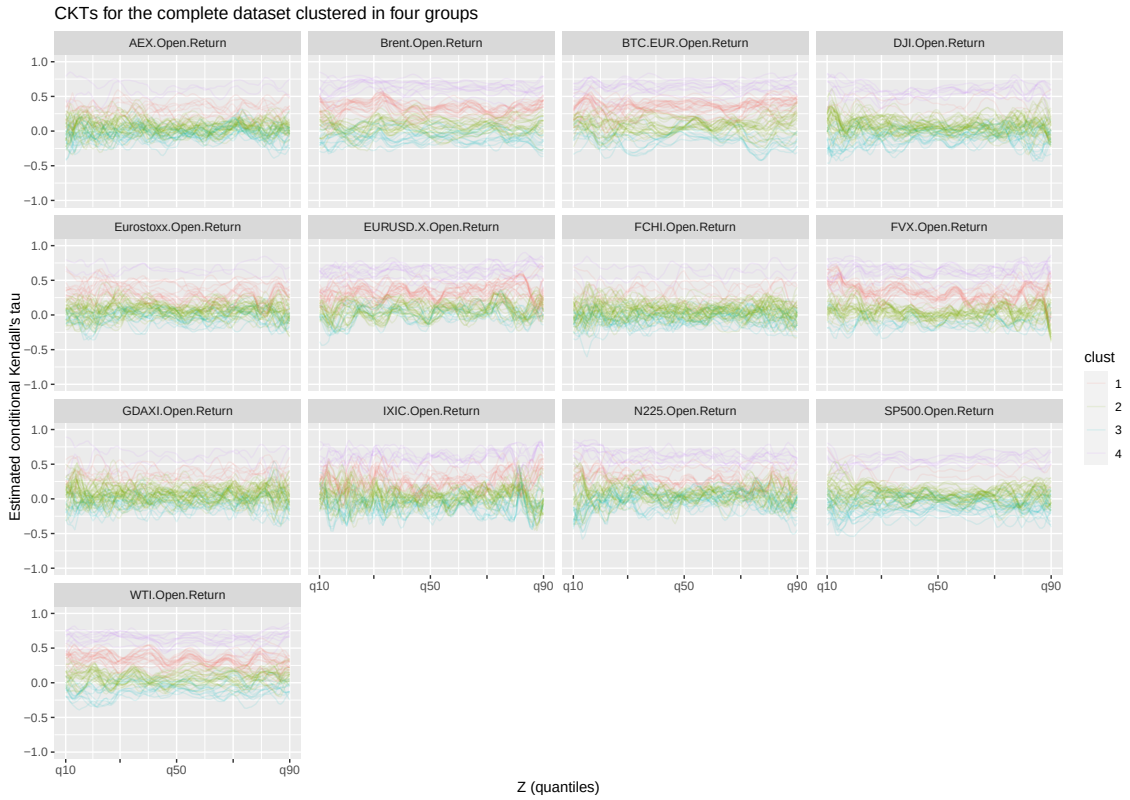


Figure 4.29: The curves of the conditional Kendalls tau for any two assets given a fixed conditioning variable using clustering.

Our previous results, reinforce Figure 4.29. For example, most of the assets are uncorrelated when conditioning on European assets (AEX, Eurostoxx, GDAXI, FCHI) since most curves correspond to cluster 2. Only a few combinations have a stronger positive correlation corresponding to cluster 1 and 4. Interestingly, only when when the combination is between US assets, conditioning on a European assets is in cluster 4.

For the other assets, the distributions of CKTs in the four different clusters do not seem to follow a pattern. This is because for every cluster there appear some curves of the CKTs per fixed conditioning variable.

5

Conclusion

In this thesis, we have examined estimation of conditional Kendall's tau (CKT). We have demonstrated the use of the estimator in a real world application of thirteen different financial assets. PCA has been used to increase the interpretability of the data of estimated CKTs. This included the use of clustering.

In general, it seems that conditional dependence is slightly larger in the tails. This means that when a conditioning variable takes large values, both negative and positive, there is a stronger correlation between two assets. Note that this holds for the positive and negative conditional dependence of the assets.

Further, it seems that scores in PCA that are not represented well by PC1 and PC2 tend to correspond to CKTs having almost no average correlation. Conversely, it seems that the better the representation of the variables by PC1 and PC2 is, the larger the average correlation. Again, this holds for both positive and negative correlation.

The influence of the conditioning event seems to strongly relate to characteristics such as the geographical properties and the type of an asset. To analyse the influence of the conditioning variable further, we used PCA clustering. We distinguish two forms of clustering. Namely, clustering with respect to X_1 and X_2 and with respect to the conditioning variable Z.

To start with the first one, it seems that given any conditioning variable, there is only positive correlation between the stock indices.

European assets (AEX, GDAXI, FCHI and Eurostoxx) have a relatively large positive correlation between each other. This holds for any conditional variable except for con-

ditioning on any European assets. Then the correlation is smaller.

US assets (DJI, IXIC and SP500) have a relatively large positive correlation between each other for any conditioning variable. Surprisingly, this does not hold for one single combination: the conditional Kendall's tau for DJI and IXIC given SP500 which has the largest negative average correlation in the complete dataset.

For the sets of assets corresponding to 'Oil prices' and 'Debt and currency' there seems no conditional dependence except for some combinations with EURUSD. In fact, EURUSD is negatively correlated with the European assets for all conditioning variables. On the other hand, there is no correlation between EURUSD and the US assets given most of the conditioning variables. For each asset Brent, BTC, FVX, and WTI there seems almost no correlation between the other assets given any conditioning variable. Only the correlation between Brent and WTI takes strongly positive values given all conditioning variables. This could be because both are futures in relation to oil prices and therefore closely linked.

To end with the second type of clustering, there does not seem a clear trend when clustering on the conditional variable. However, when we condition on European assets, the average CKT seems to be less positively correlated compared to the others. Only for US stock indices there is a strong correlation when we condition on any European assets.

6

Discussion

It is important to validate your choices and assumptions. We could have done cross-validation for our choice of bandwidth, for our choice of estimator and for choice of number of components in PCA.

Choice of bandwidths

Bandwidth selection is a key issue in density estimation [10]. We have chosen the bandwidths for each conditioning variable using visual inspection. Although this has been done carefully, it could be done more accurate. There are various ways that we could consider a next time to improve the choice of bandwidth. On the one hand, we could have used one of the existing 'rules'. These include Silverman's rule of thumb and Scott's rule of thumb. Note that these rules have some serious limitations [20]. We could also choose h as the minimizer of the cross-validation criterion, see (3.11), which has been done already for estimation of conditional Kendall's tau [8]. Lastly, we could use the Mean Integrated Squared Error (MISE). The MISE is convenient due to its mathematical tractability and its natural relation with the MSE. For bandwidth selection using MISE we refer to [10].

On the other hand, we could also have performed a cross-validation for multiple bandwidths by still using visual inspection.

Choice of estimator

We have used the estimator of (3.5) for the conditional Kendall's tau. This estimator is used before ([7], [8]). However, there are other estimators for CKT available. It would be interesting to use different estimators for the same analysis we have conducted in our

research.

Choice of number of principal components

In this thesis, we have used only the first two principal components. The explained variance of the first two components all had a percentage of at least 80%. However, we have seen in Chapter 4 that the representation of the variables by only PC1 and PC2 is not as good as for all scores. So does this percentage of at least 80% tell us choosing only PC1 and PC2 will work sufficiently? As we have mentioned before, the problem of choosing a sufficient number of PCs is still open, but there are some guidelines [1]. One method is to plot the eigenvalues according to their size and to see if there is a point in this graph such that the slope of the graph goes from 'steep' to 'flat'. Then, choose only the components before the steep decrease. This procedure is called the scree or elbow test. Note that this test has some limitations. For other procedures that could be considered, we refer to [1].

Furthermore, as choice of the financial data, we have used just the normal returns (.Open.Return) of thirteen assets to estimate the CKT. Recall that this data is not independent of time. However, the .Open.RI dataset corresponds to the ARMA-GARCH filtered innovations of the Open.Return dataset. This dataset is independent of time. The CKTs estimated from the Open.RI dataset are conceptually different from the one using the Open.Return dataset. Therefore, it is interesting to estimate CKTs of the .Open.RI dataset. Note that the distinction between Open.Return and Open.RI is unrelated to the PCA.

Lastly, in this thesis we have used thirteen different financial assets: stock indices, bonds, futures and exchange rates. It would be interesting whether adding more and different assets will reinforce our conclusions in Chapter 6. For instance, I would like to include another Asian stock index to examine the geographical influence on conditional dependence.

References

[1] H. Abdi and L. J. Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistic* 4 (2010), pp. 433–459.

[2] Andrey Akinshin. *The importance of kernel density estimation bandwidth*. 2022. URL: https://aakinshin.net/posts/kde-bw/ (visited on 06/27/2022).

[3] A. Ang and G. Bekaert. "International asset allocation with regime shifts". In: *Review of Financial Studies* (2002), pp. 1137–1187.

[4] Ashni Bachasingh. "Dependence Measures in Citation Analysis The application of parametric copulas to capture the dependence structure between the publications of a researcher and the citations of those publications." In: *BSc thesis TU Delft* (2018), p. 5.

[5] Lee J. Bain and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, 1992.

[6] A. Derumigny. *Introduction to copulas and dependence modeling*. University of Twente / Rencontres Statistiques Lyonnaises, 2019.

[7] A. Derumigny and J. Fermanian. "A classification point-of-view about conditional Kendall's tau". In: *Computational Statistics and Data Analysis* (2018), pp. 1–5.

[8] A. Derumigny and J. Fermanian. "On kernel-based estimation of conditional Kendall's tau: finite-distance bounds and asymptotic behavior". In: *Dependence Modeling* 7 (2019), pp. 292–312.

[9] Alexis Derumigny. *CondCopulas: Estimation and Inference for Conditional Copula Models*. R package version 0.1.2. 2022. URL: https://CRAN.R-project.org/package=CondCopulas.

[10] E. Garcia-Portugues. *Notes for Nonparametric Statistics*. Version 6.5.8. ISBN 978-84-09-29537-1. 2022. URL: https://bookdown.org/egarpor/NP-UC3M/.

[11] I. Gijbels, Veraverbeke N., and Omelka M. "Conditional copulas, association measures and their applications". In: *Computational Statistics and Data Analysis* (2011), pp. 1919–1932.

