Forecasting Foreign Exchange Rates

A comparison between forecasting horizons and Bayesian vs. Frequentist approaches

Max Nyström Winsa

June 12, 2014

Abstract

Forecasting foreign exchange rates and financial asset prices in general is a hard task. The best model has often been shown to be a simple random walk, which implies that the price movements are unpredictable. In this thesis models that have been somewhat successful in the past are developed and investigated for different forecasting horizons. The aim is to find models that significantly dominate the prediction performance of a random walk, and also to suggest a trading strategy that systematically can make profits using the model predictions. After investigating the data at different sampling frequencies, some significant predictive information is found for very short horizons (10 minutes) and for relatively long horizons (one week), while no useful information is found for daily data. With a forecasting horizon of 10 minutes, it is shown that a Markov model accurately predicts positive or negative returns in more than 50% of the cases for all currencies considered, with significance at the 1% level, and that the performance seems to increase with a Bayesian model. For a horizon of one week, it is shown that a Bayesian Vector Autoregressive (VAR) model outperforms the frequentist VAR model and also the random walk (although with low significance). The performance of trading strategies highly depends on the transaction costs involved. The transaction costs seem to ruin the performance on the 10 minutes horizon, while having less influence on the weekly horizon. A strategy that would have generated good profits on a weekly horizon past 2011, out of sample, is found.

Acknowledgements

Writing this thesis has been an exciting experience, and the cooperation with the hedge fund Lynx Asset Management has been very successful and truly rewarding. They have not only provided me with all the data and computational power to make this thesis possible, but also given me a lot of good advice, a great introduction to algorithmic trading as well as a very friendly and positive working environment.

Especially, I would like to thank my supervisor Tobias Rydén, Professor in Mathematical Statistics and Quantitative Analyst at Lynx, for his guidance and expertise and also because he initiated the possibility of writing this thesis at Lynx.

I would also like to thank all the staff of the Faculty of Engineering at Lund University, who has given me a great education in Engineering Physics, and especially in the truly inspiring fields of mathematical statistics and financial modeling.

Max Nyström Winsa
Stockholm, June 2014

Contents

1. Introduction

1.1 The foreign exchange market

The foreign exchange market is the largest financial market by far, with an average turnover of about $5.3 trillion each day (Bank For International Settlements, 2013). This makes the foreign exchange market very liquid, meaning that one can trade a lot and quickly without having to impact the price very much. The liquidity, or the low transaction costs implied, makes the foreign exchange market very popular to trade even at very short horizons. Besides spot exchange rates, there are a lot of different derivative assets based on foreign exchange rates, such as forward contracts, swaps, futures and options. In this thesis both spot rates and futures prices will be considered.

1.2 The futures contract

Lynx Asset Management, the hedge fund this thesis has been written in cooperation with, is often referred to as a Managed futures fund, or a Commodity trading advisor (CTA). This is because they are exclusively trading so called futures contracts, mainly on commodities, currencies, interest rates and stock indices.

A futures contract is a standardized contract between two parties, a buyer and a seller of a specified asset for a price agreed upon today, but with delivery and payment at a specified future date. In order to minimize the risk of default, the exchange institution requires both parties to put up an initial amount of cash, the so called margin. Additionally, when there is a change in the futures price, the exchange will transfer money from one of the party's margin account to the other's, equivalent to the parties' loss/profit. This is generally done each day. Thus on the delivery date, since all profits and losses already have been settled, the amount exchanged is the spot price of the underlying asset. A consequence of this is that, unlike stocks or options, a position in a futures contract does not cost anything to take. However, there is always transaction costs involved, such as brokerage fees and slippage caused by price movements when putting big orders.

The futures price on foreign exchange rates slightly differs from the spot rate, depending on the differences in interest rates. For example, consider the exchange rate JPY/USD, and assume that the interest rate is higher in USA than in Japan. If the futures price is the same as the spot price, there is an arbitrage opportunity in borrowing money in Japan, depositing the money at a US bank account, and secure the exchange rate in one year by taking a futures contract on JPY/USD.

1.3 The random walk hypothesis

Some researchers argue that foreign exchange rates and financial assets in general are best modeled by random walks. For example this is supported by the PhD thesis of Eugene F. Fama, for daily prices of the Dow Jones Industrial index (Fama, 1965), and argued in *A Random Walk Down Wall Street* (Malkiel, 1973). If the random walk hypothesis is correct, this would imply that the price movements are unpredictable, i.e. that prediction of future positive or negative returns is done at least as good by flipping a coin compared to any other model. This would imply that the work of many financial researches, and also this thesis, is completely pointless.

The random walk hypothesis has however also been rejected by many researchers, for example in *A Non-Random Walk Down Wall Street* (Lo & MacKinlay, 1999). Additional supports against the

Random walk hypothesis are the many systematic portfolio managers and hedge funds which do not apply the “buy and hold” strategy, and have proven a great success in the past.

The random walk hypothesis is also rejected in this thesis, for example we will show that the autocorrelation at lag 1 is significant for returns of exchange rates considered at 10 minute horizons. Strong evidence will be provided that one can systematically predict whether returns will be positive or negative on a 10 minute horizon, accurately in more than 50% of the cases. The results will also indicate that weekly returns are predictable to some degree, even if the support is less significant in this case.

1.4 Hypothesis and suggested models

The primary aim of the thesis is to find a model for forecasting foreign exchange rates, and which can systematically perform better than a random walk model. Another aim is to suggest a trading strategy that is able to make profits, hopefully also when transaction costs are considered. In order to achieve this, different trading horizons has been investigated, and different models have been more or less successful for different horizons.

As a first attempt, different time series models of the VAR structure were tried. A comparison between a frequentist approach, with parameters estimated by least squares, and with one of the Bayesian approaches suggested by Sune Karlsson in the working paper *Forecasting with Bayesian Vector Autoregressions* (Karlsson, 2012) was made.

The frequentist approach to time series models for foreign exchange rates has not been successful in the past, for example all linear models has been rejected for monthly exchange rates during the 70’s (Meese & Rogoff, 1983).

Bayesian VAR models have however been somewhat successful, e.g. when considering monthly samples of a broad range of currencies (Carriero, Kapetanios, & Marcellino, 2008). Carriero et al. used a Bayesian model which does not take correlations between the error terms in the VAR model into account, which possibly is explained by the computational complexity that arise, since one has to use Monte Carlo methods for evaluating predictions. In the paper by Karlsson, it is mentioned that models that take such correlations into account tend to do better than those that don’t (Karlsson, 2012, p. 14). Since a cluster of many machines have been supplied for this thesis the computational complexity is not a big issue, and the later method has therefore been chosen.

For intraday data, on short horizons, the assumption of normally distributed returns needed for the time series models is shown to be suboptimal. Therefore a discrete Markov model is tried in this case, which has been successful for high frequency data before (Baviera, Vergni, & Vulpiani, 2000). A Bayesian approach to Markov chains is also tried, which seem to increase the prediction performance.

2. The data

2.1 The different price series

Two different types of price data will be considered. For daily and intraday data futures prices will be considered, which are limited to seven currencies. For weekly data, in order to incorporate a broader range of currencies, we consider spot rates.

Futures price data for short horizons

The currencies considered on short horizons are measured as the futures price on the exchange rate to the US Dollar. Seven of the most traded FX rates in the world, and which have been supplied, are the Euro (EUR), Japanese Yen (JPY), British Pound (GBP), Australian Dollar (AUD), Swiss Franc (CHF), the Canadian Dollar (CAD) and the New Zealand Dollar (NZD). The data can be sampled on different time bars, spanning from periods of 24 hours down to as short as 5 minutes, and contain information about the High, Low, Open and Close prices, as well as the traded volume during the intervals. The different currencies are sampled at exactly the same periods, which mean that if one market is closed the data is removed for all currencies. This is important from a modeling perspective, where one wants to make sure that no information about any currency is known before another one, so that the causality assumption is appropriate.

In this thesis, futures prices are used to analyze daily returns, as well as returns on 10 minutes intervals. As the intraday market liquidity is very time dependent, price notes are considered only under times of good liquidity during the day, in this case between 14.20 and 21.00 Central European time, which gives us 38 observations of 10 minute returns each day.

Daily spot price data for broader range of currencies at longer horizons

In order to investigate model performance on a broader range of currencies, daily observations of spot rates against the US dollar (High, Low, Open and Close rates) are supplied from Bloomberg, for all currencies traded in the world since the 70’s. However, this data is not causal, which means that the different markets may close and open at different times, so that the daily price notes for one currency is not synchronized in time with the others.

The problem of causality implies that we cannot truthfully use this data to backtest predictions of the return from one day to another, using the returns of all currencies on the previous day. Instead, we assume that the weekly returns are causal, by only considering the returns between the latest known close prices on Wednesday every week.

All currencies considered are presented in table 2.1.

Table 2.1. All currencies considered, their international code names and information of what kind of price data that is supplied. The currencies are measured as the exchange rate to the US dollar.

Currency	ISO 4217 Code	Price data
Euro	EUR	Futures and Spot
Japanese Yen	JPY	Futures and Spot
British Pound	GBP	Futures and Spot
Australian Dollar	AUD	Futures and Spot
Swiss Franc	CHF	Futures and Spot
Canadian Dollar	CAD	Futures and Spot
New Zealand Dollar	NZD	Futures and Spot
Swedish Krona	SEK	Spot only
South African Rand	ZAR	Spot only
Indian Rupee	INR	Spot only
Singapore Dollar	SGD	Spot only
Thai Baht	THB	Spot only
Norwegian Krone	NOK	Spot only
Mexican Peso	MXN	Spot only
Danish Krone	DKK	Spot only
Polish Zloty	PLN	Spot only
Indonesian Rupiah	IDR	Spot only
Czech Koruna	CZK	Spot only
South Korean Won	KRW	Spot only
Chilean Peso	CLP	Spot only
Colombian Peso	COP	Spot only
Moroccan Dirham	MAD	Spot only

2.2 Transformation of the data

First of all we need to differentiate the price data, and consider the returns instead of the actual prices, which is explained by the fact that the prices do not move very drastically and could not be considered to have a constant mean, which is required for stationarity. Usually, when spot prices are considered, one uses the geometric returns:

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}},$$

or the logarithmic returns:

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right),$$

where p_t is the price at time t. However, when considering futures prices, the investor doesn't make cash investment when taking a position. Therefore one may be more interested in the arithmetic returns:

$$r_t = p_t - p_{t-1}.$$

We will use close prices for computing the returns.

The mean of the returns can with high confidence be considered constant zero, especially when considering FX rates. However, the variance cannot be considered constant. In order to assume a constant zero mean and unit variance, we need to estimate the variance during every time interval, and thereafter normalize the returns. In finance the standard deviation is often called volatility, and the choice of volatility measures is a scientific subject on its own.

As we have information about open, high, low and close prices, we can make use of all these when estimating the variance. A volatility measure that takes all this into account is the Yang-Zhang Extension of the Garman-Glass measure (Bennet & Gil, 2012, p. 10), which has been modified with a bit different weighting method. For logarithmic returns the variance is estimated as:

$$\hat{\sigma}_t^2 = \sum_{i=1}^{t-1} w_{t-1-i}\left(\left(\ln\left(\frac{O_i}{C_{i-1}}\right)\right)^2 + \frac{1}{2}\left(\ln\left(\frac{H_i}{L_i}\right)\right)^2 - (2\ln 2 - 1)\left(\ln\left(\frac{C_i}{O_i}\right)\right)^2\right),$$

where O_i, H_i, L_i and C_i are the Open, High, Low and Close prices at time interval i, and the weights $w_i = \frac{1}{\alpha}\left(1 - \frac{1}{\alpha}\right)^i$, where $\alpha > 0$. The number $\left(1 - \frac{1}{\alpha}\right)$ is often called a "forgetting factor".

This variance estimator can be modified to the case of arithmetic returns considered for our futures data. When measuring daily data, one wants to account for the open-close jumps between different days. This is not wanted for intraday data, where we are predicting only within the same day.

The Yang-Zhang volatility measure with arithmetic returns used for our daily data is:

$$\hat{\sigma}_t^2 = \sum_{i=1}^{t-1} w_{t-1-i}\left((O_i - C_{i-1})^2 + \frac{1}{2}(H_i - L_i)^2 - (2\ln 2 - 1)(C_i - O_i)^2\right).$$

For the intraday data the first term (difference in open and close between intervals) is omitted:

$$\hat{\sigma}_t^2 = \sum_{i=1}^{t-1} w_{t-1-i} \left(\frac{1}{2} (H_i - L_i)^2 - (2\ln 2 - 1)(C_i - O_i)^2 \right).$$

Transformation of futures data

As we can sample the futures data at exact synchronized intervals, both daily and for 10 minutes, and we have information about the Open, High, Low and Close prices for these intervals, we use the Yang Zhang measure to estimate the variance, and because we are considering futures prices, we use the arithmetic returns. Conclusively, the transformed quantities for the futures prices investigated and tried to predict is:

$$y_t = \frac{r_t}{\hat{\sigma}_t},$$

where r_t are arithmetic returns, daily or on 10 minutes, $\hat{\sigma}_t$ is the standard deviation estimated by the Yang-Zhang measure for daily or intraday data respectively.

As we also are given data on the traded volumes during the intervals, it can be interesting to take these into account. The volumes, like the returns, cannot be considered stationary, so we use the following transformations:

$$\Delta v_t = \frac{v_t - v_{t-1}}{v_{t-1}},$$

$$x_t = \frac{\Delta v_t - \hat{\mu}_{\Delta v_t}}{\hat{\sigma}_{\Delta v_t}},$$

where $\hat{\mu}_{\Delta v_t}$ is the estimated mean, by the average of previous volume differences, and $\hat{\sigma}_{\Delta v_t}$ the standard deviation, in this case estimated by the weighted mean of squares, which is a more simplistic estimator of non-constant variance. Conclusively:

$$\hat{\mu}_{\Delta v_t} = \frac{1}{t-1} \sum_{i=1}^{t-1} \Delta v_i,$$

$$\hat{\sigma}_{\Delta v_t}^2 = \sum_{i=1}^{t-1} w_{t-1-i} \Delta v_i^2,$$

where the weights $w_i = \frac{1}{\alpha}\left(1 - \frac{1}{\alpha}\right)^i$, $\alpha > 0$ has the same function as in the Yang Zhang measure.

In order to choose α, and thereby the forgetting factors for estimating the variances, we can use that the transformed data should have unit variance over the whole sample. In figures 2.1-4 the sample variance of the normalized quantities of returns and volume differences are plotted against different choices of α whilst estimating the variances, for data sampled daily and on 10 minutes intervals respectively.

However, we do not want α to be too large, since this would ruin the “forgetting effect” needed for stationarizing the data. The variance of 38 independent returns on 10 minute intervals is approximately the same as for the daily sampled data, and when $\frac{1}{\alpha}$ is small α can be approximated to about 38 times larger for the 10 minutes data than for the daily data. All this taken into account, it’s quite hard to choose good values for α with the help of our figures. However, since the sample variances of the normalized quantities seen in figures 2.1-4 are quite close to one for most values of α, we allow ourselves to be a bit imprecise, choosing the α:s reasonanly low as long as the sample variances are not differing too much from one. The chosen values for α is given in table 2.2.

The normalized returns for all currencies are plotted in figure 2.5-6 for the daily and 10 minutes data respectively. According to the plots the data seem to have a constant variance, and are therefore considered stationary.

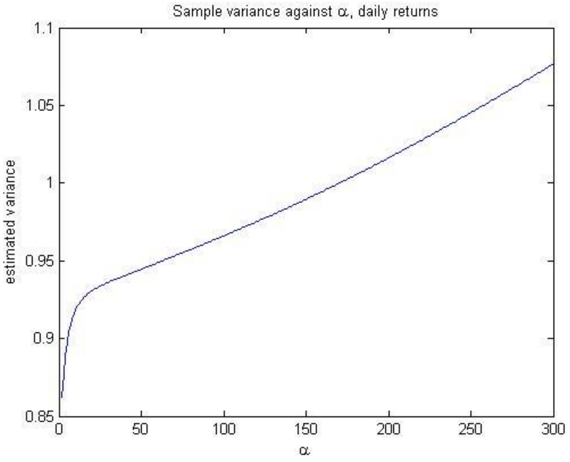


Figure 2.1. Estimated sample variance of normalized returns against α, for daily data between 2005-01-01 and 2014-01-01.

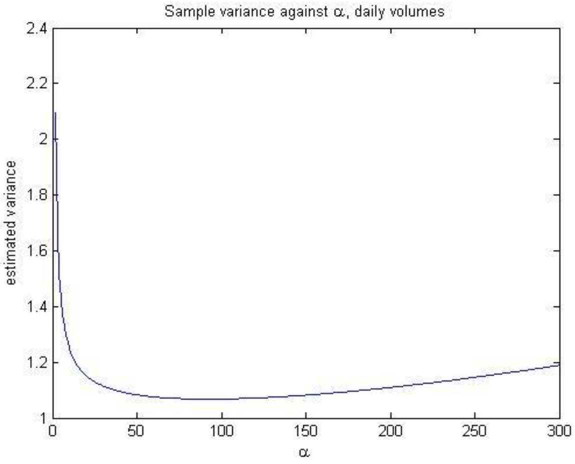


Figure 2.2. Estimated sample variance of normalized volume differences against α, for daily data between 2005-01-01 and 2014-01-01.

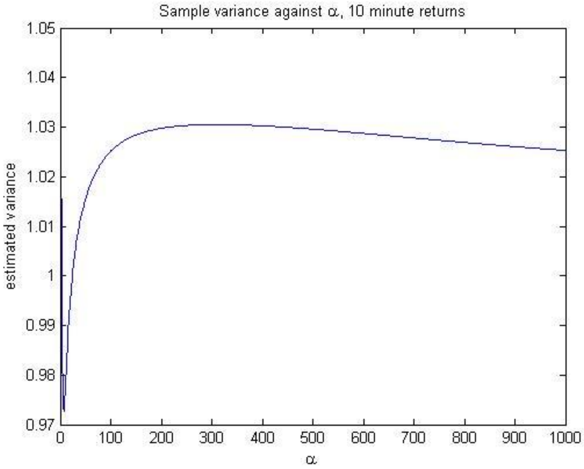


Figure 2.3. Estimated sample variance of normalized returns against α, for 10 minutes data between 2010-01-01 and 2014-01-01.

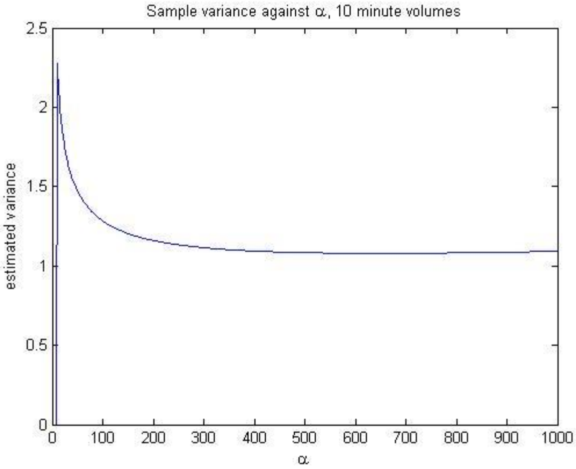


Figure 2.4. Estimated sample variance of normalized volume differences against α, for 10 minutes data between 2010-01-01 and 2014-01-01.

Table 2.2. Chosen values of α, for variance estimation on futures price returns.

	α
Daily returns	16
Daily volumes	50
10 minutes returns	600
10 minutes volumes	600

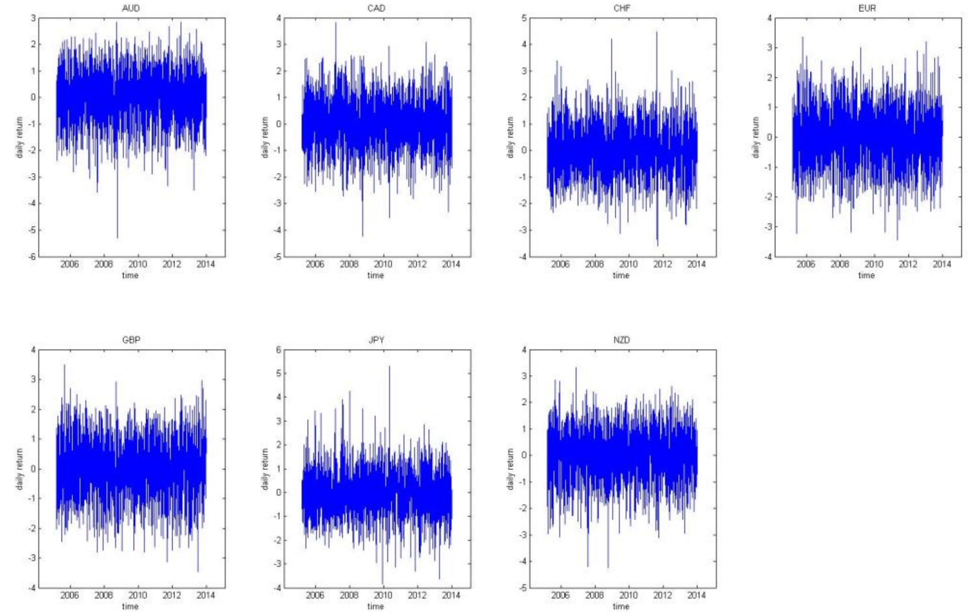


Figure 2.5. Normalized daily returns between 2005-01-01 and 2014-01-01.

