Charles University

Faculty of Social Sciences

Institute of Economic Studies

MASTER THESIS

Forecasting Term Structure of Government Bonds Using High Frequency Data

Author: **Bc. Jakub Kožíšek**

Supervisor: **PhDr. Jozef Baruník Ph.D.**

Academic Year: **2017/2018**

Declaration of Authorship

1. Hereby I declare that I have compiled this master thesis independently, using only the listed literature and sources.
2. I declare that the thesis has not been used for obtaining another title.
3. I agree on making this thesis accessible for study and research purposes.

Prague, January 5, 2018

Signature

Acknowledgments

I would like to express my sincere gratitude to my supervisor, PhDr. Jozef Baruník Ph.D., for his insight, consultations and support in helping this thesis come to life.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

Last but not least, I would like to thank my family for providing me with love and support through the course of my studies.

Abstract

This thesis investigates the use of realized volatility features from high frequency data in combination with neural networks to improve forecasts of the yield curve of government bonds. I use high frequency data on futures of four U.S. Treasury securities to estimate the Nelson-Siegel yield curve and realized variance of its parameters over the period of 25 years. The estimated parameters are used in prediction of the level, slope and curvature of the yield curve using an LSTM neural network and compared to the Dynamic Nelson-Siegel model. Results show that the use of realized variance and neural network outperforms autoregressive methods in prediction of the level and curvature in daily and monthly forecasts. The yield curve of government bonds itself has a predictive power on multiple macroeconomic variables, therefore improvements in its forecastability may have broader implications on forecasting the overall state of the economy.

JEL Classification F12, F21, F23, H25, H71, H87
Keywords neural networks, high frequency, term structure, yield curve, forecasting

Author's e-mail `kozisek.j@gmail.com`
Supervisor's e-mail `barunik@fsv.cuni.cz`

Abstrakt

Tato diplomová práce zkoumá využití kombinace realizované volatility z vysokofrekvenčních dat a neuronových sítí pro vylepšení prognóz výnosové křivky státních dluhopisů. K tomu používá vysokofrekvenční data futures kontraktů čtyř amerických cenných papírů pro odhad výnosové křivky Nelson-Siegelova modelu a denního realizovaného rozptylu jejích parametrů za období 25 let. Odhadované parametry jsou použity při predikci úrovně, sklonu a zakřivení výnosové křivky pomocí neuronové sítě LSTM a jsou porovnány s prognózou Dynamického Nelson-Siegelova modelu. Výsledky ukazují, že využití realizovaného rozptylu a neuronové sítě překonává autoregresivní metody při predikci úrovně a zakřivení křivky v denních a měsíčních prognózách. Výnosová křivka státních dluhopisů je sama o sobě využívána při prognózách makroekonomických proměnných, proto může mít zlepšení její prognózy širší dopad na předpověď celkového stavu ekonomiky.

Klasifikace JEL F12, F21, F23, H25, H71, H87
Klíčová slova neuronove site, vysokofrekvencni data

E-mail autora `kozisek.j@gmail.com`
E-mail vedoucího práce `barunik@fsv.cuni.cz`

Contents

List of Tables

List of Figures

Acronyms

ACF	Autocorrelation Function
ANN	Artificial Neural Network
AR	Autoregression
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	ARIMA with exogenous variables
CME	Chicago Mercantile Exchange
CT	Central Time
DNS	Dynamic Nelson-Siegel model
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
HDF	Hierarchical Data Format
HF	High Frequency
LSTM	Long Short-Term Memory neural network
MSE	Mean-squared-error
OLS	Ordinary Least Squares
RMSE	Root-mean-squared-error
RNN	Recurrent Neural Network
RV	Realized Variance
TDNN	Time Delay Neural Network
U.S.	United States
VAR	Vector Autoregression
YTM	Yield-To-Maturity

Master's Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University

Author:	**Bc. Jakub Kožíšek**	Supervisor:	PhDr. Jozef Baruník Ph.D.
E-mail:	kozisek.j@gmail.com	E-mail:	barunik@fsv.cuni.cz
Phone:	+420737053571	Phone:	+420776259273
Specialization:	ET&M	Defense Planned:	June 2017

Proposed Topic:

Forecasting the Term Structure of Government Bond Yields Using High Frequency Data

Motivation:

The term structure of government bonds yields is an important indicator of the performance of the economy and can be used to predict real economic activity, while having a greater predictive power than lagged growth in economic activity and index of leading indicators (Estrella, Arturo, Hardouvelis, 1992). The methods used in forecasting this indicator have been constrained by the data available, which could change with recent availability of high frequency data, opening new possibilities in forecasting the yield curve.

The unprecedented availability of high frequency data on US Treasury securities from years 1992-2010 should also allow for the use of non-parametric methods, such as neural networks, for prediction of term structure, the relationship between bond yields and its maturities. It is motivating to be one of the first to study the added value high frequency data might bring in an area with such an important predictive power on the overall state of the economy.

To test the feasibility of using high frequency data, I want to estimate the yield curve using standard parametric models and using neural networks to see which model performs better on the same data set in yield curve prediction. Neural networks have been extensively used in other areas, where large data sets had already been available, such as finance, trading or business analytics, but the availability of high frequency data could extend their use into prediction of government bond yields and the state of the economy.

The thesis should provide applicable results on its own, but also help in determining whether the use of high frequency data brings any new information in this area.

Hypotheses:

1. Term structure of government bond yields can be modeled using high frequency data, yielding results that are consistent with evaluation data.
2. High frequency data based forecasts outperform the forecasts based on daily yield curve data.
3. Forecasting the yield curve using neural networks outperforms standard parametric methods.

Methodology:

The goal of this thesis is to use new data to predict term structure of government bond yields. First, I will introduce the issue and focus on how important term structure of government bonds is. I will review existing literature on this topic and describe the data. In the empirical part, I will extract the realized volatility from the data and use it in a parametric model that has been used for such prediction, i.e. the Nelson-Siegel model. Since the data used are stochastic, the appropriate variant of this parametric model is its dynamic variant, introduced by Diebold, Li and Yue (2007). With the available data set should this dynamic parametric model yield valuable results on its own, but it will further serve as a control model to analyze whether neural networks perform better in the term structure prediction. Then I will construct a neural network, train it on a part of the data set and then use the rest of the data set to evaluate its estimation. The same will be done using the parametric model and the evaluation from the control part of the data set will be used to compare results of the two approaches.

Expected Contribution:

High frequency data have not been used extensively to model term structure of government bond yields and this thesis should show whether it is feasible to use it for this purpose, using a large dataset that was not available in the past. It should analyze whether the use of high frequency data yields precise predictions and outperforms past models that do not incorporate high frequency data.

Additionally, it should serve as an indicator whether available data on this topic can be used to train neural networks. The comparison of results of standard parametric estimation methods and the neural network approach should show whether neural networks should replace parametric models in term structure forecasting, or should act as a complement to the standard methods. If the combination of these methods tests well on evaluation data, then the forecast of the yield curve on future periods should bring valuable insight on future behavior of the market while also predicting the overall state of the economy.

Outline:

1) Introduction
2) Literature review
3) Data description
4) Methodology
5) Empirical results
6) Results interpretation
7) Conclusion

Core Bibliography:

Annaert, Jan, et al. "Estimating the yield curve using the Nelson-Siegel model: a ridge regression approach." International Review of Economics & Finance, Forthcoming (2012).

Aminian, F., Suarez, E. D., Aminian, M., & Walz, D. T. (2006). Forecasting economic data with neural networks. Computational Economics, 28(1), 71-88.

Baruník, J., & Malinska, B. (2016). Forecasting the term structure of crude oil futures prices with neural networks. Applied Energy, 164, 366-379.

Cieslak, A., & Povala, P. (2014). Information in the term structure of yield curve volatility. Journal of Finance, forthcoming.

Cox, John C., Jonathan E. Ingersoll Jr, and Stephen A. Ross. "A theory of the term structure of interest rates." *Econometrica: Journal of the Econometric Society* (1985): 385-407.

De Pooter, Michiel. "Examining the Nelson-Siegel class of term structure models: In-sample fit versus out-of-sample forecasting performance." Available at SSRN 992748 (2007).

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. Journal of econometrics, 130(2), 337-364.

Diebold, F. X., Li, C., & Yue, V. Z. (2008). Global yield curve dynamics and interactions: a dynamic Nelson–Siegel approach. Journal of Econometrics, 146(2), 351-363.

Estrella, Arturo, and Gikas A. Hardouvelis. "The term structure as a predictor of real economic activity." The Journal of Finance 46.2 (1991): 555-576.

Cheng, Wei, W. Wagner, and Chien-Hua Lin. "Forecasting the 30-year US treasury bond with a system of neural networks." NeuroVest Journal 1.2 (1996).

Koopman, S. J., M. I. Mallee, and M. van der Wel. The Dynamic Nelson-Siegel Model with Time-Varying Loadings and Volatility. Working Paper, Department of Econometrics, VU University Amsterdam, 2008.

Shin, Minchul, and Molin Zhong. "Does realized volatility help bond yield density prediction?." (2015).

Trippi, Robert R., and Efraim Turban. Neural networks in finance and investing: Using artificial intelligence to improve real world performance. McGraw-Hill, Inc., 1992.

Wegener, C., Spreckelsen, C., Basse, T., & Mettenheim, H. J. (2016). Forecasting Government Bond Yields with Neural Networks Considering Cointegration. Journal of Forecasting, 35(1), 86-92.

Author

Supervisor

Chapter 1

Introduction

The term structure and the yield curve of government bonds have been acknowledged to be a useful indicator of economic activity. It can be used to forecast the growth, recessions and even exchange rate between currencies. Compared to other methods of forecasting the state of the economy, it is praised for its simple observability and an ease of interpretation, as the sole knowledge of the slope of the yield curve can be used to indicate a growth or a downturn of the economy. The connection to other important indicators makes it viable to model and predict the term structure of government bonds. While the existing approaches to modelling the yield curve have been limited by the availability of data, the recent availability of high frequency data may change this situation, as it allows for incorporating methods which could supplement or even replace the traditional modelling techniques. This thesis analyzes the high frequency data on U.S. Treasury bonds from the past two decades, which had not been available in the past.