[12] M. Haugh. "An introduction to copulas". In: *IEOR E4602: Quantitative Risk Management (lecture notes)* (2016). http://www.columbia.edu/ mh2078/QRM/Copulas.pdf.

[13] H. Joe. *Dependence Modeling with Copulas*. Chapman and Hall/CRC, 2014.

[14] Roger B Nelsen. *An Introduction to Copulas*. second. New York, NY, USA: Springer, 2006.

[15] T. Nguyen. "Principal Component Analysis of Education-Related Data Sets". In: *BSc Thesis TU Delft* (2020), pp. 12–16.

[16] J. Karpiński O. Hryniewicz. "Prediction of reliability – the pitfalls of using Pearson's correlation". In: *Eksploatacja i Niezawodnos* (2014), pp. 473–483.

[17] Andrew Patton. "Modelling Asymmetric Exchange Rate Dependence". In: *International Economic Review* 47.2 (2006), pp. 527–556.

[18] Svetlozar T. Rachev. "Copula Concepts in Financial Markets". In: *University of Karlsruhe, KIT and University of Santa Barbara and FinAnalytica* (2009). https://statistik.econ.kit.edu/download/Copula$_C$*oncepts*$_i$*n*$_F$*inancial*$_M$*arkets*.

[19] Seb. *Principal Components Analysis Explained for Dummies*. 2022. URL: https://programmathically.com/principal-components-analysis-explained-for-dummies/ (visited on 06/14/2022).

[20] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Feb. 2018, pp. 1–175. ISBN: 9781315140919. DOI: 10.1201/9781315140919.

[21] Seema Singh. *Understanding the Bias-Variance Tradeoff*. 2018. URL: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229 (visited on 06/20/2022).

[22] R. Van der Spek. "Fast Estimation of Kendall's Tau and Conditional Kendall's Tau Matrices under Structural Assumptions". In: *TU Delft* (2022).

[23] National Institute of Standards and Technology. *NIST*. 2019. URL: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm (visited on 06/12/2022).

A

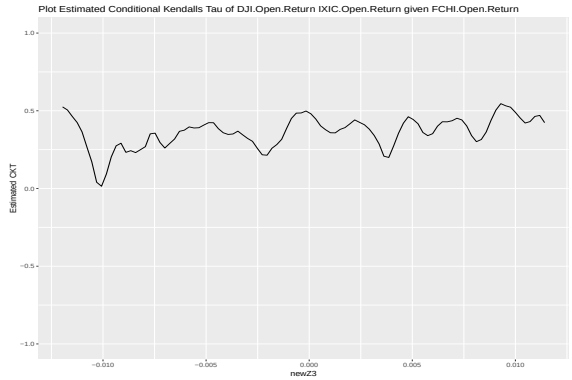
Bandwidth Selection

Bandwidth selection

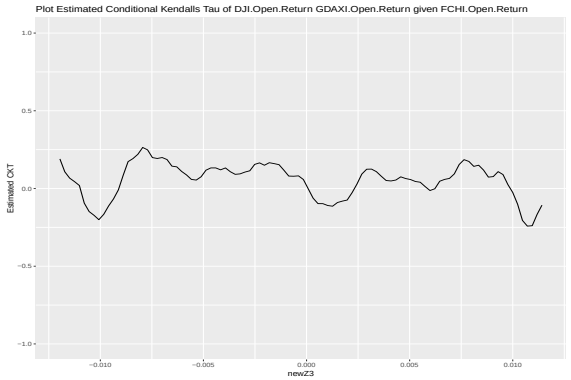
Now we will provide some curves of the conditional Kendall's tau of randomly chosen combinations of variables to reason why this bandwidth is used.

France CAC 40 Stock Index	FCHI	$h = 0.0009$

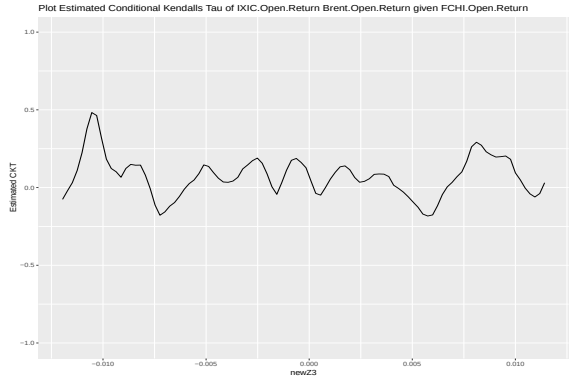
Figure A.1 shows four well-behaved curves, i.e. smooth curves that are not too flattened or too much fluctuating. Therefore, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given FCHI for other variables, we coninue seeing well-behaved curves.



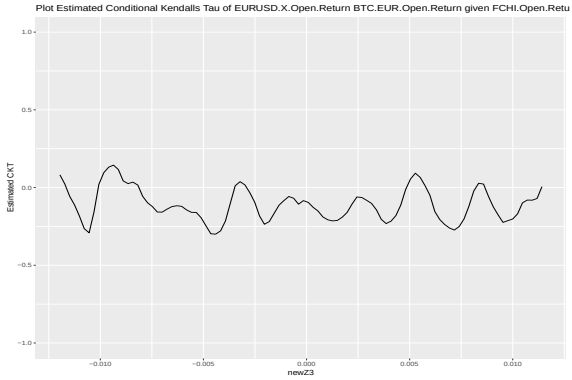
(a) CKT of DJI and IXIC given FCHI



(b) CKT of DJI and GDAXI given FCHI



(c) CKT of IXIC and Brent given FCHI

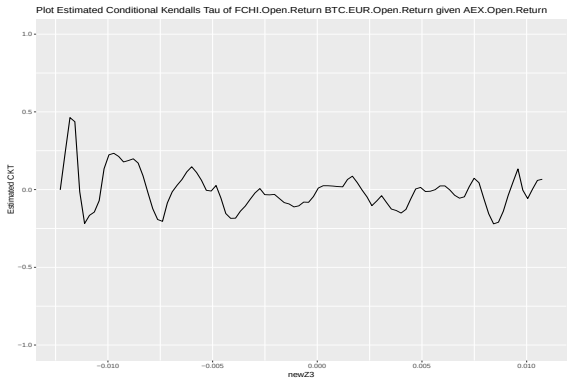


(d) CKT of EURUSD and BTC given FCHI

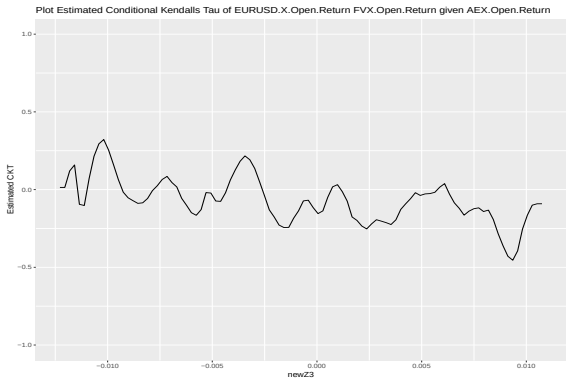
Figure A.1: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of FCHI.

Amsterdam Exchange Index	AEX	$h = 0.001$

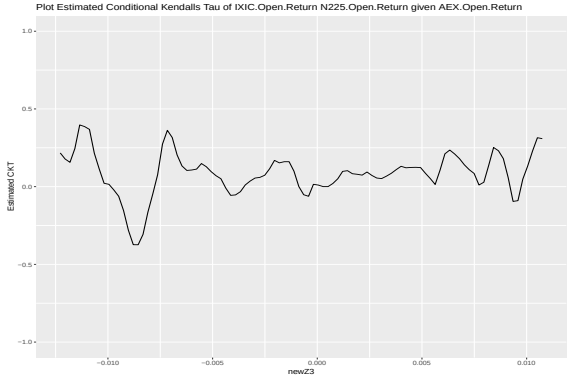
Figure A.2 shows four smooth curves. The curves are indeed not too flat or too fluctuating. Therefore, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given DJI for other variables, we continue seeing well-behaved curves.



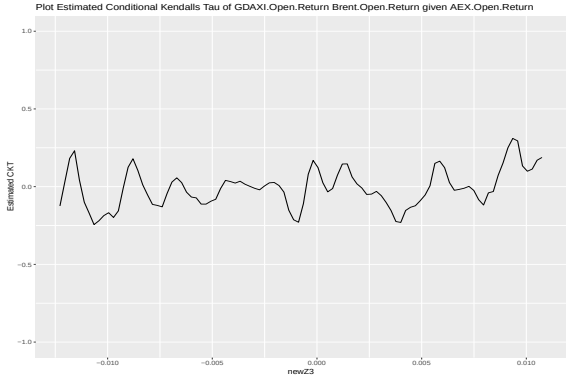
(a) CKT of FCHI and BTC given AEX



(b) CKT of IXIC and N225 given AEX



(c) CKT of EURUSD and AEX given AEX

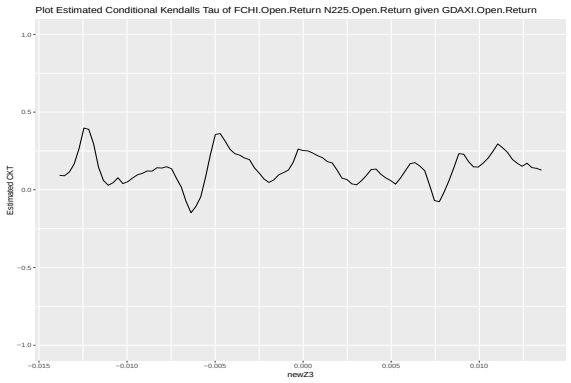


(d) CKT of N225 and FVX given AEX

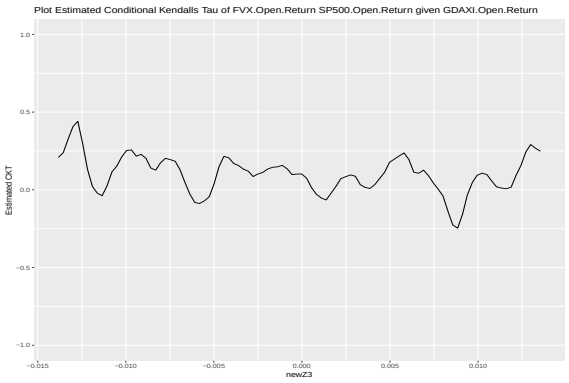
Figure A.2: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given AEX. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of AEX.