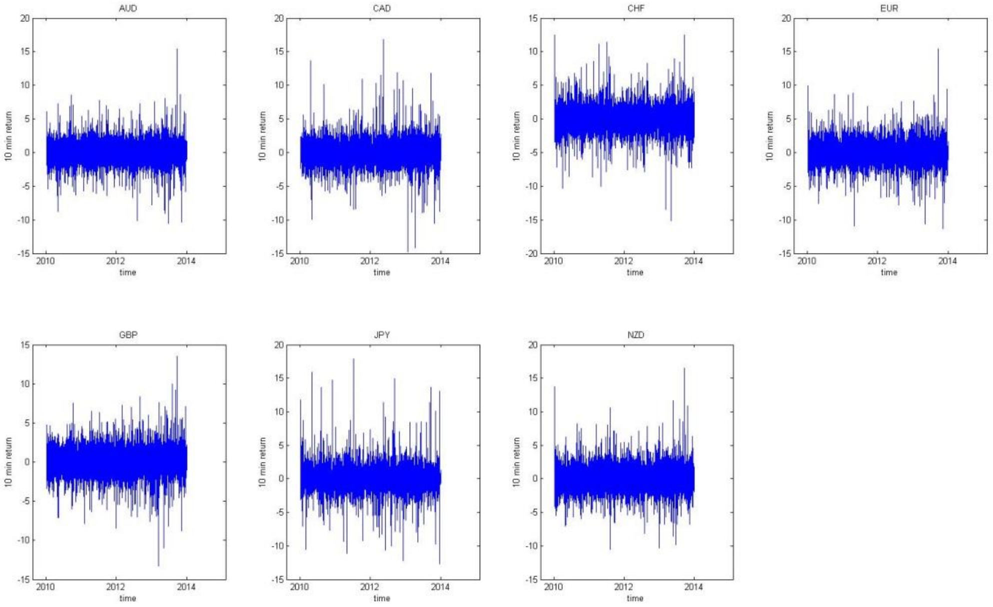


Figure 2.6. Normalized 10 minute returns between 2010-01-01 and 2014-01-01.

Transformation of spot data

As mentioned earlier, we are only considering weekly returns for the spot data in order to make the causality assumption appropriate, since the prices are not synchronized.

As these are spot prices, which require an initial investment, we use logarithmic returns. As we want to make use of our daily observations of open, high, low and close prices for estimating the variance we stationarize the daily returns and are then considering the sum of daily returns between every Wednesday. Conclusively, the quantities investigated in this case are:

$$y_t = \sum_i \frac{r_i}{\hat{\sigma}_i},$$

where r_i and $\hat{\sigma}_i$ are the daily logarithmic returns and their estimated standard deviations the week before t. Usually, when the markets are open on all weekdays, we have 5 daily observations of returns during one week. The standard deviations for daily returns are estimated by the Yang-Zhang measure for logarithmic returns.

The forgetting factor, or α, for estimating the variance is chosen in the same way as for the futures data. However, the theoretical value of the sample variance should in this case be a bit less than 5, since there usually are 5 daily observations within every week, and sometimes a little less. The sample variance against the forgetting factor, α, for weekly returns is plotted in figure 2.7. We choose $\alpha = 8$, which maximizes the sample variance around 4.55. The returns for all currencies are plotted in figure 2.8. Most currencies look stationary, whilst some still seem to have a bit non constant variance. We assume, however, that the series are stationary in further analysis.

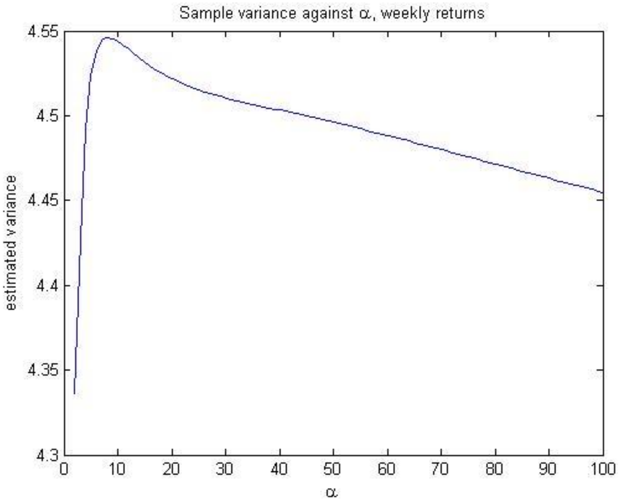
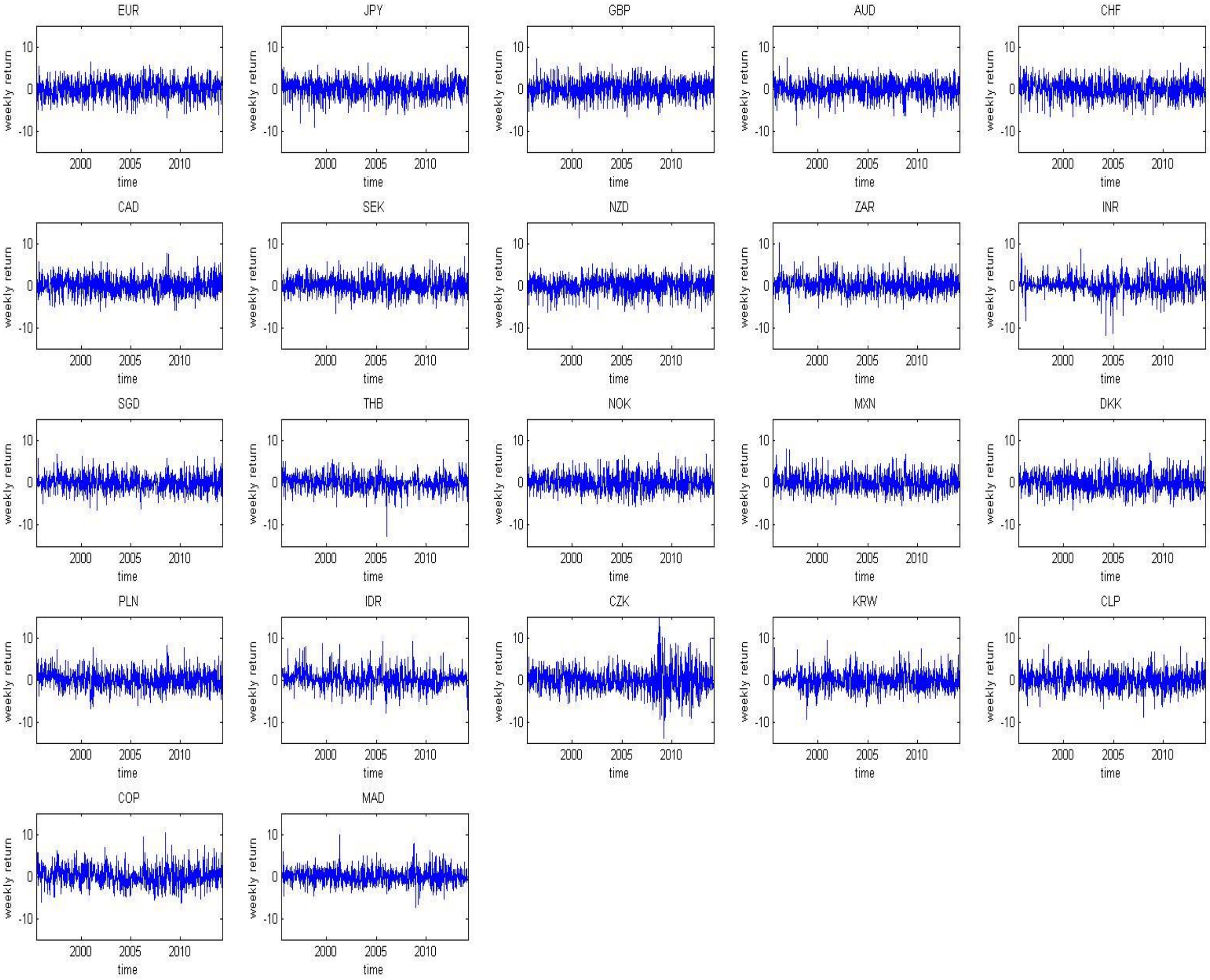


Figure 2.7. Estimated sample variance of returns against α, for weekly data between 1994-07-01 and 2014-03-01.



EUR JPY GBP AUD CHF

CAD SEK NZD ZAR INR

SGD THB NOK MXN DKK

PLN IDR CZK KRW CLP

COP MAD

Figure 2.8. Normalized weekly returns between 1994-07-01 and 2014-03-01.

2.3 Observations of predictive information

In order to find any patterns and predictive information in the data, it can be a good idea to study the autocorrelations within the series of returns, as well as the cross correlations between currencies. In order to be able to predict the return of a currency using this information, we will need significant correlations at other lags than zero. If we consider the daily normalized returns between 2005-01-01 and 2014-01-01, we cannot observe any significant predictive information for any currency. We take the New Zealand Dollar as an example, its Autocorrelation and Cross Correlation with the Japanese Yen is plotted in figures 2.9-10.

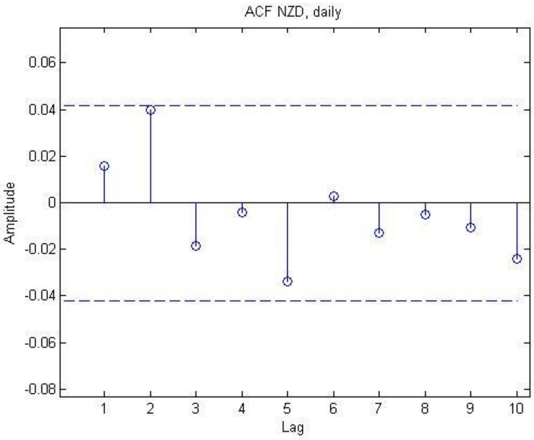


Figure 2.9. Autocorrelation for the New Zealand Dollar. Measured daily between 2005-01-01 and 2014-01-01.

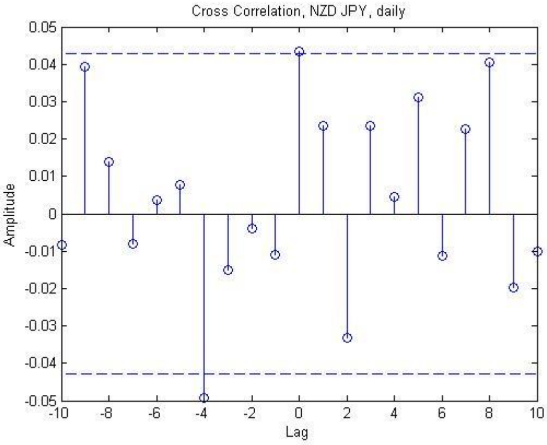


Figure 2.10. Cross Correlation between the New Zealand Dollar (positive lags on the negative axis) and the Japanese Yen (positive lags on the positive axis). Measured daily between 2005-01-01 and 2014-01-01.

The same quantities, but for weekly and 10 minutes returns respectively are plotted in figures 2.11-14. For the 10 minutes data we can observe a significant negative autocorrelation at lag 1, and a positive cross correlation with the Japanese Yen at lag -1. This means that the New Zealand dollar tends to have a positive/negative return 10 minutes after a large negative/positive one in the same series and also after a positive/negative one in the Japanese Yen. This can of course be used when speculating in the New Zealand Dollar. For the weekly returns we cannot observe any predictive information regarding the New Zealand Dollar. However, for the currencies of more developing countries, the situation seems to be different. In figure 2.15 the Cross Correlation between the New Zealand Dollar and the Thai Baht is plotted. We can observe a very significant positive cross correlation with lag 1 for the Thai Baht, which means that the Thai Baht tend to have a positive/negative return one week after a positive/negative return in the New Zealand Dollar. Similar patterns exist for several currencies of developing countries. Another example is the Autocorrelation of the Indian Rupee, plotted in figure 2.16, which is significantly positive at both lag 1 and 2.

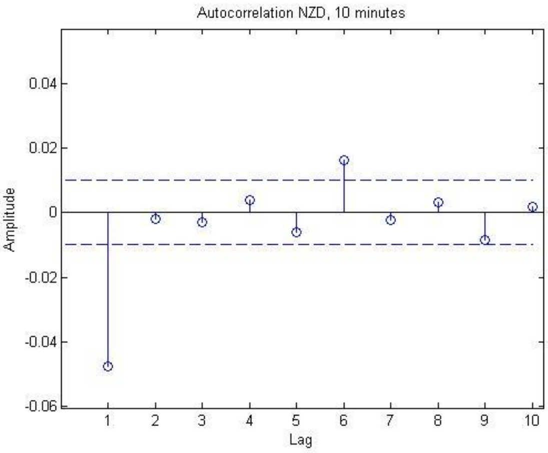


Figure 2.11. Autocorrelation for the New Zealand Dollar. Measured on 10 minutes intervals between 2010-01-01 and 2014-01-01.

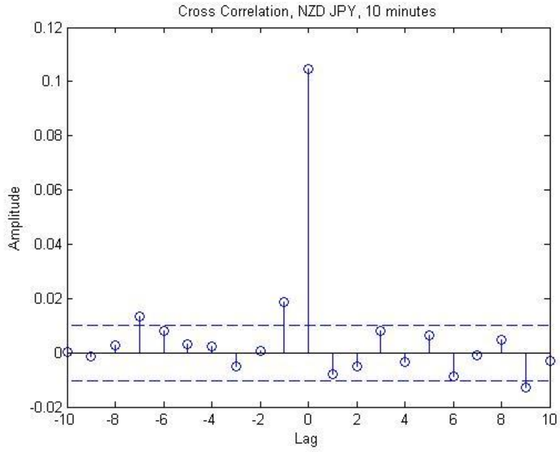


Figure 2.12. Cross Correlation between the New Zealand Dollar (positive lags on the negative axis) and the Japanese Yen (positive lags on the positive axis). Measured on 10 minutes intervals between 2010-01-01 and 2014-01-01.

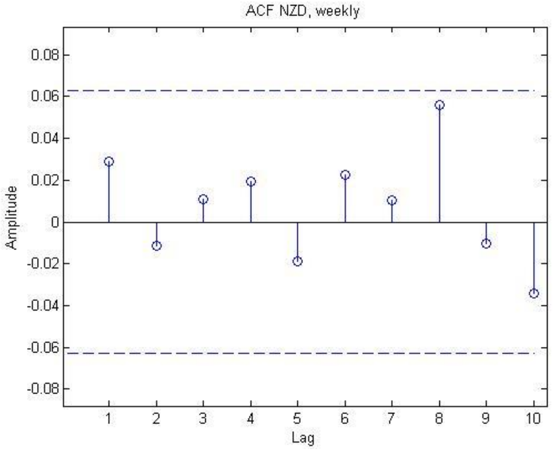


Figure 2.13. Autocorrelation for the New Zealand Dollar. Measured weekly between 1994-01-01 and 2014-03-01.

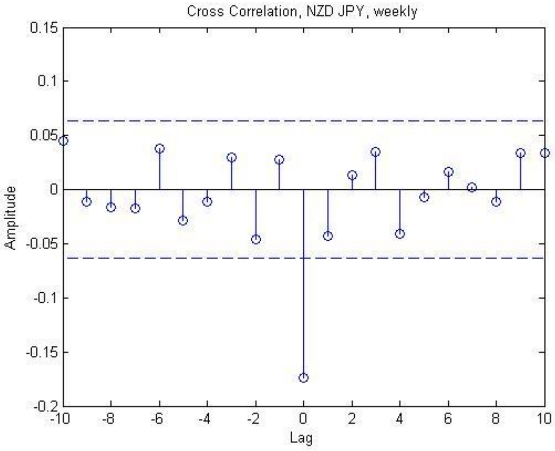


Figure 2.14. Cross Correlation between the New Zealand Dollar (positive lags on the negative axis) and the Japanese Yen (positive lags on the positive axis). Measured weekly between 1994-07-01 and 2014-03-01.

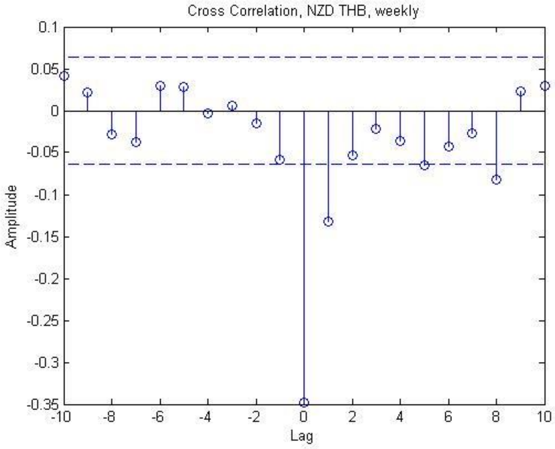


Figure 2.15. Cross Correlation between the New Zealand Dollar (positive lags on the negative axis) and the Thai Baht (positive lags on the positive axis). Measured weekly between 1994-07-01 and 2014-03-01.

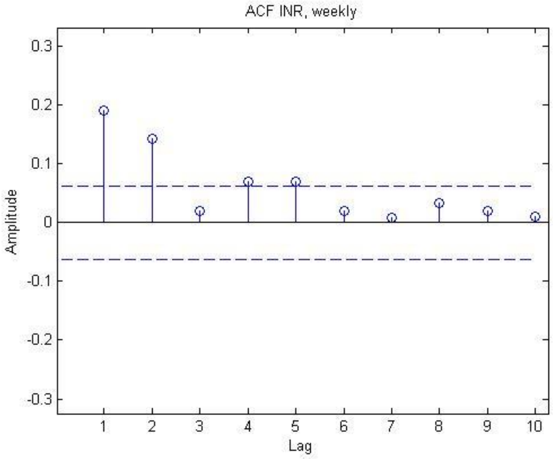


Figure 2.16. Autocorrelation for the Indian Rupee. Measured weekly between 1994-01-01 and 2014-03-01.

Another interesting case to investigate is the cross correlation between the stationarized returns and traded volumes during the intervals. For the ten minutes data the most interesting case is the cross correlations between the Euro and its traded volume, which is plotted in figure 2.17. The same quantity for the daily data is plotted in figure 2.18. We can observe a small significant negative correlation at lag 1 in the 10 minutes data. This is the only significant observation in the 10 minutes data, and there is none in the daily data at lag 1. The volume will be taken into account in one of our models.

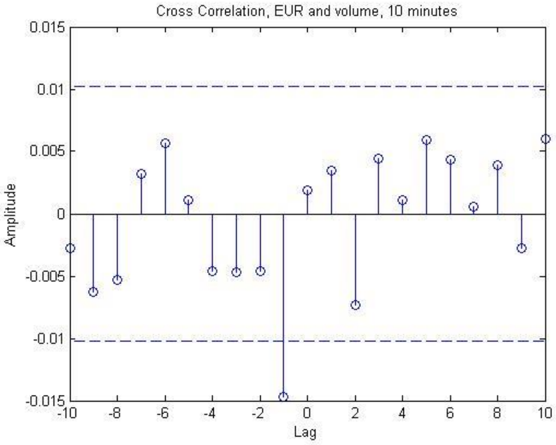


Figure 2.17. Cross Correlation between the Euro (positive lags on the negative axis) and its traded volume (positive lags on the positive axis). Measured on 10 minutes intervals between 2010-01-01 and 2014-01-01.

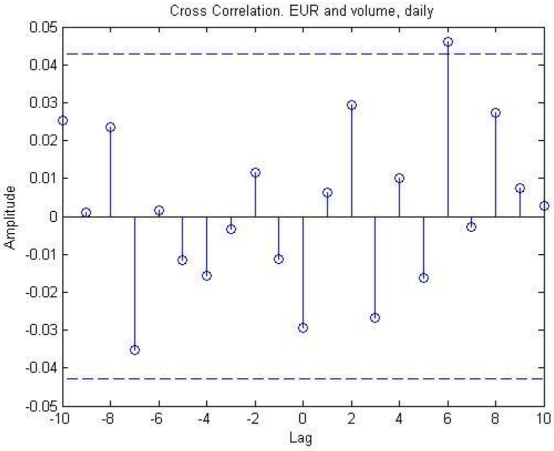


Figure 2.18. Cross Correlation between the Euro (positive lags on the negative axis) and its traded volume (positive lags on the positive axis). Measured daily between 2005-01-01 and 2014-01-01.

An insight after these observations is that there seem to be a lot more predictive information in the 10 minutes data compared to the daily data, at least for the major currencies whose data is measured in futures prices. For the weekly spot returns, there seem to be some very significant predictive information for the currencies of developing countries, but not for the major currencies. These observations motivates why we only try to model the returns on 10 minutes and on weekly basis respectively.