High frequency data are already being utilized in forecasting other indicators, such as the volatility of the S&P 500 Martens (2002) or the S&P 100 index volatility by Blair et al. (2010). Their use in the forecasting of the term structure, with the help of the traditional forecasting methods, such as the Dynamic Nelson-Siegel Model (Diebold et al. 2008), may bring additional information and improve the preciseness of the forecast, but the microstructure noise present in the data might have the opposite effect (Aït-Sahalia et al. 2011). To study the added value of the high frequency data availability in the yield curve modelling, it is necessary to review the existing literature on term structure forecasting and compare the results with the forecasts made using daily, monthly or quarterly data.

High frequency data bring additional opportunities to term structure modelling, as their size allows for the use of modern modelling techniques, such as artificial neural networks. Artificial neural networks generally need large data samples to be trained on to provide efficient out-of-sample forecasts and can model complex relationships in the data. To study the added value high frequency data may bring to yield curve modelling, a stochastic neural network will be constructed to forecast the latent factors of the Dynamic Nelson-Siegel Model. Additionally, the volatility information from intraday trading will be used as an input of the artificial neural network to evaluate its significance in high frequency data based forecasts.

With the ability of high frequency data to be aggregated to monthly or yearly data, it is possible to compare the forecasts made by the combination of an artificial neural network with the Dynamic Nelson-Siegel Model and traditional techniques on the same data and compare their predictions to real life outcomes. If the use of high frequency data and neural networks brings useful information in term structure modelling, the forecasts made using the proposed technique can yield viable alternative to forecasts made with currently used methods.

In this thesis, I 1) describe the U.S. Treasury bonds futures market 2) explain the importance of the term structure and its relation to the economic activity, 3) study high frequency data and neural networks in economics, 4) review existing methods for term structure modelling and estimate the parameters of the Nelson-Siegel yield curve, 5) construct a stochastic neural network and the dynamic variant of the Nelson-Siegel model for modelling term structure, 6) present the results of the approach and compare it to currently used methods and finally 7) summarize the findings.

Chapter 2

US Treasury Futures Market

The cornerstone of this thesis is the availability of high frequency, tick-to-tick sampled, prices of the U.S. Departement of Treasury futures. This chapter introduces the concept of futures and government bonds, description of concrete data and trading systems used in the thesis and features of future contracts sampled in high frequency.

2.1. Futures and Government Bonds

Futures, or future contracts, are legal agreements to buy the subject of the contract at a later date, for a price determined at the time of the agreement. Futures differ based on the type of the underlying subject – as such they belong to a broader group of derivative instruments, namely forward-based instruments. Futures are standardized contracts which are traded on an exchange (Heckinger et al. 2013).

Interest rate futures, such as government bond futures, are financial derivatives. A holder of government bond can either wait until the maturity of the bond and be repaid by the government in the form of the face value and coupon interest, or sell it at an earlier date. In the latter case, a government bond future is sold and the future agreement established in the previous paragraph is the face value and coupon interest which the government agrees to pay at the time of maturity of the underlying bond (Johnson et al. 2017).

U.S. Treasury bonds can be interpreted as a loan to the U.S. government, as investors buy them for a fixed price as a contract to receive the face value at the time of maturity and coupon interest over the lifetime of the bond.

The U.S. Department of Treasury currently offers Treasury Bills, Treasury Notes, Treasury Bonds and other types of securities – data on futures of three Treasury Notes (with 2-year, 5-year and 10-year maturities) and of one Treasury Bond (with 20-year maturity) are analyzed and described below, based on the specification sheet by CME Group (2013).

2-Year Treasury Note Futures

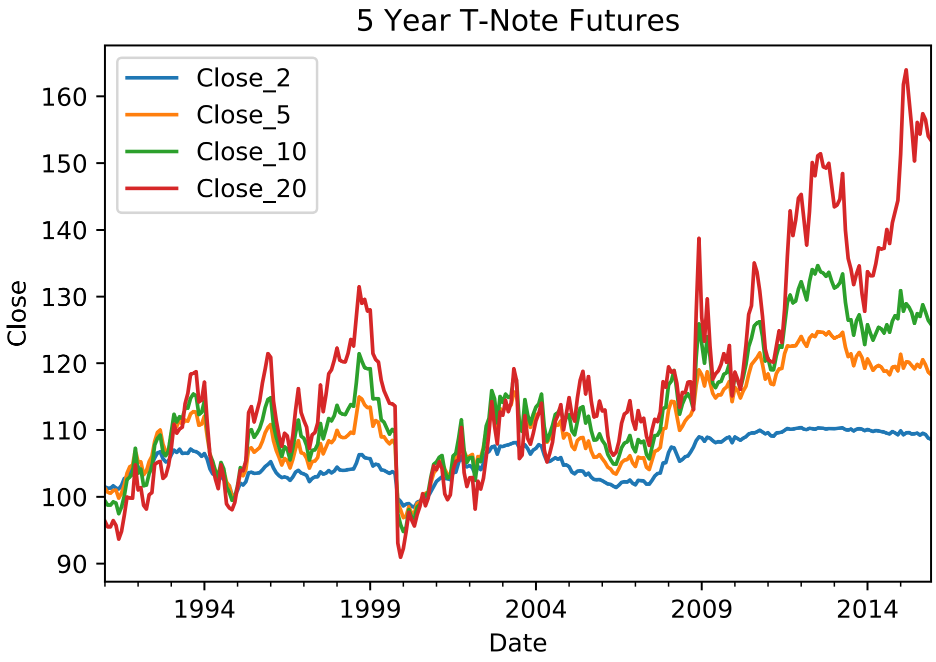

Figure 2.1.: Open Outcry Futures

Two Year Treasury Note Futures are a set of futures for U.S. Treasury Notes with an original maturity of not more than 5 years and 3 months and the remaining maturity at the time of purchase between 1 year and 9 months and 2 years. This is measured from the last day of the month of purchase/delivery. This security is traded at CME Globex under ticker symbols TU and ZT.

A single 2 year U.S. Treasury Note has a face value at maturity of $200,000 and is therefore the one with the highest face value from our data. The single price quota refers to the price of one point, which is set to $2000. It is the least traded security from our data, with the average volume of around 130,000 contracts per day in the 2004-2015 period. The tick size for this type of future is set to a $\frac{1}{4}$ of $\frac{1}{32}$ of a point ($\frac{1}{128}$ of $2000) or $15.625 per contract.

5-Year Treasury Note Futures

Five Year Treasury Note Futures are a set of futures for U.S. Treasury Notes with an original maturity of not more than 5 years and 3 months and the remaining maturity at the time of purchase no less than 4 years and 2 months. This is measured from the last day of the month of purchase/delivery. This security is traded at Open Outcry under the ticker symbol FV and at CME Globex under the ticker symbol ZT.

A single 5 year U.S. Treasury Note has a face value at maturity of $100,000. The single price quota refers to the price of one point, which is set to $1000. It is the second most traded security from our data, with the average volume of around 345,000 contracts per day in the 2004-2015 period. The tick size for this type of future is set to a $\frac{1}{4}$ of $\frac{1}{32}$ of a point ($\frac{1}{128}$ of $1000) or $7.8125 per contract.

10-Year Treasury Note Futures

Two Year Treasury Note Futures are a set of futures for U.S. Treasury Notes with an original maturity of not more than 10 years and the remaining maturity at the time of purchase of at least 6 years and 6 months, measured from the first day of the month of purchase/delivery. At the time of delivery, the security has the remaining maturity between 6.5 years and 10 years. It is traded at CME Globex under ticker symbols TY and ZN.

A single 10 year U.S. Treasury Note has a face value at maturity of $100,000. The single price quota refers to the price of one point, which is set to $1000. It is the most traded security from our data, with the average volume of more than 750,000 contracts per day in the 2004-2015 period. The tick size for this type of future is set to a $\frac{1}{4}$ of $\frac{1}{32}$ of a point ($\frac{1}{128}$ of $1000) or $7.8125 per contract.

Treasury Bond Futures

The securities in Treasury Bond Futures are divided into two categories – callable and non-callable. If a bond is callable, it is redeemable by its issuer (U.S. Department of Treasury) before it reaches its maturity. This can be beneficial to the issuer, as it does not have to pay interest for the rest of the term between the date of the call and the original maturity. This poses a risk to investors, as it is more difficult to estimate the yield without the knowledge whether (and when) the issuer decides to call the bond prematurely and possibly at a price lower than the current market price. For the yield, there are then two measures – the standard Yield to Maturity (YTM) and the

Yield to Call (YTC). The higher risk is usually accounted for by the coupon, which tends to be higher for callable bonds. Non-callable bonds may be redeemed only at the original maturity.

Treasury Bond Futures include both callable and non-callable types of bonds, but the callable bonds may not be called sooner than 15 years from the first day of the month of delivery. The non-callable bonds then have the remaining maturity of at least 15 years from the first day of the month of delivery. Between January of 1991 and February 2011, the securities in our data-set have no upper boundary of the remaining maturity. Beginning March 2011, the securities have been divided into Treasury Bond Futures with remaining maturity of not more than 25 years from the first day of the month of delivery and Ultra Treasury Bond Futures which have the remaining maturity of at least 25 years and no upper limit of the remaining maturity. Our data include the first kind, traded on CME Globex under ticker symbols US and ZB, therefore data between March 2011 and December 2015 include only securities with less than 25 years of remaining maturity. In the calculation of yield, we assume this condition to hold for the previous data as well, to obtain a single yield variable for the entire 1991-2015 period.

A single Treasury Bond has a face value at maturity of $100,000. The single price quota refers to the price of one point, which is set to $1000. It is the second least traded security in our data, with the average volume of more than 232,000 contracts per day in the 2004-2015 period. The tick size for this type of future is set to a $\frac{1}{4}$ of $\frac{1}{32}$ of a point ($\frac{1}{128}$ of $1000) or $7.8125 per contract.