German DAX Index	GDAXI	$h = 0.001$

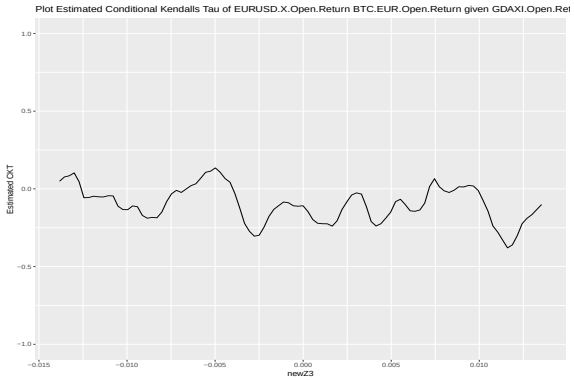
Figure A.3 shows four smooth curves. The curves are indeed not too flat or too fluctuating. Therefore, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given GDAXI for other variables, we continue seeing stable curves of the conditional Kendall's tau.



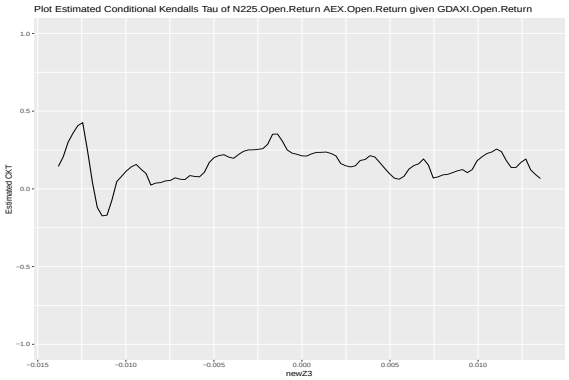
(a) CKT of FCHI and N225 given GDAXI



(b) CKT of FVX and SP500 given GDAXI



(c) CKT of EURUSD and BTC given GDAXI

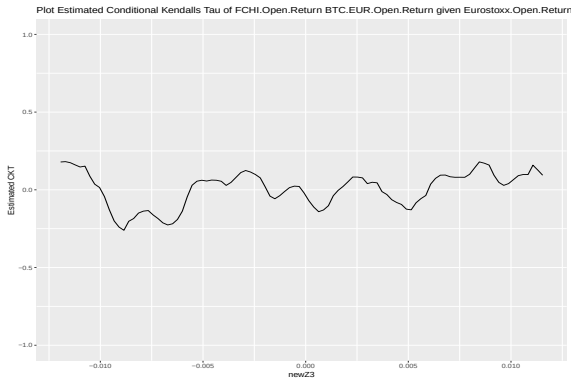


(d) CKT of N225 and AEX given GDAXI

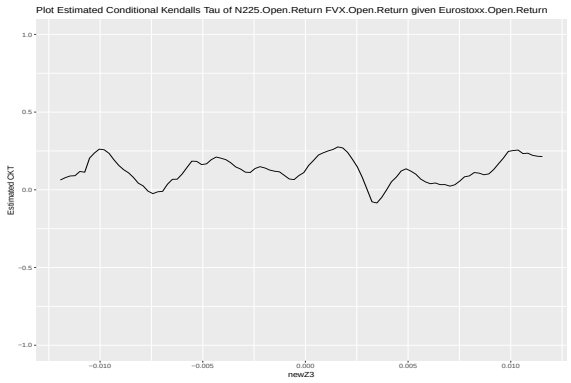
Figure A.3: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given GDAXI. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable GDAXI.

EURO STOXX 50	Eurostoxx	$h = 0.001$

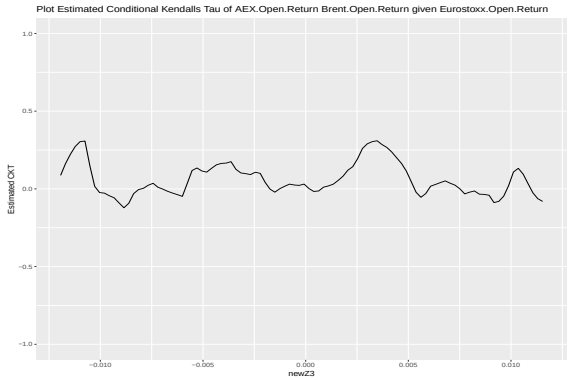
Figure A.4 shows four smooth curves without any remarkable behaviour. The curves are indeed not too flat or too fluctuating. Therefore, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given Eurostoxx for other variables, we continue seeing well-behaved curves.



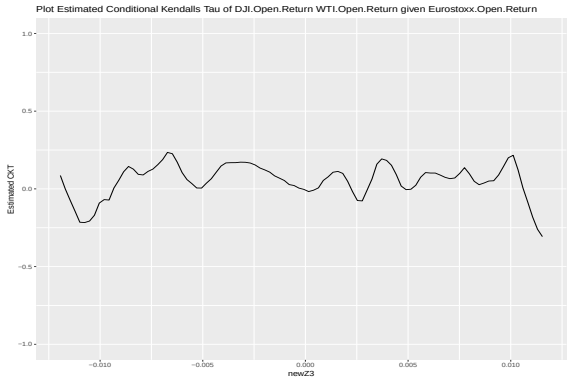
(a) CKT of FCHI and BTC given Eurostoxx



(b) CKT of N225 and FVX given Eurostoxx



(c) CKT of AEX and Brent given Eurostoxx

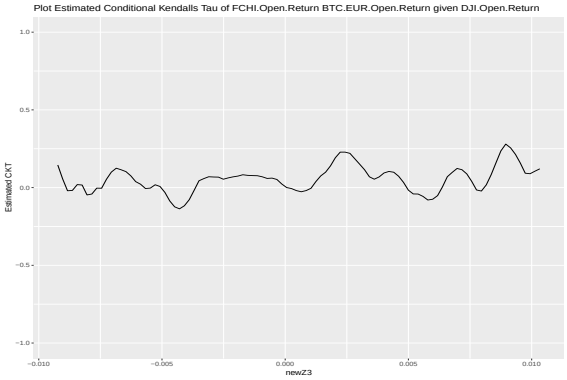


(d) CKT of DJI and WTI given Eurostoxx

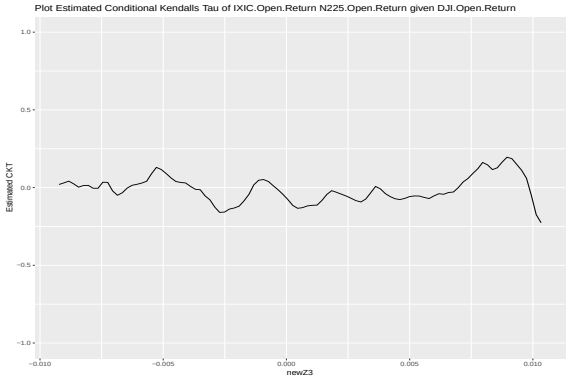
Figure A.4: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable Eurostoxx.

Dow Jones Industrial Average	DJI	$h = 0.0009$

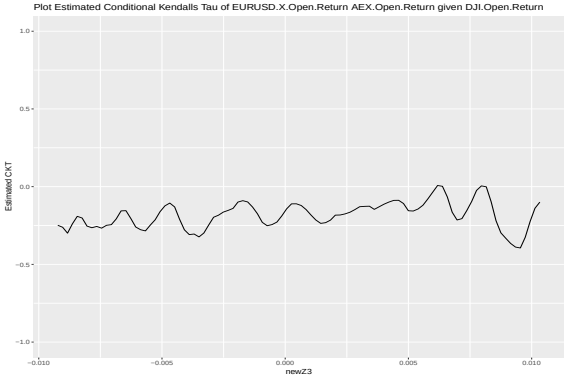
Figure A.5 shows four stable curves. The curves are indeed not too flat or too fluctuating. Therefore, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given DJI for other variables, we continue seeing well-behaved curves.



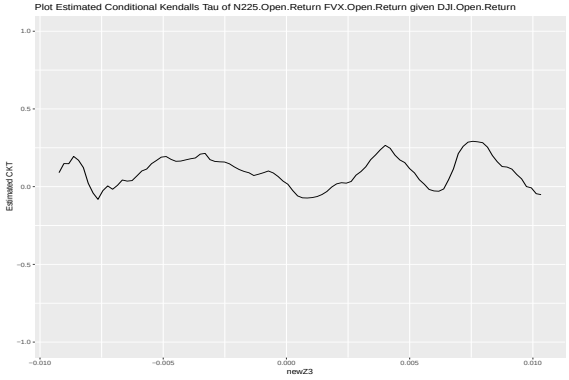
(a) CKT of FCHI and BTC given DJI



(b) CKT of IXIC and N225 given DJI



(c) CKT of EURUSD and AEX given DJI

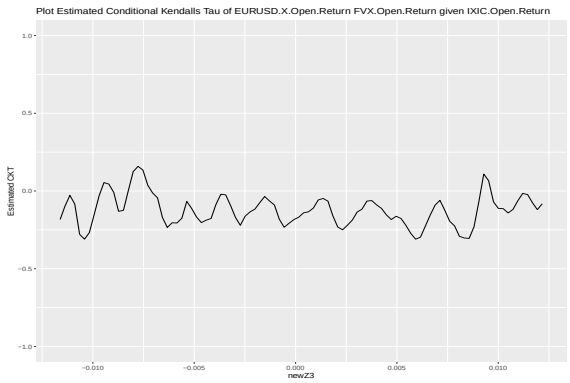


(d) CKT of N225 and FVX given DJI

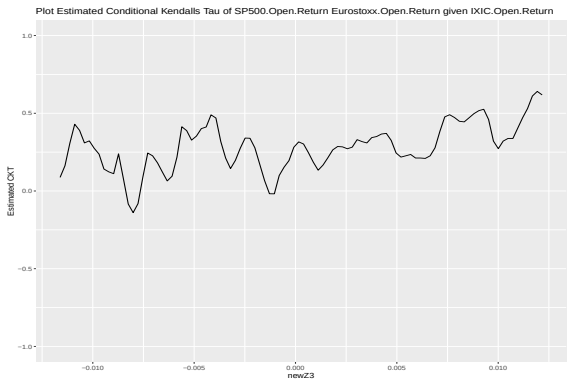
Figure A.5: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable DJI.

