LUND UNIVERSITY

MASTER THESIS

FX Trading Using Gaussian Processes

Author:
Ahmed DAADOOCH

Supervisors:
Anders VILHELMSSON
Sara MORICZ

Master of Science in Engineering, Industrial Engineering and Management

LUND UNIVERSITY

Department of Economics

June 15, 2018

LUND UNIVERSITY

Abstract

Department of Economics

Master of Science in Engineering, Industrial Engineering and Management

FX Trading Using Gaussian Processes

by Ahmed DAADOOCH

Machine learning and its application within finance have gained popularity the last decade. The traditional trading roles are changing rapidly and are being increasingly automated with algorithmic trading strategies, by proprietary trading firms, market makers, and other financial institutions. FX trading often involves strategies in the form of technical analysis – suggesting that the efficient market hypothesis might not always hold. Different machine learning techniques are often used in trading activities by Quant fund and other algorithmic and high-frequency trading firms.

In this thesis, I investigate if the Gaussian Process Regression (GPR) can predict prices on a EUR/USD FX Future from CME Globex[5]. The GPR approach has its advantages, being a non-parametric and probabilistic method, and often being much simpler to implement, in contrast to other machine learning techniques like neural networks, which might not always be easy to apply in practice. The last decades of developments within GPR has made it a solid competitor for real supervised learning applications. In this thesis the ARIMA model is used as a benchmark model for prediction.

Acknowledgements

This thesis is the final part in the MSc in Industrial Engineering and Management, being part of the specilization in Financial Engineering and Risk Management. The thesis was conducted at the Department of Economics at Lund University School of Economics and Management.

I want to thank my supervisors Anders Vilhelmsson and Sara Moricz for their support and guidance in writing this thesis. I also want to thank Niklas Höjman for taking the time to reading through the paper, and giving helpful feedback.

Contents

List of Abbreviations

EMH	Efficient Market Hypothesis
FX	Foreign Exchange
GPR	Gaussian Process Regression
GDP	Gross Domestic Product
Spot	FX market for current exchange rates
Future	An exchange-traded contract
Pip size	Smallest possible move in price of the traded asset
Market Maker	A financial institution that provides liquidity for an asset
ARIMA	Autoregressive Integrated Moving Average
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Fucntion
AIC	Akaike information criterion
BIC	Bayesian information criterion
CME Globex	An Exchange for futures and other derivatives
MSE	Mean Squared Error
SMSE	Standardized Mean Squared Error
RW	Random Walk
OTC	Over The Counter
Long	Buying the asset, speculating in an increase in valuation
Short	Selling the asset (without owning it), speculating in a decrease in valuation
Bid-Ask spread	Refers to spread between the Bid and Ask price

1 Introduction

1.1 Background

Financial markets and trading activities have gone through major changes the last decades as technology has become more sophisticated and electronic trading evolved [27]. Market transactions used to exclusively happen directly on the exchange floors in physical form. Today, trading is mainly done electronically with the help of computers [9]. In addition, developments in machine learning and easy access to data has led to new applications of these methods within finance.

This change has led to many new forms of trading strategies, including various forms of automated algorithmic trading. Today, in addition to traditional funds and asset managers, there are many Quant- and trading firms dealing with various machine learning and algorithmic trading learning techniques. Machine learning is also being used more by global investment banks in their market making and risk management activities [27].

In this paper, Gaussian Processes for machine learning is used to predict the EUR/USD future price, and an algorithmic trading strategy is built based on the predictions. The GPR method has its advantages; it is a non-parametric and probabilistic method that is easy to implement, and requires fewer assumptions compared to other machine learning methods.

1.2 Purpose

This paper look into the Gaussian Process Regression for machine learning, and how it can be used to implement trading strategies on the EUR/USD CME Globex Futures. Further, the thesis compare the Gaussian Process Regression to the benchmark ARIMA base model for predicting FX Futures prices. The purpose of this thesis is to investigate the GPR, a supervisied machine learning method. The GPR has it's advantages, being a probalistic and a non-paramteric model. The ARIMA model, for instance, assumes a linear model. The GPR can fit a smooth function to any data set, without assuimg a dimention of the function. This makes it a good method for time series modelling. The GPR model is relatively easy to implement, where the main

model selection lies in choosing a suitable covariance function. However, a drawback with the GPR is the computational time for large data sets. While this might be an issue for some applications, it is not necessarily the case here - as the number of data points to train does noes necessarily need to be very large.

1.3 Limitations

The trading is not performed live, all results are based on simulations of futures trading, more specifically the March 2018 EUR/USD CME Globex FX Future. The paper include two trading models, a simplified model, and a realistic model. The simplified strategy assumes that the Bid-Ask spread is always zero, and that there do not exist any transaction fees. To simulate a realistic trading model, transaction fees and the Bid-Ask spread have been approximated and added to the simulation in the realistic model.

1.4 Thesis Summary

In this thesis, I investigate if the Gaussian Process Regression (GPR) can be used to predict prices on a EUR/USD FX Future from CME Globex [5]. A data set consisting of around 30,000 trading minutes is used for training, validation and testing.

It seems that there are better alternatives to the linear ARIMA model. Swastanto, 2016, used the GPR model to forecast long-term time series and were able to achieve satisfactory results [26]. Many others have tried to exploit the fact that markets are not always efficient, and thus not always follow a random walk process. Gradojevic and Yang, 2006, used non-parametric artificial neural networks (ANN) to predict exchange rates and showed that their result could outperform a random walk [12]. While some have been successful in using machine learning methods such as the Gaussian Processes in predicting time series, it is not obvious if this can be applied to FX futures prices, even if FX time series are not random walk processes.

Gaussian Process Regression is a supervised machine learning method. It's a non-parametric and probabilistic method, allowing for a function to be fitted for any kind of data. Gaussian processes can be thought of as a collection of functions with infinite dimension. The idea is to condition the probability function on the training data to find a function that can map inputs to outputs. This is achieved by choosing a covariance function, and optimizing its parameters by maximizing the likelihood function.

The main part of the model selection is choosing a covariance function, also known as a kernel, which describes the correlation between data points. In this thesis, two

popular kernels, the Rational Quadratic and the Ornstein-Uhlenbeck kernels are used. The GPR models are trained by maximizing a likelihood function. Once the GPR model is trained, a validation set it used to validate the model and avoid over-fitting issues. The model is then used to forecast the next trading minute price in a test set, and the foretasted price acts as the underlying decision maker of a trading strategy.

Different sizes of training sets are tested, having 70, 100, 300, 1000, and 2000 data points. The validation set is chosen to be the same size as the training set. The test set consists of 25,000 data points.

Two trading models were built. The first one is a very simple model with no transaction fees and zero bid-ask spread. The second model is more sophisticated and realistic, it include both transaction fees and a bid-ask spread to reflect real market conditions.

The GPR model is benchmarked against the ARIMA model, compared by metrics such as the mean squared error, and mean absolute error. We show that the GPR model is slightly better than the ARIMA model, however the difference is small.

In order to predict time series, such as the FX future prices, the series must not be a random walk. We conclude that the future prices does not follow a random walk by testing the random walk hypothesis by a Ljung-Box test. While it was possible to prove that the time series is not a random walk, the GPR was not a good enough model for predicting future prices. The trading results were unsatisfactory yielding low or negative returns.

It is not very surprising that the trading simulation were not profitable. The EUR/USD futures are very liquid and are traded actively. It is, therefore, reasonable to assume that many trading firms already have implemented sophisticated algorithmic strategies, so that it is increasingly harder to find profitable algorithms.

The choice of implementing a GPR model, rather than other machine learning methods, were based on the potential strengths of the GPR model. The main strengths being that the model in non-parametric, and could therefore fit a smooth function to any type of data. In addition, it's a probabilistic method, giving a degree of certainty in the predictions. The GPR implementation requires few assumptions, where the main part of the implementation is choosing a kernel function. One drawback with the model, is the computational time, which becomes very slow with large data sets.

The GPR model proved to be in line, or slightly better, than our implemented ARIMA model, measured by our specified metrics. However, it also proved not to be a good predictor to be used for future FX trading models implemented in this paper. It's possible that the model might be improved by introducing more inputs, and/or find better suited kernels.

1.5 Thesis Outline

Section 2 provides an overview of previous literature on forecasting exchange rates. In Section 3 a brief explanation and overview of the foreign exchange market is given, along with factors affecting exchange rates. In Section 4 the theory behind the models used is explained. Section 5 is about the data used and the data characteristics. In section 6 the methodology and implementation of the models are explained. Section 7 provides a summary of the results. Finally, Section 8 provides a discussion of the results obtained.

2 Previous Literature

Many studies on forecasting foreign exchange rates have been conducted in the past. The ARIMA model has been used frequently to predict time series and has been one of the most popular methods in the past [16]. However, the ARIMA model has its flaws in assuming a linear model and therefore there might be better suited models to use [16].

Many have tried to find a better prediction model for exchange rates. There have been varied results in successfully predicting exchange rates. In 2005, Cheung, Chinn, and Pascual drew the conclusion that many of the theoretical methods such as the Purchasing Power Parity, and Interest Rate Parity are not able to consistently outperform a random walk process, but are only sometimes superior [3]. However, there are some contradicting results. Zorzi, Muck, and Rubaszek, 2015, showed in their research that PPP can be used to predict exchange rate movements better than the random walk [36]. In addition, Simpson and Grossman also used PPP to predict currencies, and according to their research they were able to correctly predict the direction of exchange rates indexes of an accuracy of up to 70%, and were able to predict better than a random walk on four out of six observed indexes [25].

Many others have tried to exploit the fact that markets are not always efficient, and thus not always follow a random walk process. Gradojevic and Yang, 2006, used non-parametric artificial neural networks (ANN) to predict exchange rates and showed that their result could outperform a random walk [12]. Dunis and Huang, 2002, used non-parametric neural network regression and recurrent neural networks (RNN) to predict GBP and YEN against the USD and achieved highest prediction accuracy with RNNs [8]. Zafari et al. showed in their paper that an evolving recurrent neural network using Cartesian genetic programming can be used to predict trends in exchange rates, and achieved a prediction accuracy of up to 98% [34].

Some have studied methods benchmarked against the ARIMA model. Kamruzzaman, 2003, used an ANN model to predict FX rates, and found that their model outperforms the ARIMA model [15]. Work by Villa and Stella, 2013, where a Bayesian network classifier was studied, showed that their model outperformed the ARIMA model on multiple performance metrics [31].

In the high-frequency area, Zeman and Maršík, 2013, showed that the random walk for EUR/USD spot rate can be rejected in 5min frequency data, but not at hourly

and 4 hours frequency [35]. Choudhry et al., 2012, used an ANN model to form a trading strategy based on second spot FX observations, and achieved a profitable result even when accounting for transaction costs [4].

Gaussian Processes has been used to predict time series in various studies. Swastanto, 2016, used the GPR model to forecast long-term time series and were able to achieve satisfactory results [26]. Mojaddady, Nabi, and Khadivi, 2011, used Twin Gaussian Processes to predict stock market prices [21]. Farrell and Correa, 2007, used Gaussian Process Regression to predict stock prices but concluded that the method were not good enough to make money in the market and pointed to issues such as computation time [10]. GRP has also been used to predict stock market volatility, Ou and Wang, 2011, used GPR and concluded that the model yields better results than the GARCH model, which is usually used when modelling volatility [22].

In summary, it seems that there are better alternatives to the linear ARIMA model. While some have been successful in using Gaussian Processes in predicting time series, it is not obvious if this can be applied to FX futures prices, even if FX time series are not random walk processes.

In this paper, I look at the machine learning method of Gaussian Process Regression (GPR) to forecast minute FX prices for the EUR/USD future.

3 The Foreign Exchange market

In this section, the theory behind the foreign exchange market and FX futures is presented to give the reader an understanding of how FX rates are decided. In section 3.1 an introduction on currencies is given. It is followed by a brief overview of the spot and futures markets in section 3.2. In section 3.3, 3.4 and 3.5, the theory behind the bid-ask spread, the Efficient Market Hypothesis, and the fundamental relationships of exchange rates is explained.

