Forecasting commodity futures using Principal Component Analysis and Copula

Martin Jacobsson

May 20, 2015

Abstract

The ever ongoing battle to beat the market is in this thesis fought with the help of mathematics with a way to reduce the information to its core. It is called PCA, Principal Component Analysis. This is used to build a model of future commodity prices. To assist PCA, Copula is used - a sort of mathematical glue which can bring multiple distributions together and represented as one.

The data used is 5 years of prices for Brent Oil, WTI Oil, Gold, Copper and Aluminium. The model parameters are fitted to 2.5 years of data and then tested on the remaining 2.5 years.

MLE, Maximum Likelihood Estimation, was used for parameter estimation and distributions that were found fitting were logistic and Student's T distribution

Cramér-von Mises tests were used to determine that T Copula was the most suitable Copula.

The main results are that the mathematical estimations fit well and profit can be generated, but with a low Sharpe Ratio.

Keywords: PCA, Copula, Mean-reversion, Momentum, Elliptical copulas, Maximum Likelihood, Cramér-von Mises, Sharpe Ratio.

Acknowledgements

I would like to thank my supervisor Associate Professor Nader Tajvidi at Lund University and my supervisor Thomas Lyse Hansen at Nordea for all their help and guidance. I also want to extend my gratitude to my family and friends for all their support. Finally, my special thanks goes to Viking Jacobsson for all the invaluable help.

Contents

Chapter 1

Introduction

Many financial traders around the world struggle with the same question: how can I beat the market? If one could come up with one easy way to do this, their financial problems would forever be gone. An approach to model the reality is in this thesis done with the help of PCA and Copula - two mathematical tools.

1.1 Purpose

The purpose of this Master's Thesis is to see if the prediction model presented will be able to generate a positive risk-adjusted absolute return or not. To measure this, Sharpe Ratio is used. Since Sharpe Ratio includes the risk-free rate r_f, we set $r_f = 0$ due to our only interest in absolute risk-adjusted return.

1.2 The commodity market

The commodity market is one of the oldest (if not the oldest) and most fundamental markets in the world. One could say that the commodity market was the start of our civilized society [Banerjee, 2013]. It was there that people first learned to trade with their own specialized good in order to procure another good which they needed. The first commodity trading activities stretch back to the ancient Sumerian civilization between 4,500 BC and 4,000 BC. They then used clay tokens to represent the number of goods to be delivered, for example the number of goats. These clay tokens were then sealed in a vessel to represent the promise they had made to deliver x number of goats.

Nowadays, the commodity market is a vast system of different markets all over the world where a future delivery of gold or corn just are some clicks on your computer away. Naturally, the greatest price factor is supply and demand, but this thesis will try to investigate if one can use mathematical tools to forecast the price movements.

1.3 Futures contract

To procure the wanted commodity, the most used way is to use so called futures. A futures contract is a standardized contract between two parties, which settles for a specified asset with a certain price today but with a future specified delivery and payment date [Hull, 2009]. Due to a default risk, both parties are required to put up a margin - the initial amount of cash. During a change in the futures price, these margins are transferred between the parties, generally once a day. This means that all profits and losses are settled continuously so at the delivery date, the exchanged amount is the spot price(the price for getting something right away) [Hull, 2009] of the underlying asset. Consequently, to take a position in a futures contract is free, excluded transaction costs such as brokerage fees.

1.4 Overview

In the forecasting model Principal Component Analysis and Copula will be used to generate buy and sell signals. These signals will then be used in different types of trading strategies where profit can be generated as the price goes up as well as when the price goes down. A total Sharpe ratio will be calculated of these strategies to evaluate the performance.

1.5 Hypothesis

The hypothesis is that the presented forecasting model will be able to generate a positive risk-adjusted return.

Chapter 2

Theory

In this chapter all the necessary theoretical background for the prediction model is presented. The chapter starts out with theory regarding log returns and PCA. Thereafter the main trading strategies are introduced - mean reversion and momentum strategies. Then more mathematical theory is presented; Copula, Elliptical Copulas and distributions. Moreover, Dependence Theory, Parameter Estimation and Goodness of Fit test are presented. Lastly, we will introduce the concept of Sharpe Ratio which is a risk adjusted performance measure.

2.1 Log returns

In this thesis log returns are used which have several advantages. Below are the main two reasons:

- There is a normalization of the variables, which means that all returns are in a comparable metric.
- Log returns are time additive, which means that to get the n-period return - we can simply add all the single period returns up to n.

To calculate the log returns the following equation is used.

$$r_{i,t} = ln\frac{P_{i,t}}{P_{i,t-1}}, i = 1, ..., n \tag{2.1}$$

where $P_{i,t}$ is the price of the futures contract - i indicates which of the commodities used at a given time t.

2.2 Principal Components Analysis (PCA)

To introduce Principal Components Analysis, we take the following excerpt from [Jolliffe, 2002]:

"The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PC's), which are uncorrelated and which are ordered so that the first few retain most of the variation present in all of the original variables."

With this concise explanation as a start, we are now ready for the definition of PCA.

***Definition* 1.** Suppose that x is a vector of p random variables. The first step is to look for a linear function $\alpha_1' x$ of the elements of x having a maximum variance, where α_1 is a vector of p constants $\alpha_{11}, ..., \alpha_{12}, \alpha_{1p}$, and "$'$" denotes transpose, so that

$$\alpha_1' x = \alpha_{11} x_1 + \alpha_{12} x_2 + ... + \alpha_{1p} x_p = \sum_{j=1}^{p} \alpha_{1j} x_j \tag{2.2}$$

Then, we look for a linear function $\alpha_2' x$, uncorrelated with $\alpha_1' x$ having maximum variance, and so on, so that the k-th stage a linear function $\alpha_k' x$ is found that has maximum variance subject to being uncorrelated with $\alpha_1' x, \alpha_2' x, ..., \alpha_{k-1}' x$. The k-th derived variable, $\alpha_k' x$ is the k-th PC and p is the number of commodities.

So, depending on what we choose k to be - we get k number of PC's. In this case we set the x in the above definition as y - the de-meaned log returns.

$$y = x - x_{mean} \tag{2.3}$$

where x_{mean} is the mean of x.