Nasdaq Composite	IXIC	$h = 0.0007$

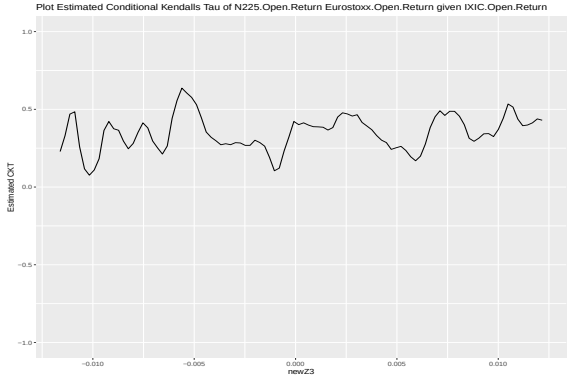
Figure A.6 shows four stable curves. Indeed, the curves are not flattened, but they seem to fluctuate a little heavily. This indicates a large bandwidth. The bandwidth is still chosen well for these four combinations of variables since the fluctuations are not to extreme. However, the bandwidth should not have been larger. Whenever we inspect plots of the CKT given IXIC for other variables, we continue seeing well-behaved curves.



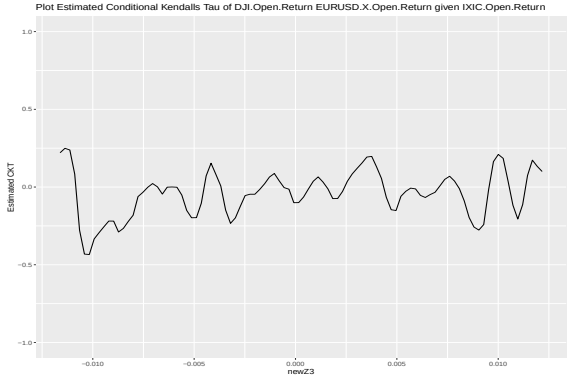
(a) CKT of EURUSD and FVX given IXIC



(b) CKT of SP500 and Eurostoxx given FCI



(c) CKT of N225 and Eurostoxx given IXIC

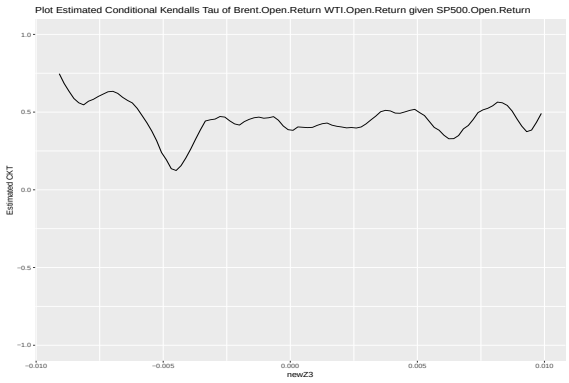


(d) CKT of DJI and EURUSD given IXIC

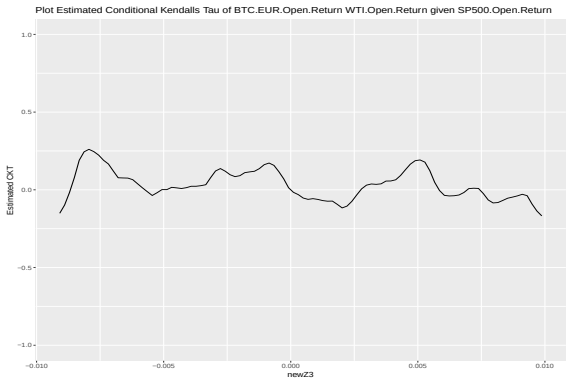
Figure A.6: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable IXIC.

SP500 Index	SP500	$h = 0.001$

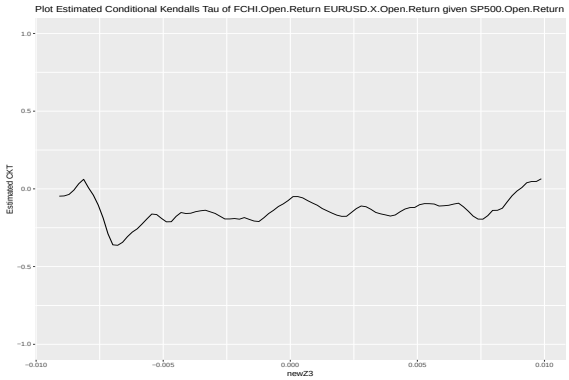
Figure A.7 shows four smooth curves that do not fluctuate heavily. Moreover, the curves are not over-smoothed, on the other hand. Hence, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given SP500 for other variables, we continue seeing stable curves.



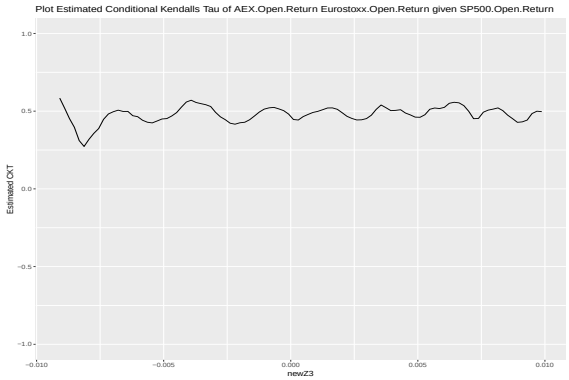
(a) CKT of Brent and WTI given SP500



(b) CKT of BTC and WTI given SP500



(c) CKT of FCHI and EURUSD given SP500

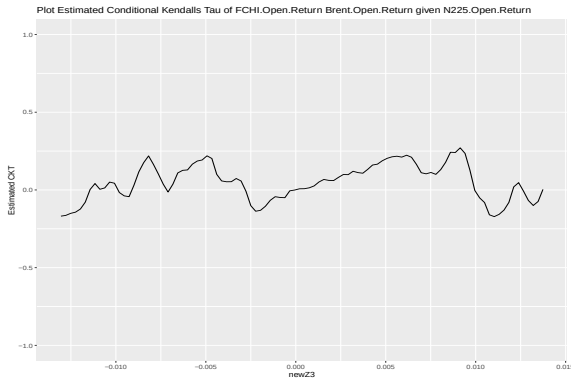


(d) CKT of AEX and Eurostoxx given SP500

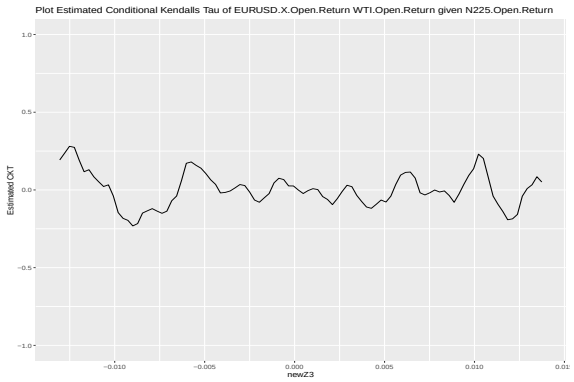
Figure A.7: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable SP500.

Nikkei Index	N225	$h = 0.001$

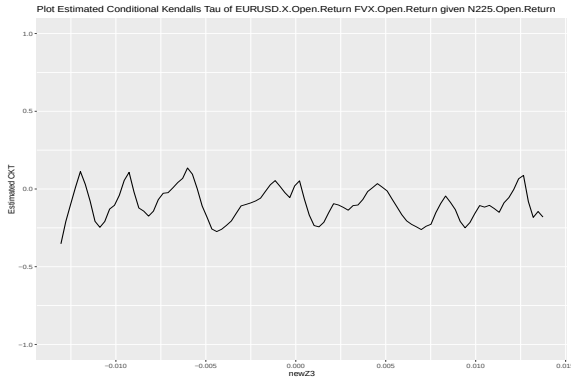
Figure A.8 shows four stable curves. Only the curve in plot (c) seems to be quite shocky. This may be due to a too small bandwidth. However, the bumps are not extreme and the CKT stays most of the time within a rather small interval $[-0.25, 0]$. Hence, the bandwidth is chosen sufficiently for these four combinations of variables. Whenever we inspect plots of the CKT given N225 for other variables, we continue seeing well-behaved curves.



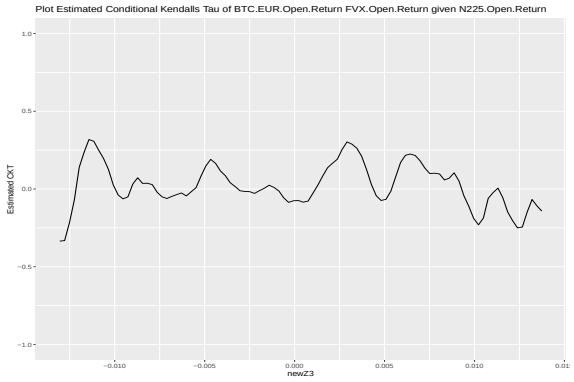
(a) CKT of FCHI and Brent given N225



(b) CKT of EURUSD and WTI given N225



(c) CKT of EURUSD and FVX given N225

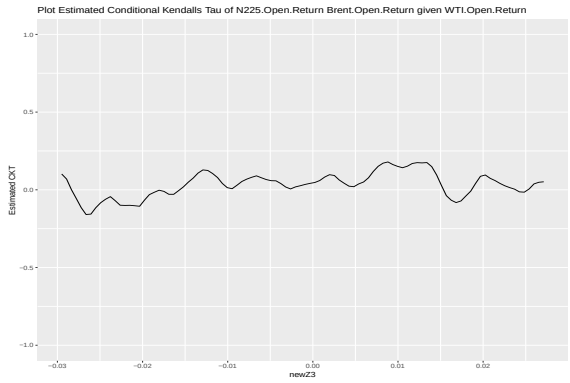


(d) CKT of BTC and FVX given N225

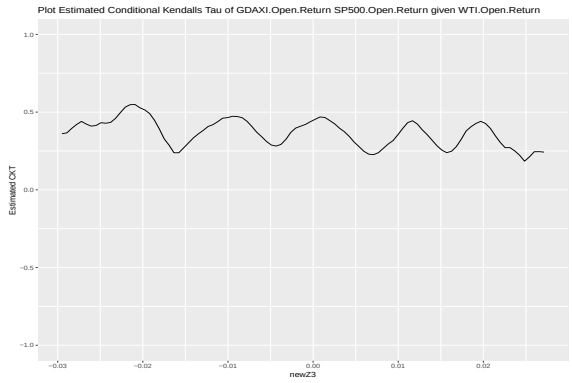
Figure A.8: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given N225. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable N225.