3.1 Introduction

Today currencies float freely and their values depend on multiple macro-economical factors as well as influence from central banks [24]. The Foreign Exchange market is today the largest and most liquid financial market, with around USD 5tn trading each day, of which around 95% is traded interbank [6] [2]. The value of currencies is expressed relative to another currency, e.g. EUR/USD, which states the amount of USD one EUR is worth.

The value of a freely floated currencies depends on the supply and demand, which is affected by many factors. The most important ones are interest rates, inflation rates, unemployment numbers, GDP growth, wage growth, and monetary and fiscal policy [24]. These statistics are usually measured and released quarterly or monthly. In addition, currencies' values also depend on other countries' inflation and interest rates.

3.2 The Spot and Futures Markets

In the Spot market currencies are traded both Over The Counter (OTC) or on certain exchanges. The Spot market represents the current exchange rates of currencies and the market is open 24 hours a day.

In addition to the spot market, there are many exchange-traded derivatives on most currencies, such as futures and options. A futures contract is a commitment to buy a currency on a future date for a predetermined price. The contract does not have to

be held until expiry, it can be sold before the expiry date. These contracts are traded on an exchange and can be used for speculative or hedging purposes.

The relationship between the spot price and futures price is given by equation 3.1. It can be seen that when the forward interest rate in the domestic currency is higher than the forward interest rate in the foreign currency, then the futures price is higher than the spot price [30].

$$F = S\frac{(1 + R_{domestic}\frac{d}{360})}{(1 + R_{foreign}\frac{d}{360})} \tag{3.1}$$

Where F is the future price, S is the spot price, R is the forward interest rate, and d is days to maturity. This relationship should hold if there is no arbitrage in the market. If for example, the futures price would be above the value given by equation 3.1, then a trader can make a profit by selling the future price and buying the spot price.

3.3 Market Makers and the Bid-Ask Spread

Market makers are banks and other financial institutions that maintain market liquidity by taking orders from investor clients. Market makers buy (sell) when clients want to sell (buy). To be compensated for their market making role, they profit from the spread between the bid and asking price, meaning that they buy at a slightly lower price than what they sell for. The Market Maker buys at the bid price P_b and sells at the asking price P_a. The Bid-Ask spread can be seen as a compensation for the market maker, it covers order processing costs, inventory holding costs, and adverse selection costs [1].

3.4 The Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) states that asset prices incorporate all available information, and that price movements are completely random [1]. The EMH occurs in three forms, the strong form, semi-strong or the weak form. The strong form is when the market is most efficient, where assets prices incorporate all information, including insider information, and all price movements are completely random, where investors can't achieve above average returns from fundamental or technical analysis. The semi-strong form is when prices incorporate all publicly available information, but not insider information. The weak form states that prices cannot be predicted based on historical prices, e.g. by technical analysis. However, in the weak form, fundamental analysis can still provide above average return [1].

However, there is some doubt on whether the EMH actually holds [17]. The EMH assumes that people act in a rational way, which is not always the case, because of emotions and other behavioral biases. To test if the EMH holds for an asset, various tests can be performed. One of these is a Variance Ratio test, described in the Theory section below.

There are typically two types of analysis that can be performed on traded financial assets, one is Fundamental Analysis and the other is Technical Analysis. In Fundamental Analysis, the asset is priced by trying to calculate its actual value, for example by estimating future cash flows for a company, and discounting them to present value. In terms of currencies, the equivalent fundamental analysis approach would be valuations based on e.g. the Interest Rate Parity.

In contrast, Technical Analysis only considers the historical data of the traded asset, and through mathematics, statistics, game theory and/or decision theory predict the future price. There exist many different technical indicators, e.g. momentum indicators, that could be helpful in these types of analysis. This means that Technical Analysis contradicts the EMH, stating that the EMH does not hold and that there might be correlations and trends in asset price movements, making it possible to predict future prices.

3.5 Fundamental Relationships of Exchange Rates

The Purchase Power Parity (PPP) states the relationship between the price of the same product in two countries is equal to the currency exchange. The idea is that a product purchased directly in e.g. the US with currency USD, should cost the same when converting the USD amount to EUR and buying the product in amounts of EUR. There are multiple versions of the equation, in 3.2 one version of the PPP is presented, assuming no transaction fees [2]. A well known example of the use of PPP is the Big Mac Index introduced by the The economist in 1986. The index measures the price ratio of a Big Mac (a hamburger from McDonald's) in different countries [7].

$$P_i^1 = SP_i^2 \ , \tag{3.2}$$

where P^1 is the price in country 1, P^2 is the price in country 2, and S is the exchange rate of the two countries' currencies.

In theory, higher interest rates should lead to a stronger currency. This is because higher interest rates attract more foreign investments, which leads to a greater demand for the domestic currency. The theoretical relationship between two countries' currencies and forward interest rates is given by the Interest Rate Parity [2].

$$(1 + i_1) = \frac{E_t(S_{t+k})}{S_t}(1 + i_2) \tag{3.3}$$

$E_t(S_{t+k})$ is the expected spot price in k periods. S_t is the current exchange rate. i_1 and i_2 are the interest rates in country 1 and country 2 respectively. In order for the parity to hold, there are some assumptions that must be true. The most important ones are capital mobility and perfect substitutability between the two countries' assets [2].

Inflation is the increase in general price levels, meaning that the value of a currency declines in real terms when inflation increases. Typically, inflation rates and interest rates are linked. For instance, low interest rates lead to more borrowing and thus more consumption. The increase in consumption leads to higher prices, thus higher inflation. In contrast, when interest rates are high, more people save their money and less borrow, which leads to lower inflation. There are several indices measuring inflation, one example is the Consumer Price Index (CPI) [33].

Most countries have a goal to keep the inflation rate at a steady level. Typically this level is around 2% p.a. for developed countries [13]. To achieve that goal, central banks can control the quantity of money in circulation and increase/decrease interest rates, known as Monetary Policy. In addition, the government can affect inflation with fiscal stimulus, e.g. by spending on infrastructure projects or adjusting tax rates, this is known as Fiscal Policy [14].

The theory behind the relationship between interest rates and unemployment is that higher unemployment rate leads to less consumption and thus the inflation rate decrease, while low unemployment leads to higher consumption and higher inflation rates. And like mentioned before, inflation is connected to interest rates, and thus affects the price of currencies. Trivially, higher wage growth leads to higher consumption and thus higher inflation [32].

3.6 Summary

Exchange rates depend on a number of factors, including macro economical data as well as theoretical price relationships to prevent arbitrage. The EMH states that all available information should be reflected in the prices if it holds. There are three different forms of the EMH, the strong, the semi-strong, and the weak.

4 Theory

This chapter starts with an introduction to time series and the theory behind them in section 4.1. In section 4.2 the theory to test for a random walk process is explained. Finally in sections 4.3, 4.4 and 4.5, the theory behind the ARIMA model, Bayesian model, and the Gaussian Process Regression is explained.

4.1 Time Series

A time series is data points collected in time-order and at the same time interval, e.g. every one minute [16]. Below are brief definitions of how some properties of observed time series are defined.

4.1.1 Mean and Variance

The sample mean $\bar{x}$, variance s^2, and covariance of a data series $X(t)$ is defined in equations 4.1, 4.2, and 4.3 [16].

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{4.1}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.2}$$

$$q_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k}) \tag{4.3}$$

4.1.2 Weakly Stationary

A time series is said to be weakly stationary if the mean, variance, and covariance is time indifferent, meaning that a shift in time won't affect the mean, variance, and covariance values [16].

4.1.3 Testing for Stationarity

Testing for stationarity can be done by the Augmented Dickey Fuller (ADF) test, or by visually looking at the data. The ADF looks at an AR model to test for stationarity [18]. For an AR(1) model, seen in equation 4.4, we can test for stationarity by testing if θ equals one. If θ equals one, then the autocorrelation is dependent on time and therefore the process is not stationary, see equation 4.5. This can be extended to AR processes of higher order, by testing if δ is equal to zero in equation 4.6.

$$\Delta y_t = \alpha + \theta y_{t-1} + \epsilon_t \tag{4.4}$$

$$\begin{aligned} cov[\Delta y_t, \Delta y_{t-k}] &= (t-k)\sigma^2 \\ \rho[\Delta y_t, \Delta y_{t-1}] &= \sqrt{\frac{t-k}{t}} \end{aligned} \tag{4.5}$$

$$\Delta y_t = \alpha + \delta y_{t-1} + \sum_{i=1}^{h} B_i \Delta y_{t-i} + \epsilon_t \tag{4.6}$$

An alternative way to check for stationarity is to plot the data, the autocorrelation function, and the sample autocorrelation function. Like mentioned earlier, a weakly stationary process should have constant mean and variance, and the autocorrelation function should decay to zero [16].

4.1.4 White Noise

A data series of uncorrelated random variables with zero mean and variance σ^2 is defined as white noise. Often, data samples of observations are assumed to include white noise [16].

4.2 Random Walk Test

To test the Efficient Market Hypothesis a Random Walk (RW) test can be performed by using a Ljung-Box Q-test. The Ljung-Box tests if the autocorrelation for a specified lag m is zero. Under the null-hypothesis of zero autocorrelation, the equation in 4.7 is χ^2 distributed with m degree of freedom [1].

$$Q'_m = T(T+2) \sum_{k=1}^{m} \frac{\rho_k^2}{(T-k)} \tag{4.7}$$

4.3 ARIMA

The ARIMA model is often used when predicting time-series, such as FX prices. ARIMA stands for Autoregressive Integrated Moving Average. It comprises of two components, an AR and MA component. In an Autoregressive (AR) process, future values depend on previous values. A model of order p is given by 4.8.

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \epsilon_t \tag{4.8}$$

In a Moving Average (MA) process the future values depend on the past random error terms, e_t. An MA(q) process is given by 4.9 below. ϵ_t is a white noise process with mean zero $\epsilon_t \sim N(0, \sigma^2)$.

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \tag{4.9}$$

The mixed ARMA(p,q) model is given by 4.10

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}. \tag{4.10}$$

One condition of the ARIMA model is stationarity. If the data is not stationary, it can be differentiated until it becomes stationary.

4.3.1 Box-Jenkins Model Selection

The Box-Jenkins methodology is an approach to finding an appropriate ARIMA model. It consists of the following steps [19].

1. Check stationarity. If the process is not stationary, differentiate it until it becomes stationary.
2. Plot the autocorrelation function ACF and the partial autocorrelation function PACF to identify the number of lags p.
3. Identify if it is an AR, MA or ARMA model according to the table 4.1 below.
4. Estimate the model. It is also a good idea to try different models and choose the one with the lowest AIC/BIC.
5. Plot residuals of ACF.

TABLE 4.1: Identifying an appropriate ARMA process.

	ACF	PACF
AR(p)	Spikes decay to zero	Spikes cutoff to zero
MA(q)	Spikes cutoff to zero	Spikes decay to zero
ARMA(p,q)	Spikes decay to zero	Spikes decay to zero

4.4 Bayesian Linear Model

Before explaining the GPR, a Bayesian linear model is first explained. Consider the linear model given by equation 4.11.Let $\mathbf{x}$ be an input vector and $\mathbf{y}$ be the output (target). Define $D = \{\mathbf{x}_i, y_i | i = 1, ..., n\}$, where n is the number of observations. Call $\mathbf{x}_1, ..., \mathbf{x}_n$ the training set, which are values already observed. Define the design matrix $\mathbf{X}$ as the D x n matrix of the vector inputs $\mathbf{x}$. Now D can be rewritten as $D = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{y}$ is the vector of targets.

$$\mathbf{y} = \mathbf{x}^T\mathbf{w} + \epsilon \tag{4.11}$$

$\mathbf{w}$ is the weights, $\mathbf{x}$ is the input variables, y is the target variable, and ϵ is Gaussian noise.