2.4 Distribution of the data

10 minute returns

The model distribution often used, for convenient reasons, is the normal distribution. However our data of normalized returns on 10 minute bars does not really seem normal distributed. The prices most often do not move very drastically on periods of 10 minutes, which commonly only gives us one or two significant figures for the returns, and also many zero-observations (often above 10 % of the cases). The normalized returns seem to have a smaller full width at half maximum (FWHM) and also heavier tails than the normal distribution. These observations might suggest either a discrete distribution or a student's t-distribution as better alternatives. See figure 2.19 and 2.20 for Quantile-Quantile plots of the seven currencies vs. the Normal distribution and the Student's t-distribution respectively. From the figures we can draw the conclusion that a Normal distribution isn't optimal, and that the Student's t-distribution indeed fits much better for all currencies.

Even if the assumption of a normal distribution seems to be suboptimal, we are still going to use it in some models. This is explained by much more convenient modeling, especially in the Bayesian case. However, this is something that could be interesting to improve in future research.

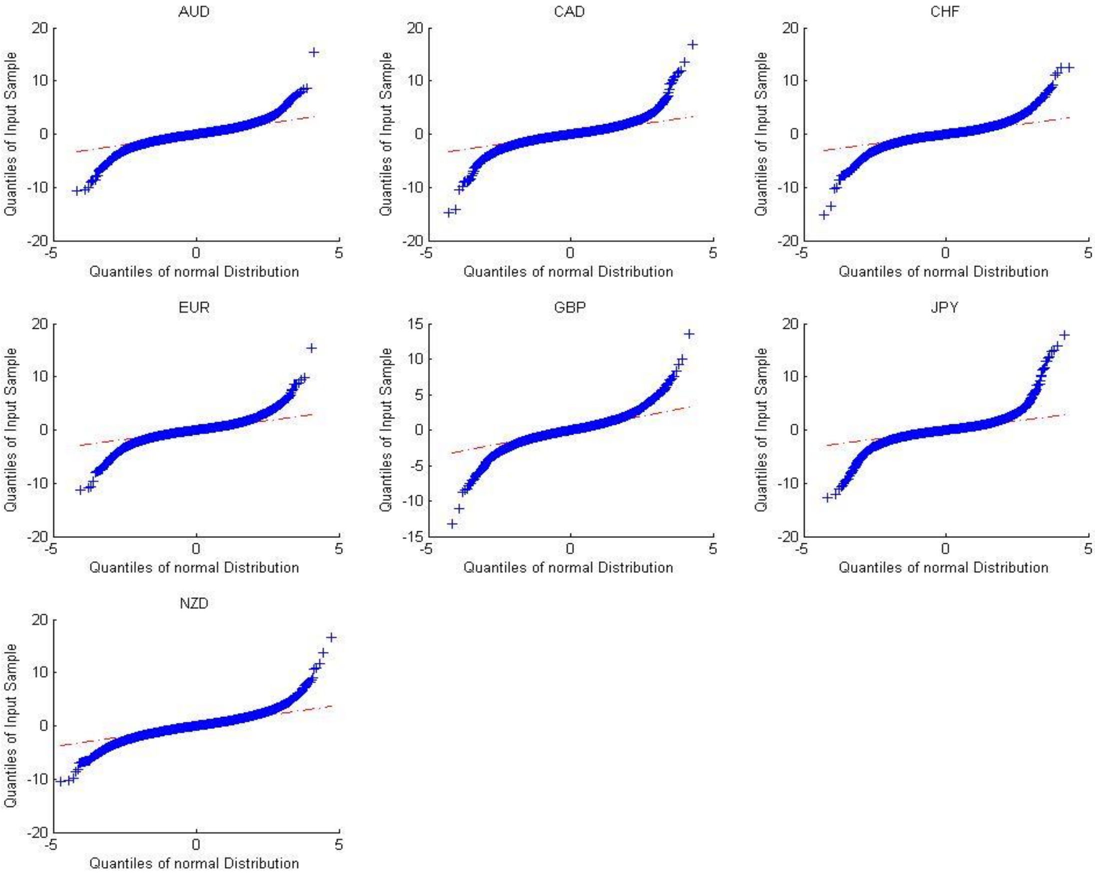


Figure 2.19. Quantile-Quantile plots for the 10 minutes returns versus the Normal distribution. Data from a perfect Normal distribution should follow the dotted line. The returns are measured on 10 minutes intervals between 2010-01-01 and 2014-01-01.

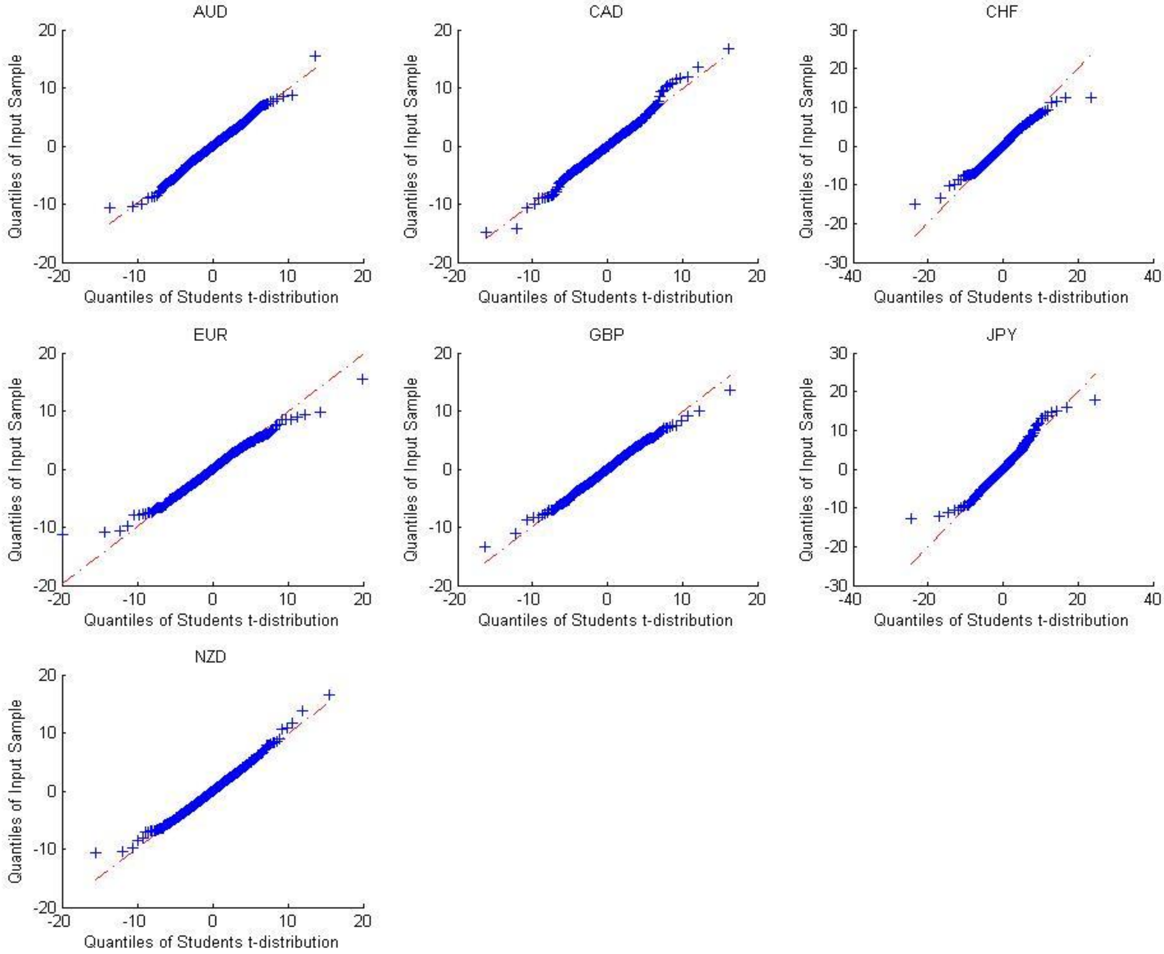


Figure 2.20. Quantile-Quantile plots for the 10 minutes returns versus the Student's t-distribution. Data from a perfect t-distribution should follow the dotted line. The returns are measured on 10 minutes intervals between 2010-01-01 and 2014-01-01. The degrees of freedom of the fitted t-distributions are: 4.31, 3.93, 3.23, 3.41, 3.87, 3.07 and 4.34 for the AUD, CAD, CHF, EUR, GBP, JPY and NZD respectively.

Weekly returns

One reason to try out weekly returns in our modeling, besides involving a broader range of currencies, is that the prices can be expected to move more on longer horizons, and thereby the continuous assumption of the returns is more appropriate. In figure 2.21 Quantile-Quantile plots for all weekly currency returns are plotted against the Normal distribution. One can see that the assumption of a Normal distribution seems much more appropriate in the case of weekly returns for most currencies. The questionable currencies are especially the Indonesian Rupiah (IDR), and maybe also the Indian Rupee (INR).

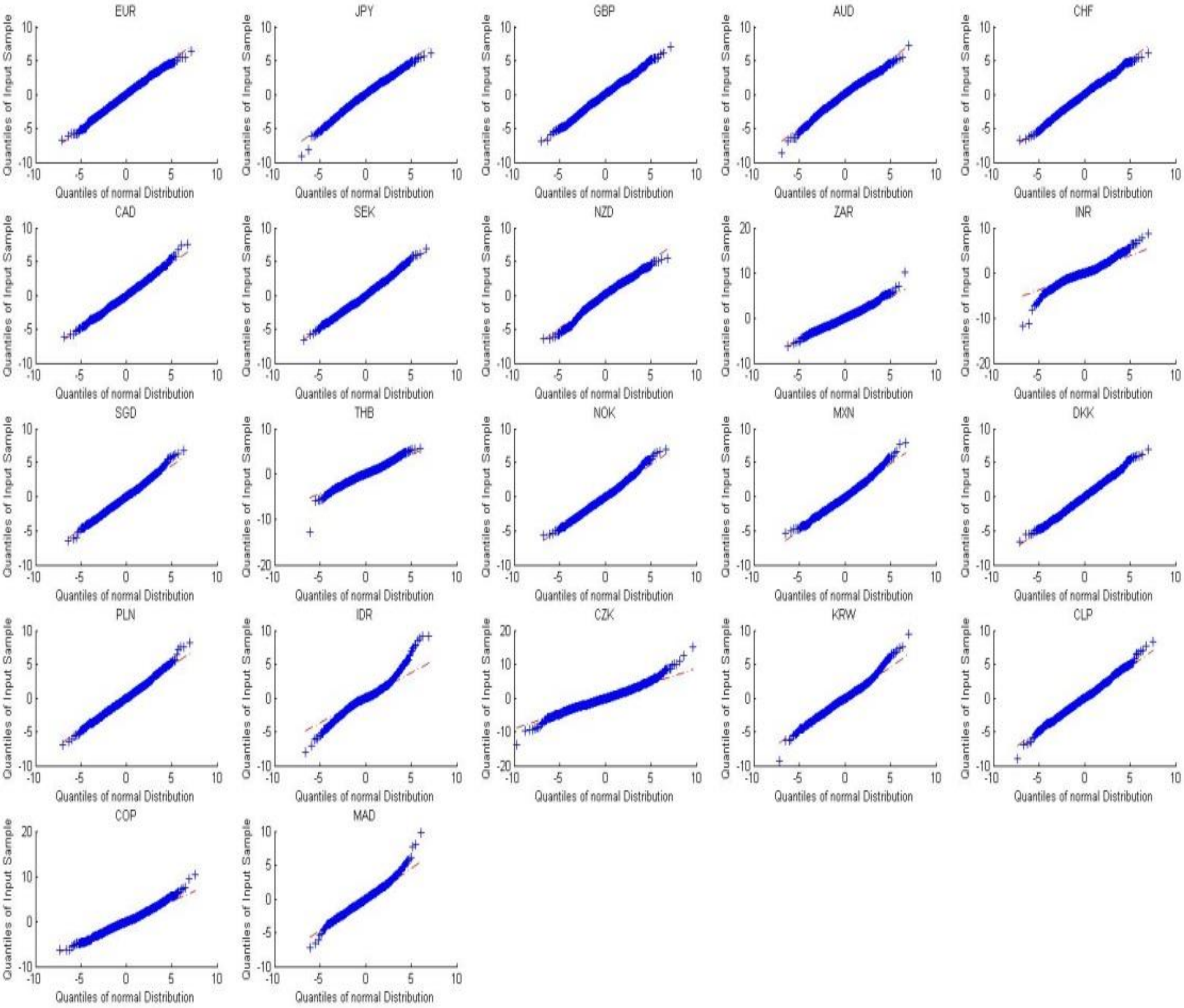


Figure 2.21. Quantile-Quantile plots for the weekly spot returns versus the Normal distribution. Data from a perfect Normal distribution should follow the dotted line. The returns are measured weekly between 1994-07-01 and 2014-03-01.

3. Forecasting models

The problem of modeling the normalized returns can be attacked in a lot of different ways. A broad range of models have been tried in order to make a comparison, and thereafter be able to choose the best performing model for further analysis.

In some models one has to make distributional assumptions on the data, and often the normal distribution is the most convenient one to use. This has been shown to be suboptimal for the 10 minute returns, but more appropriate for the weekly returns.

3.1 Bayesian modeling

If one has some prior beliefs about the data, or most importantly in order to avoid overfitting to certain training samples, one can use a Bayesian approach, and "shrink" the model parameters in the direction of those corresponding to the prior beliefs. The approach aims at reducing the total prediction error by a lower variance of the estimators, but with the price of a higher systematic error called the bias. A more throughout explanation of this is given in 5.1.

Bayesian inference has got its name after Bayes' rule:

$$P(\boldsymbol{\theta}|D) = \frac{P(D,\boldsymbol{\theta})}{P(D)} = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)},$$

where $\boldsymbol{\theta}$ and D are arbitrary random variables. When modeling, one often uses Bayes' rule with D as the observed data, $D = \{y_1, y_2, \dots, y_N\}$,and $\boldsymbol{\theta}$ as the distributional parameters which one wants to estimate. One often uses the terms "prior" for $P(\boldsymbol{\theta})$, "likelihood" for $P(D|\boldsymbol{\theta})$ and "posterior" for $P(\boldsymbol{\theta}|D)$. Notice that $P(D|\boldsymbol{\theta})$ is the likelihood one often wants to maximize in the frequentist approach to inference. $P(D)$ is not very relevant in the case of Bayesian inference, since this term is not a function of the parameters, $\boldsymbol{\theta}$, and therefore is only part of a normalizing constant in the density/probability function of the parameters, i.e. one can often derive the posterior distribution without knowledge about the unconditional distribution of the data.

In order to use Bayes' rule in this way one has to specify a prior distribution for the parameters. One can do this very freely, but there are choices that are more convenient than others. A common approach is to choose the prior such that the posterior belongs to the same family of probability distributions. These are called conjugate priors, and often make the posterior much easier to derive. (Robert, 2001, pp. 113-120). It is also quite intuitive to assume that the parameters belong to one family of distributions, which stays the same also after observations of the data. Conjugate priors are often parameterized distributions, for example: $P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}; \underline{\mu}, \underline{\sigma})$, where $\underline{\mu}$ and $\underline{\sigma}$ can be the prior mean and standard deviation. These are called hyperparameters, and need to be specified, which allow for incorporating prior beliefs, or "shrinkage".

If one does not have any prior beliefs about either distributions or hyperparameters, one can use an uninformative prior. One alternative is to choose the prior as a constant; $P(\boldsymbol{\theta}) \propto 1$, then:

$$P(\boldsymbol{\theta}|D) \propto P(D|\boldsymbol{\theta}).$$

However, this might not be invariant under reparametrization, which means that if we reparametrize the random variables $\boldsymbol{\theta}$, we might get another prior distribution than the one we suggested. One can show (by the change of variable theorem) that a prior that is invariant under reparametrization is:

$$P(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}\,,$$

where $I(\boldsymbol{\theta})$ is the Fisher Information, in matrix form:

$$\left(I(\boldsymbol{\theta})\right)_{i,j} = -E\left[\frac{\partial}{\partial\theta_i} lnP(D|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_j} lnP(D|\boldsymbol{\theta})\right].$$

This choice of prior is often referred to as the Jeffrey's prior.

Note that these uninformative priors might not satisfy the definition of a probability distribution, since the integral of their density functions on $\mathbb{R}^m, m = \dim(\boldsymbol{\theta})$, might not be equal to 1. They are therefore often called improper priors. This is something that in most cases can be overseen, as long as the posterior is a proper distribution (Robert, 2001, pp. 127-140).

When the choice of prior has been made, and the posterior distribution has been derived, inference can be made. One way to infer the parameters is by choosing the maximum a posteriori (MAP) estimates:

$$\widehat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|D)$$

Thereafter a new data observation, y, can be predicted as the maximum likelihood given the parameters:

$$\hat{y} = \max_{y} P(y|\boldsymbol{\theta})$$

However, this approach does not take the parameter uncertainty into account. Another problem with this approach is that the full joint posterior might not be possible to derive analytically, but can be sampled from through a Markov Chain Monte Carlo (MCMC) simulation. In these cases one instead tries to estimate the mean of future values given the data:

$$E[y|D] = \int yP(y|D)dy$$

The estimator used is:

$$\hat{E}[y|D] = \frac{1}{K}\sum_{i=1}^{K} \tilde{y}_i,$$

where $\{\tilde{y}_i\}_{i=1}^{K}$ are independent draws from the predictive distribution, $P(y|D)$. This estimator is known to converge to the mean with large K, by the law of large numbers.

In our case, since the marginal distribution $P(y|D)$ is not known, we will sample from the joint distribution of y and the parameters $\boldsymbol{\theta}$:

$$P(\mathrm{y}, \boldsymbol{\theta}|\mathrm{D}\,) = P(y|\boldsymbol{\theta}, D\,)P(\boldsymbol{\theta}|D).$$

Discarding the parameters gives us the sample $\{\tilde{y}_i\}_{i=1}^{K}$.

This approach does take the parameter uncertainty into account. The method of inference by draws of the predictive distribution:

$$P(y|D) = \int P(y|\boldsymbol{\theta}, D\ )P(\boldsymbol{\theta}|D)d\boldsymbol{\theta},$$

is known as Bayesian model averaging (BMA).

The main idea behind MCMC methods is to find a Markov chain with the same stationary distribution as the distribution one wants to sample from.

In our case we have the model distribution, $P(y|\boldsymbol{\theta}, D\ )$, at hand. However the posterior joint distribution for the parameters, $P(\boldsymbol{\theta}|D)$, might be unknown. Suppose that we can derive the conditional posteriors, $P(\theta_1|D, \boldsymbol{\theta}_{-1}), . . , P(\theta_m|D, \boldsymbol{\theta}_{-m})$, where m is the dimension of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_{-i} \equiv \{\theta_1, \dots, \theta_m\} \backslash \theta_i$. Then we can use a so called Gibbs sampler, given in algorithm 3.1. It can be shown that this Gibbs sampler has $P(y|D)$ as its stationary distribution. However it may take some samples for it to converge, and one should therefore use a burn in, ignoring some number of samples at the beginning. The generated samples are not independent, since the sequence of parameters has the Markov property, and do therefore depend on the most recent update. This is often ignored, or solved by only saving every kth sample, for a predetermined value of k (Robert, 2001, pp. 307-309).

Algorithm 3.1. General Gibbs Sampler, to sample from the predictive distribution:

$\boldsymbol{P(y|D) = \int P(y|\theta, D)P(\theta|D)d\theta}$

```
1. Initialize $\boldsymbol{\theta}^{(0)}$

2. For j=1, …, K:
      -  Draw $\theta_i^{(j)} \sim P\left(\theta_i \middle| \theta_1^{(j)}, \dots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \dots, \theta_m^{(j-1)}, D\right)$,
         for each $i = 1, \dots, m$.

      -  Draw $\tilde{y}_j \sim P(y|\boldsymbol{\theta}^{(j)}, D\ )$

3. Save the samples, $\{\tilde{y}_j\}_{j=1}^{K}$, and discard the parameters.
```

3.2 Time series models

General VAR model

A Vector Autoregressive (VAR) model of order p generally has the form:

$$\boldsymbol{y}_t' = \sum_{i=1}^{p} \boldsymbol{y}_{t-i}' \boldsymbol{A}_i + \boldsymbol{x}_t' \boldsymbol{C} + \boldsymbol{e}_t' = \boldsymbol{z}_t' \boldsymbol{\Theta} + \boldsymbol{e}_t' \,,$$

where $\boldsymbol{y}_t$ is the m-dimensional vector of normalized returns at time t, $\boldsymbol{x}_t$ a d-dimensional vector of exogenous variables, e.g. historical trading volumes, deterministic constants etc., $\boldsymbol{z}_t' = [\boldsymbol{y}_{t-1}', \dots, \boldsymbol{y}_{t-p}', \boldsymbol{x}_t']$ a $k = mp + d$ dimensional vector, $\boldsymbol{\Theta} = [\boldsymbol{A}_1, \dots, \boldsymbol{A}_p, \boldsymbol{C}]$ a $k \times m$ matrix and errors $\boldsymbol{e}_t \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$ a white noise process. This implies:

$$\boldsymbol{y}_t \sim N(\boldsymbol{z}_t' \boldsymbol{\Theta}, \boldsymbol{\Psi}).$$

If we collect values for all times up to T, we can write:

$$\boldsymbol{Y} \triangleq \begin{bmatrix} \boldsymbol{y}_{p+1}' \\ \vdots \\ \boldsymbol{y}_T' \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_{p+1}' \\ \vdots \\ \boldsymbol{z}_T' \end{bmatrix} \boldsymbol{\Theta} + \begin{bmatrix} \boldsymbol{e}_{p+1}' \\ \vdots \\ \boldsymbol{e}_T' \end{bmatrix} = \boldsymbol{Z\Theta} + \mathbf{E}.$$

The parameters, $\boldsymbol{\Theta}$, and the covariance matrix of the noise, $\boldsymbol{\Psi}$, can be estimated by ordinary least squares as

$$\widehat{\boldsymbol{\Theta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1} \boldsymbol{Z}'\boldsymbol{Y},$$

$$\widehat{\boldsymbol{\Psi}} = \frac{\boldsymbol{S}}{T-k},$$

where $\boldsymbol{S} = (\boldsymbol{Y} - \boldsymbol{Z}'\widehat{\boldsymbol{\Theta}})'(\boldsymbol{Y} - \boldsymbol{Z}'\widehat{\boldsymbol{\Theta}})$ is the residual sum of squares (RSS) and $T - k$ are the degrees of freedom. In non-dynamic regression problems, these estimators are known as the minimum variance unbiased estimators. In time series models they are generally not unbiased, but we can expect a much lower bias using this approach compared to the Bayesian models suggested below.