Oil prices West Texas Intermediate	WTI	$h = 0.003$

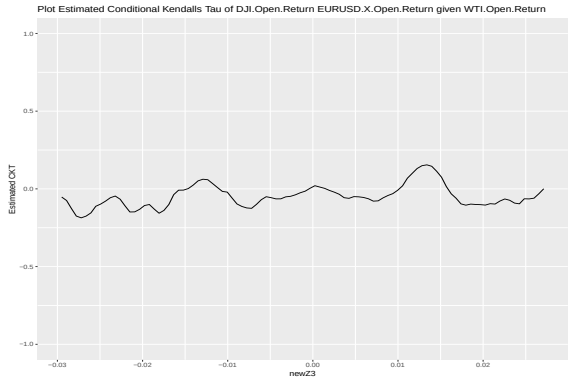
Figure A.9 shows four smooth curves without any remarkable behaviour. The curves are indeed not too flat or too fluctuating. Therefore, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given WTI for other variables, we continue seeing well-behaved curves.



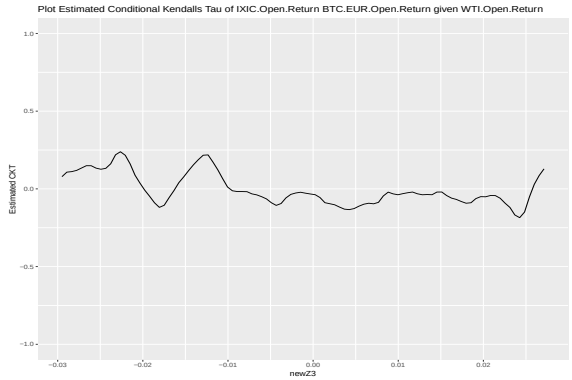
(a) CKT of N225 and Brent given WTI



(b) CKT of GDAXI and SP500 given WTI



(c) CKT of DJI and EURUSD given WTI

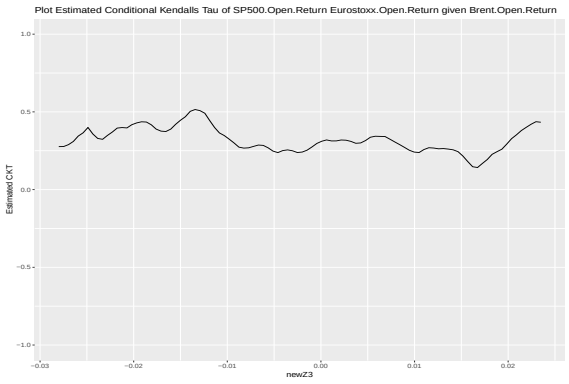


(d) CKT of IXIC and BTC given WTI

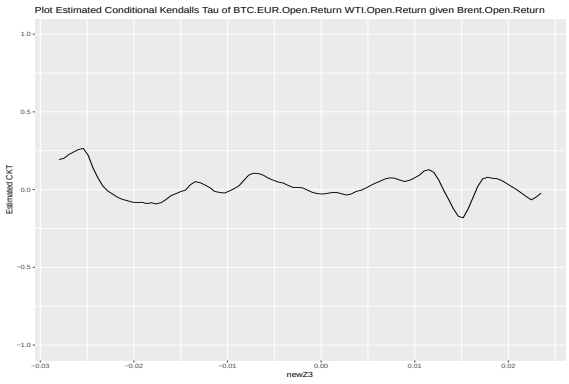
Figure A.9: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable WTI.

Brent Crude Oil	Brent	$h = 0.003$

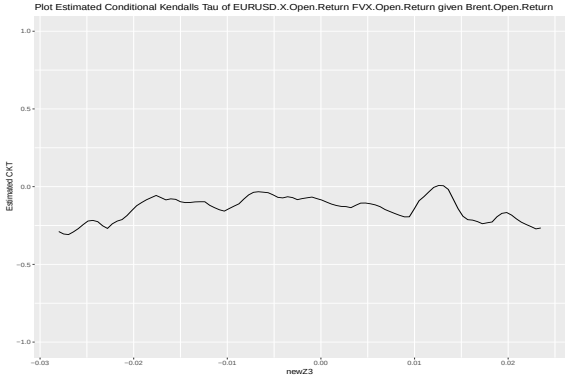
Figure A.10 shows four stable curves which are not fluctuating too heavily. Hence, the bandwidths are chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given Brent for other variables, we continue seeing stable curves.



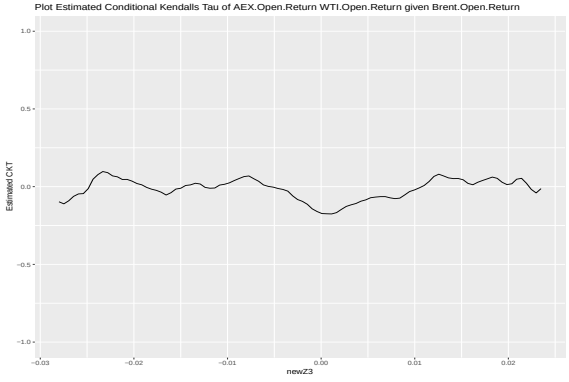
(a) CKT of SP500 and Eurostoxx given Brent



(b) CKT of BTC and WTI given Brent



(c) CKT of EURUSD and FVX given Brent

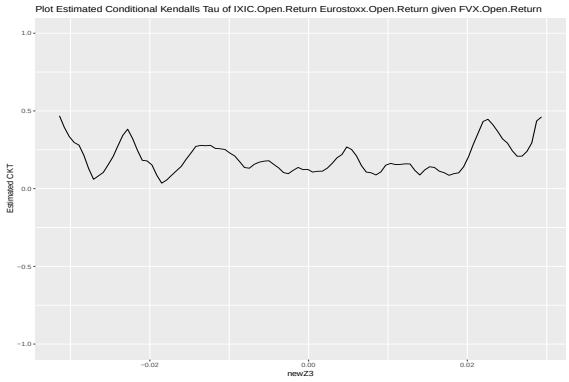


(d) CKT of AEX and WTI given Brent

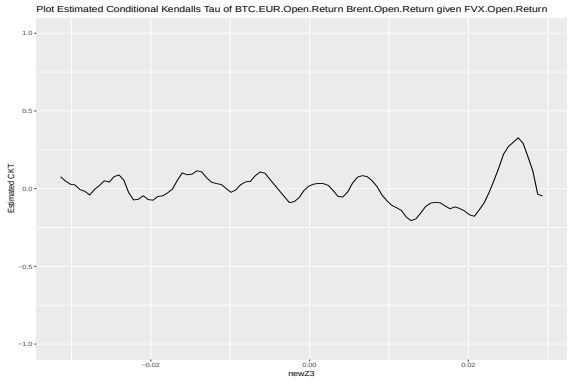
Figure A.10: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given Brent. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between interval q_{10} and q_{90} of the data of Brent.

Debt and currency 5 year US Treasury Yield	FVX	$h = 0.0025$

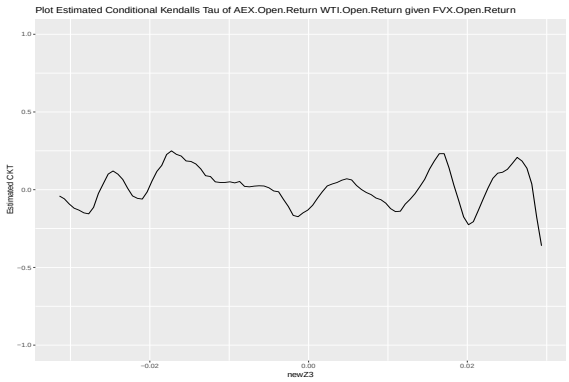
Figure A.11 shows four stable curves. Hence, the bandwidths are chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given FVX for other variables, we continue seeing stable curves.



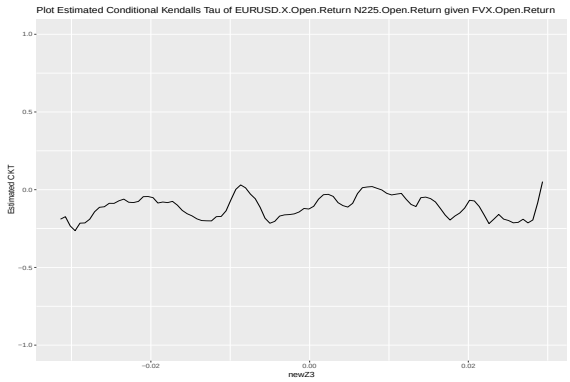
(a) CKT of IXIC and Eurostoxx given FVX



(b) CKT of BTC and Brent given FVX



(c) CKT of AEX and WTI given FVX

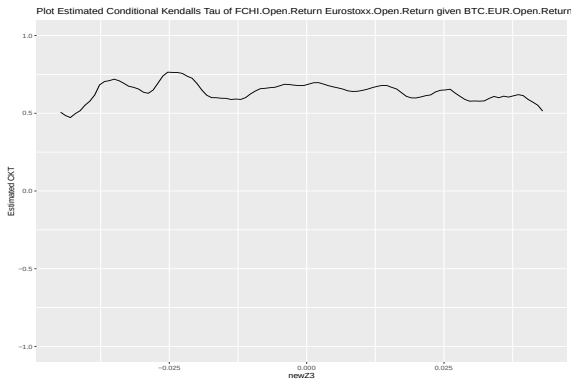


(d) CKT of EURUSD and N225 given FVX

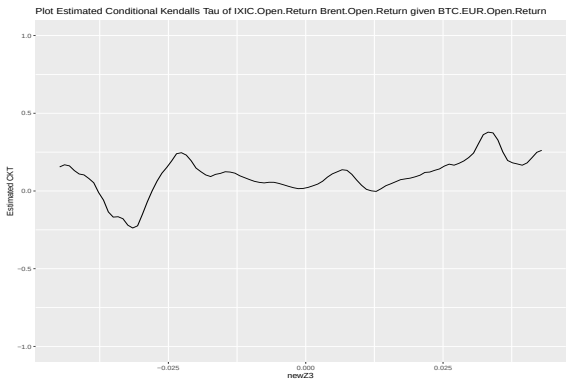
Figure A.11: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given FVX. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable FVX.