When we then project these vectors back on to our data, we get the D-matrix.

$$D = A^T \cdot y \tag{2.4}$$

where $A = (a_1, ..., a_k)$

We will later use this matrix D, which is the matrix of dimensionally reduced returns, obtained after projecting y in the principal component space.

2.3 Mean-reverting Theory

In this trading model, two main trading strategies are used. The first is called mean-reversion strategy [Investopedia, 2015] and is simply the theory that prices should move back towards their moving average(their mean calculated a constant x days back). As [Infantino et al., 2010] discusses this is perhaps the most simple of all trading strategies but it does not take the behavioural aspect of trading into account. The mean reversion theory is used in this thesis as a foundation of the PCA. Our PCA gives us a first model which tells us how the returns without noise should have been the last period. We then act accordingly - if the model tells us that the prices are too high, we sell and vice versa. Studies that mean reversion theory works and actually generates alpha in commodity prices are discussed in [Lutz, 2010] which is taken into account in this prediction model.
Main input parameters into our model are; the T number of days we are looking back at, the H future days of returns, the k number of principal components.

The following formula is the foundation of the mean reverting section.

$$r_{t+1} + ... + r_{t+H} = \beta_1 \sum_{i=0}^{H} D_{t-i,1} + ... + \beta_k \sum_{i=0}^{H} D_{t-i,k} \tag{2.5}$$

where r_t is the log return at time t, β is the factor that explains the ratio between the log returns and the matrix D. That is, the β values want to explain the connection between past periods PC's and coming periods returns.

We define a matrix B which includes all the β's with dimensions $k\times$ number of commodities.

We also define the matrix $\tilde{D}_t$ - the sum of the returns from the D-matrix:

$$\tilde{D}_t = \sum_{i=0}^{H-1} D_{t-i},\ t = (1, 2, ..., T) \tag{2.6}$$

Finally, we get our prediction of the future log return S as,

$$S_t = \tilde{D}_t \cdot B \tag{2.7}$$

So if this S is higher than the actual log return, we sell and vice versa. We get a long list of buy or sell signals, and the momentum strategy stated below can overwrite those signals.

2.4 Momentum

In this model, the behavioural aspect of trading is represented by a momentum strategy, which is the belief that if a price is in an upward trend, it will continue going up. Likewise if the price is falling, momentum strategies states that it will continue falling. In short, you could say that it is a "ride-the-wave" type of trading mindset - contradictory to the mean-reversion theory. In [Jegadeesh et al., 1999] it is shown that financial instruments with strong past performance continue to outperform those with poor past performance. This momentum theory is implemented into our prediction model.

To introduce the momentum strategy into our model, we use [Infantino et al., 2010] "the Cross Sectional Volatility of the Principal Components", which is the standard deviation $\sigma_D(t)$ of the returns, projected onto the principal components. Briefly you could say that we want to look at if the rate of change in the discrete time (Euclidean distance E_H - defined below) grows. More specifically:

$$\sigma_D(t) = \sqrt{\sum_{j=1}^{k} \frac{(d_{tj} - \hat{d}_t)^2}{k-1}},\ t = (1, 2, ..., T) \tag{2.8}$$

where k is the number of principal components, d_{tj} is the reduced-dimensionality return j at time t and $\hat{d}_t$ is the cross sectional mean:

$$\hat{d}_t = \frac{1}{k} \sum_{j=1}^{k} d_{tj} \tag{2.9}$$

We want to look at the changes in time, the derivative, to see if there are great changes in the standard deviation. To decide if there are "great

changes" we define a measure ψ for the change of the standard deviation compared to time.

$$\psi = \frac{d\sigma_D}{dt} \tag{2.10}$$

To get a distance measure we define E_H:

$$E_H(t) = \sqrt{\sum_{i}^{H} [\psi(t-i)]^2} \tag{2.11}$$

Finally, to decide if we are in a "momentum" or not and to decide if we should override the mean-reverting signal, we check:

$$E_H(t) - E_H(t-1) \tag{2.12}$$

So if this is less than or equal to zero, we continue with the mean-reverting signal. Else, we switch so that we apply the momentum strategy and continue "riding-the-wave".

2.5 Copula

To be able to study all our fitted distributions of our commodity prices, one can use copula as a glue for these distributions. This is more formally described below, from [Nelsen, 2006] first for a bivariate copula and then for the n-dimensional case.

***Definition* 2.** A two-dimensional subcopula is a function C' with following properties:

- $DomC' = S_1 \times S_2$, where S_1 and S_2 are subsets of $\mathbf{I}$ containing 0 and 1.
- C' is grounded and 2-increasing.
- For every u in S_1 and and every v in S_2,

$$C'(u,1) = u \; and \; C'(1,v) = v. \tag{2.13}$$

***Definition* 3.** A two-dimensional copula is a 2-subcopula C whose domain is $\mathbf{I}^2$.

Equivalently, a copula is a function C from $\mathbf{I}^2$ to $\mathbf{I}$ with the following properties:

- For every u, v in $\mathbf{I}$,

$$C(u,0) = 0 = C(0,u),$$
$$C(u,1) = u, C(1,v) = v.$$

- For every u_1, u_2, v_1 and v_2 in $\mathbf{I}$ such that $u_1 \leq u_2$, $v_1 \leq v_2$

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0. \tag{2.14}$$

  This is called the rectangle inequality of copula.

From [Nelsen, 2006] we find the theorems regarding Fréchet-Hoeffding bounds.