The idea is to give a prior distribution over the weights, with higher probability given to functions that are considered more likely. Usually, the prior is chosen as a Gaussian with zero mean. The prior is the probability distribution before seeing the observed data $p(w)$. We can condition the probability on the observed data. This is done by combining the prior with a likelihood function, the likelihood function is the probability of the observations conditioned on the inputs and the parameters, in this case X and the weights w. The likelihood function is $p(\mathbf{y}|X, \mathbf{w})$, according to equation 4.12 [23].

$$\begin{aligned} p(\mathbf{y}|X, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod \frac{1}{\sqrt{2\pi}\sigma_n} \exp(-\frac{(y_i - \mathbf{x}_i^T\mathbf{w})^2}{2\sigma_n^2}) = \\ \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp(-\frac{1}{2\sigma_n^2}|\mathbf{y} - X^T\mathbf{w}|^2) = N(X^T\mathbf{w}, \sigma_n^2 I) \end{aligned} \tag{4.12}$$

The prior over the weights is Gaussian with zero mean and matrix covariance Σ_p, see equation 4.13 [23].

$$\mathbf{w} \sim N(\mathbf{0}, \Sigma_p) \tag{4.13}$$

This leads to the posterior, which is the combination of the prior and the likelihood function according to Bayes rule, see equation 4.20.

$$
\begin{aligned}
p(\mathbf{w}|X,\mathbf{y}) &= \frac{p(\mathbf{y}|X,\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|X)} = \frac{p(\mathbf{y}|X,\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|X,\mathbf{w})p(\mathbf{w})d\mathbf{w}} \\
&\propto \exp(-\frac{1}{2\sigma_n^2}(\mathbf{y}-X^T\mathbf{w})^T(\mathbf{y}-X^T\mathbf{w}))\exp(-\frac{1}{2}\mathbf{w}^T\Sigma_p^{-1}\mathbf{w}) \\
&\propto \exp(-\frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}}^T)(\frac{1}{\sigma_n^2}XX^T+\Sigma_p^{-1})(\mathbf{w}-\bar{\mathbf{w}})) \\
&\mathbf{w} = \sigma_n^{-2}(\sigma_n^{-2}XX^T+\Sigma_p^{-1})^{-1}X\mathbf{y}
\end{aligned}
\tag{4.14}
$$

The expression can be expressed as a Gaussian with mean $\bar{\mathbf{w}}$ and covariance matrix A^{-1}, where A is defined as $A = \sigma_n^{-2}XX^T + \Sigma_p^{-1}$, according to equation 4.15 [23].

$$
p(\mathbf{w}|X,\mathbf{y}) \sim N(\bar{\mathbf{w}} = \frac{1}{\sigma_n^2}A^{-1}X\mathbf{y}, A^{-1}) \tag{4.15}
$$

Finally, the predictive distribution is the average, weighted over how likely they are, of all possible linear combinations of the outputs, given in equation 4.16 [23].

$$
p(f_\star|\mathbf{x}_\star,\mathbf{y}) = \int p(f_\star|\mathbf{x}_\star,\mathbf{w})p(\mathbf{w}|X,\mathbf{y})d\mathbf{w} = N(\frac{1}{\sigma_n^2}\mathbf{X}_\star^TA^{-1}X\mathbf{y}, \mathbf{x}_\star^TA^{-1}\mathbf{x}_\star). \tag{4.16}
$$

4.5 Gaussian Process Regression

Gaussian Processes Regression is one of many machine learning methods. It is a supervised learning method, mapping inputs to outputs. GPR is a non-parametric method, so the model can fit any type of function, and produces a random value when invoked. The main part of the GPR model selection is choosing a covariance function, also known as a kernel. The kernel describes the relationship between the observations. Generally, points that are very close together should have similar values. Different Kernels and their role are described in more detail below [23].

The learning process is done by giving the model a data set with observations of which we already know the values. From these observations, the model can learn how to map inputs to outputs so that when a new input is given, the model can predict the unobserved value for that input. GRP is a probabilistic method, so the prediction is given with a level of certainty. If the prediction is far away from the observed values, the uncertainty of the prediction will be large. Predictions close to the observed values will have greater certainty.

Let $\mathbf{x}$ be an input vector and $\mathbf{y}$ be the output (target). Define $D = \{\mathbf{x}_i, y_i | i = 1, ..., n\}$, where n is the number of observations. Call $\mathbf{x}_1, ..., \mathbf{x}_n$ the training set, which are

values already observed. Define the design matrix $\mathbf{X}$ as the D x n matrix of the vector inputs $\mathbf{x}$. Now D can be rewritten as $D = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{y}$ is the vector of targets. The goal is to train the GPR model using the training set, so that the model will be able to predict the output of new inputs $\mathbf{x}_\star$. The set of new inputs is called the test set and is denoted $\mathbf{X}_\star$. To be able to predict new inputs, a suitable function $\mathbf{f}$ that can map the inputs to outputs must be found.

Rasmussen and Williams, defines the Gaussian process as a collection of random variables, any finite number of which have a joint Gaussian distribution. In figure 4.1 two examples of the Gaussian distribution is plotted in the one dimensional and two dimensional space.

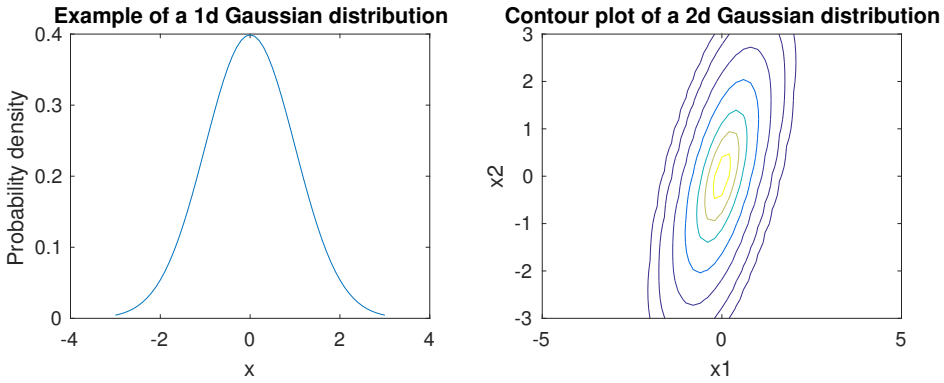

FIGURE 4.1: Example of Gaussian distributions in 1d and 2d

.

A Gaussian process can be thought of as a multivariate Gaussian distribution with infinite dimension [23]. This might be hard to imagine, and the reader might think that this would mean it is a difficult process to handle. However, because of the marginalization property in Gaussian distributions, $p(x) = \int p(x, y)dy$, it turns out that the Gaussian Process is very practical to use [23]. To demonstrate the marginalization property, imagine a two dimensional Gaussian distribution $P(x, y)$, as shown in figure 4.2. If we condition the probability on a certain point $Y = y$, we get the conditional probability $P(x|y)$. As can be seen, the distribution of $P(x|y)$ is also Gaussian.

A Gaussian Process is a distribution over functions. The process consists of a mean function $m(\mathbf{x})$, and a covariance function $k(\mathbf{x}, \mathbf{x'})$ often referred to as a kernel function. These are defines below in equation 4.17 and 4.18 [23].

$$m(\mathbf{x}) = E[f(\mathbf{x})] \tag{4.17}$$

$$k(\mathbf{x}, \mathbf{x'}) = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x'}) - m(\mathbf{x'}))] \tag{4.18}$$

So that a Gaussian Process is expressed as $\mathbf{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$.

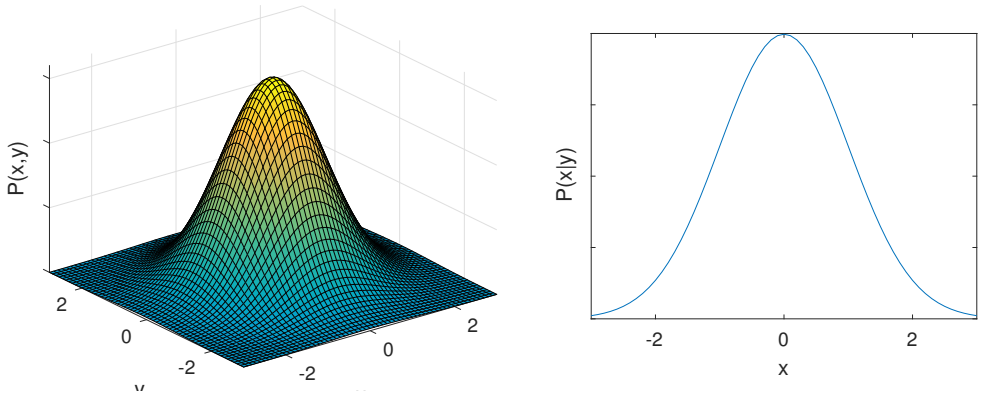

FIGURE 4.2: Marginalization property of Gaussian distributions.

To find a function that describes the data, first, a prior distribution is assumed. The prior is the best guess of the function before seeing the observations. Assuming the observations are noisy, so that $y = f(\mathbf{x}) + \epsilon$, where ϵ is IID Gaussian noise with variance σ_n^2. The prior of the observations is then given by equation 4.19 [23].

$$\mathbf{f} \sim \mathbf{GP}(0, K(X, X) + \sigma_n^2 \mathbf{I}) \tag{4.19}$$

So that,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_\star \end{bmatrix} \sim N\left(\mathbf{0}, \quad \begin{bmatrix} K(X, X) + \sigma_n^2 \mathbf{I} & K(X, X_\star) \\ K(X_\star, X) & K(X_\star, X_\star) \end{bmatrix}\right) \tag{4.20}$$

Here f is the latent variable, which equals y but without the noise, so that $y = f + \epsilon$, where $\epsilon \sim N(0, \sigma_n^2)$.

Once the observations have been seen, a conditional distribution on the observations can be obtained. The posterior distribution is obtained by conditioning the prior Gaussian distribution on the observations so that only functions that describe the observations are considered. It is obtained by Bayes rule, by combining the prior with the likelihood function.

The posterior for a Gaussian Process Regression with noisy observations is given by 4.21. The posterior is the conditional probability over functions given the observations. It gives a higher probability to functions that describe the observations well.

$$\mathbf{f}|X, \mathbf{y} \sim \mathbf{GP}(k(x, x)[k(x, x) + \sigma_n^2]^{-1}\mathbf{y}, k(x, x') - k(x, x)[k(x, x) + \sigma_n^2 \mathbf{I}]^{-1} k(x, x')) \tag{4.21}$$

The likelihood function is the probability of the observations, given the functions, expressed in equation 4.22 [23].

$$\mathbf{y}|X, \mathbf{f} \sim N(\mathbf{f}, \sigma_n^2 I) \tag{4.22}$$

Finally, the predictive distribution is given by 4.23

$$\mathbf{f}_\star|X, \mathbf{y}, X_\star \sim N(\bar{\mathbf{f}}_\star, cov(\mathbf{f}_\star)) \tag{4.23}$$

$$\bar{\mathbf{f}_\star} \triangleq E[\mathbf{f}_\star|X, \mathbf{y}, X_\star] = K(X_\star, X)[K(X, X) + \sigma_n^2\mathbf{I}]^{-1}\mathbf{y} \tag{4.24}$$

$$cov(\mathbf{f}_\star) = K(X_\star, X_\star) - K(X_\star, X)[K(X, X) + \sigma_n^2\mathbf{I}]^{-1}K(X, X_\star) \tag{4.25}$$

It should be noted that the mean value in equation 4.24 is a linear smoother, a term multiplied by y. It can be expressed as in equation 4.26 below.

$$E[\mathbf{f}_\star|X, \mathbf{y}, X_\star] = K(X_\star, X)[K(X, X) + \sigma_n^2\mathbf{I}]^{-1}\mathbf{y} = \sum_{i=1}^{n} \alpha_c k(x_\star, x_i) \tag{4.26}$$

The covariance in equation 4.25 has two terms, the first is the prior variance between the test case and the second term is a positive definite matrix. The second term is subtracted from the prior variance, based on how much the training data explained about the test case [23].

The inverse of $[K(X, X) + \sigma_n^2\mathbf{I}]$ can be computed by Cholesky decomposition.