Price in Euros of 1 bitcoin	BTC.EUR	$h = 0.005$

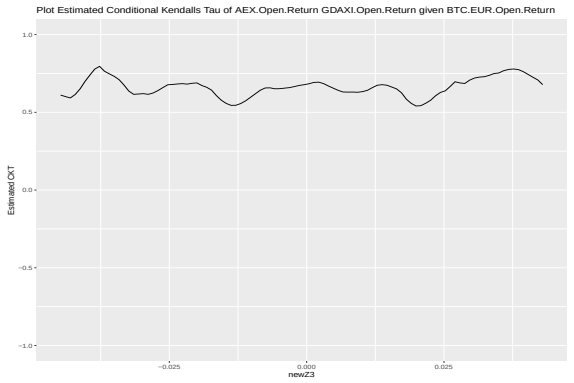
Figure A.12 shows four stable curves. Both curves in plot (a) and (b) seem to be quite flat. This may be due to a too large bandwidth. However, the curves seem to change over some time and some bumps are still visible. Hence, the bandwidth is chosen sufficiently for these four combinations of variables. Whenever we inspect plots of the CKT given BTC for other variables, we continue seeing well-behaved curves.



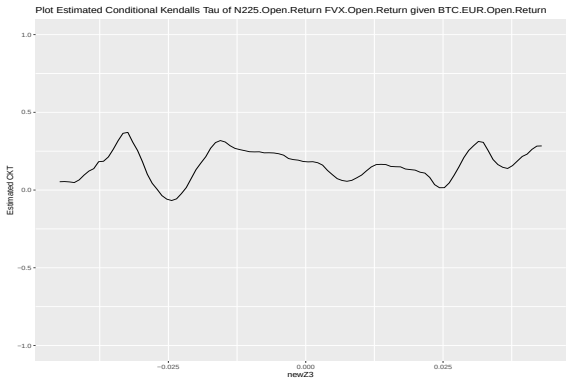
(a) CKT of FCHI and Eurostoxx given BTC



(b) CKT of IXIC and Brent given BTC



(c) CKT of AEX and GDAXI given BTC

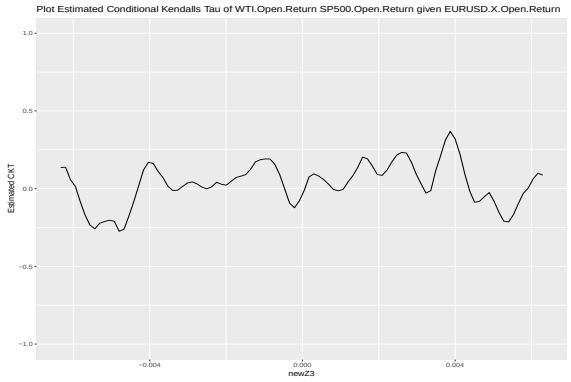


(d) CKT of N225 and FVX given BTC

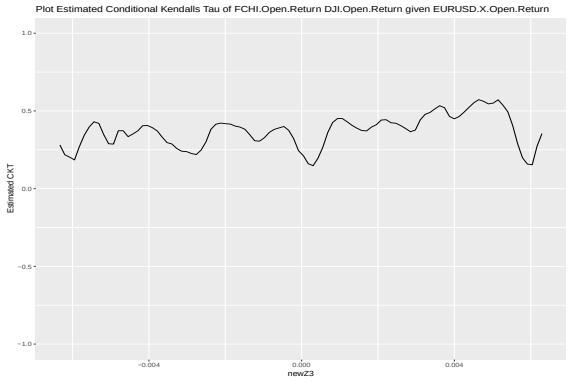
Figure A.12: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given BTC. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable BTC.

1EUR in USD	EURUSD.X	$h = 0.0005$

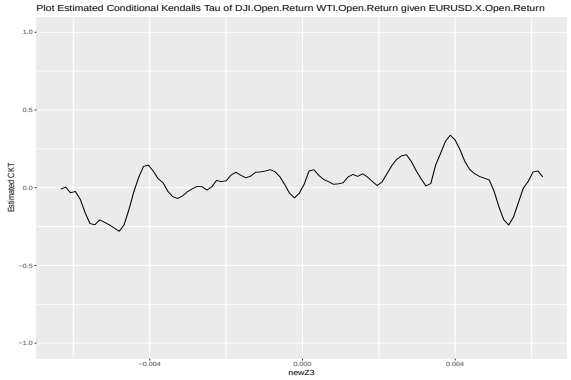
Figure A.13 shows four well-behaved curves. Indeed, they seem stable. Hence, the bandwidth is chosen well for these four combinations of variables. Whenever we inspect plots of the CKT given IXIC for other variables, we continue seeing well-behaved curves.



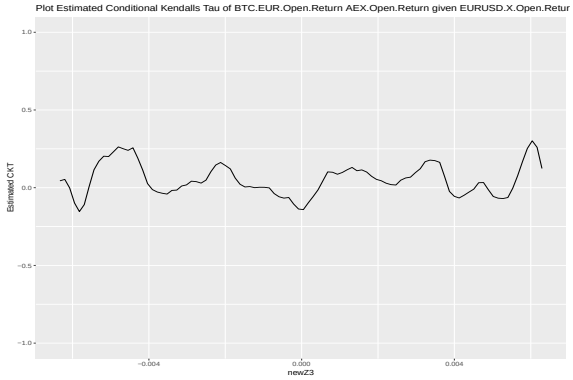
(a) CKT of WTI and SP500 given EURUSD



(b) CKT of FCHI and DJI given EURUSD



(c) CKT of DJI and WTI given EURUSD



(d) CKT of BTC and AEX given EURUSD

Figure A.13: Plots of estimates for the conditional Kendall's tau (CKT) for four different combinations of variables given EURUSD. The y-axis is the interval between $[-1, 1]$ and the x-axis is the sequence between quantiles q_{10} and q_{90} of the data of conditioning variable EURUSD.

B

Code

B.1. R Packages

```
1
2 ## LIBRARIES ESTIMATING CKT ------------------------------
3 library(CondCopulas)
4 library(ggplot2)
5 library(readxl)
6 library(readxl)
7 library(writexl)
8
9 ## LIBRARIES PCA ------------------------------
10 remotes::install_github("vqv/ggbiplot")
11 library(factoextra)
12 library(tidyverse)
13 library(dplyr)
14 library(tidyr)
15 library(ggfortify)
16 library(cluster)
17 library(corrplot)
18 library(ggforce)
19 ggforce::facet_wrap_paginate
20
21 ## LIBRARIES CLUSTERING ----------------------------
22 library(igraph)
```

B.2. Data

Importing the data

```
##DATA ----------------------------
df <- read_excel("C:/Users/jlvla/Desktop/TU Delft/BEP/03. Data/data_
   financial_market.xlsx")
financial_assets_separate <- colnames(df)
useful_assets <- df[, endsWith(financial_assets_separate, ".Open.Return")]
useful_assets_NA = sapply(useful_assets, as.numeric)
useful_assets_noNA = na.omit(useful_assets_NA)
df_useful <- data.frame(useful_assets_noNA)
index <- df$Index
```

Creating subsets for each conditioning variable

```
##CREATING DATA SET FOR EACH CONDITIONING VARIABLE-----------------

l2 = list()
l2_pca = list()

for (asset in 1:length(list_of_financial_assets)) {
  name_asset = list_of_financial_assets[asset]
  newZ3 = seq(quantile(df_useful[, name_asset], probs = 0.1),
              quantile(df_useful[, name_asset], probs = 0.9),
              length.out = 100)

  temp_df = read.csv(
    paste0(
      "C:/Users/jlvla/Desktop/TU Delft/BEP/12. Data2/CKT_conditioning_on_
         test3_",
      name_asset,
      ".csv"), header = F)

  l2[[name_asset]]  <- temp_df

  names_df <- c("V1", newZ3)
  colnames(l2[[name_asset]]) <- names_df

  l2[[name_asset]] = l2[[name_asset]] %>%
    separate(
      col = "V1",
      into = c("nameX1", "nameX2", "nameZ"),
      sep = "_")

  l2_pca[[name_asset]] <-
    prcomp(l2[[name_asset]][, -(1:3)], scale. = TRUE)
```

```
  summary(l_pca[[name_asset]])

}
```

B.3. Estimation of the conditional Kendall's tau

```
for (index1 in 1:(number_assets - 1)) {
  for (index2 in (index1 + 1):number_assets) {
    setindex3 = setdiff(1:number_assets, c(index1, index2))

    for (index3 in setindex3) {
      name1 = list_of_financial_assets[index1]
      name2 = list_of_financial_assets[index2]
      name3 = list_of_financial_assets[index3]
      newZ3 = seq(
        quantile(df_useful[, name3], probs = 0.1),
        quantile(df_useful[, name3], probs = 0.9),
        length.out = 100
      )
      hlist = c(
        0.0009,
        0.0009,
        0.0007,
        0.0005,
        0.005,
        0.0009,
        0.0009,
        0.0007,
        0.0005,
        0.005,
        0.001,
        0.001,
        0.001
      )


      estimatedCKT_kernel_1 <- CKT.kernel(
        observedX1 = df_useful[, name1],
        observedX2 = df_useful[, name2],
        observedZ = df_useful[, name3],
        newZ = newZ3,
        h = hlist[index3],
        kernel.name = "Epa"
      )$estimatedCKT
```

```

      for (asset in list_of_financial_assets) {
        if (asset == name3) {
          estimatedCKT_kernel_1 <- CKT.kernel(
            observedX1 = df_useful[, name1],
            observedX2 = df_useful[, name2],
            observedZ = df_useful[, name3],
            newZ = newZ3,
            h = hlist[index3],
            kernel.name = "Epa"
          )$estimatedCKT

          toWrite1 = paste(name1, name2, asset, sep = '_')
          toWrite2 = paste(estimatedCKT_kernel_1, sep = ',')

          matrixvalues <- matrix(toWrite2, nrow = 1)
          colnames(matrixvalues) <- newZ3
          names_datavalues = cbind(toWrite1, matrixvalues)

          write.table(
            x = names_datavalues,
            file = paste0(
              "C:/Users/jlvla/Desktop/TU Delft/BEP/12. Data2/",
              paste("CKT_conditioning_on_test3", asset, sep = '_'),
              ".csv"
            ),
            sep = ',',
            append = TRUE,
            row.names = FALSE,
            col.names = FALSE
          )


        }
      }
    }
  }
}
```