Future returns can be estimated as:

$$\widehat{\boldsymbol{y}}_{T+h}' = \mathbb{E}[\boldsymbol{y}_{T+h}'] = \sum_{i=1}^{h-1} \widehat{\boldsymbol{y}}_{T+h-i}' \widehat{\boldsymbol{A}}_i + \sum_{i=h}^{p} \boldsymbol{y}_{T+h-i}' \widehat{\boldsymbol{A}}_i + \boldsymbol{x}_{T+h}' \widehat{\boldsymbol{C}} \,,$$

or, more compact, if $h = 1$:

$$\widehat{\boldsymbol{y}}_{T+1}' = \mathbb{E}[\boldsymbol{y}_{T+1}'] = \boldsymbol{z}_{T+1}' \widehat{\boldsymbol{\Theta}}.$$

Bayesian VAR model

Currencies behave very random, which implies that our model is very sensitive to overfitting, and a too complex model can ruin the predictions completely by a too high variance. In order to improve our results we want to incorporate prior beliefs, and thereby introduce some bias in order to reduce the variance.

The Bayesian VAR model used is the one with a so called Normal-Wishart prior suggested by Karlsson in the working paper *Forecasting with Bayesian Vector Autoregressions* (Karlsson, 2012, pp. 16-17). An explanation of the model is given below.

To be able to incorporate any beliefs, we need to specify a prior distribution on the parameters $\mathbf{\Theta}$, as well as on the error covariance matrix, $\mathbf{\Psi}$. As it is hard to specify any prior beliefs about the errors, we specify an improper Jeffrey's prior for $\mathbf{\Psi}$, while we specify the natural conjugate prior for the vectorization of the parameters, $\boldsymbol{\theta} = vec(\mathbf{\Theta})$, which is the normal distribution. To be specific, we choose the prior distributions:

$$\boldsymbol{\theta} \sim N\left(\underline{\boldsymbol{\theta}}, \underline{\mathbf{\Sigma}}_\theta\right),$$

$$p(\mathbf{\Psi}) \propto \sqrt{\det I(\mathbf{\Psi})} \propto \det \mathbf{\Psi}^{-\frac{m+1}{2}},$$

where $\underline{\boldsymbol{\theta}}$ and $\underline{\mathbf{\Sigma}}_\theta$ are hyperparameters for the prior mean and covariance matrix of the parameters. We also, a priori, assume independence between $\mathbf{\Theta}$ and $\mathbf{\Psi}$.

As in the general VAR model, we have $\boldsymbol{y}_t \sim N(\boldsymbol{z}_t'\mathbf{\Theta}, \mathbf{\Psi})$. The posterior distributions are derived through Bayes theorem:

$$P(\mathbf{\Psi}, \mathbf{\Theta}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\mathbf{\Psi}, \mathbf{\Theta})P(\mathbf{\Psi}, \mathbf{\Theta})}{P(\boldsymbol{Y})} \propto P(\boldsymbol{Y}|\mathbf{\Psi}, \mathbf{\Theta})P(\mathbf{\Psi}, \mathbf{\Theta}).$$

Since we a priori assume independence:

$$P(\mathbf{\Psi}, \mathbf{\Theta}) = P(\mathbf{\Psi})P(\mathbf{\Theta}),$$

we can state the conditional posteriors:

$$P(\mathbf{\Theta}|\boldsymbol{Y}, \mathbf{\Psi}) \propto P(\mathbf{\Psi}, \mathbf{\Theta}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\mathbf{\Psi}, \mathbf{\Theta})P(\mathbf{\Theta}),$$

$$P(\mathbf{\Psi}|\boldsymbol{Y}, \mathbf{\Theta}) \propto P(\mathbf{\Psi}, \mathbf{\Theta}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\mathbf{\Psi}, \mathbf{\Theta})P(\mathbf{\Psi}).$$

For the likelihood we have:

$$
\begin{aligned}
p(\boldsymbol{Y}|\boldsymbol{\Psi},\boldsymbol{\Theta}) &= (2\pi)^{-\frac{mT}{2}}\det\boldsymbol{\Psi}^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\sum_{t=p+1}^{T}(\boldsymbol{y}_t'-\boldsymbol{z}_t'\boldsymbol{\Theta})\boldsymbol{\Psi}^{-1}(\boldsymbol{y}_t'-\boldsymbol{z}_t'\boldsymbol{\Theta})'\right\} \\
&= (2\pi)^{-\frac{mT}{2}}\det\boldsymbol{\Psi}^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}tr[(\boldsymbol{Y}-\boldsymbol{Z\Theta})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{Z\Theta})']\right\} \\
&= (2\pi)^{-\frac{mT}{2}}\det\boldsymbol{\Psi}^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}tr[\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{Z\Theta})'(\boldsymbol{Y}-\boldsymbol{Z\Theta})]\right\} \\
&= (2\pi)^{-\frac{mT}{2}}\det\boldsymbol{\Psi}^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Psi}^{-1}\left(\boldsymbol{Y}-\boldsymbol{Z}\widehat{\boldsymbol{\Theta}}\right)'\left(\boldsymbol{Y}-\boldsymbol{Z}\widehat{\boldsymbol{\Theta}}\right)\right]\right\} \\
&* \exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Psi}^{-1}\left(\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\right)'\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\right)\right]\right\},
\end{aligned}
$$

where $\widehat{\boldsymbol{\Theta}}$ is the least squares estimate, and the operation $tr(\boldsymbol{A})$ is the trace of the matrix $\boldsymbol{A}$. With $\boldsymbol{\theta} = vec(\boldsymbol{\Theta})$ and $\widehat{\boldsymbol{\theta}} = vec(\widehat{\boldsymbol{\Theta}})$, we can write:

$$tr\left[\boldsymbol{\Psi}^{-1}\left(\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\right)'\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\right)\right] = \left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)'(\boldsymbol{\Psi}^{-1}\otimes \boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}),$$

where ⊗ is the Kronecker product.

Now we get:

$$p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Psi}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)'(\boldsymbol{\Psi}^{-1}\otimes \boldsymbol{Z}'\boldsymbol{Z})\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)\right\} * \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\underline{\boldsymbol{\theta}}\right)'\underline{\boldsymbol{\Sigma}}_{\theta}^{-1}\left(\boldsymbol{\theta}-\underline{\boldsymbol{\theta}}\right)\right\}.$$

This can be written as:

$$p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Psi}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})'\overline{\boldsymbol{\Sigma}}_{\theta}^{-1}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})\right\},$$

where:

$$\overline{\boldsymbol{\Sigma}}_{\theta} = \left(\underline{\boldsymbol{\Sigma}}_{\theta}^{-1}+\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1},$$

$$\overline{\boldsymbol{\theta}} = \overline{\boldsymbol{\Sigma}}_{\theta}\left[\,\underline{\boldsymbol{\Sigma}}_{\theta}^{-1}\underline{\boldsymbol{\theta}}+(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{Z}'\boldsymbol{Z})\widehat{\boldsymbol{\theta}}\right].$$

We recognize this as the normal distribution, and we can therefore state the conditional posterior for $\boldsymbol{\theta}$, as:

$$\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\Psi} \sim \mathrm{N}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\Sigma}}_{\theta}). \qquad (1)$$

The conditional posterior for $\boldsymbol{\Psi}$ follows directly from the likelihood:

$$
\begin{aligned}
p(\boldsymbol{\Psi}|\boldsymbol{Y},\boldsymbol{\Theta}\,) &\propto \det\boldsymbol{\Psi}^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}tr[\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{Z\Theta})'(\boldsymbol{Y}-\boldsymbol{Z\Theta})]\right\} * \det\boldsymbol{\Psi}^{-\frac{m+1}{2}} \\
&= \det\boldsymbol{\Psi}^{-\frac{T+m+1}{2}}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Psi}^{-1}\overline{\boldsymbol{S}}\right]\right\}.
\end{aligned}
$$

This is recognized as the inverse Wishart distribution:

$$\boldsymbol{\Psi}|\boldsymbol{Y},\boldsymbol{\Theta} \sim \mathrm{iW}\left(\overline{\boldsymbol{S}},\mathrm{T}\right), \qquad (2)$$

$$\overline{\boldsymbol{S}} = (\boldsymbol{Y} - \boldsymbol{Z\Theta})'(\boldsymbol{Y} - \boldsymbol{Z\Theta}).$$

The inverse Wishart distribution is the multivariate extension of the inverse gamma distribution, which is the conjugate prior for the variance in a univariate normal distribution.

The predictive distribution with full posterior is:

$$p(\boldsymbol{y}_{T+1}, \dots, \boldsymbol{y}_{T+H}|\boldsymbol{Y}) = \int p(\boldsymbol{y}_{T+1}, \dots, \boldsymbol{y}_{T+H}|\boldsymbol{Y}, \boldsymbol{\Theta}, \boldsymbol{\Psi})\, p(\boldsymbol{\Theta}, \boldsymbol{\Psi}|\boldsymbol{Y}\,) d\boldsymbol{\Theta} d\boldsymbol{\Psi}.$$

Since the full posterior is unknown, we need to implement a Markov Chain Monte Carlo method, and since we have derived the conditional posteriors we can use a straight forward Gibbs Sampler, given in Algorithm 3.2.

With a sample of h-step simulations $\left\{\tilde{\boldsymbol{y}}^{(j)}_{T+h}\right\}_{j=1}^{R} \sim p(\boldsymbol{y}_{T+h}|\boldsymbol{Y})$ at hand, an estimate of $E[\boldsymbol{y}_{T+h}]$ is given by the average:

$$\widehat{\boldsymbol{y}}_{T+h} = \frac{1}{R}\sum_{j=1}^{R} \tilde{\boldsymbol{y}}^{(j)}_{T+h}$$

If one only is interested in estimating the mean, $\widehat{\boldsymbol{y}}_{T+h}$, adding the noise terms, $\boldsymbol{e}_{T+h}$,in step 3 in algorithm 3.2 is unnecessary, as it will only add variance to our predictions and thereby give a higher prediction error. Therefore we leave these out when making our predictions.

Algorithm 3.2. Gibbs sampler for a Bayesian VAR model with a Normal prior on the parameters and a Jeffrey's prior on the covariance matrix of the noise, yielding a sample $\left\{\tilde{\boldsymbol{y}}^{(j)}_{T+1}, \dots, \tilde{\boldsymbol{y}}^{(j)}_{T+H}\right\}_{j=B+1}^{B+R}$ of draws from the predictive distribution. A burn-in of at least $\boldsymbol{B} = \boldsymbol{200}$ iterations is recommended for convergence to a stationary distribution. (Karlsson, 2012, p. 18).

```
Select a starting value for the parameters, $\boldsymbol{\theta}^{(0)}$.

For $j = 1, \dots, B + R$

    1. Generate $\boldsymbol{\Psi}^{(j)}$ from the conditional posterior (2), with $\overline{\boldsymbol{S}}$ evaluated at $\boldsymbol{\theta}^{(j-1)}$.
    2. Generate $\boldsymbol{\theta}^{(j)}$ from the conditional posterior (1), with $\overline{\boldsymbol{\Sigma}}_{\theta}$ evaluated at $\boldsymbol{\Psi}^{(j)}$.
    3. If $j > B$, generate $\boldsymbol{e}^{(j)}_{T+1}, \dots, \boldsymbol{e}^{(j)}_{T+H}$ from $\boldsymbol{e}_t \sim N(\boldsymbol{0}, \boldsymbol{\Psi}^{(j)}\,)$, and calculate recursively:
       $\tilde{\boldsymbol{y}}^{(j)\prime}_{T+h} = \sum_{i=1}^{h-1} \tilde{\boldsymbol{y}}^{(j)\prime}_{T+h-i}\boldsymbol{A}^{(j)}_{i} + \sum_{i=h}^{p} \boldsymbol{y}'_{T+h-i}\boldsymbol{A}^{(j)}_{i} + \boldsymbol{x}'_{T+h}\boldsymbol{C}^{(j)} + \boldsymbol{e}^{(j)\prime}_{T+h}$
```

Choice of hyperparameters

When specifying the hyperparameters, $\underline{\boldsymbol{\theta}}$ and $\underline{\boldsymbol{\Sigma}}_{\theta}$, we want to incorporate some prior beliefs of our data. As currencies often are believed to behave like random walks, it is a good idea to shrink our model to those beliefs. This was first done (with other economic variables) by Robert Litterman (Litterman, 1979). If the currency prices behave like univariate random walks, it means that their returns behave like Gaussian noise, i.e. $\boldsymbol{y}_t = \boldsymbol{e}_t \sim N(\mathbf{0}, \boldsymbol{\Psi})$. In order to shrink our model in this direction, we put the prior mean for the elements of $\boldsymbol{\Theta}$ to:

$$(\underline{\boldsymbol{\Theta}})_{ij} = \mathbb{E}\left[(\boldsymbol{\Theta})_{ij}\right] = 0, \qquad i = 1,..,k, \qquad j = 1,\dots,m.$$

We want to apply more shrinkage (i.e. specify a smaller prior variance) to lags of independent covariate variables, as well as to larger lags than to small. A modification of the original Litterman prior is the following, for the standard deviations of the elements of $\boldsymbol{\Theta}$:

$$(\mathbf{T})_{ij} = SD\left[(\boldsymbol{\Theta})_{ij}\right] = \begin{cases} \dfrac{\pi_1}{l^{\pi_3}}, & lag\ l\ of\ the\ dependent\ variable, i = (l-1)+j \\ \dfrac{\pi_1 \pi_2 s_j}{l^{\pi_3} s_r}, & lag\ l\ of\ the\ independent\ variable\ \ r \neq j, i = (l-1)+r \\ \pi_1 \pi_4 s_j, & exogenous\ variables, i = (mp+1),\dots,k \end{cases},$$

where s_j^2 are the diagonal elements of the least squares estimate of the residual covariance matrix $\frac{\boldsymbol{S}}{T-k}$, so that s_j/s_r accounts for the different variances of the variables. π_1 is a hyperparameter for the "overall tightness". π_2, π_3 and π_4 control the tightness for independent variables, different lags and exogenous variables (e.g. traded volumes) respectively.

Then we specify $\underline{\boldsymbol{\theta}} = vec(\underline{\boldsymbol{\Theta}})$ and $\underline{\boldsymbol{\Sigma}}_{\theta}$ as the diagonal matrix of $vec(\mathbf{T}'\mathbf{T})$, i.e.:

$$\underline{\boldsymbol{\Sigma}}_{\theta} = diag\left(vec\left(\left[(\mathbf{T})_{ij}^2\right]\right)\right), \qquad i = 1,..,k, \qquad j = 1,\dots,m\,,$$

which implies that the parameters get the specified prior variances, and that we assume no prior covariance between different parameters.

3.3 Markov models

General Markov model

One way to exploit the observation about the negative autocorrelation at lag 1 (or higher lags if desired) for the 10 minutes data, without making the normality assumption, is by modeling the returns with a Markov model. The model suggested will however not take any notice to correlations between currencies or with volumes, in order to keep the number of Markov states reasonably limited.

In this model, the state at time t will represent how big/small return that is observed at time t. The positive and negative returns are divided evenly in $N/2$ different states each, i.e. as quantiles of the data (and thereby get an even number of states, N ,in total). With this approach we can estimate the transition probabilities for observing a large positive return in 10 minutes given a large negative return in the present etc. We can also just estimate the probability of a positive/negative return given the present state in general.

Define the N different states as:

$$Y = \{y_1, y_2, \dots, y_N\}.$$

In this case Y corresponds to different intervals of negative and positive returns respectively. Let the state at time t be given as Y_t. The transition probabilities for a Markov chain of order M are then defined as:

$$p_{j_M,\dots,j_1,i} = P\left(Y_t = y_i \middle| Y_{t-1} = y_{j_1}, \dots, Y_{t-M} = y_{j_M}\right),$$

where $\{j_1, \dots, j_M\} \subseteq \{1, \dots, N\}$. The corresponding $N^M \times N$ transition matrix is:

$$T = \begin{bmatrix} p_{1,\dots,1,1} & \cdots & p_{1,\dots,1,i} & \cdots & p_{1,\dots,1,N} \\ p_{1,\dots,2,1} & \cdots & p_{1,\dots,2,i} & \cdots & p_{1,\dots,2,N} \\ \vdots & \ddots & \vdots & & \vdots \\ p_{j_M,\dots,j_1,1} & \cdots & p_{j_M,\dots,j_1,i} & \cdots & p_{j_M,\dots,j_1,N} \\ \vdots & & \vdots & \ddots & \vdots \\ p_{N,\dots,N,\dots,1} & \cdots & p_{N,\dots,N,\dots,i} & \cdots & p_{N,\dots,N,N} \end{bmatrix}.$$

The transition probabilities can be estimated by maximum likelihood as:

$$\hat{p}_{j_M,\dots,j_1,i} = \frac{n_{j_M,\dots,j_1,i}}{\sum_{k=1}^{N} n_{j_M,\dots,j_1,k}},$$

where $n_{j_M,\dots,j_1,i}$ is the number of observed transitions in the order of states: $j_M, \; ..,j_1, i$.

With the estimated transition probabilities for the states of returns at hand, it is straightforward to estimate the probability of a specific future state, and also for simply a positive/negative return. The estimate for the probability of a positive return is:

$$\hat{p}_{j_M,\dots,j_1,u} = \sum_{i_u} \hat{p}_{j_M,\dots,j_1,i_u},$$

where i_u corresponds to states of positive returns. The probability of a negative return is estimated in the same way:

$$\hat{p}_{j_M,\dots,j_1,d} = \sum_{i_d} \hat{p}_{j_M,\dots,j_1,i_d},$$

where i_d corresponds to states of negative returns.

Bayesian Markov model

A Bayesian approach for general, observable, Markov chains is to assume a multinomial distribution for the occurrences, i.e. for transitions from any fixed path of length M, $j_M, \; ..,j_1$, to state i:

$$n_i \sim multinomial(\boldsymbol{p}, n_{tot}),$$

where $\boldsymbol{p} = (p_1, \dots, p_N)$ are the transition probabilities to states $1, \dots, N$.

The probability mass function of the multinomial distribution is:

$$P(n_1, \dots, n_N) = \begin{cases} \frac{n_{tot}!}{n_1! \cdots n_N!} \prod_{i=1}^{N} p_i^{n_i}, & if \sum_{i=1}^{N} n_i = n_{tot} \\ 0, & otherwise \end{cases}.$$

The conjugate prior for the probabilities in the multinomial distribution is the Dirichlet distribution:

$$p_i \sim Dir(\alpha_1, \dots, \alpha_N),$$

where the hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$ are the prior occurrences of transitions to states $1, \dots, N$ (Robert, 2001, p. 121).

The density function of the Dirichlet distribution is:

$$P(\boldsymbol{p}) = \frac{\prod_{i=1}^{N} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{N} \alpha_i\right)} \prod_{i=1}^{N} p_i^{\alpha_i - 1}.$$

It is easy to derive the posterior. Using Bayes' rule:

$$P(\boldsymbol{p}|n_1, \dots, n_n) \propto P(\boldsymbol{p}, n_1, \dots, n_n) \propto \prod_{i=1}^{N} p_i^{\alpha_i + n_i - 1},$$

which can be identified as a new Dirichlet distribution, and it is proved that the Dirichlet distribution indeed is the conjugate prior. Conclusively, the posterior is:

$$\boldsymbol{p}|n_1, \dots, n_n \sim Dir(\alpha_1 + n_1, \dots, \alpha_N + n_N),$$

with the conditional mean and variance:

$$E[p_i|n_1, \dots, n_n] = \frac{\alpha_i + n_i}{\alpha_0},$$

$$Var[p_i|n_1, \dots, n_n] = \frac{(\alpha_i + n_i)(\alpha_0 - (\alpha_i + n_i))}{\alpha_0^2(\alpha_0 + 1)},$$

where $\alpha_0 = \sum_{k=1}^{N}(\alpha_k + n_k)$.