2.1.1. Trading Regimes

The available data of fours government bonds is exported from two regimes of trading. The first regime is called Open Outcry and is open on every workday between 7:20 and 14:00 Central Time and is the sole trading regime responsible for the part of the data between 1991 and 2003. The second system is called CME Globex and it is an electronic trading system from the CME Group. As the name suggests, it provides access to electronic trading of futures and options from across the globe. As it is independent on trading hours in one particular market, it allows customers to trade virtually 24-hours per day. After the launch of the Globex system in 1987, the first futures began to trade in the system in 1992 and since then, the availability of futures and options has gradually expanded (CME Group 2017).

The U.S. government bond futures analyzed in this thesis have appeared in the system starting July 2003. Even beyond the addition of the futures, the conditions of the trade have changed during the time period of data available from the Globex system (2003-2015). Between July 2003 and December 2003, the window of trading was open between 19:40 and 16:00 Central Time (CT). The window was expanded in 2004 to 21 hours per day, being open between 19:00 and 16:00 CT.

In 2005, the window was further expanded to 22 hours per day, resulting in the effective window between 18:00 and 16:00 CT in 2005-2007.

Since 2008, the window has expanded to 23 hours per day, with the exception of the year 2010, where the available data suggests narrowing of the window to 17:30-16:00 CT. The current state, with the trade being open between 17:00 and 16:00 CT, Monday through Sunday, therefore applies to the majority of the analyzed data. It is however important to acknowledge the changes throughout the dataset, as the trading window affects variables used in inference of the yield curve. The closing time is set to 16:00 CT in the period of 2003-2015, but the off-time interval varies and it can potentially affect traders' decision to postpone their trade. Furthermore, the changing trading interval inherently affects the derivation of volatility. Several approaches to account for the problem have been proposed, such choosing a dynamic interval for the realized volatility calculation, but they could complicate generalization of the forecasting method to other series.

In order to focus effort on a single series, the Open Outcry is the sole system examined in the thesis for its duration between 1991 and 2015 and for its fixed window over the whole period. For the aforementioned reasons, the existence and structure of the second regime affects needs to be stressed, as both systems interlap in the 2003-2015 period.

2.2. Volatility in Finance

Volatility is a degree to which a financial time series varies over a certain period. It measures the level of randomness in the series. Higher volatility implies more frequent changes in price over time.

Measuring volatility is dependent on the sampling frequency and opening time of markets. According to Jones et al. (1994), the major source of short-term volatility in the market is public information and a large proportion of volatility occurs without trades, after the market is closed.

To separate different types of volatility, there is the actual volatility of the series, the implied volatility of the underlying instrument (in case of futures), historical volatility calculated from the past data and realized volatility, a measure of volatility over some period, calculated from the returns.

The relationship between the volatility itself and returns is described by the leverage and volatility feedback effects, where the leverage effect describes the effect of changes in returns to volatility and the volatility feedback describes the opposite, where volatility affects returns. Dufour et al. (2012) examine the leverage and volatility feedback effects on high frequency data and finds the volatility feedback effect to be negligable on the analyzed horizons.

Volatility is closely related to risk and overall sentiment of the market. Yu and Yuan (2011) examines the relation of returns, variance and the sentiment of the market and comes to a conclusion that the relation between returns and variance differs based on the sentiment of the market. Wang et al. (2006) find that the changes in the sentiment of the market are caused by volatility and returns rather than vice versa.

The interconnected nature of returns and volatility leads to researching volatility as a factor in predicting returns. Guo and Savickas (2006) find a predictive power of idiosyncratic volatility on stock market returns, noting that high past returns with high volatility predicts low future returns and the predictive power of average idiosyncratic volatility.

Christoffersen and Diebold (2006) describe the relationship between the direction-of-change of asset returns and asset return volatilities and find them very interconnected with the effect on the direction-of-change being smaller on daily and annual data.

In contrast, Maheu and McCurdy (2011) examine the intraday information to infer realized volatility and find its predictive power for forecasting the distribution of stock returns on the S&P 500 data. Anatolyev and Barunik (2017) propose an approach to forecasting conditional probability distributions of asset returns using past volatility.

Most importantly for the scope of this thesis, Shin and Zhong (2017) use realized volatility as a volatility proxy to improve the Dynamic Nelson-Siegel model for yield curve forecasting. He mentiones the realized volatility as being neglected in the context of the Dynamic Nelson-Siegel model, but finds that it improves density forecasts of U.S. bond yields.

Volatility is not the only feature, which can be extracted from high frequency data. Intraday structure of trading patterns can be analyzed as well, as studied by Admati and Pfleiderer (1988), but the realized volatility use in government bond yield density prediction has the theoretical basis in Shin and Zhong (2017) and can be expanded to point prediction of the term structure.

Chapter 3

Term Structure

Term structure, as understood with respect to government bonds, is the relation between their interest rate and maturity – the graphical representation of this relationship is the yield curve. The yield curve has the borrowing period, or the time to maturity on its x-axis and the yield (usually in percentage points) on its y-axis. As mentioned above, the yield curve represents the relationship between maturity and yield of a bond, which suggests that it usually has an upward sloping shape – if government bonds can be interpreted as money borrowed by the U.S. Treasury from its lenders, then it is logical to assume that the longer the money is lent, the higher is the interest rate and vice versa. It should be noted that this is not always the case, as an investment into government bonds may have other objectives than maximizing the yield. In a situation where the world markets are experiencing a decrease of confidence, investing into government bonds with a longer time to maturity may be viable if the issuer has a reliable outlook of its economy. In such case, one can see a flattening or downward sloping yield curve, as for some lenders a lower risk may be more viable than higher interest rate. Observing the shape of the yield curve is therefore an important discipline, as an inverted shape of the yield curve can indicate a worsening situation of the economy, or an economic downturn (Cwik 2005). The progression of the yield curve through time is shown in Figure 3.1.

3.1. Basic Terms

Throughout the thesis, the following terms may be used, typically without further explanation, therefore to provide their meaning beforehand, they are explained below, based on The handbook of fixed income securities by Fabozzi and Mann (2012).

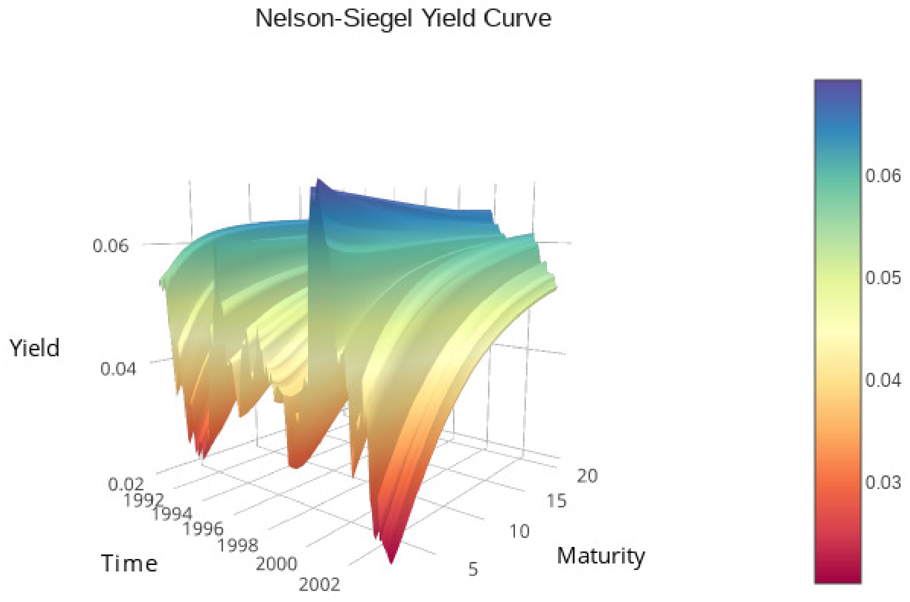

Figure 3.1.: Term structure of U.S. Treasury bond yields

Term structure

Term structure is a relation between the term of an investment (for example months to maturity of a newly issued bond) and the interest rate.

Yield curve

The graphical representation of the term structure. Typically a graph with time t on the x-axis which represents the term of the investment and the interest rate on the y-axis.

Interest rate

Interest rate is effectively how much the lender charges the borrower for the amount lent.

Arbitrage

Arbitrage is a sale that happens on different markets and takes advantage of price differences between them. Arbitrage therefore helps to remove price imbalances between markets. An example of this is buying a cheap commodity on a foreign market and instantaneously selling it on the domestic market for a higher price. With the presence of arbitrage, prices on these two markets (in reality, there can be more than two) converge.

Outstanding principal

Outstanding principal is the amount still owed from the original borrowed amount. In the case of government bonds, it is the price of the bond, which gets repaid once the bond reaches maturity, together with the interest rate.

Maturity

Maturity is the date when the outstanding principal is repaid together with the interest rate and the financial obligation ceases to exist. It typically is impossible to liquify the bond before this date, unless sold on the market. Different types of bonds are recognized by the time to their maturity.

Holding period return

Holding period return, or holding period yield is the total yield accumulated for the period an asset (a bond in our case) was held. It is usually expressed as a percentage. Similarly, depending on the type of the asset, it can be calculated as annualized.

It can be expressed by the following equation

$$HPR = \frac{\text{end value} - \text{initial value}}{\text{initial value}}$$

Coupon

Coupon is an interest rate that is paid annually. It is expressed as a percentage of the value of the particular bond.

Coupon/Zero-coupon Bonds

Coupon bonds are securities that have a coupon attached to them, so they are paid an interest rate annually.

Spot rate

Spot rate is the value of the bond on the market at the moment. Spot rate can therefore change, based on when the prices are updated.

Forward rate

In opposite to the spot rate, forward rate is the value at a future point of time. It is calculated using the spot rate and the interest rate between now and the date of the future transaction.

Discount rate

Discount rate is the rate by which the market price of a bond is lower than its face value at maturity. If one wants to sell the bond before its maturity, then the discount rate is typically the value sacrificed for liquifying the bond at such time.