The main advantages of the Gaussian Process Regression (GPR) are that it is relatively easy to implement in terms of model selection, and requires fewer assumptions in comparison to e.g. Neural Networks. It is a probalistic and non-parametric model, allowing for a function to fit any type of data without any assumpitons of the dimetionality of the data. Like mentioned earlier, the most important aspect of the model selection is the choise of a kernel function.

The main drawback of the GPR method is the increase in computation time when the trainingset becomes larger. The GPR has a time complexity of $O(n^3)$, meaning that for large training sets with n data points, the computation time will increase with a rate of n^3. This leads to very slow learning when using large training sets, which possibly could be an issue if large training data is needed [23].

4.5.1 Kernels

The choice of a covariance function is an important component of the Gaussian Process Regression. The covariance function, or kernel, describes how the data behaves, especially the covariance between pairs of data points. Every kernel has a number

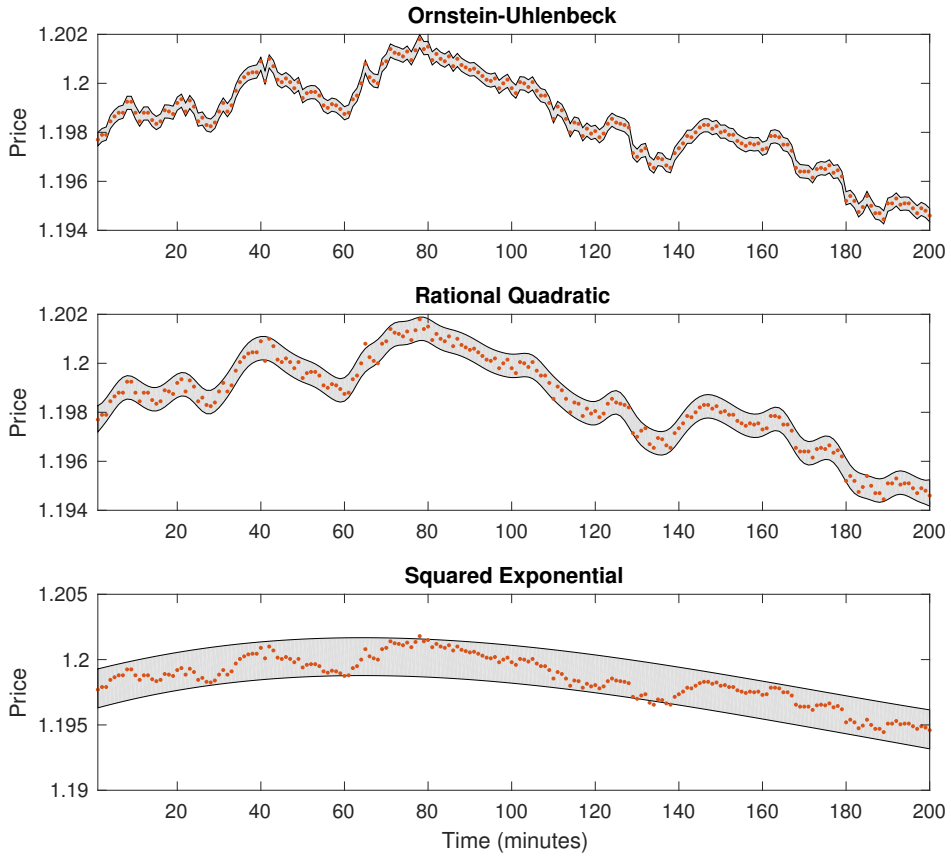

FIGURE 4.3: How different kernels fit the same data.

.

of hyperparameters that needs to be adjusted. See figure 4.3 on how different kernels fit the same data. Below the Squared Exponential, Rational Quadraric, and the Ornstein-Uhlenbeck covariance functions are described. There exist many other covariance functions, and it is also possible to combine multiple covariance functions into a new kernel. However, it should be noted that a valid covariance function must be symmetric and positive semi-definite, meaning that $v^T K v \geq 0$ for all values of $\mathbf{v}$, where K is the kernel matrix containing all the pair covariances of the inputs [23].

Squared Exponential: The Squared Exponential kernel is given by equation 4.27 below.

$$k_{SE} = \sigma_f^2 \exp(-\frac{r^2}{2l^2}) \tag{4.27}$$

l is the hyperparameter for the Squared Exponential kernel, and σ_f^2 is the variance of the signal. In this case, l represents the characteristic length-scale, which can be thought of as the distance interval where the function value does not change [23].

This kernel describes the covariance between two points, decreasing with larger distance. Meaning that for long-term prediction, the covariance will go towards zero, and the prediction will go towards the mean of the training set. This kernel is therefore appropriate for short term forecasting. The Squared Exponential is stationary and a smooth function, in fact, it is infinitely differentiable [23].

Rational Quadratic: The Rational Quadratic kernel is given by equation 4.28 below.

$$k(x, x') = \sigma^2(1 + \frac{(x - x')^2}{2\alpha l})^{-\alpha} \tag{4.28}$$

The Rational Quadratic kernel is a sum of inifitely many Squared Exponentials, with different lenght scales. As $\alpha \rightarrow$ inf, the covariance function becomes a Squared Exponential. It's differentiable and a smooth fucntion [23].

Ornstein-Uhlenbeck: The Ornstein-Uhlenbeck kernel is given by equation 4.29 below.

$$k_0(x, z) = \frac{1}{2\alpha} e^{-\alpha|t|} \tag{4.29}$$

The Ornstein-Uhlenbeck is a mathematical model describing the motion of a particle in fluid and was derived from the Brownian motion [23]. It is a mean reverting motion that has been found to describe some financial assets such as options in a good way [29]. In contrast to the Rational Quadratic and Squared Exponential kernel, the Ornstein-Uhlenbeck kernel is not differentiable.

Hyperparameters

Each kernel has a number of hyperparameters that have to be chosen. These hyperparameters affect the fit of the data [23]. To demonstrate how hyperparameters affect the data fit, an example is shown with the Squared Exponential kernel. In figure 4.4 two different lengths of hyperparameter l is chosen. One is the optimal length according to the marginal likelihood maximation, and the other is one that is to be too long in this case. When the length scale is too short, the model suffers from overfitting, when it's too long, it fits the observations poorly as shown in the figure. Therefore, a value in between is the most suitable. This value of the hyperparameter is chosen by maximizing the likelihood function, as explained in the section 4.5.2 below.

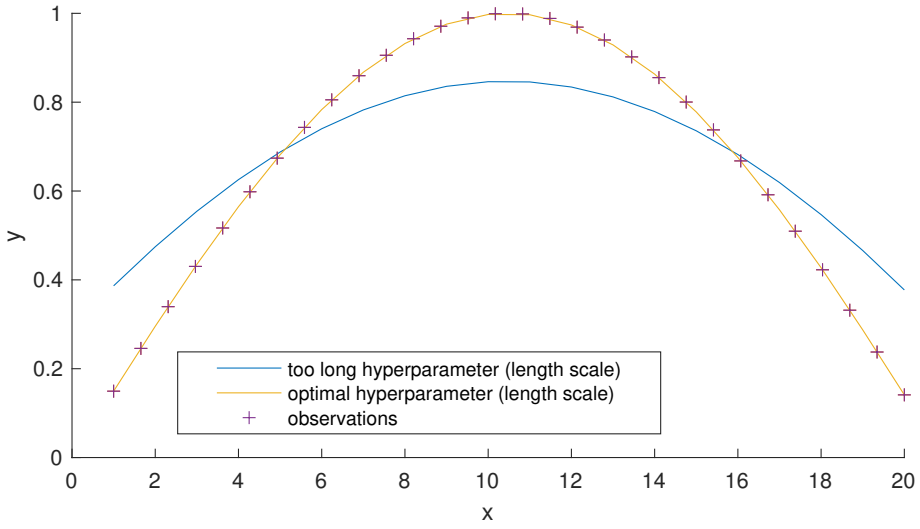

FIGURE 4.4: How the length scale of the Squared Exponential kernel effect the fit of the data

.

4.5.2 Marginal Likelihood

The marginal likelihood describes the probability of the model given the observations. The optimal hyperparameters can be found by optimizing the Marginal likelihood function. The Marginal likelihood, defined as the integral of the prior multiplied by the likelihood, is given by equation 4.30. The marginal likelihood is a trade-off between data fitness and model complexity. The function penalizes complex models. The first term in in equation is the data fit term, here when $K + \sigma_n^2 I$ increases, the first term increases, but the second term decreases. This means that when maximizing the likelihood function over the parameters, the model is robust to overfitting [23]. However, this does not mean that overfitting can't occur. There could still be problems with overfitting a GPR model, therefore model selection should still be done with care, and by using a validation set. Mohammed and Cawley explores the model GPR selection in more depth, including avoiding overfitting [20].

$$\begin{aligned} p(\mathbf{y}|X) &= \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \\ \log p(\mathbf{y}|X) &= -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \end{aligned} \tag{4.30}$$

The optimal hyperparameters is found by the partial derivatives w.r.t the hyperparameters, see equation 4.31. To maximize the function, an inverse of the matrix $K + \sigma_n^2 I$ have to be computed, which takes time $O(n^3)$.

$$\frac{\partial \log p(\mathbf{y}|X,\theta)}{\partial \theta} = \frac{1}{2}\mathbf{y}^T K^{-1}\frac{\partial K}{\partial \theta}K^{-1}\mathbf{y} - \frac{1}{2}trace(K^{-1}\frac{\partial K}{\partial \theta}) \tag{4.31}$$

4.5.3 Occam's Razor

The reader might wonder why too complex models lead to overfitting problems. Occam's Razor explains the situation [23]. In figure 4.5 the marginal likelihood is on the y-axis, and "all sets of observations" on the x-axis. Imagine three types of models according to the figure, a too simple model, a good model, and a too complex model.

Choosing a too simple model to describe our data D is not a good choice, as the probability of the observations is zero for the data. An example is choosing a linear model to describe a non-linear process, it would, however, have a high probability if the process were linear. The too complex model, on the other hand, can describe our data D, but also a wide range of other sets of observations. However, the probability of the observations from the too complex model is low, and therefore not a good choice as it can describe almost any type of observations. The good model, is the "just right" model, describing our observations with a high probability. To avoid overfitting when choosing a GPR model, a validation set is used after training to asses the model, before applying it to the test set.

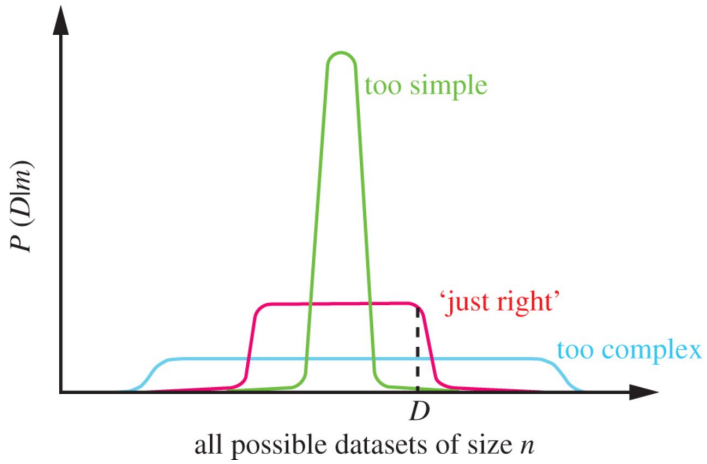

FIGURE 4.5: Graphic illustration of Occam's Razor. Figure is from Ghahramani [11].

.

4.5.4 Other Likelihood Functions

The likelihood function does not have to be a Gaussian. There are multiple other options when choosing Likelihood function. However, if a non-Gaussian likelihood function is chosen, exact inference is not possible. Instead, approximation methods for inference must be used.