This approach allows us to shrink the transition probabilities towards 1/N, which means that all states are equally probable. This is done by choosing $\alpha_1 = \alpha_2 = \cdots = \alpha_N$, the larger we choose them, the more shrinkage is applied.

When modeling, the transition probabilities are simply estimated by their posterior means:

$$\hat{p}_i = \frac{\alpha_i + n_i}{\alpha_0}.$$

Note that this is done for all possible paths, $j_M, \, .., j_1$, to state i, those are just left out from the sub-indexes for a more convenient notation.

As we are only interested in shrinking the model towards equally probable states, we only get one hyperparameter:

$$\alpha = \alpha_1 = \alpha_2 = \cdots = \alpha_N$$

3.4 Making predictions

For the VAR model, predictions can be made in a very straightforward way, since we are actually estimating the mean of future returns, $\hat{\boldsymbol{y}}_t$. However, if we are only interested in predicting whether the return is going to be positive or negative, and only want to have an opinion about this at times where we consider ourselves reasonably certain about the outcome, for example in order to make a trade. Then we can try to predict the sign of the returns as follows:

$$\widehat{sign}(y_t) = \begin{cases} 1, & if\ \hat{y}_t > \kappa \\ -1, & if\ \hat{y}_t < -\kappa \\ no\ opinion, & otherwise \end{cases},$$

where $\kappa \geq 0$ is a threshold that the predictions need to exceed in order to have an opinion about a positive/negative return. Only the case with $\kappa = 0$ will be considered under the results in 7.1.

For the Markov models, since we do not have any predicted values of returns, but only transition probabilities corresponding to different states of positive/negative returns, a similar approach is conducted in all cases:

$$\widehat{sign}(y_t) = \begin{cases} 1, & if\ \hat{P}(y_t > 0|y_{t-1},..,y_{t-M}) \geq \eta \\ -1, & if\ \ \hat{P}(y_t < 0|y_{t-1},..,y_{t-M}) > \eta \\ no\ opinion, & otherwise \end{cases},$$

where $\eta \geq 0.5$ is a threshold that the transition probability corresponding to a positive/negative state needs to exceed in order to have an opinion.

3.5 Updating the models

Even if we have assumed stationarity, the correlation structure within or between the return series might be time dependent. We want dynamic models that are able to adapt to these changes over time, therefore we are not only training the models on specific previous samples of data, but are updating them over time, when new samples are observed. To do this, we choose a number of previous samples that our model is trained on and either update our model at every iteration, or choosing a frequency of how often the model is updated, i.e. the model is updated every k_u:th iteration using β previous data points. The choice of updating frequency depends on the computational complexity of the model. Updating the models at every iteration can take a lot of time, especially when validating the models in order to specify hyperparameters. The number of previous samples to be used for training, β, is regarded as a hyperparameter chosen by out of sample validation.

4. Trading strategies

When having predictions of future returns we need to specify when we want to trade, and what currencies to trade during certain time intervals. An intuitive way to trig the trades is to specify suitable thresholds for the predicted return or (for the Markov models) a threshold for the estimated probability of negative/positive return in the next state.

When we've come up with suitable FX rates to trade during a certain interval, then we need to know how much we should go long/short in each currency. To do this we use the modern portfolio theory (MPT) suggested by Harry Markovitz (Wikipedia, MPT, 2014).

4.1 Portfolio theory

The main idea of MPT is to maximize our portfolio's expected return:

$$E[R_p] = \boldsymbol{\mu}'\boldsymbol{w},$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$, $\mu_i = E[r_i]$, are the expected returns on the traded markets, and $\boldsymbol{w}$ is the weights for the assets of the portfolio.

At the same time we don't want to take too high risks, which mean that we need a limit for the variance of the portfolio:

$$Var[R_p] = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}\,,$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the returns.

This can be stated as the following optimization problem:

$$\max_{\boldsymbol{w}} \boldsymbol{\mu}'\boldsymbol{w}\,, \qquad s.t\; \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \leq C.$$

This problem can be rewritten as:

$$\max_{\boldsymbol{w}} L(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad L(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu}'\boldsymbol{w} - \lambda(\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - C)\,,$$

where $\lambda > 0$ is a Lagrange multiplier. The gradient w.r.t the weights is:

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu} - 2\lambda\boldsymbol{\Sigma}\boldsymbol{w}\,.$$

Setting the gradient equals to zero gives us the solution:

$$\boldsymbol{w}^* = \frac{1}{2\lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\,.$$

The analysis will be based on the normalized returns, which each should follow processes of zero mean and unit variance, which implies that the covariance matrix is the same as the correlation matrix. We use the weights $\boldsymbol{w}_t = \widehat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\mu}_t$, where $\boldsymbol{\mu}_t$ will be our predictions of normalized returns at time t, and $\widehat{\boldsymbol{\Sigma}}_t$ the sample covariance matrix for the previously observed normalized returns, which with zero mean is:

$$\widehat{\boldsymbol{\Sigma}}_t = \frac{1}{(t-1)-1}\sum_{i=0}^{t-1} \boldsymbol{y}_i \boldsymbol{y}_i{}',$$

Where $\boldsymbol{y}_i$ are the vectors of normalized returns at time i.

With weights at hand we can test our model and estimate the expected returns and standard deviation for the portfolio. To do this, we should also take the transaction costs into account. If we assume that the transaction costs are constant, C for all FX rates, we get the portfolio return at time t as:

$$R_{p,t} = \sum_{i=1}^{N} w_{t,i} y_{t,i} - C \sum_{i=1}^{N} \left| w_{t,i} - w_{t-1,i} \right|.$$

VAR models

For the VAR models we can use the MPT very straightforward as:

$$\boldsymbol{\mu}_t = \widehat{\boldsymbol{y}}_t$$

$$\boldsymbol{w}_t = \widehat{\boldsymbol{\Sigma}}_t{}^{-1} \boldsymbol{\mu}_t$$

Where $\widehat{\boldsymbol{y}}_t$ is our predictions of stationarized returns at time t, when all information up to time $t-1$ can be observed.

Markov models

In the Markov models we estimate transition probabilities from the current state to a state corresponding to a positive/negative return. Therefore we will not have predictions of the mean of the future returns, and cannot use MPT in the same straightforward way as for the VAR models. However, we can make predictions of the sign of returns, and are simply putting:

$$\boldsymbol{\mu}_t = \begin{cases} 1, & if\ \hat{P}(y_t > 0 | y_{t-1}, .., y_{t-M}) \geq 0.5 \\ -1, & if\ \hat{P}(y_t < 0 | y_{t-1}, .., y_{t-M}) > 0.5 \end{cases}$$

With the sign-predictions at hand, the weights are computed as before:

$$\boldsymbol{w}_t = \widehat{\boldsymbol{\Sigma}}_t{}^{-1} \boldsymbol{\mu}_t \,.$$

5. Performance measures

In order to validate our models and be able to tell whether their predictions are good or not, we will need performance measures. The ones used are the root mean squared error (RMSE), when we are predicting values of the returns as in the VAR models, and the hit rate when we only try to predict whether future returns are expected to be positive or negative. In order to evaluate the trading strategies, we use the Sharpe ratio as the performance measure.

5.1 Root mean squared error

The most common performance measure in regression is the mean squared error (MSE), and its square root (RMSE). The MSE is often also referred to as the "loss function" one want to minimize by the regression. The MSE for the prediction $\hat{y}$ of the dependent variable y is defined as:

$$MSE(\hat{y}) = E[(\hat{y} - y)^2].$$

In order to minimize the MSE, one is often referring to the problem of finding an optimal trade-off between bias and variance. This is where the Bayesian modeling comes into place, where one incorporate prior beliefs for the model parameters. This results in a lower variance of the estimator, but also introduces a systematic error between the estimator and the dependent variable, i.e. the bias. With this approach one can reduce the total expected error compared to the unbiased estimators, e.g. ordinary least squares in the case of regression.

To get a better understanding about the tradeoff between bias and variance, we prove the fact that the MSE can be decomposed in three terms: the variance of the estimator, the squared bias of the estimator and the variance of the so called innovation.

One can think of $\hat{y}$ as an estimate of $E[y]$, the true mean value of y given all information known at the moment when the estimate is made. The bias can then be defined as the difference between the mean of the chosen estimator and $E[y]$, i.e. a systematic error in the model. Consider the term $\epsilon = y - E[y]$, generally called the innovation. It has zero mean and is uncorrelated with $\hat{y}$ and $E[y]$.

The decomposition of the MSE can be done as follows:

$$MSE(\hat{y}) = E[(\hat{y} - y)^2]$$

$$= E[(\hat{y} - E[y] - \epsilon)^2]$$

$$= E[\epsilon^2] + E[(\hat{y} - E[y])^2]$$

$$= Var(\epsilon) + E[(\hat{y} - E[y])^2],$$

since the innovation, ϵ, has zero mean and is uncorrelated with $(\hat{y} - E[y])$. In the second term, we add and subtract $E[\hat{y}]$, which is the true mean of the prediction, i.e. the average prediction with infinitely many replications of the training data.

$$E[(\hat{y} - E[y])^2]$$

$$= E[(\hat{y} - E[\hat{y}] + E[\hat{y}] - E[y])^2]$$

$$= E[(\hat{y} - E[\hat{y}])^2] + E[(E[\hat{y}] - E[y])^2] + 2E[(\hat{y} - E[\hat{y}])(E[\hat{y}] - E[y])]$$

$$= E[(\hat{y} - E[\hat{y}])^2] + (E[\hat{y}] - E[y])^2$$

$$= Var(\hat{y}) + Bias(\hat{y})^2,$$

since $(E[\hat{y}] - E[y])$ is a constant and $E\left[\hat{y} - E[\hat{y}]\right] = E[\hat{y}] - E[\hat{y}] = 0$. Conclusively:

$$MSE(\hat{y}) = Var(\epsilon) + Var(\hat{y}) + Bias(\hat{y})^2.$$

With the observed values: $y_1, \ldots, y_N$, and the corresponding predictions: $\hat{y}_1, \ldots, \hat{y}_N$, the MSE is estimated as:

$$MSE(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2.$$

In order to get the same scale as the quantity predicted, we use the square root:

$$RMSE(\hat{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}\ .$$

A common way to get an idea of the prediction accuracy for financial data is to compare the result with the one of a random walk-model for the prices, i.e. setting the mean of the returns to zero, and we get the RMSE:

$$RMSE_0 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i)^2}.$$

5.2 Hit Rate

In some cases it can be more interesting to investigate how often we are able to predict the correct sign of the returns. In these cases we use the hit rate as the performance measure:

$$HR = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{\{sign(y_i)=\widehat{sign}(y_i)\}}\ .$$

Since we have a lot of zero observations of returns (especially in the 10 minutes data), we only count those that are strictly positive or negative when computing the hit rate. This is reasonable, since we only try to predict positive and negative signs, and we doesn't really lose money on a zero return, at least not if the transaction costs are omitted.

For the hit rate we can estimate a confidence interval to investigate if the hit rate is significantly better than 50%, i.e. if we can predict the signs of positive/negative returns better than random.

With the assumption that the hit rate is independent in time, we can put:

$$\sum_{i=1}^{N}\mathbb{I}_{\{sign(y_i)=\widehat{sign}(y_i)\}} \sim Bin(N,p)\ ,$$

which gives us:

$$HR = \hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\{sign(y_i)=\widehat{sign}(y_i)\}} \,.$$

The standard deviation of the hit rate is estimated as:

$$\hat{\sigma}_{HR} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \;:$$

Then a confidence interval can be estimated as:

$$HR \pm z_{\alpha/2}\hat{\sigma}_{HR} \,,$$

where $z_{\alpha/2}$ is the $100\left(1-\frac{\alpha}{2}\right)$:th percentile of the standard normal distribution, 1.645, 1.96 and 2.58 for the 95, 97.5 and 99.5 percentiles respectively.

5.3 Sharpe ratio

The Sharpe ratio will be used as the performance measure for our trading strategies, and is indeed used in reality by investors comparing the performance of different portfolio managers.

As mentioned in 4.1, a good trading strategy yields a high positive return, but also doesn't take too much risk, measured by the standard deviation (volatility). We define the (yearly) Sharpe ratio as:

$$S_p = \frac{E[R_p]}{\sigma_p},$$

where $E[R_p]$ and its standard deviation, σ_p, usually are measured on a yearly basis, and can simply be estimated by the average and the sample variance respectively. In many definitions the Sharpe ratio also contains a term for the risk free interest rate, which is omitted in this case. For 10 minutes/weekly data, if one assumes that the returns are uncorrelated during different intervals, we get the yearly Sharpe ratio:

$$S_p = \frac{N_{\Delta t}E\left[R_p\right]}{\sqrt{N_{\Delta t}}\sigma_p} = \sqrt{N_{\Delta t}}\frac{E\left[R_p\right]}{\sigma_p},$$

where the returns and standard deviation in this case are measured on 10-minutes or weekly intervals, and $N_{\Delta t}$ is the number of 10-minutes/weekly intervals traded during a year. In our case 38 10 minutes intervals during approximately 252 trading days makes $N_{\Delta t} = 9576$ for the 10 minutes data. As there are 52 weeks per year and we are trading every week, $N_{\Delta t} = 52$ for the weekly data.

6. Model specifications

There are a lot of different choices that have to be made when specifying the models. The orders of the VAR models and Markov chains have to be selected, the different hyperparameters have to be chosen etc. There are theoretical ways to specifying the model orders, while the hyperparameters often need to be specified through empirical studies of the results while back testing.

6.1 Training, validation and test sets

In order to not overfit our specifications to one sample of data points we use a limited data set for the specifications, in order to save a sample for evaluating our models when all specifications have been made. Usually one divides the data in three different data sets: training, validation and test sets. The training set often refers to a set where one “trains” a model using a specific model specification. Thereafter one validates the model on the validation set, and chooses the specification that gives the best results. In order to not overfit the model to the validation set one evaluates the model on the previously unused test set. If the model performs well on the test set, one expects the model to perform well even in the future, for presently unobserved data.

As we have chosen a dynamical approach to update our models the training sets will change over time, while new observations are made. However, we divide our data in a sample for training and validation, and a test set for evaluation and comparison of the models.

For the data of 10 minute returns we choose the train/validation set as 25 000 observations between 2010-06-01 and 2012-08-01, and the test set as 13549 observations between 2012-08-01 and 2013-12-30.

For the data of weekly returns we have a lot less observations. We choose the train/validation set as 700 observations between 1995-06-09 and 2008-10-31, and the test set as 280 observations between 2008-11-07 and 2014-03-14.

All data in the train/validation set might not always be used for training and validation, since the model complexity might be too high for validating several different choices of model specifications. The most important part is that we want to separate the test and train/validation sets in order to avoid overfitting and be able to test our models on a previously unused sample of data.

6.2 Order Selection

To select the order of the time series and Markov chain models, we use the Bayesian information criterion (BIC) as well as the Akaike information criterion (AIC):

$$BIC = -2\ln\hat{L} + p\ln n,$$

$$AIC = -2\ln\hat{L} + 2p,$$

where $\hat{L}$ is the estimated maximum likelihood of the model, p is the number of free parameters to be estimated and n is the number of data points. (Wikipedia, AIC, 2014), (Wikipedia, BIC, 2014).

The idea of both measures is to select an optimal model with respect to goodness of fit as well as the complexity of the model. Having a lot of parameters make the model more complex, which might result in overfit to the training data.

The model to select is the one with the smallest AIC/BIC. However, these measures can be in conflict with each other, BIC generally tends to penalize many parameters more heavily than AIC.

General VAR

In the general VAR model, we have that:

$$\boldsymbol{y}_t \sim N(\boldsymbol{z}_t'\boldsymbol{\Theta}, \boldsymbol{\Psi}).$$

The corresponding multivariate normal density function is:

$$\frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Psi})}} \exp\left\{-\frac{1}{2}[\boldsymbol{y}_t - \boldsymbol{z}_t'\boldsymbol{\Theta}]'\boldsymbol{\Psi}^{-1}[\boldsymbol{y}_t - \boldsymbol{z}_t'\boldsymbol{\Theta}]\right\}.$$

Taking the product for all t up to T, evaluating at the maximum likelihood estimates given under the general VAR model and then taking the logarithm, yields the maximum log likelihood:

$$\ln\hat{L} = -\frac{T}{2}\left(m\ln 2\pi + \ln\det(\widehat{\boldsymbol{\Psi}})\right) - \frac{1}{2}\sum_{t=1}^{T}\left[\boldsymbol{y}_t - \boldsymbol{z}_t'\widehat{\boldsymbol{\Theta}}\right]'\widehat{\boldsymbol{\Psi}}^{-1}\left[\boldsymbol{y}_t - \boldsymbol{z}_t'\widehat{\boldsymbol{\Theta}}\right]$$

The number of free parameters is the sum of the elements in $\boldsymbol{\Theta}$ and the free parameters of $\boldsymbol{\Psi}$:

$$p = k * m + \frac{m}{2}(m+1)$$

The AIC and BIC for a general VAR model of order 1-4 is presented in table 6.1 for 10 minutes returns, and in table 6.2 for the weekly returns. The models do not contain any exogenous variables, such as volumes. For the weekly returns both AIC and BIC suggests order 1, but for the 10 minutes returns AIC suggests order 2 while BIC suggests order 1. We will only investigate the case of order 1, since the margin between order 1 and 2 for BIC is considered large.

Table 6.1. AIC and BIC for 10 minutes returns modeled by a general VAR model of orders 1-4.

Order	AIC	BIC
1	399 403	400 122
2	399 362	400 500
3	399 372	400 930
4	399 386	401 363

Table 6.2. AIC and BIC for weekly returns modeled by a general VAR model of orders 1-4.

Order	AIC	BIC
1	56 217	59 927
2	56 604	62 679
3	56 960	65 401
4	57 368	68 174

Bayesian VAR

The AIC and BIC tests for the general VAR gives us a hint of what orders to use also in the Bayesian case. However, as we are able to shrink parameters corresponding to lags of higher orders more than those of low orders; we at least want to try one model of a bit higher order than one. We also want to investigate the results when the traded volumes are taken into account. The Bayesian VAR models investigated, both for weekly and 10 minute returns, are:

1. Bayesian VAR of order 1 without exogenous variables, called VAR1.
2. Bayesian VAR of order 1 with traded volumes as exogenous variables, called VARX1.
3. Bayesian VAR of order 4 without exogenous variables, called VAR4.

Markov Models

For a first order Markov chain the likelihood function is derived as:

$$P(Y_T = y_T, \dots, Y_1 = y_1) = P(Y_1 = y_1)P(Y_2 = y_2|Y_1 = y_1) \cdots P(Y_T = y_T|Y_{T-1} = y_{T-1})$$
$$= P(Y_1 = y_1)\prod_{t=2}^{T} p_{y_{t-1},y_t} = P(Y_1 = y_1)\prod_{j=1}^{N}\prod_{i=1}^{N} p_{j,i}^{n_{j,i}}.$$

This gives us the maximum log likelihood:

$$\ln\hat{L} = \ln\hat{P}(Y_1 = y_1) + \sum_{j=1}^{N}\sum_{i=1}^{N} n_{j,i}\ln\hat{p}_{j,i},$$

where $\hat{p}_{j,i}$ are the maximum likelihood estimates of the transition probabilities. The term $\ln\hat{P}(Y_1 = y_1)$ is often left out, since it’s small compared to the other term when we have a lot of observations.

For a Markov chain of order M, the log likelihood is:

$$\ln\hat{L} = \ln\hat{P}(Y_1 = y_1) +, \dots, + \ln\hat{P}(Y_M = y_1) + \sum_{j_1=1}^{N}, \dots, \sum_{j_M=1}^{N}\sum_{i=1}^{N} n_{j_1,\dots,j_M,i}\ln\hat{p}_{j_1,\dots,j_M,i},$$

where the terms $\ln \hat{P}(Y_1 = y_1) +, \dots, + \ln \hat{P}(Y_M = y_1)$ can be left out if we have a lot of observations.

The number of free parameters is:

$$p = N^M(N-1),$$

since we have N^M rows in the transition matrix, and N possible transitions with the constraint $\sum_{i=1}^{N} p_{j_1,\dots,j_M,i} = 1$.

Before choosing the order of the Markov chains, we need to specify how many states we should have. This can also be done by AIC/BIC, but since we want to get as significant probabilities for positive/negative returns as possible, we shouldn't have less than four states (two for positive/negative returns respectively). The AIC/BIC measures for Markov chains with four and six states, for all currencies, of order 0-3 are presented in table 6.3-9.