Yield to maturity

Yield to maturity is the discount rate for the case when the investor holds the bond until its maturity.

Rate of return

Rate of return is the rate of the bond at maturity. If the investor holds the bond until its maturity, then the rate of return is equal to the yield to maturity. If more than zero, it is called the capital gain, if less than zero, it is called the capital loss.

Term spread

Term spread, or yield spread is the difference between the yields of two bonds, typically a short term one and long term one. In recessions, the difference generally rises, as investors are more risk averse and prefer short term securities. This needs to be reflected in the yield of the long term securities to motivate lenders.

3.2. Yield Curve Features

The list of typical shapes of the yield curve can help in better recognition of different states of the economy.

Normal (upward-sloping) curve

The usual shape of the yield curve, as mentioned above, is upward-sloping.

Inverted (downward-sloping) curve

The inverted shape of the yield curve is being linked to upcoming recession, as further studied in the section on the importance of the term structure.

Hump-shaped curve

Also known as bell-shaped curve. According to Vayanos and Vila (2009), such hump can arise in a response to a decrease in demand. Hump-shaped curve could therefore signalize a slowing down economy. Its inverted counterpart is the **through-shaped** (or inverted hump-shaped) curve

Variance

In general, even though both long-term and short-term yields vary (both their means and standard deviations), it can be observed that the short-term yields move more than their long-term counterparts, as reported by Campbell (1995).

3.3. Theory

The shape of the yield curve is not trivial to interpret, but there are three main theories that explain the relationship between yields and different maturities.

3.3.1. Expectation Hypothesis

The expectation hypothesis explains the relationship between the short term and the long term interest rates. It has the following form

$$i_t^{(n)} = \frac{1}{k} \sum_{i=0}^{k-1} E_t \left[i_{t+mi}^{(m)} \right] + c^{(n,m)}$$

where m is the period of the short term bond and n is the period of the long term bond. $c^{(n,m)}$ is then the difference between their premiums (or a term premium between them), i_t is the interest rate belonging to the respective period and $k = \frac{n}{m}$, $k \in Z$ (Della Corte et al. 2008).

In other words, the interest rate of the long term bond is equal to the ratio of $\frac{m}{n}$ times the sum of the expected interest rates at time $t + mi$ plus the term premium.

According to the expectation hypothesis, an investor should be indifferent between holding one long term bond or subsequently holding a number of short term bonds, in case the constant term $c^{(n,m)}$ is equal to zero. The only difference between these two scenarios is then the constant term, given it is non-zero. The constant term is equal to zero only under the pure expectation hypothesis (Della Corte et al. 2008).

The yield curve of the term structure described in this hypothesis will be either flat (in the pure version of the expectation hypothesis), or upward sloping, if the constant term is larger than zero.

3.3.2. Liquidity Preference Theory

The liquidity preference theory differs from the expectation theory by the assumption that the investors are risk-averse and to invest in bonds with longer term maturities, they need to be motivated by a premium. If the premium is not offered, they choose bonds with short term maturities instead (Gibson et al. 2010). In short, agents in this model are risk averse and prefer liquidity.

The yield curve in the liquidity preference theory is expressed by the following equation

$$R(t,T) = \frac{1}{T-t} \left[\int_t^T E_t(r(s))\, ds + \int_t^T L(s,T)\, ds \right]$$

where t is the time, T is the time of the maturity of a particular bond and $L(t,T) > 0$ is the instantaneous term premium at time t (Gibson et al. 2010).

The yield curve of the term structure described by the aforementioned equation is upward sloping, as investors need to be motivated by higher premiums in order to invest into bonds with longer time to maturity. The yield will be higher for longer term investments.

3.3.3. Preferred Habitat Theory

In the preferred habitat theory, the yield curve has the same equation as in the liquidity preference theory

$$R(t,T) = \frac{1}{T-t}\left[\int_t^T E_t\left(r(s)\right)ds + \int_t^T L(s,T)\,ds\right]$$

The main difference is that different investors have different time preferences and the instantaneous term premium $L(t,T)$ is not bounded by zero, but can be even negative or equal to zero (Gibson et al. 2010). With the no-arbitrage restrictions to this theory, it explains demand and supply shocks as driving forces behind term structure movements and therefore assumes the monetary policy role of the Central Bank to be able to affect the yield curve shapes. This assumption is important in the yield curve estimation and forecasting, as the monetary policy may create exogenous shocks, impossible to explain by the standard forecasting models.

The resulting yield curve can therefore have different shapes – upward sloping, flat or even downward sloping.

3.4. The Importance of Term Structure

The term structure, or the yield curve is one of the leading economic indicators for its strong relationship with the performance of the economy. While stock prices are usually a better indicator for short term predictions, any forecast more than one quarter ahead is better suited for the yield curve, which performs better than any other single economic indicator (Estrella and Mishkin 1998).

3.4.1. Forecasting Recessions

The importance of the slope of the yield curve can be shown on its ability to forecast recessions, as studied by Wright (2006). He acknowledges that downward-sloping yield curves do not necessarily mean a recession is imminent (as shown by the inverted slopes of yield curves of Australia or the United Kingdom), but shows that using a simple probit regression model and incorporating both the federal funds rate and term spread can be used to forecast the odds of economic recession. In the past, Wright (2006) and Estrella and Mishkin (1996) have shown that the steepness of the yield curve can be a valuable indicator of an upcoming recession. Using the term spread in the model alone can supplement more complex economic models and even surpass them for its simplicity and easy interpretation. It can also behave as a double check mechanism to verify the more complicated models in case of inconsistencies between them and the yield curve and therefore act as an indicator of a possible errors in more traditional forecasting methods. Finally, it can serve as a predictor of its own, if used in a probit regression model (Estrella and Mishkin 1996). Similarly, Dueker (1997) presents a robust evidence that the yield curve itself is a useful recession predictor, using a probit model on monthly data. Among the tested variables (the change in the index of leading indicators, real M2 money supply growth, the percentage change in the S&P 500, term spread and the yield curve slope), the yield curve slope turned out to be the single best predictor of recession (Dueker 1997).

With the presence of structural breaks in the data (with uncertain dates), the interpretation of the probit models using yield curve may be tricky, as shown by Chauvet and Potter (2002) – when comparing models with and without structural breaks in the data (with unknown breakpoints), the prediction may differ significantly. It should be therefore noted that one must take into account the specifics of the given economy.

3.4.2. Forecasting Growth

As a generalization of the previously mentioned recession forecast, the yield curve has been used to predict the future economic activity as well, substituting a binary response variable for a continuous one. Nevertheless, according to Haubrich and Dombrosky (1996), it has become more difficult to use the term structure for this purpose, as the relationship of the real output growth and the yield curve has changed and despite their closed line up, it diminishes the possibility to use the yield curve to forecast future growth. The authors argue that while the yield curve may not replace more complex forecasting methods, it can serve as a check of the more sophisticated forecasts, as long as its worsening performance as a sole predictor is taken into account (Haubrich and Dombrosky 1996).

A revision of the overview made 8 years later by Ang et al. (2006) shows that using a Vector Autoregression (VAR) and multiple factors of the yield curve, such as the short rate and term spreads, one can efficiently predict GDP growth. The slope of the curve (in the form of the maximal maturity difference in their model) is therefore not the only a factor, but an important one in GDP forecasting. In their no-arbitrage model, the short rate has the highest explanatory power and an overall conclusion is that the yield curve model produces efficient out-of-sample forecasts of the GDP. Furthermore, they propose including additional macroeconomic variables into the yield-curve GDP forecast model, to attain more efficient forecasts (Ang et al. 2006). It is important to note that including the same macroeconomic variables both into forecasting the term structure and into the subsequent yield-curve based GDP growth modelling may introduce collinearity in the former model and should be avoided if not accounted for.

The study of stability of forecasting growth using the yield curve by Giacomini and Rossi (2006) offers a counterargument against using the yield-curve/growth relationship, because this relationship changes based on different monetary regimes and cannot be therefore applied universally.

3.4.3. Forecasting Exchange Rate

With the ability to observe the term structure of government bonds in different countries, it is also possible to forecast the exchange rate of their respective currencies. As the exchange rate is related to the expectations of the future economic activity, the difference of their yield curves can be used to predict the relative value of their currencies (Chen and Tsang 2013). Being able to forecast the exchange rate is useful in predicting the foreign trade, which may bring valuable information on the development of the economy, especially for import/export oriented countries.

3.4.4. Criticisms

The instability of models based on the term structure with different monetary regimes is not the only criticism of using the term structure to forecast economic activity.

Berk (1998) argues that the explanatory power of the yield curve in inflation forecasting models should not be overstated, as it is unclear how much of the future inflation can be attributed to the interest rate. While it is still a valuable indicator of the growth in the economy, with the presence of structural shocks, its use in monetary policy should be reviewed if the goal is to forecast inflation.

Chapter 4

Neural Networks and High Frequency Data

4.1. High Frequency Data

High frequency data differs from their traditional counterparts by the frequency of their sampling. The name high frequency data is relative, as in the past, even daily observations were considered high frequency. Today, with the availability of vast storage space and distributed computing and with the rising importance of big data, the interpretation of high frequency data varies. The disciplines of finance and trading have the advantage of having each transaction logged, yielding tick-to-tick data. The problems that arise from the availability of such data are the sampling frequency, where some data points can be spread across several minutes, but others can differ by mere milliseconds. To account for these discrepancies, it is important to choose a common denominator, an interval that can be observed across the whole data set. Similarly, with the availability of intra-day observations, what needs to be taken into account are the periods of inactivity, such as the closing hours of stock markets.

In the high-frequency data, it is possible to find periodicities across weeks, days, hours or even minutes, which cannot be observed in monthly or quarterly data (Andersen and Bollerslev 1997a). Additionally, it can also carry long-memory and long-run dependencies (Andersen and Bollerslev 1997b).