Theorem 1 (Fréchet-Hoeffding bounds in 2-dimensions). *Let C' be a subcopula. Then for every (u,v) in $DomC'$*

$$max(u+v-1,0) \leq C'(u,v) \leq min(u,v) \tag{2.15}$$

The n-dimensional case follows from [Nelsen, 2006]

Theorem 2 (Fréchet-Hoeffding bounds in n-dimensions). *Let C be a copula, then the following inequality is satisfied,*

$$max(u_1 + u_2 + ... + u_n - d + 1, 0) \leq C(\mathbf{u}) \leq min(u_1, u_2, ..., u_n). \tag{2.16}$$

where $\mathbf{u} \in [0,1]^n$

Theorem 3 (Sklar's theorem).*Let H be an n-dimensional distribution function with margins $F_1, F_2, ..., F_n$. Then there exists a n-copula C such that for all x in $\mathbb{R}^n$,*

$$H(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)). \tag{2.17}$$

If $F_1, F_2, ..., F_n$ are all continuous, then C is unique; otherwise C is uniquely determined on Ran $F_1\times$ Ran $F_2\times$ ... $\times$ Ran F_n. Conversely, if C is a n-copula and $F_1, F_2, ..., F_n$ are distribution functions, then the function H defined by (2.16) *is a n-dimensional distribution function with margins $F_1, F_2, ..., F_n$.*

2.6 Elliptical Copulas

One can describe the elliptical copulas (named so for their elliptical contour shape) as the most basic of copulas. One advantage that is used in this thesis is that they can handle both positive and negative dependence.

To describe the dependence structure for elliptical copulas, the notation of Σ is used to represent the correlation matrix. Elliptical copulas have the following relationship between its dependence parameter and Kendall's Tau.

$$\tau_K = \frac{2}{\pi} arcsin\rho \tag{2.18}$$

where ρ is the corresponding "off-diagonal" parameter of dependence in Σ.

2.6.1 Gaussian Copula

As defined in [Bouyé, 2000] the definition for multivariate gaussian copula (MVN) is the following.

***Definition* 4.**
Let ρ be a symmetric, positive definite matrix with $diag\ \rho = 1$ and Φ_ρ the standardized multivariate normal distribution with correlation matrix ρ. The multivariate gaussian copula is the defined as follows:

$$\mathbf{C}(u_1, ..., u_n, ..., u_N; \rho) = \Phi_\rho(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n), ..., \Phi^{-1}(u_N)) \tag{2.19}$$

The density of the gaussian copula is then defined as follows.

***Definition* 5.**

$$\mathbf{c}(u_1, ..., u_n, ..., u_N; \rho) = \frac{1}{\mid \rho \mid^{\frac{1}{2}}} exp^{(-\frac{1}{2}\zeta^T(\rho^{-1}-\mathbb{I})\zeta)} \tag{2.20}$$

where $\zeta_n = \Phi^{-1}(u_n)$

2.6.2 Student's T Copula

Continuing with the writings from [Bouyé, 2000], the multivariate Student's T Copula (MVT) is defined as follows.

***Definition* 6.**

$$\mathbf{C}(u_1, ..., u_n, ..., u_N; \rho, \nu) = T_{\rho,\nu}({t_\nu}^{-1}(u_1), ..., {t_\nu}^{-1}(u_n), ..., {t_\nu}^{-1}(u_N)) \tag{2.21}$$

with ${t_\nu}^{-1}$ as the inverse of the univariate Student's T distribution.

Corresponding density is

***Definition* 7.**

$$\mathbf{c}(u_1, ..., u_n, ..., u_N; \rho, \nu) = \mid \rho \mid^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+N}{2})[\Gamma\frac{\nu}{2}]^N)(1 + \frac{1}{\nu}\zeta^T\rho^{-1}\zeta)^{-\frac{\nu+N}{2}}}{[\Gamma(\frac{\nu+1}{2})]^N\Gamma(\frac{\nu}{2})\prod_{n=1}^{N}(1 + \frac{{\zeta_n}^2}{\nu})^{-\frac{\nu+1}{2}}} \quad (2.22)$$

2.7 Distributions

Below are the distributions that were fitted to the log returns of the commodities. Both logistic distribution and Student's T distribution are known to have fatter tails than the normal distribution, which suited the log returns.

2.7.1 Generalised Logistic Distribution

In [Shao, 2002] the generalised logistic distribution is described with the following density function.

$$f_{GL1}(x; \theta, \sigma, \alpha) = \frac{\alpha}{\sigma} * \frac{e^{-(x-\theta)/\sigma}}{(1 + e^{-(x-\theta)/\sigma})^{\alpha+1}}, \quad (2.23)$$

where θ is the location parameter, $\sigma > 0$ is the scale and $\alpha > 0$ is the shape parameter.

2.7.2 Student's T Distribution

As described in [Jackman, 2009] the Student's T distribution is defined as follows.

***Definition* 8.**
If x follows a (standardized) Student's T density with $\nu > 0$ degrees of freedom, conventionally written $x \sim t_\nu$, then

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}\ (1 + \frac{x^2}{\nu})^{-(\nu+1)/2} \quad (2.24)$$

and has mean 0 and variance $\nu/(\nu-2)$.
In unstandardised form, the Student's T density is

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sigma\sqrt{\nu\pi}}\ (1+\frac{1}{\nu}(\frac{x-\mu}{\sigma})^2)^{-(\nu+1)/2} \tag{2.25}$$

and is conventionally written $x \sim t_\nu(\mu, \sigma^2)$, where μ is a location parameter, $\sigma > 0$ is a scale parameter and $\nu > 0$ is a degree of freedom parameter.

- The standardized version of the T density is and unstandardised T density with $\mu = 0$ and $\sigma = 1$.
- Provided $\nu > 1$, $E(x) = \mu$ and $V(x) = \frac{\nu}{\nu-2}\sigma^2$.
- As $\nu \to \infty$, $p(x)$ tends to the normal density.

2.8 Dependence Theory

There are different types of dependence measures but the most common is called linear correlation, or more formally - Pearson's correlation coefficient. This is calculated through,

$$\rho_{X,Y} = \frac{cov[X,Y]}{\sqrt{Var[X]Var[Y]}} \tag{2.26}$$

But this dependence measure has three main disadvantages.

- It requires that mean and variance exists, else it is useless.
- It can only measure linear dependence between X and Y, so a modification in X must have a likewise constant proportional modification in Y.
- It is not invariant to non-linear transformations.

So this is called correlation coefficient but when we are talking about scale-invariant measures of dependence we call them measures of association. One of the main measure of association - Kendall's Tau, is presented below (Spearman's Rho is not used in this thesis, hence the lack of definition).