In this thesis, in addition to the Gaussian likelihood, the use of student's t likelihood is evaluated, presented in equation 4.32.

$$p(\mathbf{y}|\mathbf{f}) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma\frac{v}{2}\sqrt{v\pi\epsilon_n^2}} \cdot (1 + \frac{(\mathbf{f}-\mathbf{y})^2}{v})^{-\frac{v+1}{2}} \tag{4.32}$$

To perform inference, there exist multiple approximation methods. Below the Laplace's approximation for inference with a student's t likelihood is presented 4.33. Laplace's approximation requires calculations of up to the third derivative, given in equation 4.33.

$$\begin{aligned}
&r = y - f \\
&\ln p(\mathbf{y}|\mathbf{f}) = \ln\Gamma(\frac{v+1}{2}) - \ln\Gamma(\frac{v}{2}) - \frac{1}{2}\ln v\pi\sigma_n^2 - \frac{v+1}{2}\ln(1 + \frac{r^2}{v\sigma_n^2}) \\
&\frac{\partial \ln p}{\partial f} = (v+1)\frac{r}{r^2 + v\sigma_n^2} \\
&\frac{\partial^2 \ln p}{(\partial f)^2} = (v+1)\frac{r^2 - v\sigma_n^2}{(r^2 + v\sigma_n^2)^2} \\
&\frac{\partial^3 \ln p}{(\partial f)^3} = 2(\partial+1)\frac{r^3 - 3rv\sigma_n^2}{(r^2 + v\sigma_n^2)^3} \\
&\frac{\partial \ln p}{\partial v} = \frac{v}{2}\frac{d\ln\Gamma(\frac{v+1}{2})}{d\ln v} - \frac{v}{2}\frac{d\ln\Gamma\frac{v}{2}}{d\ln v} - \frac{1}{2} - \frac{v}{2}\ln(1 + \frac{r^2}{v\sigma_n^2}) + \frac{v+1}{2}\frac{r^2}{r^2 + v\sigma_n^2} \\
&\frac{\partial^3 \ln p}{(v\ln\partial)(\partial f)^2} = v\frac{r^2(r^2 - 3(v+1)\sigma_n^2) + v\sigma_n^2}{(r^2 + v\sigma_n^2)^3} \\
&\frac{\partial \ln p}{\partial \ln\sigma_n} = (v+1)\frac{r^2}{r^2 + v\sigma_n^2} - 1 \\
&\frac{\partial^3 \ln p}{(\partial\ln\sigma_n)(\partial f)^2} = 2v\sigma_n^2(v+1)\frac{v\sigma_n^2 - 3r^2}{(r^2 + v\sigma_n^2)^3}
\end{aligned} \tag{4.33}$$

4.6 Definitions of Performance Metrics

To assess and compare the results of the predictions from different models, a number of performance metrics are used. These are defined in table 4.2 below.

TABLE 4.2: Performance metrics

Name	Definition
Mean Squared Error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(y_\star - \bar{f}(\mathbf{x}_\star))^2$
Standardized Mean Squared Error (SMSE)	$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_\star - \bar{f}(\mathbf{x}_\star))^2}{\frac{1}{n}\sum_{i=1}^{n} var(\bar{f}(\mathbf{x}_\star))}$
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_\star - \bar{f}(\mathbf{x}_\star)\rvert$

It can be seen from table 4.2 that a good model would strive towards a low MSE, an SMSE close to 1, and a small MAE.

4.7 Summary

A time seris can be predictible if it can be concluded that the series is not a random walk. To test for a random walk process, a Ljung-Box test can be conducted, where the null hypothesis is that the autocorrelation for a specified lag is zero.

The GPR model is a non-parametric and probabilistic supervised machine learning method. Gaussian processes can be thought of as a collection of function with infinite dimension. The idea is to condition the function on the data to be able to map inputs to outputs. The GRP model selection consists of choosing an appropriate kernel function to describe the data. It is then used to train the model over a training set, by maximizing the likelihood function. In order to avoid over-fitting, the model is tested on a validation set, before being applied on the test set for prediction. The main drawback of the GPR model is it's computational time, which increases by an order of n^3 for a training set with n data points.

5 Data

This chapter presents the data that were used in the thesis, together with an overview of its characteristics.

5.1 Data Set

The data contains information on trading prices between Dec 2017 and Jan 2018 of the March 2018 EUR/USD FX Future from CME Globex [5]. Each contract is worth EUR 125,000. Trading hours are Sunday - Friday 6pm to 5pm ETC New York Time. The pip size is 0.0001, meaning that the smallest possible move in price is 0.0001. All data is obtained from the Bloomberg Terminal. The pair EUR/USD is chosen because it is one of the most liquid currency pair in the foreign exchange market [28].

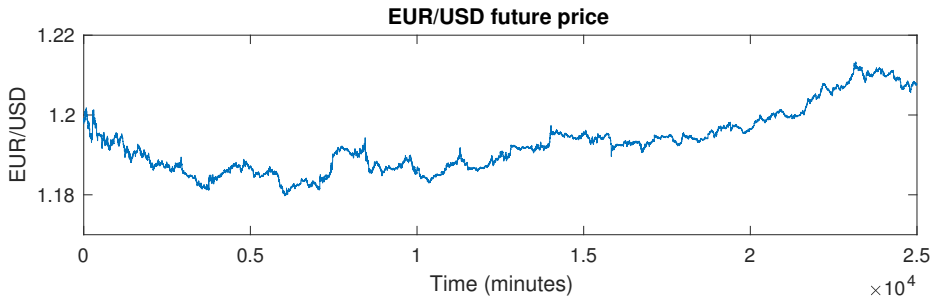

FIGURE 5.1: EUR/USD CME future price data.

The Data consists of around 30,000 trading minutes, see figure 5.1. Each minute contains information on closing price. The training and test data are pairs of $(t_0, y_0), \ldots, (t_n, y_n)$, where t is the time and y is the future price at that time. How the data is divided into training and test sets is explained in Section 6.

5.2 Data Characteristics

When looking at the autocorrelation function of the sample data, we see that the functions decay towards zero, therefore we here assume that the data set is stationary. The data is distributed similarly to a student t distribution because of the fat

tails seen in figure 5.2. This is expected due to periods of excessive price movements, leading to excess kurtosis with fatter tails than a normal distribution. Therefore the Student t distribution is often used when modelling financial data. See figure 5.2 below for an overview of the data characteristics.

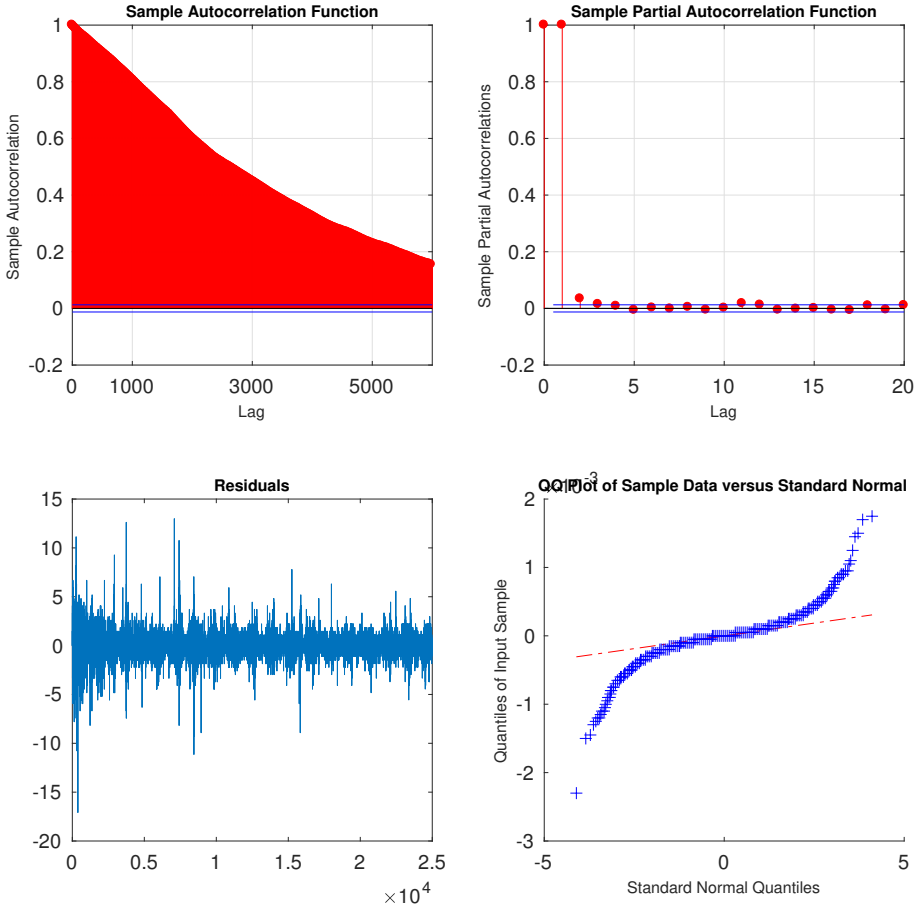

FIGURE 5.2: Data characteristics of the first 400 observations

5.3 Training Set and Test Set

Different data points are used for training, validation, and testing. Typically, a training set is defined, which is used to train the model, and a validation set of the same size is used to validate the model, and later a test set containing new data points is passed through the model to predict the outputs. More details on the choice of training, validation, and test sets are provided in section 6.

6 Methodology and Model Implementations

6.1 Trading Strategies

Two trading models have been developed. The first model assumes no transaction fees, while the second incorporate transaction fees and the Bid-Ask spread in a realistic way. Both models are described in more detail below.

6.1.1 Model 1: Simplified Trading Model

This strategy is very simple. If the next predicted minute price is higher than the current, the future is bought and a long position is taken, if it's lower, the future is sold and a short position is taken. A net position of long or short is always held throughout the trading period so that when a position is exited, a position in the opposite direction is entered.

It is assumed that transaction fees are nonexistent and that the Bid-Ask spread is zero. The results here are not very realistic due to transaction fees and the Bid-Ask spread in real trading. This strategy is used as a first indication of how well different models perform, by looking at the cumulative return. The trading strategy is described in figure 6.1.

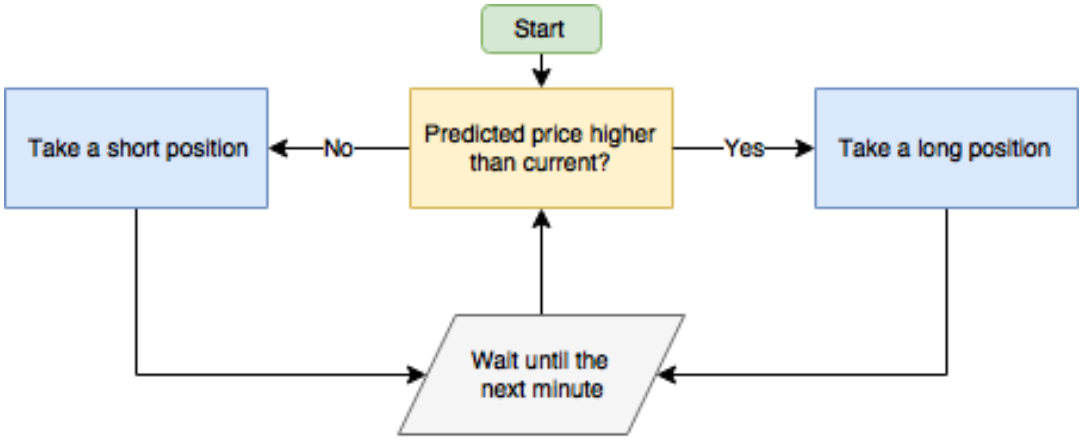

FIGURE 6.1: Flow chart of Trading Model 1

6.1.2 Model 2: Realistic Trading Model

In this strategy, a more realistic approach is used by including transaction fees and an additional fee representing the Bid-Ask spread.

In this strategy, the Bid-Ask spread is assumed to be 1 pip, and an additional transaction fee of 0.0012% every time a position is entered or exited. For every trade (entry and exit) the following fees occur:

1. Enter trade: 0.0001 (1 pip) due to Bid-Ask spread and 0.0012% transaction fee to the broker
2. Exit trade: 0.0001 (1 pip) due to Bid-Ask spread and 0.0012% transaction fee to the broker
3. Total for one trade: 0.0024% + 0.0002 (2x1 pips)

A trade is only entered if the next foretasted price equals a higher return than the total transaction fees. If a trade is already entered, and the next predicted price is in the same direction, the position is held. In addition, if a trade is already entered, and the next predicted price is in the opposite direction, then the trade is exited if the return between current price and the next predicted price is less than the total transaction fees. However, if the next predicted return from the next predicted price in higher than the total transaction fees, then in addition to the exit trade, a short position is taken. The trading strategy is described in figure 6.2.