4.1.1. High Frequency Data in Economics and Finance

High frequency data has been used in economics and finance for a few decades, as the computers allowed collection of observations on tick-to-tick basis. A prime example of this is the foreign exchange rate market, where the tick-by-tick collection began in the early nineteen-eighties. The main characteristic of the high frequency foreign exchange rate data is the constant availability of observations. While in other markets the collection is limited by the opening trading hours (such as in the futures markets analyzed in this thesis), currencies can be traded non-stop. The reason for this difference is the existence of many markets over the world which essentially trade the same product (same currencies) while their opening hours differ according to their local time zone. Time zone differences then create patterns in the data which can be traced back to the opening hours of markets, as different currencies have trading peaks at different times of the 24-hour cycle.

4.2. Artificial Neural Networks

Artificial neural networks are a type of a mathematical system, whose model has been inspired by the design of the human brain. It structures the system into neurons which are connected to each other. Each neuron is given a different weight (coefficient) and each connection between neurons is given such weight as well. As the neural network is trained (imagine training a regression model such as OLS), the weights of neurons and their connections change, which can be interpreted as learning (Klerfors and Huston 1998). While artificial neural networks are being used in many fields, in the particular field of time series forecasting, they can be compared to parametric models. The neural network created for regression of dependent and explanatory variables is then trained on a set of data, consequently validated on the testing data set and then it can serve as a prediction engine on its own.

4.2.1. Types of Artificial Neural Networks for time series forecasting

4.2.1.1. Radial basis function networks

A type of artificial neural network that uses radial basis functions, which are used for interpolating in a high dimensional space. This reduces a non-linear problem to a linear least squares optimization (Lowe and Broomhead 1988).

4.2.1.2. Stochastic neural networks

Stochastic neural network is a stochastic process that uses time series to predict the values at time $t_0 + \Delta$ based on the variables from time $t \leq t_0$ (Zhang 2013). With its ability to forecast time series data, it is useful in volatility forecasting, which is crucial for estimation of the Dynamic Nelson-Siegel model.

4.2.1.3. Time delay neural networks

modelling high frequency data may pose as a challenge to standard methods, as the availability of large quantities of frequently sampled observations allows to track and model longer time periodicities. An approach to modelling longer temporal contexts has been proposed by Peddinti et al. (2015) in the form of using Time-Delay Neural Networks (TDNN). The history of time-delay neural networks reaches back to 1989, when Waibel et al. (1989) used it for phoneme recognition. TDNN is a feed-forward neural network and consists of multiple layers where units of each layer are connected to all units of the layer below. Each unit in the lowest layer represents a time period of an observation, therefore the higher the information flows through the layers, the more complex and wider temporal relationships between periods can be captured. Activations are computed between each layer on their output – their hyperparameters. During the network training, the layers go through a back-propagation in which the hyperparameters are updated. The architecture allows to capture longer temporal dynamics of the training data, such as seasonalities. Increasing width of the input window may nevertheless decrease the performance of the network, as according to Clouse et al. (1997), time-delay neural networks perform poorly on wide input windows or on data with little repetition.

Focused Time-Delay Neural Networks (FTDNN) have been used by Baruník and Malinska (2016) to forecast the Dynamic Nelson-Siegel parameters, modelling the term structure of crude oil futures prices. In forecasting the term structure, the proposed regression framework based on the neural networks outperformed both autoregressive models and vector autoregressive models which served as a benchmark. Other uses of TDNNs include the aforementioned speech recognition and video analysis.

4.2.1.4. Recurrent neural networks

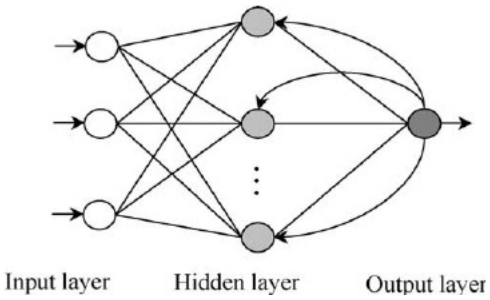

Figure 4.1.: RNN Architecture (Chiang et al. (2004))

Unlike front-feeding neural networks, recurrent neural networks are designed specifically to model sequential or time-varying patterns and dynamics of time series. Their architecture is based on interconnected nodes which can be connected to each other and to themselves. In fully connected RNNs, all nodes are connected to each other (including themselves) with bidirectional connections, while in partially connected networks some connections or direction can be omitted. The design allows to learn non-linear temporal relationships using back-propagation in continuous-time. The extension to discrete-time data modelling has been proposed by dos Santos and Von Zuben (2001), introducing delays of one period in each feedback path.

Originally created by Rumelhart et al. (1985) for character strings modelling, their use has expanded to other sequential datasets, such as speech and handwriting recognition or financial time series forecasting in their discrete-time application (Medsker and Jain 2001).

4.2.1.5. LSTM

Long short-memory neural network is a special case of the recurrent neural network, which tries to capture long term relationships in time series. It was initially developed by Hochreiter and Schmidhuber (1997a), but it lacked the core components of today‘s LSTMs, such as forget gates. The core idea of contemporary LSTMs is a memory cell guarded by gates, which decide how the cell is updated, if the information in the cell is kept, or how it is withdrawn.

The RNN diagram in Figure 4.2 compares the architecture of a simple recurrent network (SRN) cell and an LSTM block. As with other recurrent neural networks, a cell output is connected to its input and all other cells, but it adds gates, which filter

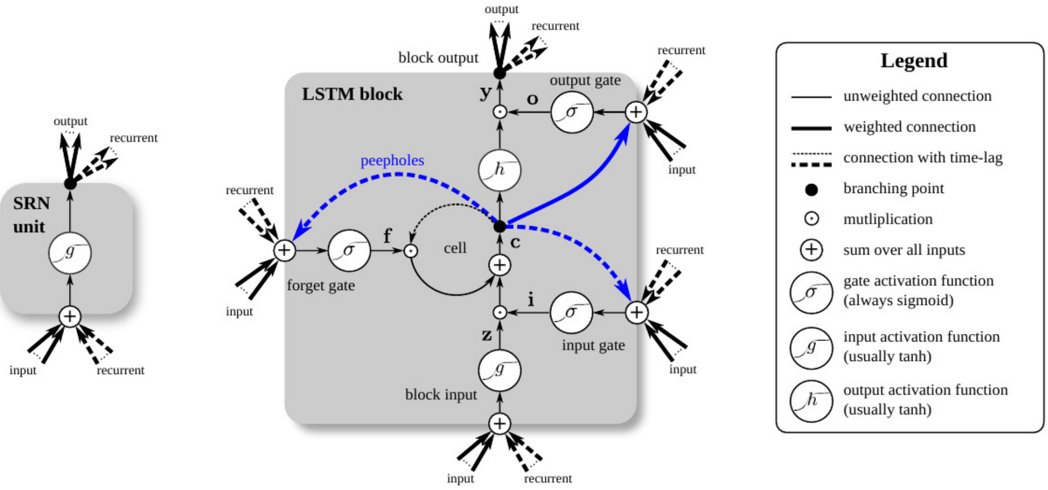

Figure 4.2.: Simple recurrent network vs LSTM (Greff et al. 2016)

and regulate the cell input, allowing the cell to retain its state over multiple periods. Peephole connection (blue on the diagram) were added by Gers et al. (2002) to regulate the continuous nature of the network in pursuit of better recognition of rhythmic sequences in the data. The forget gate evaluates the input (and outputs from other cells) and using its own activation function decides when the cell contents should be reset, effectively forgetting the time information which is no longer necessary. The selective nature of the time learning process can be more effective than learning in simple recurrent networks, as only the important information is kept in the network. With enough training data, this allows the LSTMs to be trained on longer time windows while only retaining the relevant signals (Hochreiter and Schmidhuber 1997b). LSTMs are being used on supervised an unsupervised problems, such as sequence classification, speech recognition, or time series forecasting (Greff et al. 2016). In addition to being commonly used in classification problems, they have been shown to outperform auto-regressive integrated moving average (ARIMA) models in real-time traffic prediction (Zhao et al. 2017).

4.2.2. Advantages of Artificial Neural Networks

The use of artificial neural networks can not only greatly complement the use of standard parametric methods, can surpass them in some disciplines, leading to their expansion in the recent years. Among the advantages of their use, as analyzed by Tu (1996) is the ability to capture complex relationships between seemingly unrelated variables and weigh the relationships to attain the best possible model. Especially in their multi-layer form, they allow for an effective dimensionality reduction in an automated manner. While other dimensionality reduction techniques, such as the principal component analysis require the knowledge of the correlation between variables in order to correctly interpret their results, neural networks behave as a “black box” which adapts to the data fed in the training phase. The training abilities are a significant feature on its own, since it is possible to use multiple training mechanisms to utilize a training data set, which results in a more precise forecast capabilities.

4.2.3. Disadvantages of Artificial Neural Networks

There are naturally also drawbacks and criticisms to the artificial neural network approach, as compared to standard parametric methods. Some of the criticisms are mentioned by Tu (1996).

The main criticism of artificial neural networks is their low explanatory value coming from their complexity. While their predictive value may be superior to other methods, researchers have limited insight into the relationships of the explanatory variables and the underlying structure of the network. The “black box” nature of artificial neural networks makes the interpretation of the results difficult, as identifying significant independent variables in the network can be non-trivial (Olden and Jackson 2002).

Another criticism is targeted at their high computational requirements. The progress in the field of computer science not only helped to spread the use of neural networks, but it also lead to diminishing importance of such criticism . Therefore one of the reasons for the recent expansion of neural networks into new fields of use is the increasing availability of accessible computational power, which allows for faster training and more complex models, such as deep neural networks (DNN).

Neural networks can be also prone to overfitting which poses a significant problem in their use. The reason for such proneness to overfitting is the sampling noise existant in training the neural network which may not appear in the data used for forecasting. Overfitting is a problem present in regression analysis as a whole and can be explained as including unnecessary terms in the model, making the model too complicated and therefore violating parsimony. According to Hawkins (2004), an example of overfitting is making the model too flexible or adding irrelevant predictors.