2.8.1 Concordance

To define Kendall's Tau, we need to explain what concordance is, which is presented in [Nelsen, 2006],

Briefly, you could say that a pair of random variables are concordant if large values of one tend to be associated with large values of the other. Likewise, small values of one tend to be associated with small values of the other. This is presented more accurately below.

Let (x_i, y_i) and (x_j, y_j) denote two observations from a vector (X, Y) of continuous random variables. We say that (x_i, y_i) and (x_j, y_j) are *concordant* if $x_i < x_j$ and $y_i < y_j$, or if $x_i > x_j$ and $y_i > y_j$. Similarly, we say that (x_i, y_i) and (x_j, y_j) are *discordant* if $x_i < x_j$ and $y_i > y_j$, or if $x_i > x_j$ and $y_i < y_j$.

2.8.2 Kendall's Tau

We are now ready to define the measure of association Kendall's Tau in terms of concordance and discordance.

***Definition* 9.**
Let $[(x_1, y_1), (x_2, y_2), ..., (x_j, y_j)]$ denote a random sample of n observations from a vector (X,Y) of continuous random variables. There are $\binom{n}{2}$ distinct pairs (x_i, y_i) and (x_j, y_j) of observations in the sample, and each pair is either concordant or discordant. Let c denote the number of concordant pairs and d denote the number of discordant pairs. Then Kendall's Tau is defined as,

$$\tau_K = \frac{c-d}{c+d} = (c-d)/\binom{n}{2} \tag{2.27}$$

2.9 Parameter Estimation

To estimate the parameters of both the margins of commodity prices and the copulas Maximum Likelihood Estimation (MLE) is used.

2.9.1 Maximum Likelihood Estimation

In [Myung, 2003] maximum likelihood estimation is the process of finding a value θ which is the estimation that makes the underlying data as plausible as possible. Generally, let $x_1, ..., x_n$ be i.i.d observations of a random variable X with density function $f(x;\theta)$, where θ is an unknown parameter in the space Ω_Θ. Then, the corresponding likelihood function $L_x(\theta)$ is defined as follows,

***Definition* 10.**

$$L_x(\theta) = \prod_{k=1}^{n} f(x_k;\theta). \tag{2.28}$$

and the MLE is the parameter $\hat{\theta}_{mle}$ which maximizes $L_x(\theta)$. Hence,

$$\hat{\theta}_{mle} = argmax_\theta \; L_x(\theta). \tag{2.29}$$

2.10 Goodness of Fit

To determine if you have found a fitting model or not, several tests can be done. In order to determine if we had fitted our probability distributions well, Quantile-Quantile plots are used. And to test whether a copula C can be represented in a multivariate distribution, the Cramér-von Mises Method is used. More on this below.

2.10.1 Quantile-Quantile Plot

The Quantile-Quantile Plot (QQ-plot called henceforth) is a graphical test method to see whether our data fits to a distribution or not. The empirical quantiles are plotted against the quantiles of the fitted theoretical distribution, the points will lie on the line of a 45 degree slope.

2.10.2 Cramér-von Mises Method

To test the goodnes of fit for a copula, one can use the Cramér-von Mises method. This method [Genest et al., 2009] is originally based on the "empirical copula" as

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} l(U_{i1} \leq u_1, ..., U_{id} \leq u_d), \tag{2.30}$$

where $\mathbf{u} = (u_1, ..., u_d)$ in $[0,1]^d$.

Then, the goodness of fit test process is based on the empirical process

$$\mathbb{C}_n(\mathbf{u}) = \sqrt{n}(C_n - C_{\theta,n}) \tag{2.31}$$

where $C_{\theta,n}$ is an estimator of C, obtained under the null-hypothesis $H_0 : \ C \in C_0$ for a class C_0 of copulas. θ_n is the estimation of θ which is derived from pseudo-obeservations.

Moreover, the test statistic S_n of Cramér-von Mises method can be calculated based on the empirical process to

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\mathbf{u}^2)\, \mathrm{d}C_n(\mathbf{u}). \tag{2.32}$$

where a large value of the test statistic S_n leads to rejection of the null-hypothesis.

2.11 Portfolio Return

Given we have our buy and sell signals, return can be generated both by going long (you buy the asset and generate return when selling if the asset has risen in price) and by going short (you sell the asset and generate return when buying back the asset if the asset has fallen in price). Whilst going long, one unit of the asset is purchased and then sold the next trading day. Whilst going short, one unit of the asset is sold and the bought the next trading day. All of the portfolio returns are calculated in their real value, but the calculations are made with the log returns.

2.12 Sharpe Ratio

Finally, to determine if our prediction model had performed well during our data time frame, we measure the Sharpe Ratio. This is a measure to see if we performed well compared to the risk taken. Defined [Investopedia, 2015] as follows,

$$S_r = \frac{\hat{r}_p - r_f}{\sigma_p} \tag{2.33}$$

where $\hat{r}_p$ is the portfolio return, r_f is the return of the risk free asset and σ_p is the standard deviation of the portfolio.

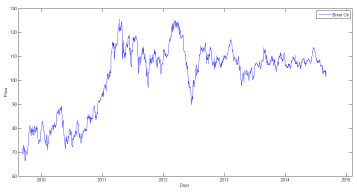
Chapter 3

Implementation and Results

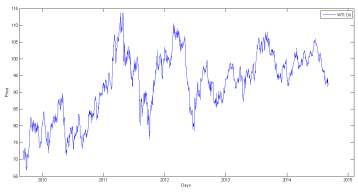
The programs used for implementation of the model were Matlab and R. Packages used in R were *MASS, copula* and *glogis*.

3.1 The Data Set

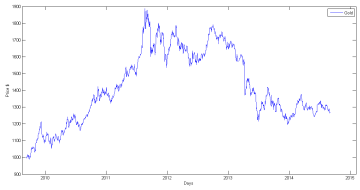
The data set consists of five different commodities; Brent Oil, West Texas Intermediate (WTI) Oil, Gold, Copper and Aluminium. We have five years of old 3-months futures prices stretching from September 2009 to September 2014. It is always the daily closing price that have been used. It should be noted that different commodities have been trading different days, due to holiday days and other occurrences where the market was closed. This have been corrected such that the data used was only when all commodities were trading at the same date. The original daily closing prices for the different commodities are presented graphically below.