6.2 GPR Model Selection

6.2.1 Different Kernels

When choosing a kernel, the most popular kernels were investigated. It is reasonable to think that kernels such as the Rational Quadratic could describe the data, this is because of its ability to describe different lengths with multiple hyperparameters so that different trends are captured [23]. In addition to the Rational Quadratic, the Ornstein-Uhlenbeck were also tested because of its popularity within finance [29].

After trying the Rational Quadratic and Ornstein-Uhlenbeck kernels, it was found that the Ornstein-Uhlenbeck gave the best result when fitted to the data. See figure 6.3 for a comparison between different kernels with a training set of 200 observations. When performing the trading simulation, both the Rational Quadratic and Ornstein-Uhlenbeck are tested with different training points.

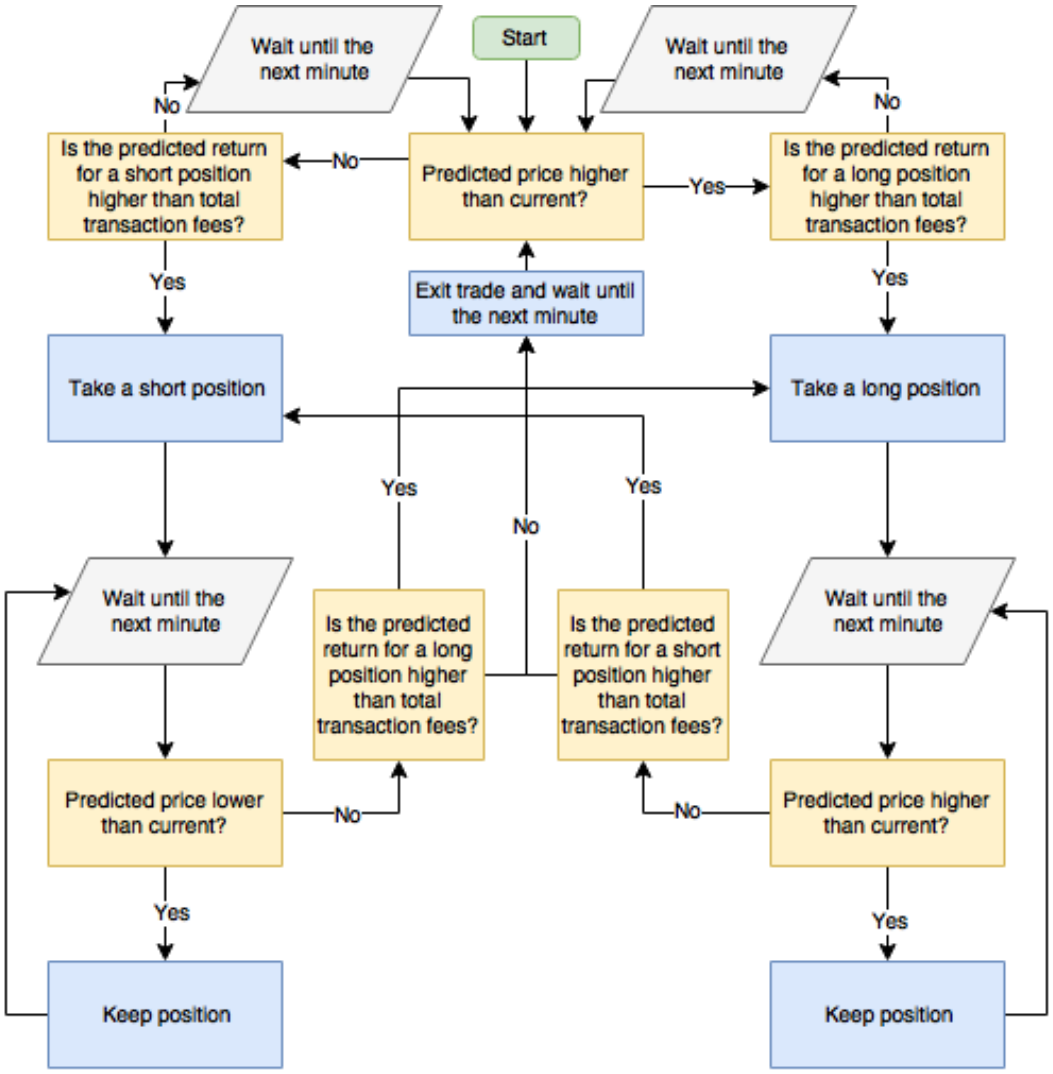

FIGURE 6.2: Flow chart of Trading Model 2

6.2.2 Choosing Number of Training Points

Different lengths of training, validation, and test sets were tested. The number of training points tested were 70, 100, 200, 300, 1000, and 2000, along with a validation test of the same length. The test set contained 25,000 data points. Figure 6.4 gives a visual description of the division of the data.

When different number of training points were tested, a trade-off between computation time and data fit was considered. After trying different training sets sizes, it was found that a cieling of 2000 training points gave a good result for a reasonable computation time. This choice is somewhat arbitrary. To assess whether larger training sets give better result, separate tests with 70, 100, 200, 300, and 1000, and 2000 training points was conducted when running the trading simulations. The results are presented in the Result section below.

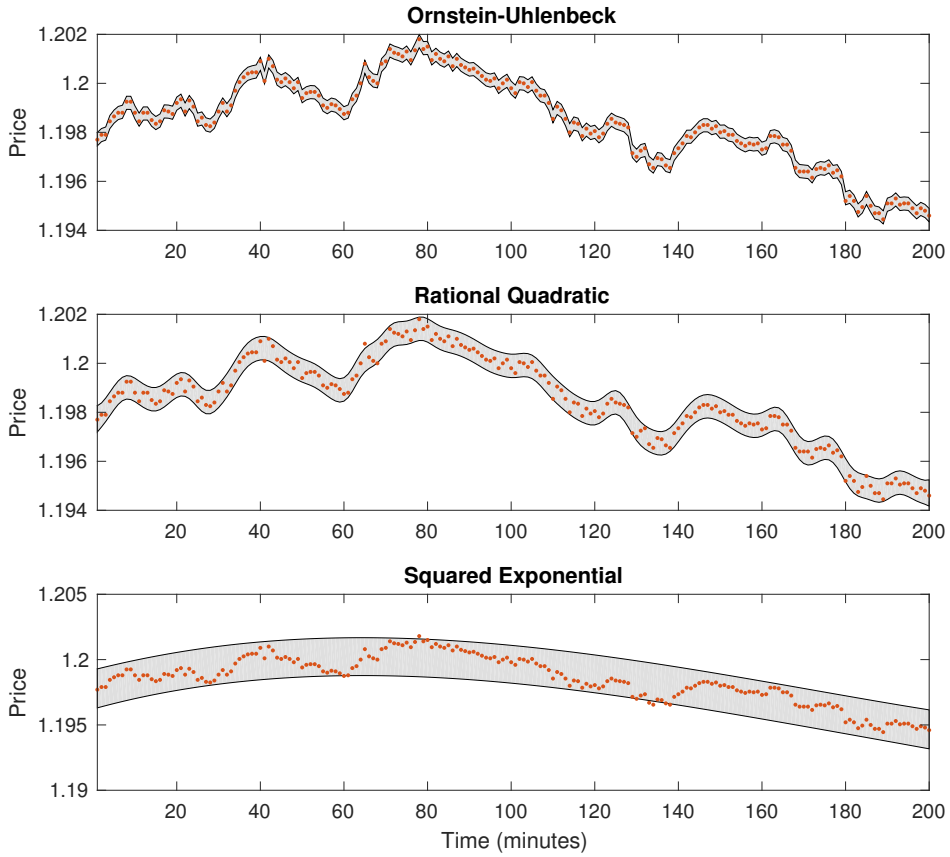

FIGURE 6.3: Data fit with different kernels and 200 training points.

See figure 6.5 of how the data is fitted with 100, 1000, and 2000 training points and an Ornstein-Uhlenbeck kernel.

6.2.3 Student t Compared To Gaussian Likelihood

So far, the Gaussian Likelihood function has been used. However, because our data has fat tails, a student t distribution might be a better fit for the observations. When comparing forecasting from student t with Laplace inference, and the Gaussian likelihood with Gaussian inference, it was found that the student's t gave a worse result when measured by the performance metrics, see table 7.3 below. In addition, because the student t likelihood can't use exact inference, and has to be approximated, it is more time consuming. Therefore, the Gaussian likelihood was preferred in our model.

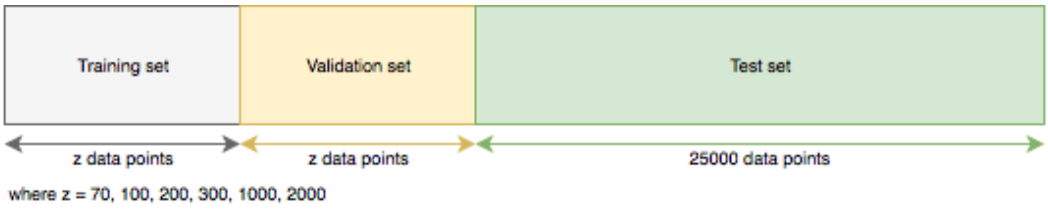

FIGURE 6.4: Overview of how the data set is divided

6.2.4 Final GPR Model

The final model is a Gaussian Process Regression model with a Gaussian likelihood function and a Rational Quadratic or Ornstein-Uhlenbeck. The number of observations in the training that are applied are 70, 100, 200, 300, 1000, and 2000. The size of the trainingset and validation sets are the same, while the test set consists of 25,000 data points. The model uses the observed closing minutes prices of the same size as the trainingset as inputs, to predict the next minute's closing price, once the next minute's real price is observed it is used as an input when the following minute price is predicted. This is illustrated in figure 6.6.

6.3 ARIMA

The ARIMA model based on 1000 observations is presented in table 6.1. According to the Box-Jenkins model selection, because the ACF spikes decay to zero, and the SACF cut off to zero after 1 lags, the model that was chosen was an AR(1).