Other criticisms stem from the empirical nature of the development of neural networks and is understandably connected to the “black box” nature. As every neural network is different and its fit is attained by training it on real data, it is difficult to theoretically support and anticipate their results. It is therefore important to choose a family of neural networks suited for the intended use and then dedicate enough data to the training and evaluation. With a small training data set, the prediction may behave unexpectedly. Additionally, the prediction results cannot be extrapolated, which poses a problem in their generalization and use beyond the type of data set it has been trained on.

4.2.4. Neural Networks in Economics and Finance

Artificial neural networks have found their way into econometrics, as alternatives to traditional regression models.

They have been used to forecast the exchange rate, stock performance, but its use in forecasting the government bond interest rate term structure with high frequency data is an approach that has not been featured extensively in the literature. To compare the performance of neural networks and standard statistical models, Toulson (1996) has compared the performance of neural networks and statistical techniques on a high frequency sample of exchange rate data. While the meaning of high frequency has changed over time, she was able to study 5-minute intervals of the USD/DEM (Deutsche Mark) exchange rate, a frequency that is considered high frequency in the contemporary research as well. After forecasting the volatility with both the neural network and a stochastic volatility model, expressed in space state form as

$$log\left(v_t^2\right) = -1.27 + h_t + \varepsilon_t^*$$

$$h_{t+1} = \gamma + \phi h_t + \eta_t$$

where ε_t^* is a Gaussian white noise process, h_t is the volatility, v_t is the variance and $0 \leq \phi \leq 1$.

The forecast made using the aforementioned model performed similarly to a neural network based forecast. It is important to note that forecasts made using the combination of these two approaches performed significantly better than either of the single models.

Chapter 5

Yield Curve Estimation

The purpose of this thesis is to test three hypothesis related to forecasting of the yield curve of government bonds.

- To explore the possibility of using features from high frequency data in yield curve forecasts.
- To test whether the inclusion of features of high frequency data helps to improve accuracy of yield curve forecasts.
- To compare forecasts made with neural networks to forecasts from common parametric methods already used in yield curve modelling and show that neural networks yield higher accuracy on the same testing data.

In this chapter, the practical possibility of using high frequency features is presented with the use of intraday volatility of Nelson-Siegel yield curve parameters. In extracting the volatility information from the tick-by-tick data, the 5-minute sampled variance as an estimator of the volatility to mitigate the presence of microstructure noise as described (Aït-Sahalia et al. 2011). The choice of this particular method is justified by Liu et al. (2015). In a comparison to more than 400 other realized measures on 31 different financial assets, they conclude it is difficult to significantly outperform the 5-minute realized variance measure.

To determine the usefulness of the extracted realized variance of Nelson-Siegel parameters, the Granger causality method was selected as the deciding criteria. Proposed by Granger (1969), the test compares two vector autoregressive models and inferes which variables “Granger-cause” the other. While the Granger causality should not be taken as causality in the traditional sense, it uses autoregressive methods for the causality inference and it is therefore related to the Dynamic Nelson-Siegel model where the parameters follow the autoregressive process as well. The proposal and comparison of forecasting methods is described in the next chapter.

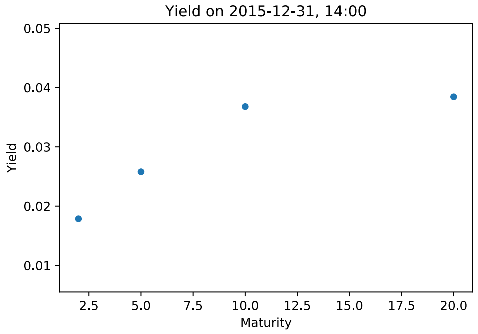

Figure 5.1.: Yield before fitting the yield curve

5.1. Yield Curve

The yield curve is created by connecting the points of yield of securities with various maturities. In the case of this thesis, it is a yield-to-maturity of four U.S. Treasury bonds with maturities of 2, 5, 10 and 20 years. An example of yield-to-maturity before fitting a yield curve in one particular moment can be seen in Figure 5.1.

5.2. Term Structure Models

Being able to forecast the yield curve of government bonds has been important not only for predicting the economic activity, but also for investors to manage their portfolio and help with decision making with regards to government bond investments. It is possible to model the yield curve using methods estimating the closest fit to yield across maturities.

The main family of the forecast models is the class of no-arbitrage affine models, which impose the no-arbitrage restriction. Notable models of the yield curve include McCulloch (1971), which uses polynomial splines, the Vasicek and Fong (1982) model with polynomial splines or single factor models of Vasicek (1977) or Cox et al. (1985). The aforementioned models were later expanded by Hull and White to allow for stochastic changes in the model parameters, therefore opening the possibility of yield curve forecasting.

The most notable family of models for the use in this thesis is the Nelson-Siegel model and its successors.

The widely used **Nelson-Siegel model** for fitting the term structure was first presented by Nelson and Siegel (1987) and further enhanced to its dynamic variant by

Diebold et al. in 2008. The original single-country model is constructed as follows:

$$y_i(\tau) = l_i + s_i\left(\frac{1-e^{-\lambda_i\tau}}{\lambda_1\tau}\right) + c_i\left(\frac{1-e^{-\lambda_i\tau}}{\lambda_1\tau} - e^{-\lambda_i\tau}\right) + \nu_i(\tau)$$

where $y_i(\tau)$ is the continuously-compounded zero-coupon nominal yield, τ is the time to the maturity of the particular bond (it is for example a τ-month bond), ν_i is a disturbance term and l_i, s_i and c_i are parameters (Diebold et al. 2008). While some models use the τ variable to signify months, others use years as the measure of time to maturity.

The generalized variant which allows the parameters to vary in time (a dynamic variant of the original model) is presented by Diebold as

$$y_{it}(\tau) = l_{it} + s_{it}\left(\frac{1-e^{-\lambda_{it}\tau}}{\lambda_{it}\tau}\right) + c_{it}\left(\frac{1-e^{-\lambda_{it}\tau}}{\lambda_{it}\tau} - e^{-\lambda_{it}\tau}\right) + \nu_{it}(\tau)$$

and the parameters are interpreted as latent factors of the curve – its level, slope and curvature.

5.2.1. Interval Prediction in the Nelson-Siegel Model

While this thesis is focused on point and direction prediction of Nelson-Siegel model parameters and subsequently the yield curve of the zero-coupon bond, it is important to also mention interval prediction. Interval prediction is a tool which allows to estimate the error of the point prediction and therefore predict the preciseness of the point estimation. The usual approach to interval prediction is estimating parameters of the distribution of the error term in the specified model.

Given an assumption of a normally distributed error term, it is possible to estimate its parameters – the mean μ and the variance σ^2 in the distribution specification $N(\mu, \sigma^2)$. These parameters are often unknown and therefore a sample mean and variance can be used for their estimation. The sample mean can be calculated as all the errors divided by their count

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

and the sample variance as

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

where $\bar{X}_n$ is the sample error mean.

DNS – Realized Volatility Model

The Dynamic Nelson Siegel – Realized Volatility model extends the stochastic model with an addition of a measurement equation for inference of the measurement error with respect to realized volatility of the underlying data. It augments the Nelson-Siegel model of the yield curve with realized volatility of each security (a bond, for example) and explores its relationship to the estimation errors.

The model proposed by Shin and Zhong (2017) uses high frequency data to estimate the realized volatility of the Dynamic Nelson-Siegel predictions with respect to each security and uses it to estimate the error interval.

5.3. Methodology

In order to estimate the Nelson-Siegel parameters of the yield curve, several transformations of the underlying data are necessary. As the main data set of the thesis is an export of prices for four U.S. government bond futures in tick-frequency certified by the Chicago Board of Trade (CBOT), the following procedure has been constructed to estimate the Nelson-Siegel yield curve parameters and their realized variance.

1. Resample each tick-by-tick series to 5-minute samples for the selected trading regime (Open Outcry) and fill the missing observations with forward fill (assuming a price in a given moment is equal to the last observed price).
2. Merge the series in one dataset with a common index.
3. For each closing price in the 5-minute sampled dataset, generate yield-to-maturity according to a formula from the U.S. Treasury Futures Conversion Factor Look-Up Tables by CME Group (2011).
4. Divide data in equal sized parts and estimate Nelson-Siegel yield curve in parallel from yield-to-maturity for each 5-minute sampled observation.
5. For each of the Nelson-Siegel β factor create a new first-differenced time series and multiply it with itself (each difference to the power of 2) and call it 'squared

returns'. The λ variable is not included, as it lacks stochastic properties in the Dynamic Nelson-Siegel model by Diebold et al. (2008).

6. Resample the data into daily observed frequency (with the value of the day being the last value before market close – 14:00 ET for the Open Outcry regime). For each Nelson-Siegel factor, the last value is selected. As for the 'squared returns' a sum is taken over each day, creating a realized variance variable for each of the latent factors.
7. Resample the data into monthly, quarterly and annually observed frequencies. For each of the Nelson-Siegel model factors a last value is selected. As for the daily realized variances, a mean is calculated over the respective period, creating a mean daily realized variance, which is independent of the number of days in each period.

The data used for forecasting the Nelson-Siegel yield curve is described in detail in previous chapters. This chapter describes the steps starting with the yield-to-maturity estimation and leading to the yield curve estimation itself.

5.3.1. Yield To Maturity

Yield to maturity is the annualized yield of a bond. For the yield curve estimation, it is necessary to first convert the price of the future to a particular bond price.

$$\text{Invoice amount} = \text{Futures Settlement Price}$$

As the available conversion factors account for every possible coupon, zero-coupon bond will be assumed, to ease the calculation of the yield to maturity and to simplify the interpretation of the future contract in terms of government bonds. While different types of bonds could be taked into account in the data, it would be a pure speculation on the nature of each contract, as detailed data on each transaction is not available. Furthermore, it would not change the calculated yield to maturity, as the conversion factors account for the heterogeneity of contracts in treasury futures. Therefore homogeneous contracts composed of only zero-coupon bonds are assumed.