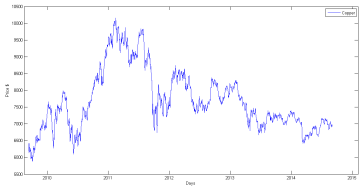
(a) 3-months futures - Brent Oil



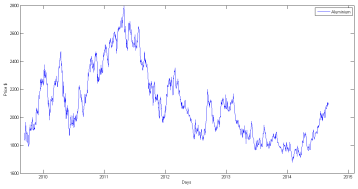
(b) 3-months futures - WTI Oil



(c) Spot price for Gold



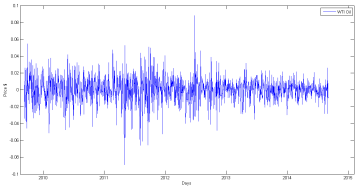
(d) 3-months futures - Copper



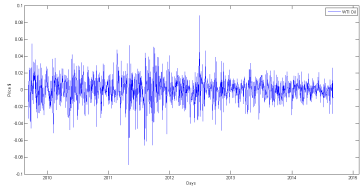
(e) 3-months futures - Aluminium

Figure 3.1: Commodity prices in USD 2009-2014 - not adjusted for inflation.

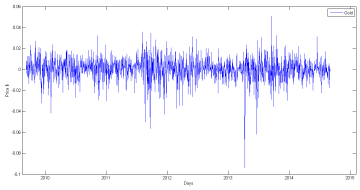
The log returns of these are given below.



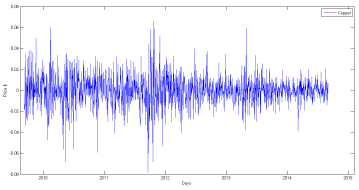
(a) Log returns - Brent Oil



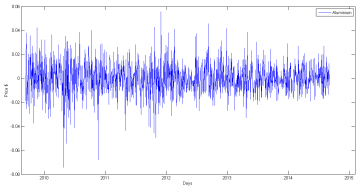
(b) Log returns - WTI Oil



(c) Log returns for Gold



(d) Log returns - Copper



(e) Log returns - Aluminium

Figure 3.2: Log returns for commodity prices 2009-2014 - not adjusted for inflation.

Brent Oil, WTI Oil, Copper and Aluminium all are priced in their three months contract, which means they are initially priced three months before the date of delivery. Once this contract expires the next three months contract is used. However, Gold is priced on the spot price, which means that the buyer will send someone to pick up the Gold, unless the contract is sold before that.

3.2 Implementation

The implementation phase was divided into three parts, the first one only the PCA with the mean reversion and momentum strategies was taken into account. The second part only takes the Copula analysis into account. The third and last part combines the two first parts for the final complete model.

3.3 Part One - PCA

The main model for the initial PCA was built in Matlab. To start out, the log returns were calculated. Then, a PCA is conducted accordingly to the Theory chapter. To determine how many principal components that should be used to explain the variance a target is set to have the number of principal components that explains close to 80 % of the variance. This results in around 3 principal components, which will henceforth be the number of principal components we use.

After this, both the mean-reverting strategy and the momentum strategy are taken into account. As described in the Theory chapter, mean reversion gives us a list of buy or sell signals. Some signals in this list will then be overwritten if we are in a momentum - euclidean distance today compared with euclidean distance yesterday. Hence, we end up with a long list of buy or sell signals, where both mean reversion and momentum strategies are taken into account.

To do a proper check if our model is sound so far, the 5-year data is divided in half. So we apply our PCA-model to the first half and then check if our parameters could be successful on the second half.

In the next step, an extensive testing is carried out. To decide which T and H that should be used, multiple regressions to get the highest returns with the lowest standard deviation were made with values ranging from $T = (7, 8, ..., 206)$ where $H = (3, 4, ..., 103)$. Here the value $T = 7$ corresponds to $H = 3$ and $T = 8$ corresponds to $H = 4$ and so forth. The results varied for different commodities and the best T and H for the first 2,5 years are presented below in Table 3:1.

Commodity	T	H	*Mean*	*Standard Deviation*	*Sharpe Ratio* ($r_f = 0$)
Brent Oil	124	62	0.2470	1.684	0.1467
WTI Oil	88	44	0.2084	1.655	0.1259
Gold	26	13	1.658	17.24	0.0962
Copper	10	5	16.94	141.6	0.1197
Aluminium	35	17	4.671	34.95	0.1336

Table 3.1: Best T and H - all commodities during first 1:616 trading days.

It should be noted that all the calculations are done using log returns but when we get the respective buy or sell signal - the original prices are used to calculate the Mean, Standard Deviation and Sharpe Ratio.

So if we would sum all of these Sharpe Ratios during the first 616 trading days, we would get a Sharpe Ratio for the whole portfolio of 0.622. However, what is interesting here is our T and H values. We use these values on our second part of the data, the last 2,5 years, and see if we can generate alpha.

The results for this is presented below in Table 3:2.

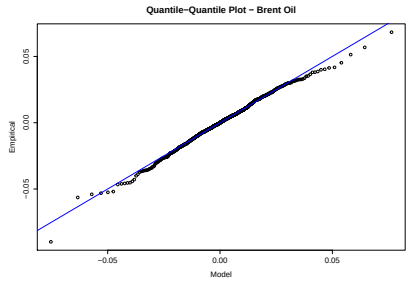
Commodity	T	H	*Mean*	*Standard Deviation*	*Sharpe Ratio* ($r_f = 0$)
Brent Oil	124	62	−0.0190	1.0985	−0.0173
WTI Oil	88	44	−0.0460	1.0906	−0.0422
Gold	26	13	0.1646	15.8010	0.0104
Copper	10	5	1.4020	83.6219	0.0168
Aluminium	35	17	−1.2569	21.4226	−0.0587

Table 3.2: Best T and H - all commodities last 617:1233 trading days.