TABLE 6.1: ARIMA statistics

Parameter	Value	Standard Error	t statistics
Constant	0.000102467	0.000144139	0.71089
AR(1)	0.999914	0.000120882	8271.84
Variance	1.81169e-08	8.2952e-09	2.18402

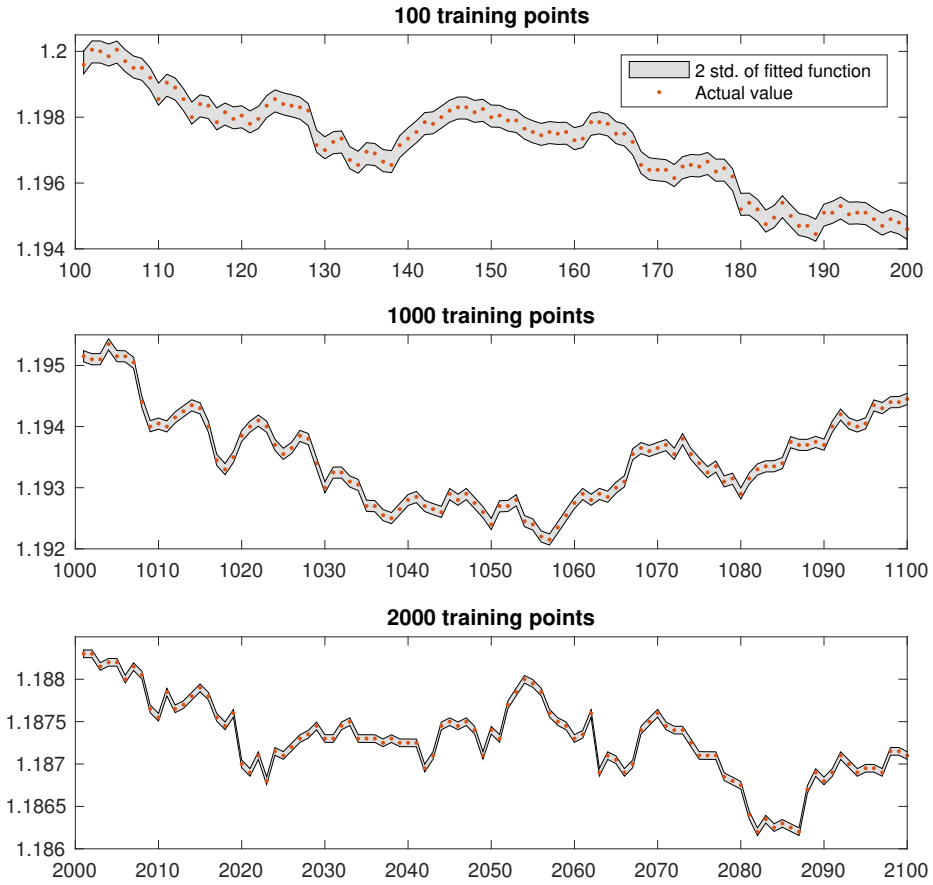

FIGURE 6.5: Data fit for different size of training sets

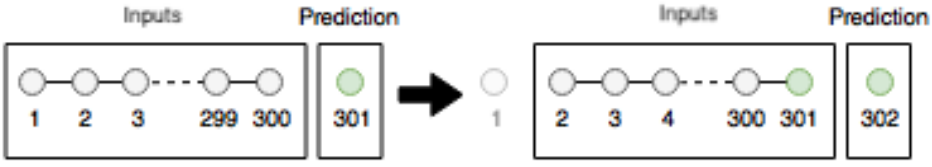

FIGURE 6.6: An illustration of the iterative prediction model.

7 Results

7.1 Random Walk Test

Below the Box-Ljung test has been performed on different lags to test for a random walk process, see table 7.1. The result shows that the random walk can be rejected at a 5% significance level for lags up to 150. This suggests that the price movements are not random, and therefore might be able to be predicted. The results are expected, due to the fact that Zeman and Maršík were also able to reject the random walk process when evaluating spot EUR/USD with 5min frequency [35].

TABLE 7.1: Random walk test for different periods at 5% significance level

Lag	Reject RW?	p-value
2	Yes	0.0021e-07
3	Yes	0.0096e-07
4	Yes	0.0330e-07
5	Yes	0.1183 e-07
10	Yes	0.1259e-07
50	Yes	0.1799e-07
100	Yes	0.0021e-07
150	Yes	0.0002e-07

7.2 GPR Compared ARIMA

The GPR with Gaussian likelihood performed slightly better the ARIMA(1,0,0) model measured by MAE and SMSE, and MSE, see table 7.3. A graphical representation of the two models' predictions are given in figure 7.1. The GPR with a student t likelihood gave worse result compared to the Gaussian likelihood.

TABLE 7.2: Comparison between Gaussian Process Regression and an ARIMA(1,0,0) model with 300 training points (observations) and 200 predicted minutes (forecast)

	GPR	GPR	GPR	ARIMA(1,0,0)
Kernel	Orns.Uhlen.	Orns.-Uhlen.	Rational Quad.	-
Likelihood	Gaussian	Student t	Gaussian	-
MSE	8.5088e-08	5.1409e-07	8.5660e-08	8.5992e-08
SMSE	0.0286	0.2368	0.0300	0.0294
MAE	1.7170e-04	5.4795e-04	1.8013e-04	1.7813e-04

TABLE 7.3: Comparison between Gaussian Process Regression and an ARIMA(1,0,0) model with 1000 training points (observations) and 200 predicted minutes (forecast)

	GPR	GPR	GPR	ARIMA(1,0,0)
Kernel	Orns.Uhlen.	Orns.-Uhlen.	Rational Quad.	-
Likelihood	Gaussian	Student t	Gaussian	-
MSE	3.0465e-08	1.8394e-07	3.1877e-08	3.0494e-08
SMSE	0.0429	0.4809	1.0200	0.0471
MAE	1.2648e-04	3.4687e-04	1.3143e-04	1.2682e-04

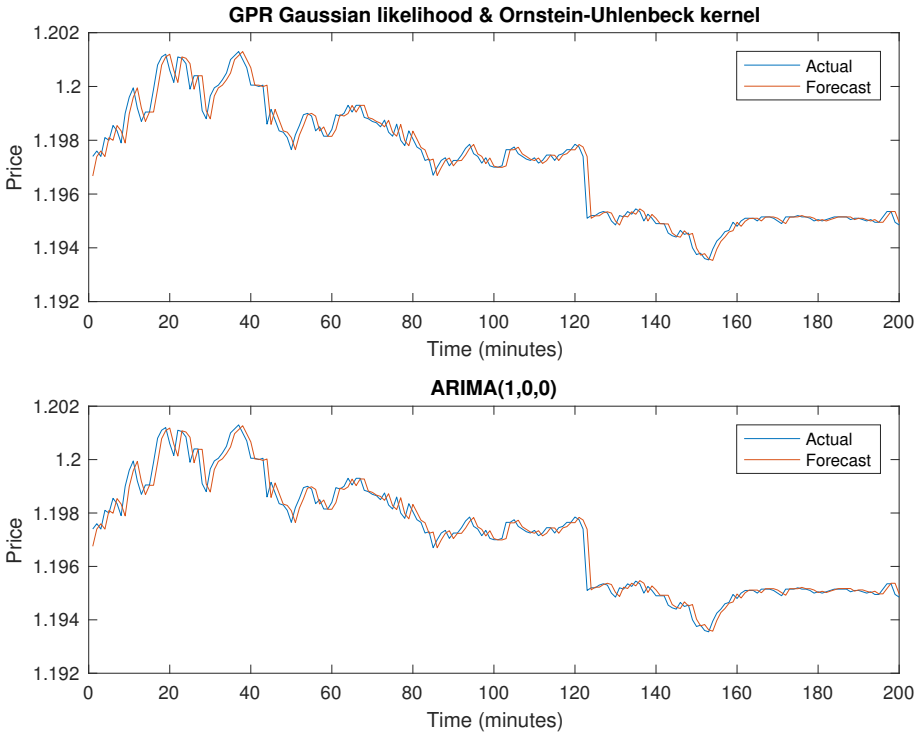

FIGURE 7.1: Comparison between GPR and ARIMA with 300 training points and 200 predicted minutes

7.3 Trading Results

7.3.1 Trading Model 1

The trading models were run for 25,000 trading minutes. The results showed that the trading model where barely profitable on the EUR/USD Future chosen. The model resulted in poor results of returns around -0.5% $-$ 2.1%. The results of the trading simulations are presented in table 7.4 and 7.5 below. All results showed that the trading model performed poorly.

TABLE 7.4: Results of Trading Model 1 during 25,000 trading minutes

# training points	kernel	return	# trades	return/trade	variance
70	Orns.Uhlen.	-0.1006%	5600	-1.7971e-07%	1.4663e-08
100	Orns.Uhlen.	1.5537%	4870	3.1904e-06%	1.2052e-08
200	Orns.Uhlen.	0.5878%	2265	2.5953e-06%	1.1673e-08
300	Orns.Uhlen.	2.1040%	9493	2.2164e-06%	1.1720e-08
1000	Orns.Uhlen.	-0.5074%	5938	-8.5454e-07%	1.0381e-08

TABLE 7.5: Results of Trading Model 1 during 25,000 trading minutes

# training points	kernel	return	# trades	return/trade	variance
70	Rational Quad.	1.2624%	5400	2.3377e-06%	1.1834e-08
100	Rational Quad.	1.5706%	5151	3.0492e-06%	1.1778e-08
200	Rational Quad.	-0.1728%	4992	-3.4612e-07%	1.1932e-08
300	Rational Quad.	0.1425%	1896	7.5177e-07%	1.1143e-08
1000	Rational Quad.	1.6626%	2219	7.4925e-06%	9.3240e-09

7.3.2 Trading Model 2

Naturally, trading model 2 also gave poor results. Below only two examplse of our simulations are presented, see table 7.6.

TABLE 7.6: Result of trading model 2 during 25,000 minutes

# training points	kernel	return	# trades	return/trade	variance
1000	Rational Quad.	-2.1823%	191	-1.1426e-04%	1.3444e-09
1000	Orns.Uhlen.	0%	0	0%	0

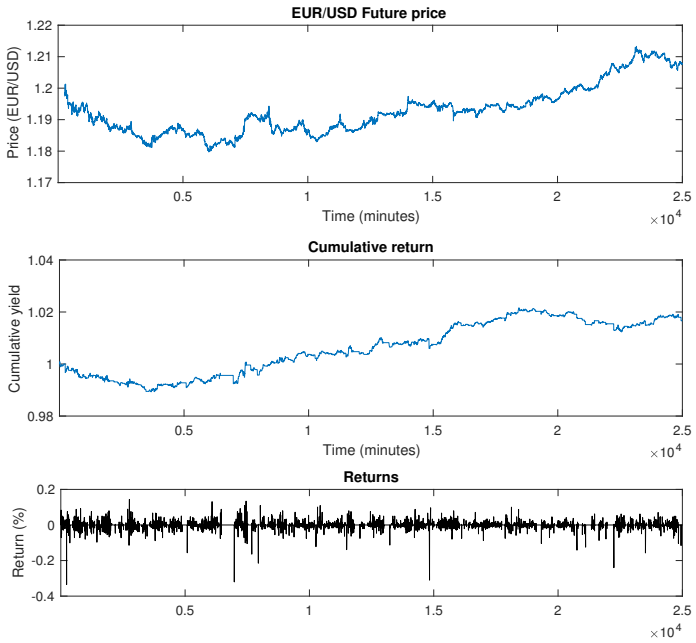

FIGURE 7.2: Result of trading model 1 with a Rational Quadratic Kernel and a training set of 1000 data points, during 25000 minutes

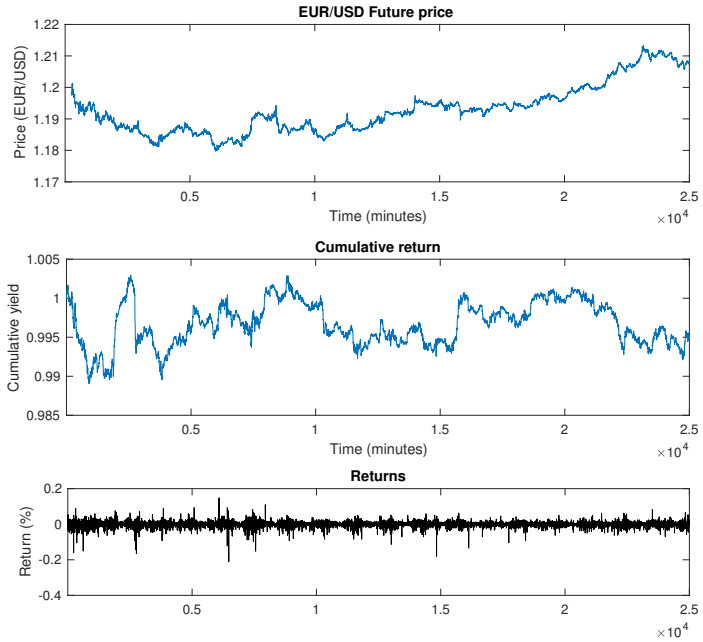

FIGURE 7.3: Result of trading model 1 with a Ornstein-Uhlenbeck Kernel and a training set of 1000 data points, during 25000 minutes