B.4. MinMax

```
## MIN/MAX -----------------

df_min_findata <- apply(df_useful, 2, FUN = min, na.rm = TRUE)
```

```
df_max_findata <- apply(df_useful, 2, FUN = max, na.rm = TRUE)
df_minmax_findata <- cbind(df_min_findata, df_max_findata)

ggplot(df_minmax_findata, aes(x = df_min_findata, y = df_max_findata)) +
  geom_point()


##PREPARING MIN/MAX applied to PCA--------------------------

df_min <- apply(df_pca_2[,-1], 1, FUN = min, na.rm = TRUE)
df_max <- apply(df_pca_2[,-1], 1, FUN = max, na.rm = TRUE)
df_names_pca <- df_pca[1:3]
df_minmax_pca <- cbind(df_names_pca, df_min, df_max)


ggplot(df_minmax_pca, aes(x = df_min, y = df_max, colour = nameZ)) +
  geom_point() +
  xlim(-1,1) +
  ylim(-1,1) +
  ggtitle("Minimum Maximum Plot of the principal components clustered by
      colouring the conditioning variables") +
  xlab("Minimum") + ylab("Maximum") +
  # geom_abline(slope=1, intercept=0) +
  # geom_abline(slope=-1, intercept=0)
```

B.5. Principal Component Analysis

PRCOMP

```
df_pca <- read.csv(file = "C:/Users/jlvla/Desktop/TU Delft/BEP/12. Data2/
    001. CKT_ALL_NAMES.csv",
                     header = F)

df_pca[,1] <- str_replace_all(df_pca$V1, ".Open.Return", "")

Useful_assets_pca <- prcomp(df_pca[,-1], scale. = TRUE)
Useful_assets_pca_2 <- prcomp(df_pca_2[,-1], scale. = TRUE)
row.names(df_pca_2) <- df_pca_2[,1]
summary(Useful_assets_pca)

df_pca = df_pca %>%
  separate(col = "V1",
           into = c("nameX1", "nameX2", "nameZ"),
           sep = "_")
```

Dataframe and plot using pivot_longer

```
df_pivot = df_pca %>%
  pivot_longer(cols = starts_with("V"),
               names_to = "cond_value",
               values_to = "CKT") %>%
  mutate(cond_value = as.numeric( substring(cond_value, 2) ) - 1 ,
         id = paste(nameX1, nameX2, nameZ, sep = "_"),
         cond_value_2 = newZ3[cond_value])

ggplot(df_pivot, aes(x = cond_value, y = CKT, group = id, color = clust))
   +
  geom_line(alpha = 0.1) +
  ylim(-1,1) +
  xlim(-1,1) +
  ggtitle("Plot of all conditional Kendall's tau estimates for the
     complete dataset") +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100),
                     labels = c("q10", "", "q50", "", "q90")) +
  xlab("Z (quantiles)") +
  ylab("Estimated conditional Kendall's tau")

ggsave(allCKT, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/
   All_CKT.pdf",
       width = 20, height = 12, units = "cm")
```

Plots - complete dataset

```

##00. COLOURING PER CONDITIONING VARIABLE ------------------(2x)

pwhole_cond1 <- autoplot(Useful_assets_pca_2, data = df_pca, colour = "
   nameZ", scale = 0,
         label = TRUE, label.size = 3,
         main = "PCA Plot - First two principal components grouped per
            conditioning variables")
ggsave(pwhole_cond1, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13.
   Pictures2/PCA Whole Data set/ Whole Data Set 1 coloured per
   conditioning variable NO NAMES.pdf",
       height = 12, width = 25, units = "cm")

pwhole_cond2 <- autoplot(Useful_assets_pca_2, data = df_pca, colour = "
   nameZ", scale = 0,
         main = "PCA Plot - First two principal component grouped per
            conditioning variables for complete dataset")+
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = 0, linetype = "dashed")
```

```
ggsave(pwhole_cond2, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PCA Whole Data set/ WholeCondC.pdf",
       height = 12, width = 25, units = "cm")


plot(x = Useful_assets_pca_2$x[,1], y = Useful_assets_pca_2$x[,2])


##01. PC1 and PC2--------------------------(5x)

pwhole_pc1 <- ggplot(Useful_assets_pca_2$rotation, aes(x = indexZ_whole, y = Useful_assets_pca_2$rotation[,1])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC1 for entire financial data set")) +
  xlab("Z") +
  ylab("PC1")
ggsave(pwhole_pc1, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PCA Whole Data set/ PCA 1 Whole Data set.pdf",
      height = 12, width = 20, units = "cm")

pwhole_pc2 <- ggplot(Useful_assets_pca_2$rotation, aes(x = indexZ_whole, y = Useful_assets_pca_2$rotation[,2])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC2 for entire financial data set")) +
  xlab("Z") +
  ylab("PC2")
ggsave(pwhole_pc2, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PCA Whole Data set/ PCA 2 Whole Data set.pdf",
       height = 12, width = 20, units = "cm")

pwhole_pc3 <- ggplot(Useful_assets_pca_2$rotation, aes(x = indexZ_whole, y = Useful_assets_pca_2$rotation[,3])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC3 for entire financial data set")) +
  xlab("Z") +
  ylab("PC3")
ggsave(pwhole_pc3, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PCA Whole Data set/ PCA 3 Whole Data set.pdf",
       height = 12, width = 20, units = "cm")

pwhole_pc4 <- ggplot(Useful_assets_pca_2$rotation, aes(x = indexZ_whole, y = Useful_assets_pca_2$rotation[,4])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC4 for entire financial data set")) +
  xlab("Z") +
```

```
51   ylab("PC4")
52 ggsave(pwhole_pc4, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13.
       Pictures2/PCA Whole Data set/ PCA 4  Whole Data set.pdf",
53         height = 12, width = 20, units = "cm")
54
55 pwhole_pc5 <- ggplot(Useful_assets_pca_2$rotation, aes(x = indexZ_whole, y
       = Useful_assets_pca_2$rotation[,5])) +
56   geom_line() +
57   ggtitle(paste("PCA Plot - PC5 for entire financial data set")) +
58   xlab("Z") +
59   ylab("PC5")
60 ggsave(pwhole_pc5, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13.
       Pictures2/PCA Whole Data set/ PCA 5 Whole Data set.pdf",
61         height = 12, width = 20, units = "cm")
62
63 pwhole_pc6 <- ggplot(Useful_assets_pca_2$rotation, aes(x = indexZ_whole, y
       = Useful_assets_pca_2$rotation[,6])) +
64   geom_line() +
65   ggtitle(paste("PCA Plot - PC6 for entire financial data set")) +
66   xlab("Z") +
67   ylab("PC6")
68
69 ggsave(pwhole_pc6, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13.
       Pictures2/PCA Whole Data set/ PCA 6 Whole Data set.pdf",
70         height = 12, width = 20, units = "cm")
71
72
73 pwhole_combi <- ggarrange(pwhole_pc1, pwhole_pc2, pwhole_pc3, pwhole_pc4,
       pwhole_pc5, pwhole_pc6,
74                           labels = c("A", "B", "C", "D", "E", "F"),
75                           ncol = 2, nrow = 3)
76 ggsave(pwhole_combi, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13.
       Pictures2/PCA Whole Data set/ PCA 1-6 Whole Data set.pdf",
77         height = 15, width = 25, units = "cm")
78
79
80 ##02. PERCENTAGES EACH PC-------------------------------
81
82 pwhole_1 <- fviz_eig(Useful_assets_pca_2, addlabels = TRUE, ncp = 20,
83                 main = paste("PCA Plot - Percentage of explained variance",
84                 "per principal component for entire financial data set",
                        sep = " "),
85                 xlab = "Principal Components")
86 ggsave(pwhole_1, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2
       /PCA Whole Data set/ Whole Data set - Percentage each PC 1-20.pdf",
```

```
87        height = 12, width = 20, units = "cm")
88
89
90 ##03. Individuals-------------------------(2x)
91
92 pwhole_2 <- fviz_pca_ind(Useful_assets_pca_2, repel = F) +
93   geom_point(colour = 'red') +
94   ggtitle(paste("PCA Plot - Individuals with name for entire financial
       data set")) +
95   xlab("Principal Component 1") +
96   ylab("Principal Component 2")
97 # xlim(-35, 20) +
98 # ylim (-5,5)
99 ggsave(pwhole_2, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2
     /PCA Whole Data set/ Whole Data set - Individuals WITH name.pdf",
100        height = 12, width = 20, units = "cm")
101
102 pwhole_3 <- fviz_pca_ind(Useful_assets_pca_2 , repel = F, geom = "point")
     +
103   ggtitle(paste("PCA Plot - Individuals with name for entire financial
       data set")) +
104   xlab("Principal Component 1") +
105   ylab("Principal Component 2")
106 ggsave(pwhole_3, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2
     /PCA Whole Data set/ Whole Data set - Individuals WITHOUT names.pdf",
107        height = 12, width = 20, units = "cm")
108
109
110
111 ##07. COS2-----------------------------------------------(4x)
112
113 pwhole_7 <- fviz_pca_ind(Useful_assets_pca_2, col.ind="cos2", geom = c("
     point")) +
114   scale_color_gradient2(low="white", mid="cyan",
115                         high="red", midpoint = 0.4) +
116   ggtitle(paste("PCA Plot - Contribution cos^2 for the complete dataset"))
       +
117   xlab("Principal Component 1") +
118   ylab("Principal Component 2")
119 ggsave(pwhole_7, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2
     /PCA Whole Data set/ contributioncos2.pdf",
120        height = 12, width = 20, units = "cm")
121
122
123 ##08. CLUSTERING-----------------------------------(4x)
```

```
124
125 pwhole_8 <- fviz_cluster(pam(df_pca_2[,-1], 4), ellipse.type = "convex",
        main = "Cluster plot", geom = "point") +
126   ggtitle("PCA Plot - Clustering for entire financial data set") +
127   # xlab("Principal Component 1") +
128   # ylab("Principal Component 2") +
129   geom_hline(yintercept = 0, linetype = "dashed") +
130   geom_vline(xintercept = 0, linetype = "dashed")
131 ggsave(pwhole_8, file = "C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2
        /PCA Whole Data set/ Whole Data set - Clustering.pdf",
132        height = 12, width = 20, units = "cm")
```