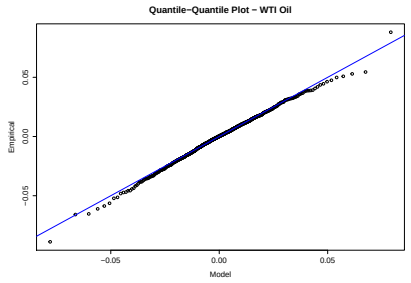
The sum of all Sharpe Ratios are now down to −0.0910. Which means that we would actually lose money using this strategy with these values of T and H. Hence, we hope that with the aid of the copula part that we will be able to generate alpha.

3.4 Part Two - Copula

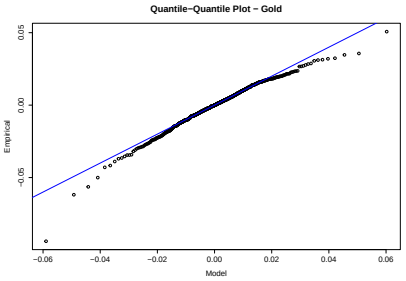
First of all - to start the copula part, the different log returns of the commodities had to be fitted to distributions. To decide which distributions that would be fitting, QQ-plots (see Theory) were used. The best plots of the Student's T distribution are presented below.



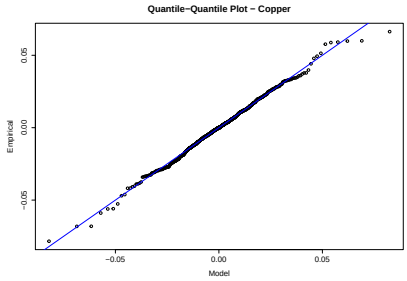
(a) QQ-Plot Brent Oil(T-distr.)



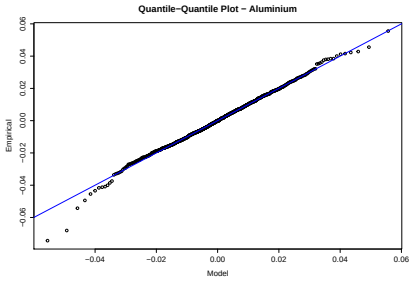
(b) QQ-Plot WTI Oil(T-distr.)



(c) QQ-Plot Gold(T-distr.)



(d) QQ-Plot Copper(T-distr.)



(e) QQ-Plot Aluminium(T-distr.)

Figure 3.3: Commodity prices 2009-2014

So all of these QQ-plots confirms that Student's T distribution works fine for the commodities Brent Oil, WTI Oil, Copper and Aluminium. Gold, however, does not fit as good in the tails. Due to that, a logistic distribution was fitted for the Gold.

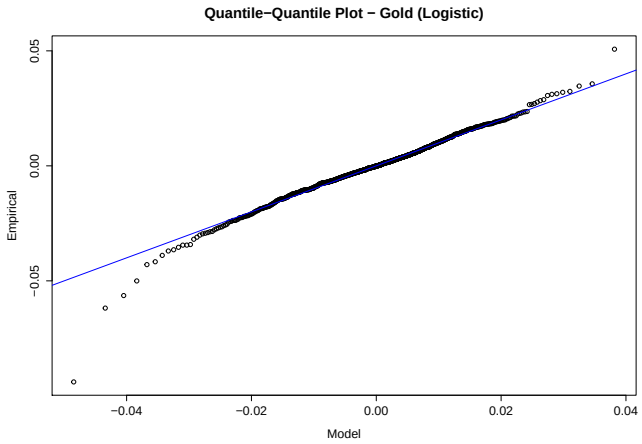


Figure 3.4: QQ-Plot for Gold(Logistic-distribution)

The lower tail is however still a bit off but the upper tail is much better. So we now have decided our margins for our coming multivariate distribution (these are presented in Table 3.3).

The parametric marginal distributions are estimated with Maximum-Likelihood. This is presented below.

Commodity	*Distribution*	*Location*	*Scale*	*Degrees of Freedom/Shape*
Brent	Student's T	0.000 534 9	0.011 19	4.624
WTI	Student's T	0.000 608 6	0.012 43	4.924
Gold	Student's T	0.000 649 6	0.008 404	4.470
Copper	Student's T	0.000 206 2	0.011 46	4.263
Aluminium	Student's T	0.000 079 61	0.011 84	8.079
Gold	Logistic	0.003 376	0.005 141	0.7056

Table 3.3: Parametric marginal distribution

However, to create a multivariate distribution function, we need to establish the dependence between the commodities. Hence, we use Kendall's Tau for this. The Kendall's Tau for the whole data set (5 years) is presented below.

Kendall's	*Brent*	*WTI*	*Gold*	*Copper*	*Aluminium*
Brent	1	0.694 884 7	0.194 701	0.283 692 6	0.243 487 8
WTI	0.694 884 7	1	0.211 222 1	0.313 573 4	0.287 275 9
Gold	0.194 701	0.211 222 1	1	0.261 361 7	0.240 895 8
Copper	0.283 692 6	0.313 573 4	0.261 361 7	1	0.514 945 8
Aluminium	0.243 487 8	0.287 275 9	0.240 895 8	0.514 945 8	1

Table 3.4: The Kendall's Tau for the whole five year data set.

Remark: The largest dependence is between Brent Oil and WTI Oil - since they are both oil and fungible commodities.

To decide which copula to use, the Cramér-von Mises method (see Theory) was used. A wide Goodness-of-Fit test was conducted and the results are given in Table 3.5 and 3.6 below.

Dimensions	*Test statistic*	$P - value$	H_0 *rejected*
2	0.0281	0.005 495	Yes
3	0.0347	0.063 44	No
4	0.0605	0.003 497	Yes
5	0.0595	0.000 499 5	Yes

Table 3.5: Goodness of fit for Normal Copula

where Dimensions are the number of commodities used.