We notice that the BIC always suggests the model of order 1, while AIC suggests the model of order 2 in all cases but for the Swiss franc. Since the difference in BIC between order 1 and 2 is more significant than for AIC, and because of the convenience of having the same order for all currencies, we choose to model all currencies with the first order Markov chain.

If we do the same procedure for weekly returns, BIC suggests order 0 for all currencies and AIC suggests order 1 only for a few. This indicates that the Markov models are not appropriate for the weekly returns. We will therefore only consider 10 minute returns in the Markov models.

Conclusively we choose the following Markov models, both in the general and in the Bayesian case:

1. 1:st order Markov model of 4 states.
2. 1:st order Markov model of 6 states.

Table 6.3. Markov model for AUD. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	**Order 1**	**Order 2**	**Order 3**
AIC	**4 states**	64 710	64 541	64 491	64 578
	6 states	83 640	83 422	83 409	84 203
BIC	**4 states**	64 734	64 638	64 878	66 125
	6 states	83 680	83 664	84 860	92 906

Table 6.4. Markov model for CAD. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	**Order 1**	**Order 2**	**Order 3**
AIC	**4 states**	64 042	63 863	63 782	63 835
	6 states	82 775	82 477	82 444	83 238
BIC	**4 states**	64 066	63 959	64 168	65 380
	6 states	82 815	82 718	83 893	91 929

Table 2.5. Markov model for CHF. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	Order 1	Order 2	Order 3
AIC	**4 states**	64 363	64 072	64 020	64 089
	6 states	83 191	82 743	82 750	83 512
BIC	**4 states**	64 387	64 168	64 406	65 635
	6 states	83 231	82 985	84 200	92 208

Table 6.6. Markov model for EUR. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	Order 1	Order 2	Order 3
AIC	**4 states**	65 333	65 089	64 969	65 031
	6 states	84 444	84 121	84 055	84 812
BIC	**4 states**	65 357	65 186	65 356	66 580
	6 states	84 484	84 363	85 507	93 525

Table 6.7. Markov model for GBP. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	Order 1	Order 2	Order 3
AIC	**4 states**	65 156	64 898	64 830	64 867
	6 states	84 215	83 868	83 853	84 697
BIC	**4 states**	65 180	64 994	65 218	66 415
	6 states	84 256	84 110	85 305	93 407

Table 6.8. Markov model for JPY. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	Order 1	Order 2	Order 3
AIC	**4 states**	63 227	62 818	62 688	62 715
	6 states	81 721	81 017	80 939	81 578
BIC	**4 states**	63 251	62 914	63 074	64 258
	6 states	81 762	81 258	82 385	90 256

Table 6.9. Markov model for NZD. AIC and BIC for 4 and 6 states, of orders 0-3.

		Order 0	Order 1	Order 2	Order 3
AIC	**4 states**	64 152	63 931	63 888	63 956
	6 states	82 917	82 559	82 600	83 480
BIC	**4 states**	64 176	64 028	64 275	65 502
	6 states	82 958	82 800	84 049	92 173

6.3 Specifying hyperparameters

In the Bayesian models, there are a lot of different hyperparameters that need to be specified. As mentioned earlier, some are chosen by prior beliefs about the data, while some have to be chosen in some more empirical way.

In order to specify the hyperparameters we validate the model on samples of data within the training/validation data sets. This can take a lot of time depending on the computational complexity of the model. In some cases we choose to only update the model with a specified iteration frequency, in order to reduce computation time. This might not give us as good prediction results as when updating at every iteration, but should be sufficient to indicate preferred values on the hyperparameters. Later, when testing our specified models, we will update the models at every iteration.

Another approach to save time is to run the algorithms in parallel, using the computational power of several machines. This is done for the Bayesian VAR models, where we use a parallel loop to evaluate the models for different combinations of hyperparameters.

General VAR

In the general VAR model the parameters are estimated by least squares, and do therefore not require any hyperparameters. However, we need to specify how large training sample to use. To save time, we choose to update our model every 10:th iteration for the weekly data, and every 50:th iteration for the 10 minute data.

For the weekly data, we validate the model for different values of β on the last 200 observations in the train/validation set, and for the 10 minutes data we validate on the last 1000 observations of the train/validation set. The average MSE for the different return series against β is plotted in figure 6.1-2 for the weekly and 10 minutes returns respectively. In the figures, we notice that the MSE is decreasing with β. We can draw the conclusion that one should use as large training sample as possible in the general VAR model, i.e. we choose to train on all previous observations.

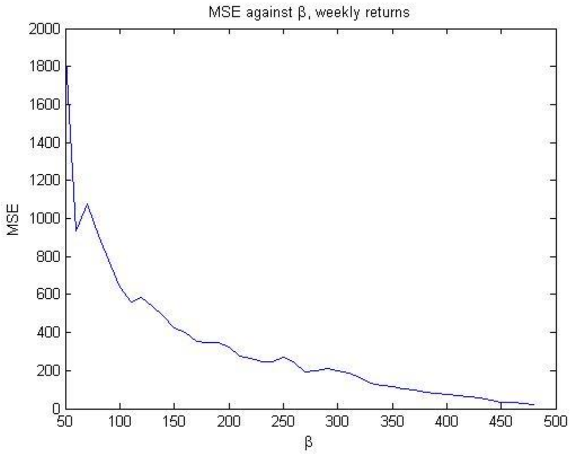


Figure 6.1 Average MSE against the number of training samples used, β, when validating on 200 observations of weekly returns.

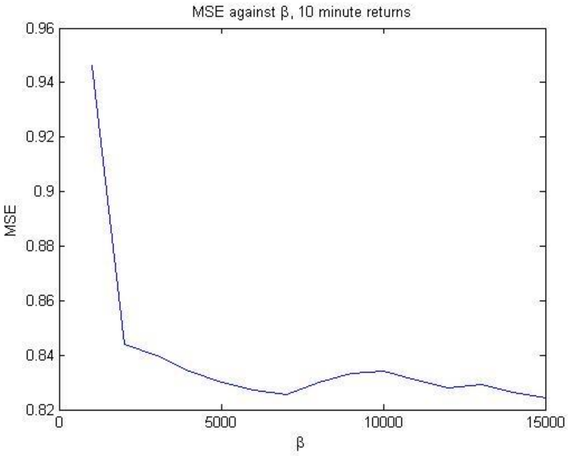


Figure 6.2 Average MSE against the number of training samples used, β, when validating on 1000 observations of 10 minute returns.

Bayesian VAR
In the Bayesian VAR models we need to specify the hyperparameters π_1, π_2, π_3 and π_4 mentioned under the description of the Bayesian VAR model. We also need to specify the number of previous observations to train our models on, β.

The chosen approach is to randomly pick 1000 combinations of the hyperparameters at reasonably specified intervals, and validate the models for 4 randomly chosen currencies (in each iteration) on the 500 and 200 latest observations in the train/validation set for the 10 minutes and weekly observations of returns respectively. Then we use Nadaraya-Watson kernel regression to fit a curve for the average MSE of all currencies against the different hyperparameters. With this approach we can plot the average MSE for the currencies involved against any individual hyperparameter, and also under smaller intervals of other hyperparameters.

Nadaraya-Watson estimates the conditional mean of a random variable Y, given covariates X:

$$E[Y|X=x]=\int yf(y|x)dy=\int y\frac{f(x,y)}{f(x)}dy,$$

where $f(x)$ is the marginal density function of X, and $f(x,y)$ is the joint density function of X and Y. (Powell, 2008). In our case Y is the MSE and X is the hyperparameters. The density functions are estimated by so called kernel density estimation, where we use a Gaussian kernel and a suggested optimal bandwidth (Bowman & Azzalini, 1997, p. 31).

With the plots of MSE's against hyperparameters at hand we want to choose the hyperparameters so that the MSE is minimized. However, as will be noted, there is a lot of variance in our data, which often implies that there is no perfect minimum, and that the choices might be quite arbitrary.

When validating the models, we use a cluster of machines to parallelize the MSE calculations for different combinations of parameters. With the configuration used it still takes several hours to perform the validation for 1000 combinations. We are updating the models in each iteration, since this step takes far less time than the Monte Carlo simulation performed in the prediction step.

10 minute returns
For the 10 minute data, we will select the hyperparameters in the following order:

1. Select π_1
2. Select π_2, π_3 and π_4 under a shortened interval of π_1
3. Select β from a new validation under the other specified hyperparameters.

It is quite intuitive that we first select π_1, which accounts for the "overall tightness", and then select the parameters controlling the tightness of independent variables, different lags and exogenous variables (e.g. traded volumes) respectively, which indeed should be dependent of π_1, but approximately independent of each other. Lastly we select β, which should be dependent on how much shrinkage that is applied to the parameters. With a large β the model mostly rely on patterns within the whole sample to not overfit the model, while we with a high regularization and a smaller β might be able to catch time dependent patterns and dependencies without overfitting the model.

In figure 6.3, 6.5, and 6.8 the average MSE's are plotted and regressed against π_1 for the different Bayesian VAR models. We will choose π_1 as the minimizing values, however the decisions might be very vague due to a lot of variance in most cases.

In figure 6.4, 6.6, 6.7, 6.9 and 6.10 the average MSE's are plotted against π_2, π_3 and π_4 for the different models under smaller intervals for π_1, around its chosen value. We again choose the minimizing value for π_2, π_3 and π_4 in all cases.

When the values of π_1, π_2, π_3 and π_4 has been chosen, we run a new validation in order to choose the number of previous samples to train on, β. The result of the average MSE's against β for the different models are plotted in figures 6.11-13.

Finally, all chosen values of hyperparameters are presented in table 6.1

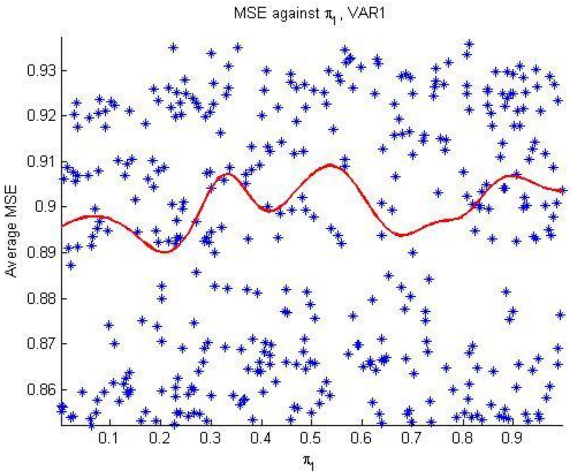


Figure 6.3. MSE against π_1 in the VAR1 model, for 1000 randomly chosen combinations of π_1, π_2 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

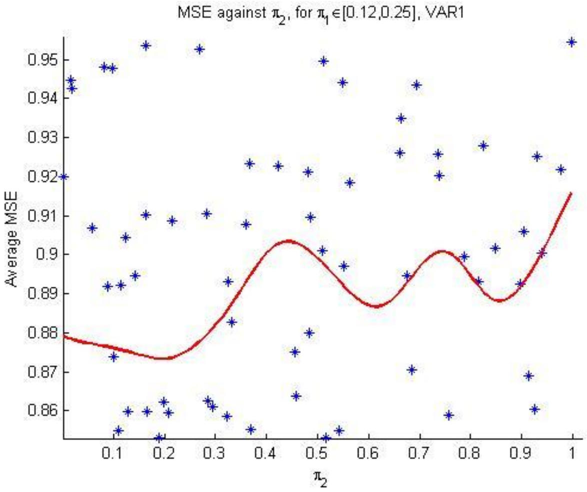


Figure 6.4. MSE against π_2 when $\pi_1 \in [0.12, 0.25]$, in the VAR1 model, for 1000 randomly chosen combinations of π_1, π_2 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

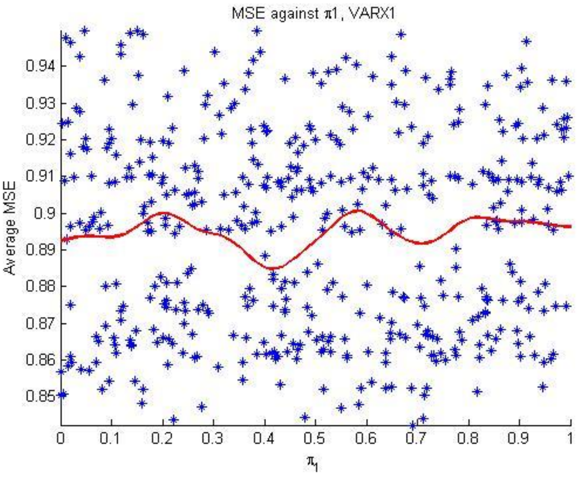


Figure 6.5. MSE against π_1 in the VARX1 model, for 1000 randomly chosen combinations of π_1, π_2, π_4 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

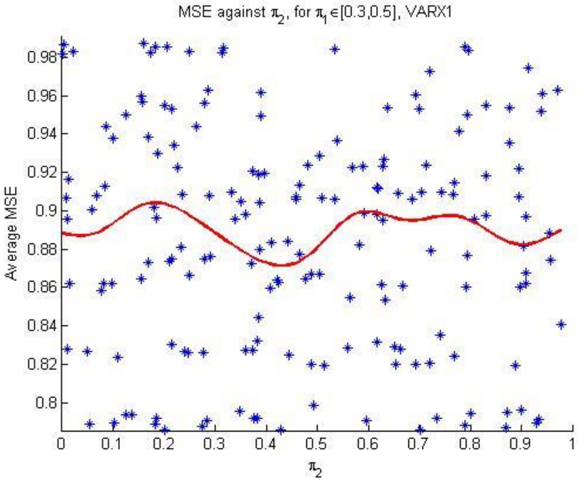


Figure 6.6. MSE against π_2 when $\pi_1 \in [0.3, 0.5]$, in the VARX1 model, for 1000 randomly chosen combinations of π_1, π_2, π_4 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

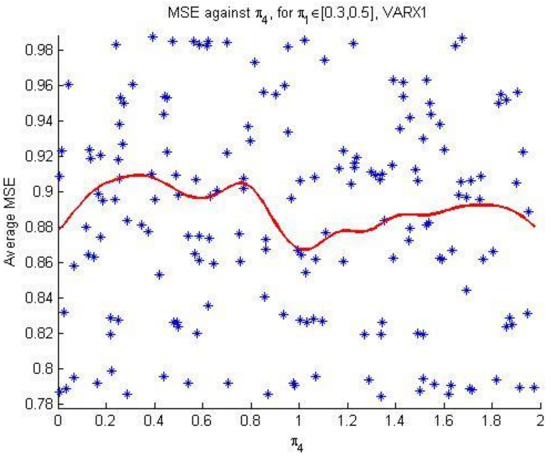


Figure 6.7. MSE against π_4 when $\pi_1 \in [0.3, 0.5]$, in the VARX1 model, for 1000 randomly chosen combinations of π_1, π_2, π_4 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

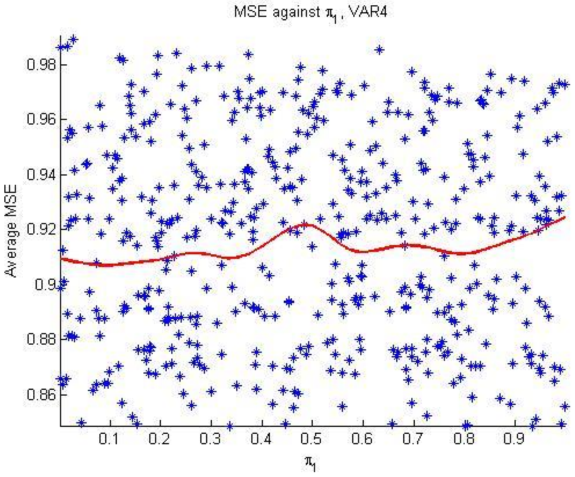


Figure 6.8. MSE against π_1 in the VAR4 model, for 1000 randomly chosen combinations of π_1, π_2, π_3 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

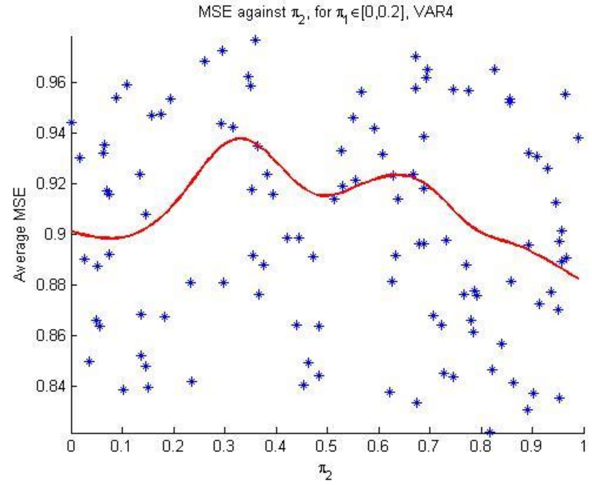


Figure 6.9. MSE against π_2 when $\pi_1 \in [0, 0.2]$, in the VAR4 model, for 1000 randomly chosen combinations of π_1, π_2, π_3 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

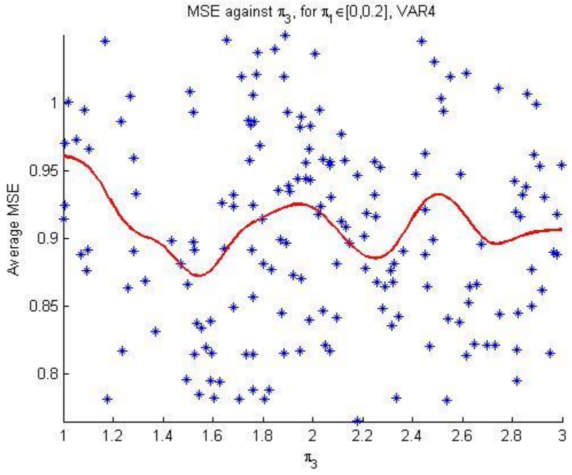


Figure 6.10. MSE against π_3 when $\pi_1 \in [0, 0.2]$, in the VAR4 model, for 1000 randomly chosen combinations of π_1, π_2, π_3 and β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

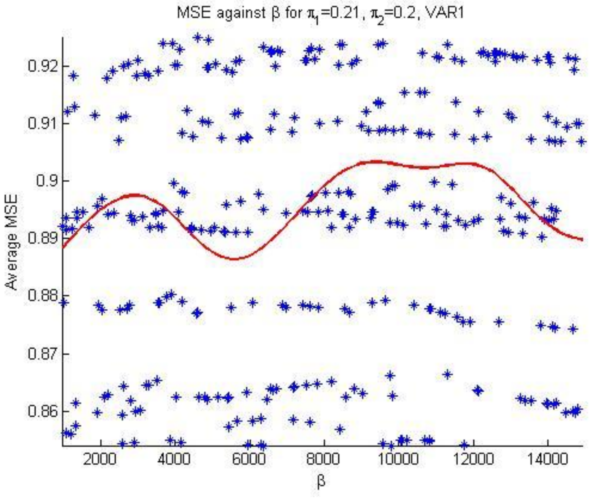


Figure 6.11. MSE against β when $\pi_1 = 0.21$ and $\pi_2 = 0.20$, in the VAR1 model, for 1000 randomly chosen values of β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

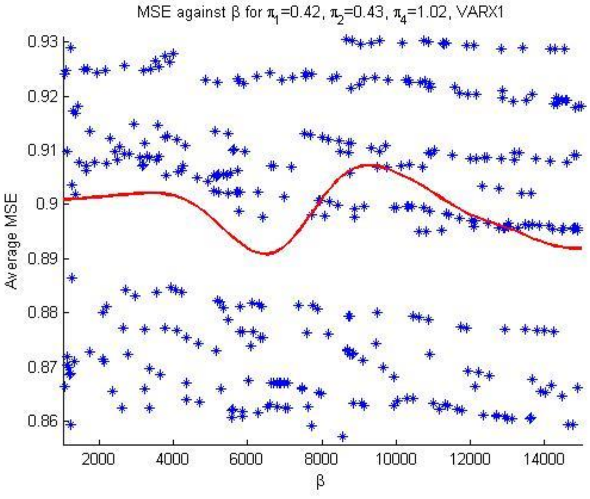


Figure 6.12. MSE against β when $\pi_1 = 0.42$, $\pi_2 = 0.43$ and $\pi_4 = 0.20$, in the VARX1 model, for 1000 randomly chosen values of β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

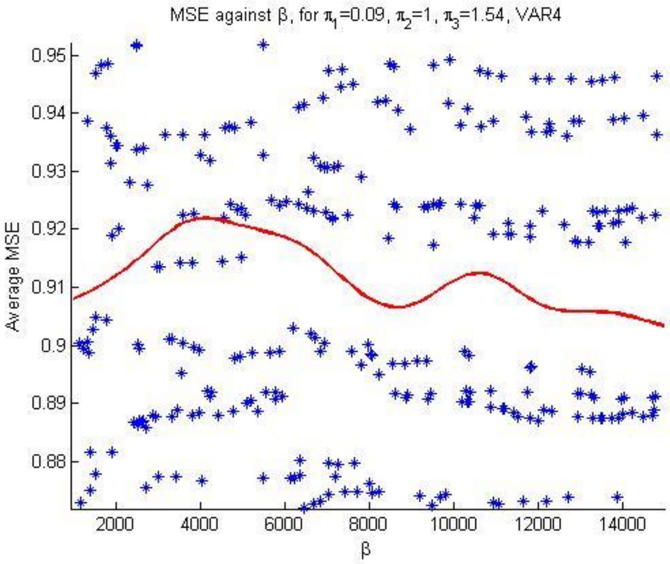


Figure 6.13. MSE against β when $\pi_1 = 0.09$, $\pi_2 = 1$ and $\pi_3 = 1.54$, in the VAR4 model, for 1000 randomly chosen values of β. The model is validated on a sample of 500 returns on 10 minute intervals. The regression line (red) is computed by Nadaraya-Watson kernel regression.