Plots per conditioning variable

```
1
2 ##PCA Plots Per Conditioning variable-----------------------------
3
4 for (asset in 1:length(list_of_financial_assets)){
5
6   name_asset = list_of_financial_assets[asset]
7   indexZ = seq(quantile(df_useful[,name_asset], probs = 0.1),
8                quantile(df_useful[,name_asset], probs = 0.9), length.out =
                     100)
9
10  #   ##Indivuduals--------------------------
11  # autoplot(l2_pca[[name_asset]], label = TRUE, label.size = 3,
12  #          main = paste("PCA Plot - First two principal component for
        data conditioned on", name_asset, sep = " "))
13
14
15  ##01. PC1 and PC2----------------------(5x)
16  p_pc1 <- ggplot(l2_pca[[name_asset]]$rotation, aes(x = indexZ, y = l2_
        pca[[name_asset]]$rotation[,1])) +
17    geom_line() +
18    ggtitle(paste("PCA Plot - PC1 for", name_asset, sep = " ")) +
19    xlab("Z") +
20    ylab("PC1")
21  ggsave(p_pc1, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
        Pictures2/PLOTS on cond variable ",
22                              name_asset, "/ Principal Component 1 for
                                    values of Z for data of conditioning
                                    variable ", name_asset, ".pdf", sep = "")
                                    ))
23
24
25  p_pc2 <- ggplot(l2_pca[[name_asset]]$rotation, aes(x = indexZ, y = l2_
```

```
    pca[[name_asset]]$rotation[,2])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC2 for", name_asset, sep = " ")) +
  xlab("Z") +
  ylab("PC2")
ggsave(p_pc2, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
    Pictures2/PLOTS on cond variable ",
                            name_asset, "/ Principal Component 2 for
                                values of Z for data of conditioning
                                variable ", name_asset, ".pdf", sep = "")
                                ),
       height = 12, width = 20, units = "cm")

p_pc3 <- ggplot(l2_pca[[name_asset]]$rotation, aes(x = indexZ, y = l2_
    pca[[name_asset]]$rotation[,3])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC3 for", name_asset, sep = " ")) +
  xlab("Z") +
  ylab("PC2")
ggsave(p_pc3, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
    Pictures2/PLOTS on cond variable ",
                            name_asset, "/ Principal Component 3 for
                                values of Z for data of conditioning
                                variable ", name_asset, ".pdf", sep = "")
                                ),
       height = 12, width = 20, units = "cm")

p_pc4 <- ggplot(l2_pca[[name_asset]]$rotation, aes(x = indexZ, y = l2_
    pca[[name_asset]]$rotation[,4])) +
  geom_line() +
  ggtitle(paste("PCA Plot - PC4 for", name_asset, sep = " ")) +
  xlab("Z") +
  ylab("PC2")
ggsave(p_pc4, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
    Pictures2/PLOTS on cond variable ",
                                name_asset, "/ Principal Component 4
                                    for values of Z for data of
                                    conditioning variable ", name_
                                    asset, ".pdf", sep = "")),
       height = 15, width = 20, units = "cm")


p_combi <- ggarrange(p_pc1, p_pc2, p_pc3, p_pc4,
                     labels = c("A", "B", "C", "D"),
                     ncol = 2, nrow = 2)
```

```
ggsave(p_combi, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PLOTS on cond variable ",
                                    name_asset, "/ PC1-4 in one figure for data of conditioning variable ", name_asset, ".pdf", sep = "")),
       height = 10, width = 30, units = "cm")

##02. Screeplot------------------------(1x)
p1 <- fviz_eig(l2_pca[[name_asset]], addlabels = TRUE, ncp = 10,
               main = paste("PCA Plot - Percentage of explained variance per principal component for data of conditioning variable ", name_asset, sep = " "),
               xlab = "Principal Components")
ggsave(p1, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PLOTS on cond variable ",
                                name_asset, "/ PCA Plot Percentage PC - conditioning on ", name_asset, ".pdf", sep = "")),
       height = 12, width = 20, units = "cm")

##03. Individuals------------------------(2x)

p2 <- fviz_pca_ind(l2_pca[[name_asset]], repel = F, geom = "text") +
  geom_point(colour = 'red') +
  ggtitle(paste("PCA Plot - Individuals for data of conditioning variable", name_asset, sep = " ")) +
  xlab("Principal Component 1") +
  ylab("Principal Component 2")
# xlim(-35, 20) +
# ylim (-5,5)
ggsave(p2, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PLOTS on cond variable ",
                         name_asset, "/ PCA Plot - Individuals - conditioning on ", name_asset, ".pdf", sep = "")),
       height = 12, width = 20, units = "cm")

p3 <- fviz_pca_ind(l2_pca[[name_asset]], repel = F, geom = "point") +
  ggtitle(paste("PCA Plot - Individuals for data of conditioning variable", name_asset, sep = " ")) +
  xlab("Principal Component 1") +
  ylab("Principal Component 2")
ggsave(p3, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13. Pictures2/PLOTS on cond variable ",
```

```
86                                  name_asset, "/ PCA Plot Individuals NO
                                        NAME - conditioning on ", name_asset,
                                         ".pdf", sep = "")),
87          height = 12, width = 20, units = "cm")
88
89
90   05. Contributions-----------------------(4x)
91  # var <- get_pca_var(Useful_assets_pca)
92  # corrplot(var2$contrib, is.corr=FALSE)
93
94
95  p4 <- fviz_contrib(l2_pca[[name_asset]], choice = "ind", axes = 1:2, top
        = 15) +
96    ggtitle(paste("PCA Plot - Contribution to PC1 and PC2 for data of
          conditioning variable", name_asset, sep = " "))
97  ggsave(p4, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
      Pictures2/PLOTS on cond variable ",
98                           name_asset, "/ PCA Plot - Contribution
                                Histogram - conditioning on ", name_asset, "
                                .pdf", sep = "")),
99          height = 12, width = 20, units = "cm")
100
101
102
103 p5 <- fviz_pca_ind(l2_pca[[name_asset]], col.ind="contrib",  repel = F,
      geom = "point") +
104   scale_color_gradient2(low="white", mid="cyan",
105                         high="red") +
106   ggtitle(paste("PCA Plot - Contribution for data of conditioning
          variable", name_asset, sep = " ")) +
107   xlab("Principal Component 1") +
108   ylab("Principal Component 2")
109 ggsave(p5, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
      Pictures2/PLOTS on cond variable ",
110                           name_asset, "/ PCA Plot - Contribution Names -
                                conditioning on ", name_asset, ".pdf", sep =
                                 "")),
111         height = 12, width = 20, units = "cm")
112
113 p5v2 <- fviz_pca_ind(l2_pca[[name_asset]], col.ind="contrib",  repel = T
      ) +
114   scale_color_gradient2(low="white", mid="cyan",
115                         high="red", midpoint=1) +
116   ggtitle(paste("PCA Plot - Contribution for data of conditioning
          variable", name_asset, sep = " ")) +
```

```
117     xlab("Principal Component 1") +
118     ylab("Principal Component 2")
119   ggsave(p5v2, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
        Pictures2/PLOTS on cond variable ",
120                                name_asset, "/ PCA Plot - Contribution
                                       Names (No overlap) - conditioning
                                      on ", name_asset, ".pdf", sep = "")
                                      ),
121           height = 12, width = 20, units = "cm")
122
123   p6 <- fviz_pca_ind(l2_pca[[name_asset]], col.ind="contrib", geom = "
        point") +
124     scale_color_gradient2(low="white", mid="cyan",
125                           high="red", midpoint=1) +
126     ggtitle(paste("PCA Plot - Contribution for data of conditioning
          variable", name_asset, sep = " ")) +
127     xlab("Principal Component 1") +
128     ylab("Principal Component 2")
129   ggsave(p6, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
        Pictures2/PLOTS on cond variable ",
130                                name_asset, "/ PCA Plot - Contribution
                                    Scatter - conditioning on ", name_
                                    asset, ".pdf", sep = "")),
131           height = 12, width = 20, units = "cm")
132
133
134
135   ##07. COS2-----------------------(1x)
136   p7 <- fviz_pca_ind(l2_pca[[name_asset]], col.ind="cos2", geom = c("point
        ")) +
137     ggtitle(paste("PCA Plot - Representation using cos^2 for data of
          conditioning variable", name_asset, sep = " ")) +
138     xlab("Principal Component 1") +
139     ylab("Principal Component 2") +
140     scale_color_gradient2(low="white", mid="cyan",
141                           high="red", midpoint = 0.4)
142   ggsave(p7, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
        Pictures2/PLOTS on cond variable ",
143                                name_asset, "/ PCA Plot - COS2 -
                                    conditioning on ", name_asset, ".pdf"
                                    , sep = "")),
144           height = 12, width = 20, units = "cm")
145
146
147   ##08. Clustering-----------------(1x)
```

```
148
149   p8 <- fviz_cluster(pam(l2[[name_asset]][,-(1:3)], 3), ellipse.type = "
          convex", main = "Cluster plot", geom = "point") +
150     ggtitle(paste("PCA Plot - Clustering for data of conditioning variable
            ", name_asset, sep = " ")) +
151     xlab("Principal Component 1") +
152     ylab("Principal Component 2") +
153     geom_hline(yintercept = 0, linetype = "dashed") +
154     geom_vline(xintercept = 0, linetype = "dashed")
155   ggsave(p8, file = (paste("C:/Users/jlvla/Desktop/TU Delft/BEP/13.
          Pictures2/PLOTS on cond variable ",
156                                    name_asset, "/ PCA Plot - Clustering -
                                        conditioning on ", name_asset, ".pdf"
                                        , sep = "")),
157          height = 12, width = 20, units = "cm")
158
159
160 }
```

B.6. Clustering

```
1  ##INFORMATION CLUSTERING COMPLETE DATASET ----------------------
2  dataclust1 = pwhole_8$data %>%
3    separate(col = "name",
4             into = c("nameX1", "nameX2", "nameZ"),
5             sep = "_") %>%
6    mutate(nameX1X2 = paste(nameX1, nameX2, sep = "_")) %>%
7    filter(cluster == 3)
8
9
10 ##INFORMATION CLUSTERING SUBSET CONDITIONING VARIABLE DJI
       -------------------
11 dataclust_DJI = p8$data %>%
12   separate(
13     col = "name",
14     into = c("nameX1", "nameX2", "nameZ"),
15     sep = "_") %>%
16   mutate(nameX1X2 = paste(nameX1, nameX2, sep = "_")) %>%
17   filter(cluster == 1)
18
19 tables_couples = table(dataclust_DJI$nameX1X2)
20 df_tables_couples = data.frame(nameX1X2 = names(tables_couples),
21                                nlinks = as.numeric(tables_couples))  %>%
22   separate(col = "nameX1X2",
```

```
            into = c("nameX1", "nameX2"),
            sep = "_")


mygraph = graph_from_data_frame(df_tables_couples,
                                directed = FALSE)

plot(mygraph, main = "Connected graph of variables in cluster 1 for subset
    conditioning DJI")
```