Once the price of the theoretical bond behind the future is calculated, yield to maturity can be calculated using the price, face value of the bond and years to maturity, with the following formula

	5 min	daily	monthly	quarterly	annual
count	515808	6320	300	100	25
mean	0.0326	0.0322	0.0254	0.0240	0.0169
std	0.0154	0.0153	0.0130	0.0125	0.0024
min	0.0097	0.0097	0.0098	0.0102	0.0102
max	0.0706	0.0704	0.0594	0.0536	0.0203

Table 5.1.: 2-year Treasury Note Yield (Open Outcry)

$$\text{Yield} = \sqrt[\text{Years}]{\frac{\text{Face value}}{\text{Price}}} - 1$$

2-year Treasury Note Yield

Yield to maturity of the 2-year treasury note is calculated from price of the 2-year treasury note futures using the CME conversion factor lookup tables. The 6% conversion factor of a zero-coupon bond with maturity of 2 years and 0 months is 0.8885, the accrued interest is 0 and the contract multiplier is \$2000, therefore the 2-year treasury note price is (in U.S. dollars)

$$\text{Price}_2 = [(\text{Close}_{TU} \times 0.8885) + 0] \times 2000$$

As the face value of the 2 year treasury note is \$200,000, its yield to maturity is

$$\text{Yield}_2 = \sqrt[2]{\frac{200000}{\text{Price}}} - 1$$

and the resulting yield is described in Figure 5.1.

5-year Treasury Note Yield

Yield to maturity of the 5-year treasury note is calculated from price of the 5-year treasury note futures using the CME conversion factor lookup tables. The 6% conversion factor of a zero-coupon bond with maturity of 5 years and 0 months is 0.7441, the accrued interest is 0 and the contract multiplier is \$1000, therefore the 5-year treasury note price is (in U.S. dollars)

	5 min	daily	monthly	quarterly	annual
count	515808	6320	300	100	25
mean	0.0395	0.0391	0.0333	0.0329	0.0252
std	0.0128	0.0128	0.0112	0.0113	0.0030
min	0.0146	0.0147	0.0150	0.0152	0.0160
max	0.0685	0.0684	0.0613	0.0613	0.0330

Table 5.2.: 5-year Treasury Note Yield (Open Outcry)

$$\text{Price}_5 = [(\text{Close}_{FV} \times 0.7441) + 0] \times 1000$$

As the face value of the 5 year treasury note is \$100,000, its yield to maturity is

$$\text{Yield}_5 = \sqrt[5]{\frac{100000}{\text{Price}}} - 1$$

and the resulting yield is described in Figure 5.2.

10-year Treasury Note Yield

Yield to maturity of the 10-year treasury note is calculated from price of the 10-year Treasury Note Futures using the CME conversion factor lookup tables. The 6% conversion factor of a zero-coupon bond with maturity of 10 years and 0 months is 0.5537, the accrued interest is 0 and the contract multiplier is \$1000, therefore the 10-year treasury note price is (in U.S. dollars)

$$\text{Price}_{10} = [(\text{Close}_{TY} \times 0.5537) + 0] \times 1000$$

As the face value of the 10 year treasury note is \$100,000, its yield to maturity is

$$\text{Yield}_{10} = \sqrt[10]{\frac{100000}{\text{Price}}} - 1$$

and the resulting yield is described in Figure .5.3

	5 min	daily	monthly	quarterly	annual
count	515808	6320	300	100	25
mean	0.0478	0.0476	0.0433	0.0423	0.0369
std	0.0085	0.0085	0.0081	0.0078	0.0024
min	0.0293	0.0293	0.0298	0.0307	0.0316
max	0.0678	0.0675	0.0666	0.0636	0.0460

Table 5.3.: 10-year Treasury Note Yield (Open Outcry)

	5 min	daily	monthly	quarterly	annual
count	515808	6320	300	100	25
mean	0.0529	0.0527	0.0479	0.0467	0.0400
std	0.0066	0.0067	0.0080	0.0079	0.0036
min	0.0341	0.0344	0.0350	0.0350	0.0384
max	0.0671	0.0670	0.0650	0.0644	0.0535

Table 5.4.: 20-year Treasury Bond Yield (Open Outcry)

20-year Treasury Bond Yield

Yield to maturity of the 20-year treasury bond is calculated from price of the 20-year Treasury Bond Futures using the CME conversion factor lookup tables. The 6% conversion factor of a zero-coupon bond with maturity of 20 years and 0 months is 0.3066, the accrued interest is 0 and the contract multiplier is $1000, therefore the 20-year treasury bond price is (in U.S. dollars)

$$\text{Price}_{20} = [(\text{Close}_{US} \times 0.3066) + 0] \times 1000$$

As the face value of the 20 year treasury bond is $100,000, its yield to maturity is

$$\text{Yield}_{20} = \sqrt[20]{\frac{100000}{\text{Price}}} - 1$$

and the resulting yield is described in Figure 5.4.

5.4. Nelson-Siegel parameters estimation

The Nelson-Siegel model was chosen as the estimation curve of the yield curve for its wide use, especially in modelling the yield curve of U.S. Treasury bonds.

The dynamic variant of the Nelson-Siegel by Diebold differs from the original model in the addition of the transition equation which describes the dynamics of Nelson-Siegel parameters in time. For the initial estimation of Nelson-Siegel parameters, the transition equation can be omitted and the parameters are estimated with the sole use of the measurement equation. This is effectively the first step of the Two-Step DNS estimation, proposed by Diebold et al. (2008), which divides the approach into the measurement/estimation step and the forecast step.

In this particular case, the measurement equations are as follows

$$\text{Yield}_2 = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix}' \begin{bmatrix} 1 \\ \left(\frac{1-e^{-2\lambda}}{2\lambda}\right) \\ \left(\frac{1-e^{-2\lambda}}{2\lambda} - e^{-2\lambda}\right)\end{bmatrix}$$

$$\text{Yield}_{10} = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix}' \begin{bmatrix} 1 \\ \left(\frac{1-e^{-10\lambda}}{10\lambda}\right) \\ \left(\frac{1-e^{-10\lambda}}{10\lambda} - e^{-10\lambda}\right)\end{bmatrix}$$

$$\text{Yield}_{20} = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix}' \begin{bmatrix} 1 \\ \left(\frac{1-e^{-20\lambda}}{20\lambda}\right) \\ \left(\frac{1-e^{-20\lambda}}{20\lambda} - e^{-20\lambda}\right)\end{bmatrix}$$

Which yields three equations for four unknown parameters. The λ parameter can be fixed, but since it can be interpreted as the location of the “hump” on the yield curve and the β_2 factor is interpretable as the curvature of the curve, the following optimization problem is solved instead:

λ takes discrete values for each 6 months of a maturity ($\lambda = 1, 1.5, 2, \ldots 19.5, 20$)

$$\lambda = \underset{\lambda \in [2,20]}{\operatorname{argmax}} \left(\frac{1-e^{-\tau\lambda}}{\tau\lambda} - e^{-\tau\lambda}\right)$$

for each maturity τ to find the optimal placement of the curvature in the yield curve. This yield a three-equation problem to estimate three parameters.

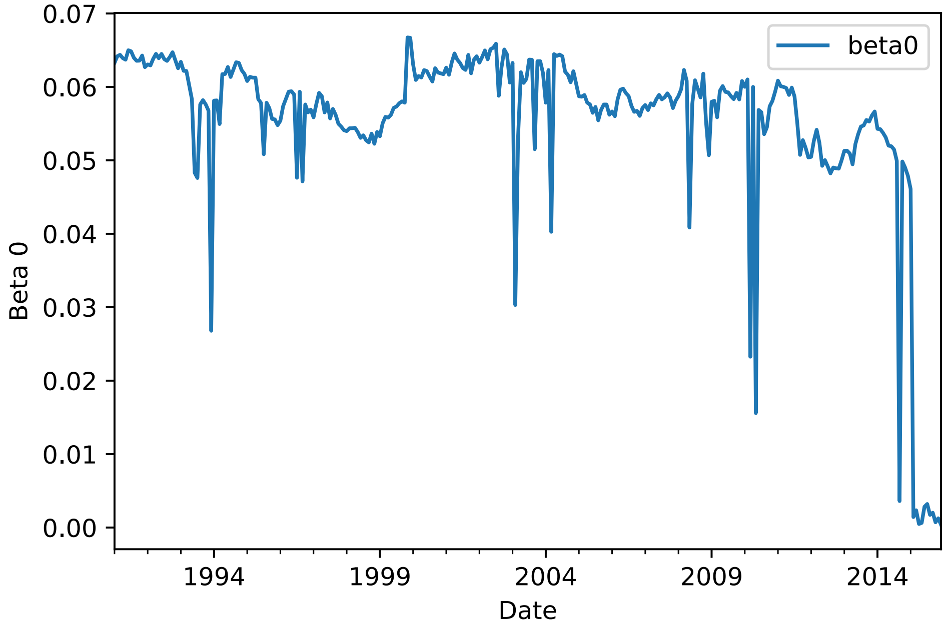

Figure 5.2.: β_0 parameter

The β_0, β_1 and β_2 parameters are then estimated using an ordinary least squares (OLS) model in which the first Nelson-Siegel expression serves as the constant term for the β_0 estimation. The final estimates (from the optimal models in each maturity) are decided by choosing the model with the lowest sum of squared residuals (SSR) which translates to the best fit of all three points on the curve.

This procedure is iterated on all 515 808 observations of the 5-minute frequency for the Open Outcry regime. While this procedure does not smooth the dynamic of the Nelson-Siegel curve in time, it is as close to observed yields as the model allows and the following parameters in 5-minute intervals are obtained.

β_0 – parameter

can be interpreted as the level of the yield curve. The level of the government bonds yield curve estimated by the Nelson-Siegel model can be seen in Figure 5.2. It is interesting to note the very end of the data, where the level sharply declines, while the curvature of the curve, β_2, shown figure 5.4 matches the decline with its sharp rise. It is as apparent on the 3D model of the curve in time in figure 3.1, but it can be traced to a sudden rise of the price of the Treasury Bond at the time, visible in 2.1.