Dimensions	*Test statistic*	$P - value$	H_0 *rejected*
2	0.0169	0.1304	No
3	0.0207	0.5759	No
4	0.0335	0.1543	No
5	0.0449	0.0415	Yes

Table 3.6: Goodness of fit for T-Copula

In Table 3.6 we can see that the Student's T Copula was the best up to $dim = 4$ (since $p > 0.05$ for $dim = 2, 3, 4$), after that - the p-value is too low to accept H_0.

To counter this problem, we decide to divide the commodities into two groups where Aluminium is in an own group.

So, the first four commodities (Brent Oil, WTI Oil, Gold and Copper) are considered into one group. The same Kendall's Tau as in table 3:3 is used to decide the dependence structure(but without the Aluminium part). A Student's T Copula is constructed with these dependences using the function *tCopula* (Please see Theory regarding Student's T Copula) in R from the package *Copula*. The distributions are then fitted to their respective distribution that we found fitting (see Table 3:4). We then use the function *mvdc* in R to construct our multivariate distribution function. Finally, we calculate the probability that the return is less than or equal to 0. This is done looking back T number of days (hence constructing a multivariate distribution function looking back T days.) In our case, a probability plot is constructed looking back 76 days (chosen since it is a bit larger than the mean of all T's from the PCA part). So the first multivariate distribution function uses the data from 76 days - starting on day 1, the second looks at the 76 days - starting on day 2 and so forth. To explain, when we calculate the probability from the first 76 days we get a value indicating how big is the chance that all commodities used are going to have negative or zero return. This means that checking how this develops, we can see which kind of "mood" the market is in at that time - comparing to the other days. When can then represent this graphically - see Figure 3:5 below.

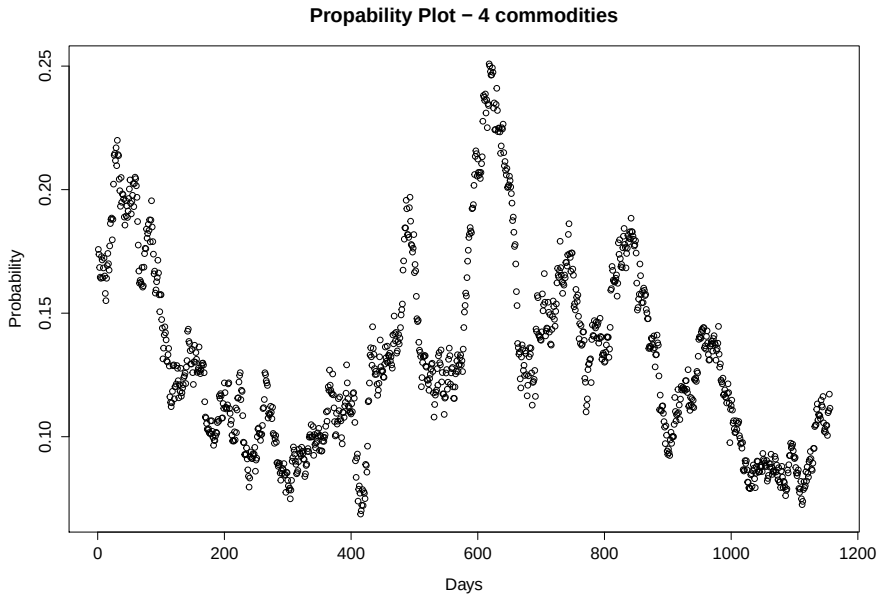


Figure 3.5: Probability Plot that Brent, WTI, Gold and Copper have a return ≤ 0 when $T = 76$. This can be seen as a "indicator" on how those markets are going where a high probability means that the markets are going down and a low probability that the markets are supposedly going up.

3.5 Part Three - PCA and Copula Combined

To summarize, in the PCA part we use the mean-reversion strategy and that is then overridden by the momentum strategy. In the Copula part, we use our results to calculate the probability that the returns for four of our five commodities are less than or equal to zero. So what happens when we combine these parts? Let us find out.

The main idea is that the Copula part should override the two first strategies when the probability is large or small enough. Hence, let us try a trading model where we buy all commodities when the probability that the returns are less than or equal to zero is lower than 0.09 ($mean of the probabilities - 1$ *standard deviation*). And then sell when the probability is higher than 0.17 ($mean of the probabilities + 1$ *standard deviation*). The mean and 1 standard deviation is chosen to have a standard reference. This is also acting accordingly to the model described in Part Two - Copula above and with $T = 76$. With this probabilities, only $197 * 2 = 394$ trades were made out of 1156 ($1232 - 76 = 1156$) possible days. Results are as follows:

Commodity	*Mean*	*Standard deviation*	*Sharpe Ratio* ($rf = 0$)
Brent	−0.045 89	1.511	−0.030 37
WTI	−0.039 64	1.511	−0.026 23
Gold	0.1320	16.17	0.008 166
Copper	5.665	122.5	0.046 23

Table 3.7: Test results for copula probability model - selling at $P \geq 0.17$

Commodity	*Mean*	*Standard deviation*	*Sharpe Ratio* ($rf = 0$)
Brent	0.1129	1.460	0.0773
WTI	0.1072	1.399	0.076 63
Gold	0.4979	15.59	0.031 92
Copper	−1.655	101.2	−0.016 35

Table 3.8: Test results for copula probability model - buying at $P \leq 0.09$

The Sharpe Ratios in total 0.167 which means that it could generate alpha.

So, let us combine this with the results that was given in Part One - PCA. We trade all commodities with both mean reversion and momentum with T and H as presented in Table 3:1. Then the copula part overrides the trading days where the probability is large or small enough ($mean \pm standard\ deviation$) for Brent, WTI, Gold and Copper but not Aluminium. Looking at the last 617:1233 trading days we get the following results.

Commodity	*Mean*	*Standard deviation*	*Sharpe Ratio* ($rf = 0$)
Brent	0.0517	1.2044	0.042 94
WTI	0.023 95	1.1980	0.019 99
Gold	0.3274	16.05	0.020 39
Copper	−0.7098	84.03	−0.008 447

Table 3.9: Test results for both PCA and Copula last 617:1233 trading days.

Which gives us a positive total Sharpe Ratio of 0.075.