Table 6.1 Chosen values for the hyperparameters in the Bayesian VAR models for 10 minute returns.

	VAR1	VARX1	VAR4
π_1	0.21	0.42	0.09
π_2	0.20	0.43	1
π_3	–	–	1.54
π_4	–	1.02	–
β	5600	6500	15000

Weekly returns

Since the sample of weekly data is much smaller than for the 10 minute data, we expect that one should train on all observations at hand. We therefore select β in the first stage for the weekly data, together with π_1. Conclusively we choose the hyperparameters in the following order:

1. Select β and π_1
2. Select π_2 and π_3 under a new validation under specified values of β and π_1.

In figures 6.14 and 6.16 the average MSE's are plotted against π_1 for the VAR1 and VAR4 model respectively. By a first inspection it seems like the regressed curves are increasing for all values of π_1, but if we zoom in for very small values (see figures 6.15 and 6.17), we can find min values for both models, even if the significance of the regression may be very low.

In figures 6.18 and 6.19 the average MSE's are plotted against β. The plots confirm our hypothesis that we should train on all previous observations, as the curves are decreasing for all values of β.

In figures 6.20-22 the average MSE's against π_2 and π_3 in the two different models are plotted, under a new validation where π_1 and β are specified. In figure 6.21, the plot is increasing for all values of π_2, which implies that we choose a very low value; $\pi_2 = 0.01$.

In table 6.2 all chosen values of the hyperparameters are presented.

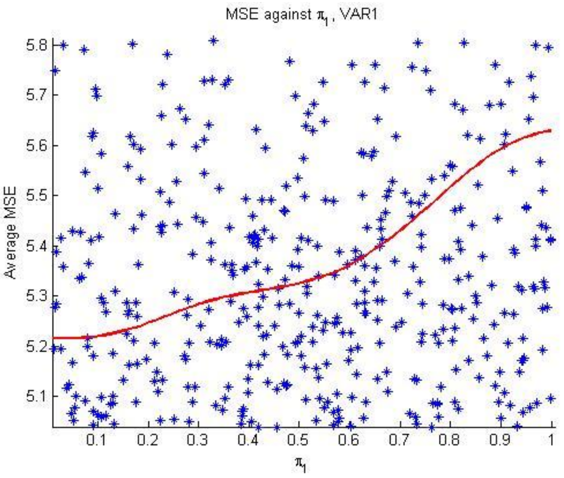


Figure 6.14. MSE against π_1 in the VAR1 model, for 1000 randomly chosen combinations of π_1, π_2 and β. The model is validated on a sample of 200 weekly returns.. The regression line (red) is computed by Nadaraya-Watson kernel regression.

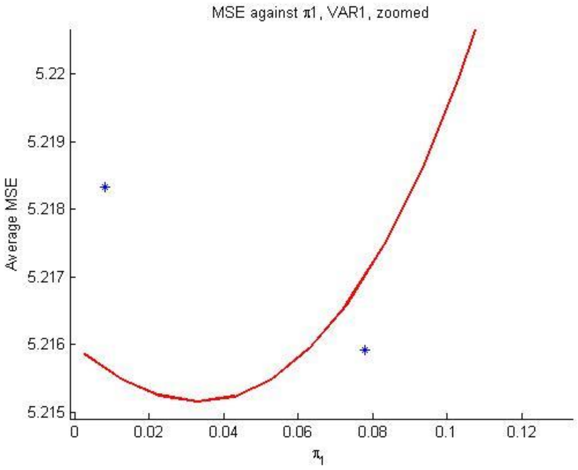


Figure 6.15. A Zoom of figure 6.14, for small values of π_1.

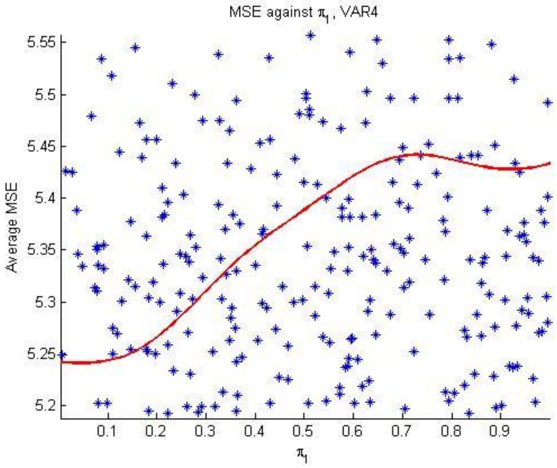


Figure 6.16. MSE against π_1 in the VAR4 model, for 1000 randomly chosen combinations of π_1, π_2, π_3 and β. The model is validated on a sample of 200 weekly returns.. The regression line (red) is computed by Nadaraya-Watson kernel regression.

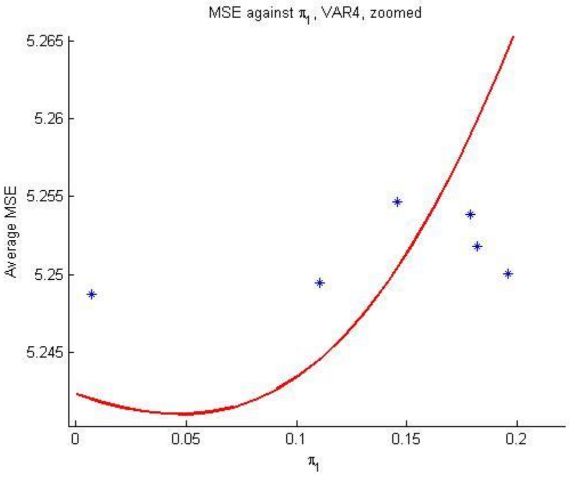


Figure 6.17. A Zoom of figure 6.16, for small values of π_1.

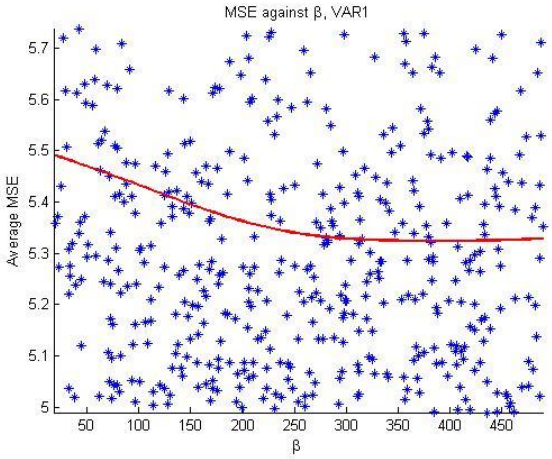


Figure 6.18. MSE against β in the VAR1 model, for 1000 randomly chosen combinations of π_1, π_2 and β. The model is validated on a sample of 200 weekly returns.. The regression line (red) is computed by Nadaraya-Watson kernel regression.

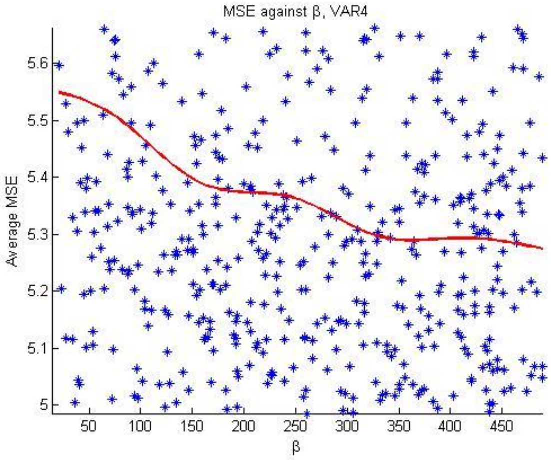


Figure 6.19. MSE against β in the VAR4 model, for 1000 randomly chosen combinations of π_1, π_2, π_3 and β. The model is validated on a sample of 200 weekly returns.. The regression line (red) is computed by Nadaraya-Watson kernel regression.

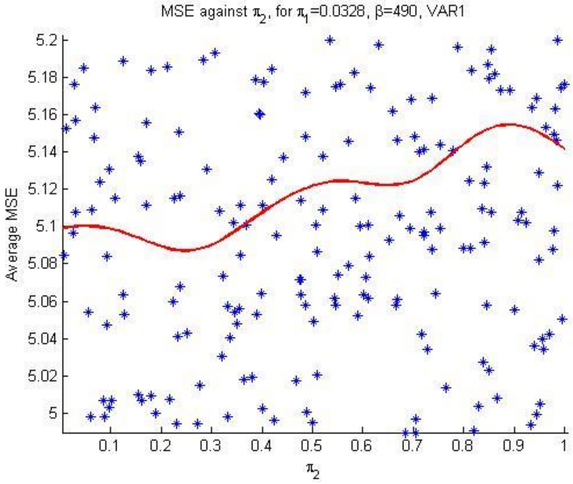


Figure 6.20. MSE against π_2 when $\pi_1 = 0.0328$ and $\beta = 490$, in the VAR1 model, for 1000 randomly chosen values of π_2. The model is validated on a sample of 200 weekly returns. The regression line (red) is computed by Nadaraya-Watson kernel regression.

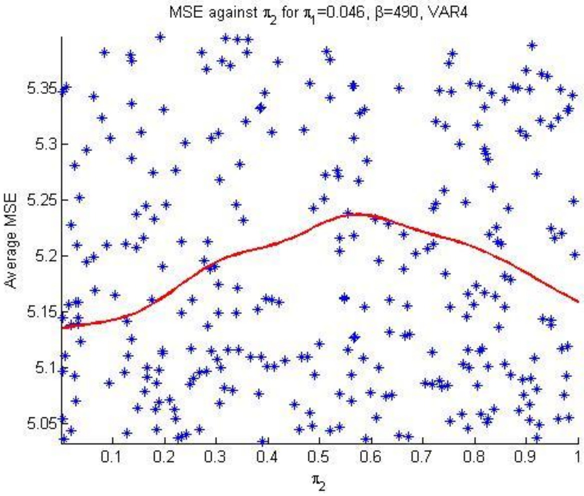


Figure 6.21. MSE against π_2 when $\pi_1 = 0.046$ and $\beta = 490$, in the VAR4 model, for 1000 randomly chosen values of π_2 and π_3. The model is validated on a sample of 200 weekly returns. The regression line (red) is computed by Nadaraya-Watson kernel regression.

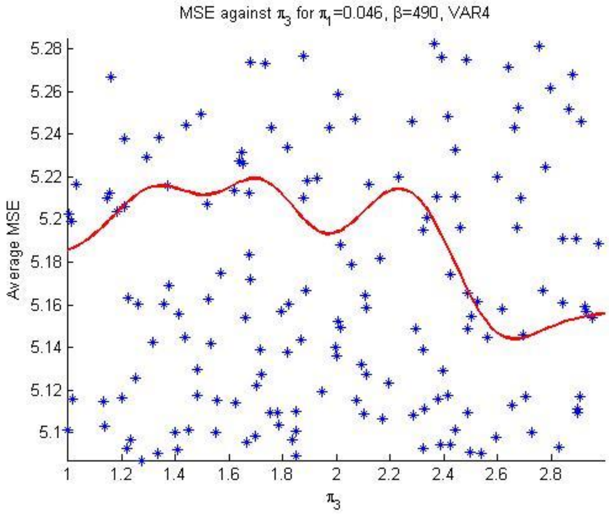


Figure 6.22. MSE against π_3 when $\pi_1 = 0.046$ and $\beta = 490$, in the VAR4 model, for 1000 randomly chosen values of π_2 and π_3. The model is validated on a sample of 200 weekly returns. The regression line (red) is computed by Nadaraya-Watson kernel regression.

Table 6.2. Chosen values for the hyperparameters in the Bayesian VAR models for weekly returns.

	VAR1	VAR4
π_1	0.0328	0.046
π_2	0.25	0.01
π_3	–	2.66
β	All previous samples	All previous samples

General Markov model

In the general Markov Model we also need to specify the number of previous samples to train on, β. We are validating the model on the 5000 last observations of nonzero 10 minute returns in the train/validation set. We are updating the model every 50:th iteration, and evaluate the hit rate when we classify the returns as positive/negative with respect to a probability above 0.5 for the two events respectively. The hit rate over all currencies against β is plotted in figure 3.4-5, for the 4 states and 6 states models respectively. We notice that the hit rate seem highest for large values of β, and we therefore choose $\beta = 15000$.

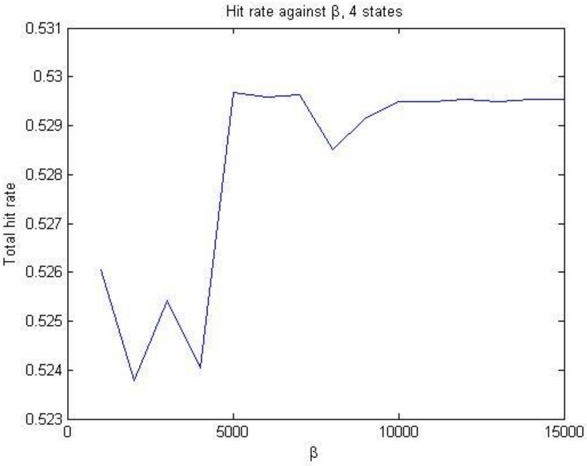


Figure 6.23. Hit rate against β, for a 4 states Markov model, on 5000 10 minute returns.

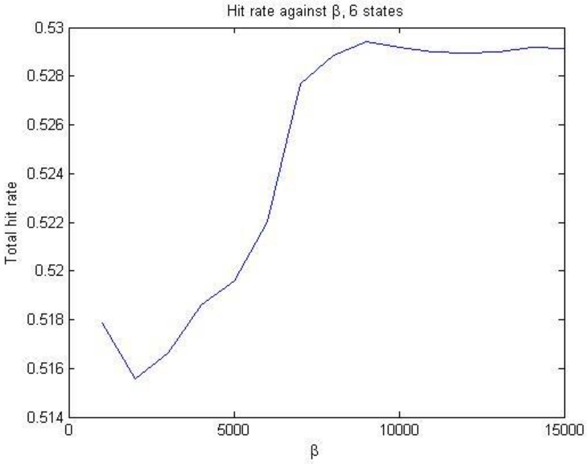


Figure 6.24. Hit rate against β, for a 6 states Markov model, on 5000 10 minute returns.

Bayesian Markov model

When specifying the regularizing hyperparameter in the Bayesian Markov model, α, it doesn't make sense to consider the hit rate under a probability threshold of 0.5, as above. If the threshold is exceeded with no regularization, i.e. $\alpha = 0$, it will still be exceeded no matter how much regularization that is applied, since the probabilities of positive/negative returns will be shrinked towards precisely 0.5. The Bayesian Markov model will therefore not be used for testing our trading strategy. However, in order to compare its predictive power, we will consider the model with a fixed threshold of 0.51, and with $\beta = 15000$ the same as in the general Markov model. In figure 6.25-26 the hit rate is plotted against α for the 4 and 6 state models respectively, when the model is validated on the 5000 last nonzero observations in the train/validation set for the 10 minutes data, and updated every 50:th iteration.

As seen in the figures, we manage to obtain a hit rate above 0.55. The higher we choose α, the less predictions will be made, so for further analysis we choose the α's as the smallest value where the hit rate seem stabilized at a high level. For the 4 states model we choose $\alpha = 1600$, resulting in 12867 predictions. For the 6 states model we choose $\alpha = 800$, which results in 8298 predictions. As we are observing 7 different currency returns at 5000 time stamps, there are in total $7 * 5000 = 35000$ occasions where predictions can be made.

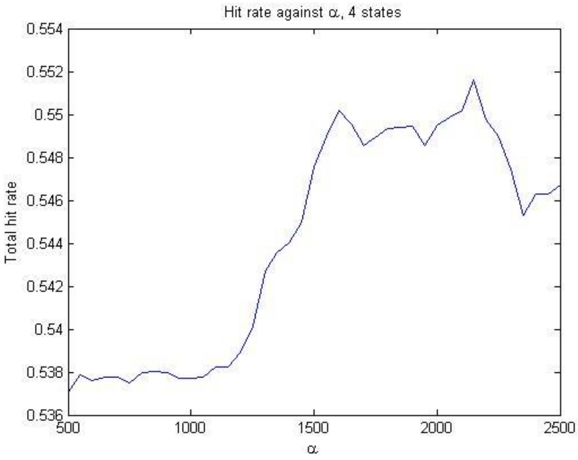


Figure 6.25. Hit rate against α for a 4 states Bayesian Markov model, on 5000 10 minute returns.

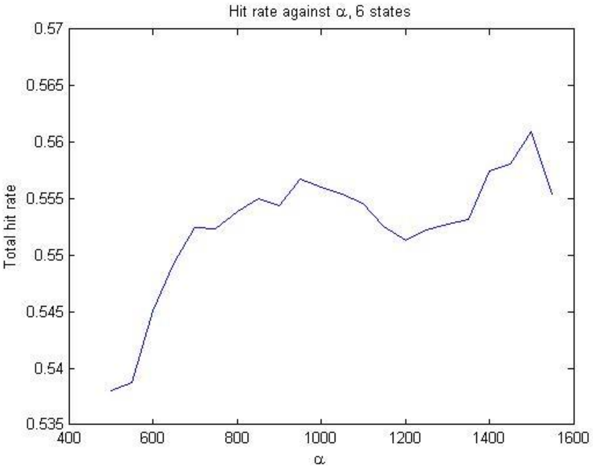


Figure 6.25. Hit rate against α for a 6 states Bayesian Markov model, on 5000 10 minute returns.

7. Results

The results will be presented as the performances on the test sets defined in 6.1. For the 10 minute data we validate the models on the 10 000 latest observations in the test set, and for the weekly data on the whole test set i.e. 280 observations. In section 7.1 we will investigate the predictive power of the models using the RMSE and the hit rate as performance measures. In section 7.2 we will investigate the performance of our trading strategies using the Sharpe ratio as performance measure.

7.1 Results of predictive power

When considering the hit rates of classified positive/negative returns, it is crucial to investigate the percentage of positive/negative returns in the whole data samples. These numbers are presented for all currencies respectively, for the 10 minute data in table 7.1 and for the weekly data in table 7.2. We are not counting the zero returns, since these are not taken into account when computing the hit rates.

As seen in table 7.1, the percentage of positive/negative returns in the 10 minute data seem very concentrated around 50%, and we therefore believe that significant results above 50% in this case do indicate predictive power in our models.

In table 7.2 for the weekly data, the percentages differs quite a lot from 50% for some currencies. For example 55% of the returns of the Swedish krona are positive. One should keep this in mind when judging the resulting hit rates.

Table 7.1 Percentages of positive/negative returns, for all currencies on the whole data set of nonzero 10 minute returns (38 000 observations).

Currency	**Percentage of positive returns**	**Percentage of negative returns**
AUD	50.32%	49.68%
CAD	49.74%	50.26%
CHF	49.99%	50.01%
EUR	50.36%	49.64%
GBP	50.1%	49.9%
JPY	49.68%	50.32%
NZD	50.31%	49.69%
*Tota*l	50.07%	49.93%

Table 7.2. Percentages of positive/negative returns, for all currencies on the whole data set of weekly nonzero returns (980 observations).

Currency	**Percentage of positive returns**	**Percentage of negative returns**
EUR	50.1%	49.9%
JPY	52.76%	47.24%
GBP	52.35%	47.65%
AUD	53.67%	46.33%
CHF	50.51%	49.49%
CAD	48.06%	51.94%
NZD	47.76%	52.24%
SEK	55%	45%
ZAR	52.14%	47.86%
INR	50.1%	49.9%
SGD	48.98%	51.02%
THB	49.90%	50.1%
NOK	48.67%	51.33%
MXN	49.29%	50.71%
DKK	49.90%	50.1%
PLN	49.39%	50.61%
IDR	53.78%	46.22%
CZK	46.33%	53.67%
KRW	47.04%	52.96%
CLP	50.61%	49.39%
COP	51.43%	48.57%
MAD	48.42%	51.58%
Total	50.28%	49.72%

General VAR

In table 7.3 the RMSE for the general VAR1 model divided by the RMSE for a random walk model (when all returns are predicted zero) and the hit rates (where positive predicted values above zero corresponds to predictions of positive returns and vice versa), for all currencies, are presented for the 10 minutes data. Notice that all currencies are predicted worse using the VAR model compared to the random walk, with respect to the RMSE. However, some currencies seem to have hit rates significantly greater than 0.5 at the 1% level.