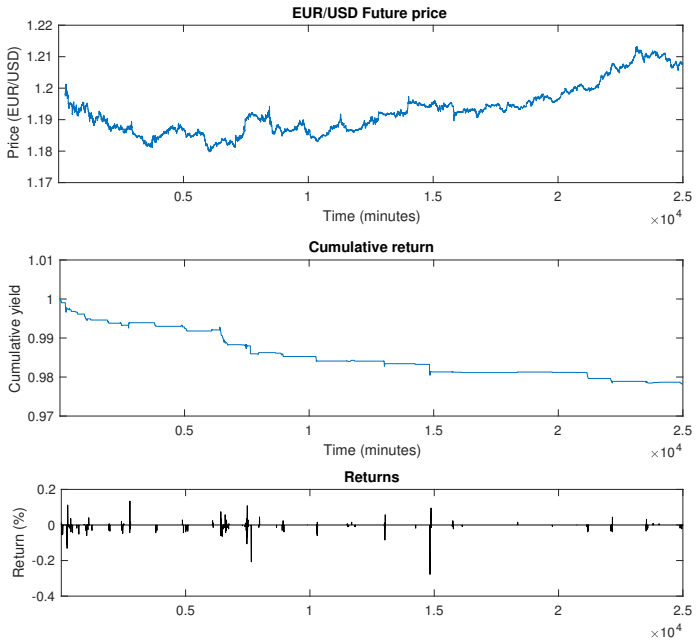

FIGURE 7.4: Result of trading model 2 with a Rational Quadratic Kernel and a training set of 1000 data points, during 25000 minutes

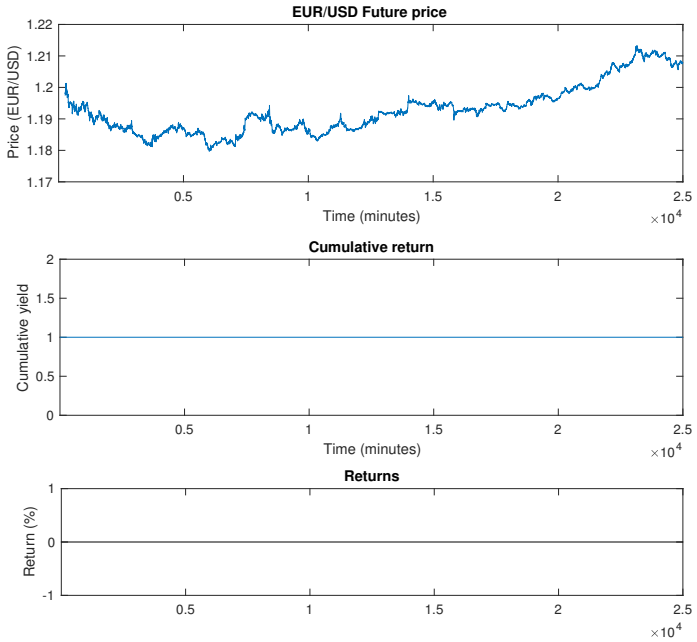

FIGURE 7.5: Result of trading model 2 with a Ornstein-Uhlenbeck Kernel and a training set of 1000 data points, during 25000 minutes

8 Discussion

The results showed that a real live implementation of the trading models, together with the GPR prediction model, would not be profitable. While it was possible to show that the futures price is not a random walk process, the implemented GPR model is not predicting the time series good enough. One important aspect is the direction of the predictions. Since the implemented trading models in this paper, rely on whether the next predicted price is higher or lower than the current, it is essential that the direction predicted is correct. Even if the mean squared error is very low, but the direction of the price is wrongly predicted, that would lead to a non-profitable trading result.

It is not very surprising that the trading simulation were not profitable. The EUR/USD futures are very liquid and are traded actively. It is, therefore, reasonable to assume that many quant funds and other trading firms, already have implemented sophisticated algorithmic strategies, so that it is increasingly harder to find profitable algorithms. The models implemented in this paper is rather simple, and it would therefore make sense to expect models implemented by established quant funds to already have exploited potentially profitable simple models.

The GPR model with a student t likelihood performed even worse than the Gaussian GPR model. This could be due to the fact that the training set fits a Gaussian better than a student t distribution, even though there were signs of excess kurtosis, it does not necessarily mean the student t is a better fit for the data. In addition, the student t likelihood leads to Laplace approximations in the inference, which also could contribute to a poor result.

The implemented models in this paper could be further developed by trying to find a better kernel describing the data and potentially by using more training points with or without some approximation methods. Other potential areas of improvements is multiple input observations. In this paper we only use historic prices, which is a rather simple model, only relying on autocorrelations of the prices. Inputs from other data series could be useful in predicting the future price. These could, for example, be different technical indicators such as moving averages of the prices and volume size of the trades.

There might also be possibilities to combine several prediction models so that one predicts longer than one minute. This becomes especially important in the second

trading model, where only the next minute is predicted, and the model enters a trade only if the return for the price difference is higher than the total transaction fees. If longer periods can be predicted accurately, that would lead to better trades and possibly a higher return. Both implemented models also assume that the time of buying and selling can be done at the end of every trading minute. This might not always be the case in live trading, for example issues with latency can be problematic. In addition, the Bid-Ask spread is assumed to be constant at one pip, this is considered to be a realistic representation of the real Bid-Ask spread, at least during liquid hours. However, the Bid-Ask spread could increase in less active trading hours.

The choice of implementing a GPR model, rather than other machine learning methods, were based on the potential strengths of the GPR model. The main strengths being that the model in non-parametric, and could therefore fit a smooth function to any type of data. In addition, it's a probabilistic method, giving a degree of certainty in the predictions. Furthermore, the GPR implementation requires few assumptions, where the main part of the implementation is choosing a kernel function. The GPR model proved to be in line, or slightly better, than our implemented ARIMA model, measured by our specified metrics. However, it also proved not to be a good predictor to be used for future FX trading models implemented in this paper.

8.1 Conclusion

While it was possible to show that the random walk could be rejected at the 1min frequency for the EUR/USD futures, this does not necessarily mean that the GPR is good enough to predict future movements. In fact, the GPR barely performed better than the ARIMA model implemented. This resulted in a poor performance in trading simulations. It's possible that the model might be improved by introducing more inputs, and/or find better suited kernels. However, the GPR model implemented in this paper was not good enough to obtain a profitable trading strategy.

Bibliography

[1] Hossein Asgharian. "Empirical Finance Lecture Notes - Lund School of Economics". In: (2016). URL: https://www.cmegroup.com/education/files/understanding-fx-futures.pdf.

[2] Hans Byström. "Lecture Notes course code NEKGN81 at Lund School of Economics". In: (Mar. 2018).

[3] Yin-Wong Cheung, Menzie D. Chinn, and Antonio Garcia Pascual. "Empirical exchange rate models of the nineties: Are any fit to survive?" In: (2005).

[4] Taufiq Choudhry et al. "HIGH-FREQUENCY EXCHANGE-RATE PREDICTION WITH AN ARTIFICIAL NEURAL NETWOR". In: (2012).

[5] *CME Group*. Visited 2018. URL: http://www.cmegroup.com/trading/fx/g10/euro-fx.html.

[6] *Daily FX trading volume falls 5.5 pct to $5.1 trillion*. Mar. Visited 2018. URL: https://www.reuters.com/article/bis-currency/daily-fx-trading-volume-falls-5-5-pct-to-5-1-trillion-bis-idUSL8N1BC4PL.

[7] D.H. and R.L.W. "The Big Mac index". In: (Jan. Visited 2018). URL: https://www.economist.com/content/big-mac-index.

[8] Christian L. Dunis and Xuehuan Huang. "Forecasting and Trading Currency Volatility: An Application of Recurrent Neural Regression and Model Combination". In: (2002).

[9] *Evolution Of The Marketplace: From Open Outcry To Electronic Trading*. Visited 2018. URL: https://www.fxcm.com/insights/evolution-of-the-marketplace-from-open-outcry-to-electronic-trading/.

[10] M. Todd Farrell and Andrew Correa. "Gaussian Process Regression Models for Predicting Stock Trends". In: (2007).

[11] Zoubin Ghahramani. "Bayesian non-parametrics and the probabilistic approach to modelling". In: (2012).

[12] Nikola Gradojevic and Jing Yang. "Non-linear, non-parametric, non-fundamental exchange rate forecasting". In: (2006).

[13] Neil Irwin. "Of Kiwis and Currencies: How a 2Global Economic Gospel". In: (Dec. 2014). URL: https://www.nytimes.com/2014/12/21/upshot/of-kiwis-and-currencies-how-a-2-inflation-target-became-global-economic-gospel.html.

[14] Otmar Issing. *The role of fiscal and monetary policies in the stabilisation of the economic cycle*. Nov. 2005. URL: https://www.ecb.europa.eu/press/key/date/2005/html/sp051114.en.html.

[15] Joarder Kamruzzaman. "Forecasting of currency exchange rates using ANN: a case study". In: (2003).

[16] Georg Lindberg, Holger Rootzén, and Maria Sandsten. *Stationary Stochastic Processes For Scientists and Engineers*.

[17] Burton G. Malkiel. "The Efficient Market Hypothesis and Its Critics". In: (2003). URL: https://eml.berkeley.edu/~craine/EconH195/Fall_14/webpage/Malkiel_Efficient%20Mkts.pdf.

[18] Mathworks. *Augmented Dickey-Fuller test*. Visited 2018. URL: https://se.mathworks.com/help/econ/adftest.html.

[19] Mathworks. *Box-Jenkins Methodology*. Visited 2018. URL: https://se.mathworks.com/help/econ/box-jenkins-methodology.html.

[20] Rekar O. Mohammed and Gavin C. Cawley. *Over-Fitting in Model Selection with Gaussian Process Regression*. URL: https://www.springerprofessional.de/over-fitting-in-model-selection-with-gaussian-process-regression/12497380.

[21] Mohammad Mojaddady, Moin Nabi, and Shahram Khadivi. "Stock Market Prediction using Twin Gaussian Process Regression". In: (2011).

[22] ichHang Ou and Hengshan Wang. "Modeling and Forecasting Stock Market Volatility by Gaussian Processes based on GARCH, EGARCH and GJR Models". In: (2011).

[23] Carl Edward Rasmussen and Christopher K.I. Williams. "Gaussian Processes for Machine Learning". In: (2006).

[24] P. Samarasiri. "How do central banks manage exchange rates?" In: (). URL: http://www.cbsl.gov.lk/pics_n_docs/11_edu/_docs/Articles_How_Centralbanks%20manage%20exchange%20rates.pdf.

[25] MARC W. Simpson and AXEL Grossman. "CAN A RELATIVE PURCHASING POWER PARITY-BASED MODEL OUTPERFORM A RANDOM WALK IN FORECASTING SHORT-TERM EXCHANGE RATES?" In: (2011).

[26] Bagas Abisena Swastanto. "Gaussian Process Regression for Long-Term Time Series Forecasting". In: (2016).

[27] *The Future of Computer Trading in Financial Markets*. Visited 2018. URL: http://www.cftc.gov/idc/groups/public/@aboutcftc/documents/file/tacfuturecomputertrading1012.pdf.

[28] "The Most Traded Currency Pairs in the Forex Market in 2016 and Why You Should Choose STO". In: (Visited 2018). URL: https://www.stofs.co.uk/en/newsroom/entry/GENERAL/the-most-traded-currency-pairs-in-the-forex.

[29] Dr Christian Thierfelder. "The trending Ornstein-Uhlenbeck Process and its Applications in Mathematical Finance". In: (2015).

[30] *Understanding FX Futures*. Mar. Visited 2018. URL: https://www.cmegroup.com/education/files/understanding-fx-futures.pdf.

[31] S. Villa and F. Stella. "A Continuous Time Bayesian Network Classier for Intraday FX Prediction". In: (2013).

[32] *Wage Push Inflation*. Visited 2018. URL: https://www.investopedia.com/terms/w/wage-push-inflation.asp.

[33] *Worlbank - Consumer Price Index*. Visited 2018. URL: https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?view=chart.

[34] Faheem Zafari et al. In: ().

[35] Petr Zeman and Martin Maršík. "HIGH-FREQUENCY DATA AND THE EFFECTIVENESS OF THE SPOT EXCHANGE RATE EUR/USD". In: (2013).

[36] Michele Ca' Zorzi, Jakub Muck, and Michal Rubaszek. "Real Exchange Rate Forecasting and PPP: This Time the Random Walk Loses". In: (2015). URL: https://www.dallasfed.org/-/media/documents/institute/wpapers/2015/0229.pdf.