However, if we include the Aluminium PCA trading part (results of this can be found in Table 3:2), we get a total Sharpe Ratio of $0.075 - 0.0587 = 0.0163$.

Chapter 4

Discussion

4.1 Summary

The purpose of this master's thesis was to determine whether PCA and Copula could be used to predict future commodity prices and hence generate alpha. The PCA was used on the first 2.5 years out of 5 years of data to see if the results could be used on the last half to generate alpha. That could not be done with our method. Then the returns were fitted to distributions using Maximum-Likelihood. Student's T distribution and Logistic distribution was found to be fitting for the different returns.

Then, the dependence between the commodities was decided using Kendall's Tau.

To decide which copula to use, a Goodness-of-Fit test was conducted using the Cramér-von Mises method. This showed us that a T-copula could be fitted well up to four dimensions.

A probability plot was then constructed based on our $dim = 4$ T-copula where we saw the probability that all first four commodities have a return less than or equal to zero.

Lastly, the PCA part and the Copula part were combined and a positive Sharpe Ratio was given. However, the low Sharpe Ratio of 0.0163 is not a great result. This is due to mainly two reasons,

- We have assumed that the risk-free rate r_f to be 0 which means that we barely are gaining on investing in this prediction model.
- We have not assumed any brokerage fees which must be taken into account since every trade costs - further reducing our positive return.

What should be noted is that we have a good result with the PCA part the first 2.5 years, but those T and H parameters does not provide any profit the last 2.5 years. If we had chosen a different criteria for T and H, maybe the model results would have been much better. These values are very important for the model.

In the description of Maximum Likelihood an assumption is that the random variables should be i.i.d but this can pretty easily be seen in Figure 3:2 that this is not the case. Instead perhaps a GARCH-model could be used to model the observed time series.

Moving on to comment the Copula part - the fitted distributions worked very well as could be seen in the QQ-plots. Also, the p-values for up to dim 4 was very good using T Copula. So why does the model not generate more alpha? Well, perhaps the way we used our results from the Copula part should have been used in a different way. Maybe more focus on the Copula part could have given better results. The largest part of the trading comes from the PCA part, it would be interesting to see what would happen if the Copula part was the largest part.

Also, it is possible to use the copula distribution in many different ways. Perhaps one could look at expectations or conditional distributions instead. This could have given an even better result and a model which could generate a better Sharpe Ratio.

4.2 Tradeable Results

This thesis does not cover the extra costs that comes with trading such as hedging, brokerage fees, initial investments(margins) and salaries. But with that said, the main purpose to see if we could generate a positive risk-adjusted return with this model is proven.

One should also note that aluminium perhaps was not the best commodity to exclude in the Copula part. Perhaps gold would be the best commodity to exclude since it is a more defensive asset compared to the other commodities used in the thesis. This since investors tend to buy gold when the economy is going down as a "safe harbour". The other four commodities used, are more connected to good times in the economy. Hence, gold should maybe have been excluded.

The model only covers two alternative states, in a buying mode or in a selling mode. Perhaps a third state should have been introduced where one simply does nothing. This could have be used in very volatile times when it is very hard to predict where the market is going. The varying volatility can easily be seen when the log-returns are presented in Figure 3:2.

If one looks at this in a larger scale, it is a very narrow field in which the investigation was conducted. Nothing says that commodities is the most preferred tradeable asset and the model could have been used on assets such as stocks, foreign exchange or bonds. Another thing to consider is that very few commodities were used. If more assets would have been used, maybe the model would be more profitable.

4.3 The Future

Another thing to consider is if it in this case was the optimal solution to optimize the parameters on the first 2.5 years. Maybe a more continual optimization would be more fitting. That perhaps one always look back at the latest year and optimize on that. It would be interesting to see since especially the Copula part shows much promise (the Cramér-von Mises tests were sound up to dim 4).

4.4 Conclusion

To conclude, we can generate a positive risk-adjusted return from this prediction model but with a very low Sharpe Ratio.

Chapter 5

References

5.1 Reference List

[Banerjee, 2013] Banerjee, J. Origins of Growing Money (January 2015) (Available at forbesindia.com/printcontent/34515)

[Bouyé, 2000] Bouyé, E (2000) Copulas for Finance A Reading Guide and Some Applications. Financial Econometrics Research Centre, City Universe Business School, London.

[Hull, 2009] John C. Hull (2009) Options, futures and other derivatives.

[Infantino et al., 2010] L.R. Infantino and S.Itzhaki (2010) Developing High-Frequency Equities Trading Models. Massachusetts Institute of Technology.

[Investopedia, 2015] Definition of Mean Reversion (January, 2015). (Available at www.investopedia.com/terms/m/meanreversion.asp)

[Investopedia, 2015] Definition of Sharpe Ratio (January, 2015). (Available at www.investopedia.com/terms/s/sharperatio.asp)

[Jackman, 2009] Jackman, S. (2009) Bayesian Analysis for the Social Sciences. Page 507. John Wiley and sons, Ltd

[Jegadeesh et al., 1999] Jegadeesh, N. and Titman, S. Profitiability of Momentum strategies: An evaluation of alternative explanations. National Bureau of Economic Research, Cambridge MA.

[Jolliffe, 2002] Jolliffe, I. T., 2002, Principal Component Analysis, 2nd edition, Springer.

[Krishnamoorthy, 2006] Krishnamoorthy, K. (2006) Handbook of Statistical Distributions with Applications. Chapman and Hall, Boca Ration, Florida.

[Lutz, 2010] Lutz, B (2010). Pricing of Derivatives on Mean-Reverting Assets, Chapter 2 - Mean reversion in commodity prices. Springer-Verlag Berlin Heidelberg.

MATLAB Version R2011a

[Myung, 2003] Jae Myung, 2003, Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology.

[Nelsen, 2006] Roger B. Nelsen (2006) An Introduction to Copula. Springer

R Package version 1.0-2

[Seber, 2007] George A.F. Seber. (2007) A Matrix Handbook for Statisticians. John Wiley and Sons, Inc., Publication

[Shao, 2002] Shao, Q. (2002) Maximum Likelihood Estimation for generalised logistic distributions. Marcel Dekker, Inc. New York.