In table 7.4 the RMSE for the general VAR1 model divided by the RMSE for a random walk model and the hit rates, for all currencies, are presented for the weekly data. Note that all currencies are predicted worse compared to the random walk, some with more than the double RMSE. However, the total hit rate is significantly above 0.5 at the 10% level.

Table 7.3. RMSE for the VAR1 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	$RMSE/RMSE_0$	Hit rate
AUD	1.00312	0.5021
CAD	1.00151	0.5213^{***}
CHF	1.00330	0.5077
EUR	1.01337	0.4793
GBP	1.00169	0.5204^{***}
JPY	1.00325	0.5050
NZD	1.00614	0.5212^{***}
Total	1.00462	0.5080^{***}

Table 7.4. RMSE for the VAR1 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 280 observations of weekly returns. *, ** and * denotes a hit rate significantly larger than 0.5 at the 1%, 5% and 10% level respectively.**

Currency	$RMSE/RMSE_0$	Hit rate
EUR	1.33504	0.5143
JPY	1.07044	0.4607
GBP	1.25937	0.4893
AUD	1.13782	0.5500^{*}
CHF	1.01591	0.5214
CAD	1.07874	0.5393
NZD	1.80891	0.5179
SEK	1.02509	0.4821
ZAR	1.07140	0.5036
INR	1.22946	0.4750
SGD	1.04086	0.4821
THB	1.02336	0.5679^{**}
NOK	2.11257	0.4679
MXN	1.23580	0.4893
DKK	1.89838	0.5214
PLN	1.33631	0.4929
IDR	1.01772	0.5857^{***}

CZK	1.07113	0.5286
KRW	1.01392	0.5536^{*}
CLP	1.01377	0.5036
COP	1.00159	0.5036
MAD	2.41829	0.5071
Total	1.31225	0.5117^{*}

Bayesian VAR

10 minute returns

In tables 7.5-7 the RMSE for the Bayesian VAR models divided by the RMSE for a random walk model and the hit rates, for all currencies, are presented for the 10 minutes data. Notice that all currencies get a higher RMSE with our model compared to the random walk model. However, the total hit rate for the VAR1 model is significantly greater than 0.5 at the 1% level, which indicates that the model has a little predictive power.

Table 7.5. RMSE for the VAR1 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	$\boldsymbol{RMSE/RMSE_0}$	**Hit rate**
AUD	1.00254	0.5002
CAD	1.00128	0.5219^{***}
CHF	1.00010	0.5272^{***}
EUR	1.00470	0.4805
GBP	1.00009	0.5230^{***}
JPY	1.00286	0.5046
NZD	1.00428	0.5257^{***}
Total	1.00234	0.5118^{***}

Table 7.6. RMSE for the VARX1 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	$RMSE/RMSE_0$	**Hit rate**
AUD	1.00557	0.5154^{***}
CAD	1.00178	0.5078
CHF	1.00793	0.5011
EUR	1.01199	0.5042
GBP	1.00670	0.4905
JPY	1.00532	0.4940
NZD	1.00537	0.4975
Total	1.00628	0.5014

Table 7.7. RMSE for the VAR4 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 10 000 observations of 10 minute returns. ** and * denotes a hit rate significantly larger than 0.5 at the 5% and 10% level respectively.

Currency	$RMSE/RMSE_0$	**Hit rate**
AUD	1.01068	0.5093^{*}
CAD	1.00266	0.5124^{**}
CHF	1.00015	0.5108^{**}
EUR	1.00474	0.4876
GBP	1.00359	0.4947
JPY	1.00308	0.5049
NZD	1.01821	0.4872
Total	1.00702	0.5009

Weekly returns

In tables 7.8-9 the RMSE for the Bayesian VAR models divided by the RMSE for a random walk model and the hit rates, for all currencies, are presented for the weekly data. The VAR1 model performs slightly better than the random walk for most currencies, and gets a bit smaller total RMSE (only 0.04%). Best predicted is the Thai Baht with more than 1% lower RMSE than the random walk, and a hit rate of 0.5964. The VAR4 model does not perform well at all, with a total hit rate below 0.5.

Table 7.8. RMSE for the VAR1 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 280 observations of weekly returns. * and ** denotes a hit rate significantly larger than 0.5 at the 1% and 5% level respectively.**

Currency	$RMSE/RMSE_0$	**Hit rate**
EUR	1.00378	0.4821
JPY	0.99860	0.5250
GBP	0.99944	0.4893
AUD	1.00057	0.4571
CHF	1.00231	0.4964
CAD	0.99915	0.5071
NZD	0.99737	0.5000
SEK	1.00079	0.4857
ZAR	1.00070	0.5214
INR	0.99233	0.5214
SGD	1.00020	0.4821
THB	0.98927	0.5964^{***}
NOK	0.99833	0.4892
MXN	1.00021	0.5214
DKK	1.00446	0.5000
PLN	1.00012	0.4785
IDR	0.99365	0.5250
CZK	1.00262	0.4893
KRW	0.99642	0.5643^{**}
CLP	0.99937	0.4928
COP	0.99729	0.4786
MAD	1.00402	0.5036
Total	0.99960	0.5049

Table 7.9. RMSE for the VAR4 model divided by the RMSE for a random walk model, and the hit rate when validating the model on 280 observations of weekly returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	$RMSE/RMSE_0$	**Hit rate**
EUR	1.00547	0.5036
JPY	1.00144	0.4464
GBP	1.00017	0.5179
AUD	0.99969	0.5250
CHF	1.00749	0.4607
CAD	1.00112	0.5071
NZD	1.00894	0.4750
SEK	1.00012	0.4964
ZAR	0.99902	0.5000
INR	1.00155	0.4500
SGD	1.00008	0.5107
THB	1.00843	0.4750
NOK	0.99738	0.5036
MXN	0.99921	0.5107
DKK	0.99867	0.5071
PLN	0.99927	0.5321
IDR	1.00360	0.5357
CZK	0.99905	0.6000^{***}
KRW	1.00356	0.4500
CLP	1.00132	0.4714
COP	1.01015	0.4500
MAD	0.99981	0.4821
Total	1.00181	0.4959

General Markov

In tables 7.10-11 the hit rates (with a threshold of 0.5 on the transition probabilities) are presented for the 10 minute returns using a Markov model of 4 and 6 states respectively. All hit rates are significantly greater than 0.5 at the 1% level for both models, which strongly indicates that the models have predictive power.

Table 7.10. Hit rates for the 4 states Markov model when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	**Hit rate**
AUD	0.5282^{***}
CAD	0.5279^{***}
CHF	0.5301^{***}
EUR	0.5446^{***}
GBP	0.5226^{***}
JPY	0.5291^{***}
NZD	0.5265^{***}
Total	0.5299^{***}

Table 7.11. Hit rates for the 6 states Markov model when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	**Hit rate**
AUD	0.5278***
CAD	0.5244***
CHF	0.5274***
EUR	0.5446***
GBP	0.5207***
JPY	0.5291***
NZD	0.5254***
Total	0.5285***

Bayesian Markov

In tables 7.12-13 the hit rates are presented for the 10 minute returns using the Bayesian Markov model of 4 and 6 states respectively, with a threshold of 0.51 on the transition probabilities. Note that for the 6 states model all hit rates are greater than for the general Markov models, which indicates that we can increase the predictive power by shrinking the transition probabilities. This can be useful when one only wants to trade at certain times, e.g. for reducing the transaction costs, but these strategies will not be further discussed in this thesis.

Table 7.12. Hit rates for the 4 states Bayesian Markov model when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	**Hit rate**
AUD	0.5371***
CAD	0.5359***
CHF	0.5439***
EUR	0.5528***
GBP	0.5336***
JPY	0.5487***
NZD	0.5246***
Total	0.5404***

Table 7.13. Hit rates for the 6 states Bayesian Markov model when validating the model on 10 000 observations of 10 minute returns. * denotes a hit rate significantly larger than 0.5 at the 1% level.**

Currency	**Hit rate**
AUD	0.5427***
CAD	0.5452***
CHF	0.5535***
EUR	0.5617***
GBP	0.5360***
JPY	0.5443***
NZD	0.5359***
Total	0.5453***

7.2 Results of trading performance

To investigate whether it is possible to make any profit using our models, we validate our performance with respect to the Sharpe ratios using the trading strategies described in 4.1. The portfolio returns are computed by the formula (also given in 4.1):

$$R_{p,t} = \sum_{i=1}^{N} w_{t,i} y_{t,i} - C \sum_{i=1}^{N} \left| w_{t,i} - w_{t-1,i} \right|.$$

Notice that the parameter C, denoting the transaction costs, will be given in terms of daily standard deviations for the weekly data and standard deviations on 10 minutes for the 10 minute data.

First we will compare the different model performances when no transaction costs are taken into account. Then, we will investigate what transaction costs that can be allowed while still making profit, when using the best performing models for weekly and 10 minute returns respectively.

Without transaction costs

In tables 7.14-15 the Sharpe ratios for the models are presented for 10 minute and weekly returns respectively, when not taking any transaction costs into account. In table 7.14 we notice that all Sharpe ratios are positive for 10 minute returns, and that the Markov models dominate the others by far with Sharpe ratios above 9. This was expected since the Markov models dominated the others also with respect to the hit rates.

In table 7.15, for the weekly returns, we notice that the only model that produces a positive Sharpe ratio is the Bayesian VAR1 model, which also was the one that produced the best prediction results with respect to the RMSE.

Table 7.14. Sharpe ratios for the different models for 10 minute returns. Validated on 10 000 samples of 10 minute returns.

Model	**Sharpe ratio**
General VAR1	0.328
Bayesian VAR1	5.28
Bayesian VARX1	0.994
Bayesian VAR4	3.43
Markov, 4 states	9.62
Markov, 6 states	9.18

Table 7.15. Sharpe ratios for the different models for weekly returns. Validated on 280 samples of weekly returns.

Model	**Sharpe ratio**
General VAR1	-0.262
Bayesian VAR1	0.789
Bayesian VAR4	-0.211

The total yields in terms of daily and 10 minute standard deviations for the best performing models on 10 minutes and weekly returns, i.e. the 4 states Markov model and the Bayesian VAR1 model, are plotted in figures 7.1-2. No transaction costs are taken into account. While the Markov model for 10 minute returns seem to perform well constantly over time, the Bayesian VAR1 model for weekly returns seem to underperform during 2009 and 2010, but thereafter perform very well.

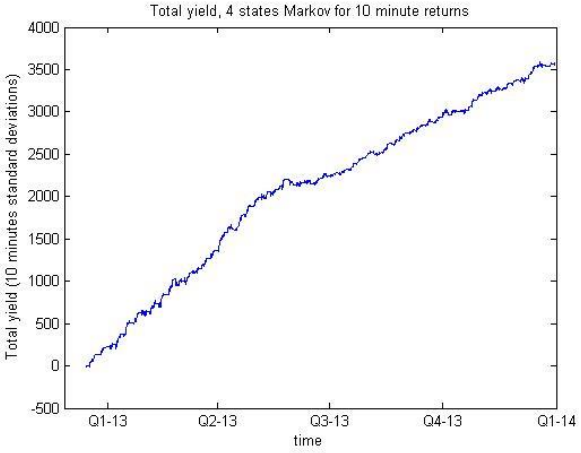


Figure 7.1. Total yield over time for the 4 states Markov model for 10 minute returns, when no transaction costs are taken into account. The model is validated on 10 000 observations of 10 minute returns.

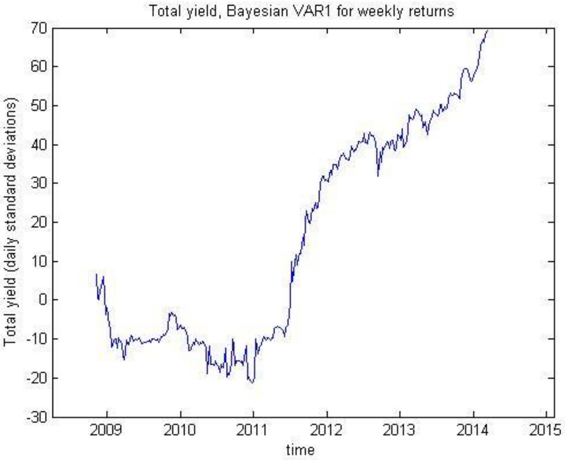


Figure 7.2. Total yield over time for the Bayesian VAR1 model for weekly returns, when no transaction costs are taken into account. The model is validated on 280 observations of weekly returns.

With transaction costs

In figure 7.3 the Sharpe ratio against the average transaction cost is plotted for the 10 minute returns using the best performing model, i.e. the 4 states Markov model. Notice that the Sharpe ratio becomes negative with average transaction costs of 0.021 standard deviations on 10 minutes, or about $\frac{0.021}{\sqrt{38}} = 0.0034$ daily standard deviations.

In figure 7.4 the Sharpe ratio is plotted against the average transaction cost for the weekly returns, using the Bayesian VAR1 model (which performed best without transaction costs). The Sharpe ratio is positive for transaction costs below 0.0415 daily standard deviations, considerably larger than for the 10 minutes data. It should however be noted that the 10 minute data is measured as arithmetic returns, whilst the weekly data are logarithmic returns. One can approximate the resulting normalized quantities of arithmetic and geometric returns as the same, resulting in approximately the same transaction costs measured in standard deviations.

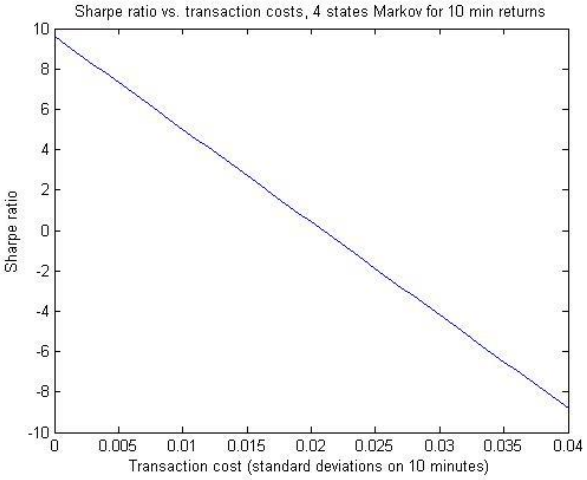


Figure 7.3. Sharpe ratio for the 10 minute returns against transaction costs in terms of standard deviations on 10 minutes, for the 4 states Markov model validated on a sample of 10 000 observations.

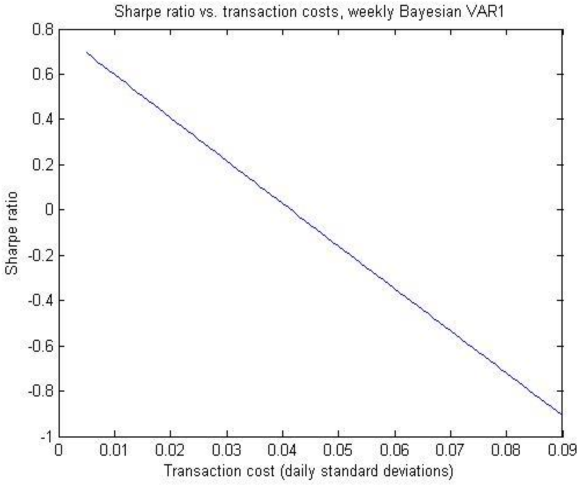


Figure 7.4. Sharpe ratio for the weekly returns against transaction costs in terms of daily standard deviations, for the Bayesian VAR1 model validated on a sample of 280 weekly observations.

The transaction costs involved varies in time and between market participants, depending on the spread between bid and ask prices, what amounts that are traded and what brokerage fee that is offered. However, it can be considered hard to get as low transaction costs as needed for the Sharpe ratio to be positive when trading at a 10 minute horizon, i.e. 0.0034 daily standard deviations, while it should be very possible to get lower transaction costs than 0.0415 daily standard deviations, as needed for the weekly trading horizon.

8. Conclusions

An aim of this thesis was to find models that could systematically predict future returns of foreign exchange rates better than a random walk model. This is something that has been achieved, with high significance for the 10 minute returns.

Another aim was to find a strategy that, by making use of the forecasting models, can yield a profit. This highly depends on the transaction costs involved, but we have shown that a positive Sharpe ratio had been achieved using the Bayesian VAR1 model for weekly returns out of sample, even for moderately high transaction costs.

8.1 The 10 minute horizon

For data sampled on 10 minutes, we found that a Markov model can predict whether the return on the next 10 minutes interval will be positive or negative accurately in significantly more than 50% of the cases, and that the predictive power seemed to increase even more with a regularized Bayesian Markov model. This indeed means that the random walk model can be rejected.

However, there might be an intuitive explanation of the positive prediction results, which would ruin the possibility of making profits using this model. Even if the foreign exchange market is very liquid, there will always be a spread between the bid and ask prices. If a market participant choose to cross the spread, i.e. buy at the ask price or sell at the bid price, the latest noted price will temporarily be higher/lower, and thereafter decrease/increase to somewhere in the middle of the spread. This might explain the negative autocorrelation at lag 1 for the 10 minute returns, and would imply that one cannot expect to be able to trade at the same price as the latest noted and thereby that this pattern cannot be taken into advantage for making profits.

Another aspect is the transaction costs. The simple trading strategy investigated, where one is trading on every 10 minute interval and updating the portfolio weights totally based on the prediction of the return on the next 10 minute interval, has shown to be very successful with a Sharpe ratio above 9 out of sample if no transaction costs are taken into account. However, the Sharpe ratio becomes negative even when very small transaction costs are considered. The dramatic negative effect of the transaction costs depends a lot on the fact that the portfolio weights change drastically on every 10 minute interval with the strategy considered. However, there are ways to reduce the variability of the weights, and thereby reduce the transaction costs. This is indeed something that would be interesting to investigate in future research.

8.2 The weekly horizon

The Bayesian VAR1 model for weekly returns performed slightly better than the random walk model out of sample with respect to the RMSE, and the hit rates were shown to be significantly above 50% for two currencies, the Thai Baht (at the 1% level) and the South Korean Won (at the 5% level). This indicates that the random walk hypothesis can be rejected, even if the arguments are somewhat vaguer than in the case of 10 minute returns. It was also shown that the RMSE can be decreased a lot using a Bayesian model compared to the frequentist approach.

For the weekly returns we made the assumption of causality, after adding non-causally sampled daily returns together at a weekly basis. One should keep this in mind, even if it is quite intuitive to assume that a few hours of knowledge about one currency not influences the future weekly return of another currency.

The trading strategy suggested yields a positive Sharpe ratio out of sample, even when moderately high transaction costs are considered. If one only had considered the period past 2011, the Sharpe ratio would be even higher, and it is very interesting to have found a trading strategy which actually had performed well out of sample in very recent times.

9. Bibliography

Bank For International Settlements. (2013, September 5). *Triennial Central Bank Survey of foreign exchange turnover in April 2013*. Retrieved June 04, 2014, from www.bis.org: http://www.bis.org/press/p130905.htm

Baviera, R., Vergni, D., & Vulpiani, A. (2000). Markovian approximation in foreign exchange markets. *Physica A 280*, 566-581.

Bennet, C., & Gil, M. A. (2012). *Measuring Historical Volatility.* Santander Investment Bolsa.

Bowman, A. W., & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis.* Oxford Science Publications.

Carriero, A., Kapetanios, G., & Marcellino, M. (2008). *Forecasting Exchange Rates with a Large Bayesian VAR.* European University Institute.

Fama, E. (1965). The Behavior of Stock-Market Prices. *The Journal of Business Vol. 38, No.1*, 34-105.

Karlsson, S. (2012). *Forecasting with Bayesian Vector Autoregressions.* Örebro University.

Litterman, R. B. (1979). *Techniques of forecasting using vector autoregressions.* Federal Reserve Bank of Minneapolis.

Lo, A. W., & MacKinlay, A. C. (1999). *A Non-Random Walk Down Wall Street.* Princeton University Press.

Malkiel, B. (1973). *A Random Walk Down Wall Street.* W. W. Norton & Company Inc.

Meese, R. A., & Rogoff, K. (1983). Empirical Exchange Rate Models of the Seventies. *Journal of International Economics 14*, 3-24.

Powell, J. L. (2008). *Notes On Nonparametric Regression Estimation.* Retrieved June 4, 2014, from University of California, Berkeley: http://eml.berkeley.edu/~powell/e241a_sp06/nrnotes.pdf

Robert, C. P. (2001). *The Bayesian Choice, 2nd Ed.* Springer.

Wikipedia, AIC. (2014, May 20). *Akaike information criterion*. Retrieved June 6, 2014, from www.wikipedia.org: http://en.wikipedia.org/wiki/Akaike_information_criterion

Wikipedia, BIC. (2014, April 9). *Bayesian information criterion*. Retrieved June 4, 2014, from www.wikipedia.org: http://en.wikipedia.org/wiki/Bayesian_information_criterion

Wikipedia, MPT. (2014, June 2). *Modern portfolio theory*. Retrieved June 4, 2014, from www.wikipedia.org: http://en.wikipedia.org/wiki/Modern_portfolio_theory