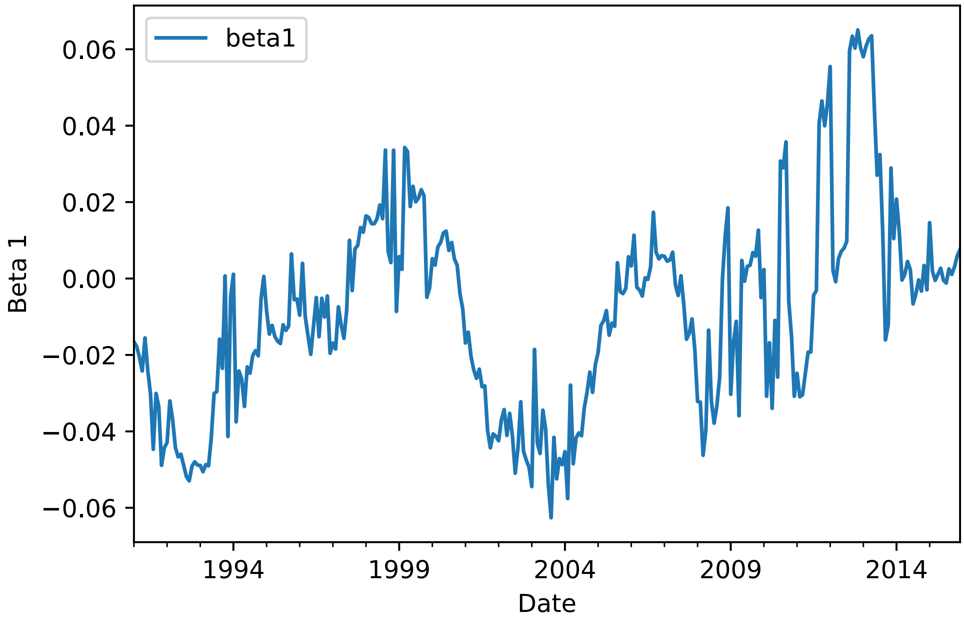

Figure 5.3.: β_1 parameter

β_1 – parameter

can be interpreted as the slope of the yield curve. According to Ang et al. (2006), it is connected to GDP growth and can be therefore used in forecasts. Wright (2006) shows the ability of the slope parameter to predict recessions – in Figure 5.3, the relation of negative β_1 (and therefore the downward sloping yield curve) can be seen right before the 2000 and 2008 recessions, but the decrease below zero in 2013 does not seem to support the hypothesis.

β_2 – parameter

can be interpreted as the curvature of the yield curve. The hump-shaped curve and inverted hump-shaped curves are affected by this parameter and it should therefore serve as a signal for the momentum of the economy according to Vayanos and Vila (2009). It is visible in Figure 5.4.

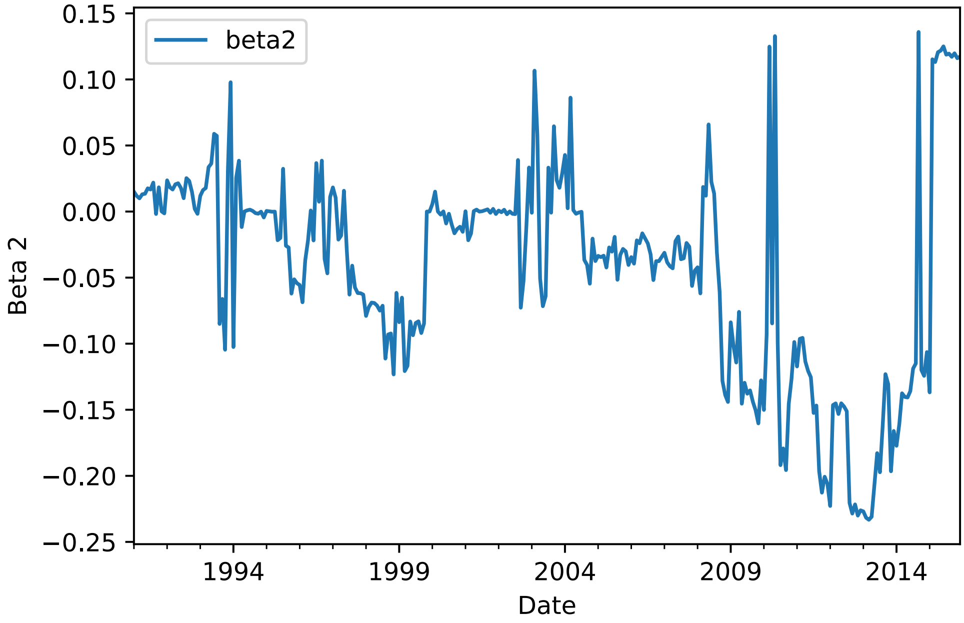

Figure 5.4.: β_2 parameter

λ -parameter

can be interpreted as the position of the “hump” on the yield curve and is shown in Figure 5.5.

5.4.1. Realized Variance of Nelson-Siegel Parameters

To capture the volatility in the high frequency data, daily realized volatility is often calculated as a representation of the changes in the data. While this applies to stock and other market prices, to capture the changes in the level, slope and curvature of the yield curve during the day, I choose the realized variance, or the sum of squared returns as an estimate of the realized volatility.

$$RV_{\beta}^{D} = \sum_{t \in D} (\beta_t - \beta_{t-1})^2$$

The resulting data set is described in Figure 5.5 and seen in Figure 5.6. From the graphical representation of the realized variance, we can clearly see response to recessions on the parameters.

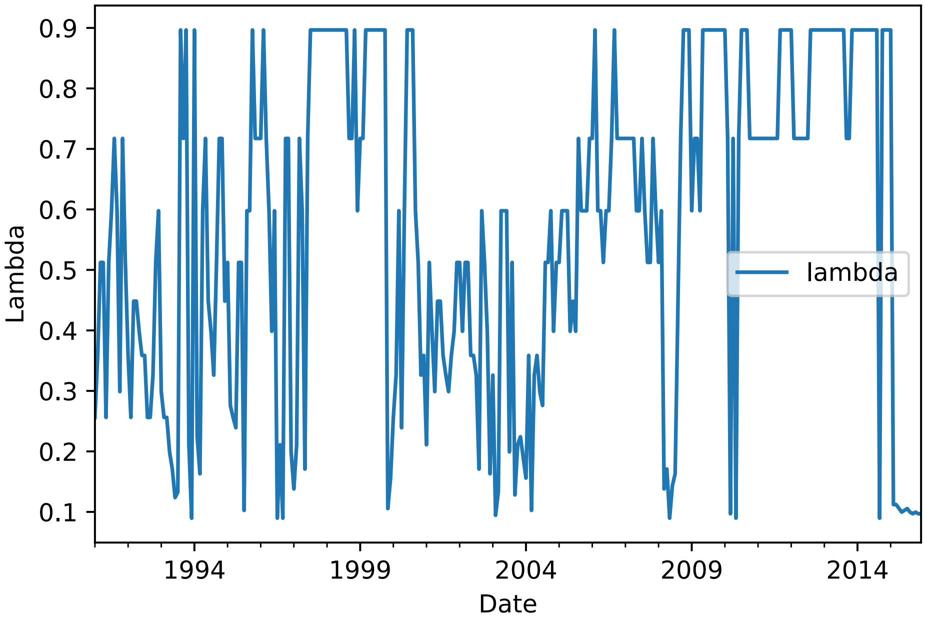

Figure 5.5.: λ factor loading

	β_0^{RV}	β_1^{RV}	β_2^{RV}
count	6326	6326	6326
mean	3.558852e-04	9.744184e-04	1.617242e-02
std	2.234680e-03	7.911060e-03	6.939006e-02
min	8.466261e-10	2.110415e-09	3.689672e-09
max	4.950641e-02	2.397779e-01	1.387199e+00

Table 5.5.: Realized Variance of Nelson-Siegel yield curve parameters

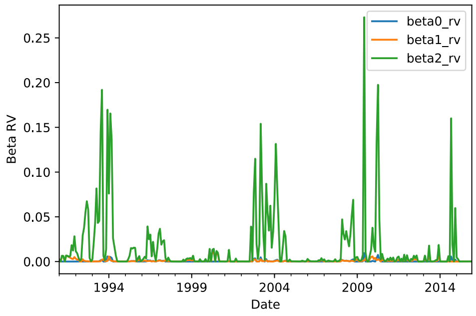

Figure 5.6.: Realized Variance of Nelson-Siegel yield curve

Chapter 6

Forecasting Methodology

The different methods of forecasting the yield curve of government bonds will be summarized in this chapter, together with some competing methods, such as different types of neural networks. The last part of the chapture is dedicated to technical details of yield curve estimation and forecast from high frequency data, as the advances in computer science are a factor in an increasing use of both high frequency data and machine learning methods.

To evaluate the effect of including high frequency data features (daily realized variance of each latent factor of the Nelson-Siegel yield curve) in yield curve forecasting, it is important to choose a baseline to compare the findings with the proposed methods.

1. The benchmark approach to forecasting the yield curve is the Dynamic Nelson Siegel model, where each of the latent factors of the Nelson-Siegel equation is modeled as an autoregressive process of order 1.
2. This approach was subsequently expanded by Diebold et al. (2006) with an introduction of a vector autoregressive model to model all three latent factors once stationarity of the vector autoregressive process is ensured (Koopman et al. 2010) and this model serves as the second baseline for the forecast. For reference, unit root tests for stationarity of the underlying data can be found in the Appendix.
3. The first proposed model is an extension to the work by Diebold et al. (2006), with the each realized variance of each latent factor being used as a predictor in the vector autoregressive process.
4. Finally a Long Short-Term Memory (LSTM) neural network is constructed as the second proposed model, with the same set of variables as the vector autoregressive model proposed above, but only one response variable respective to

each latent factor. In practice, it results in three deep neural network models, one for each latent factor.

6.1. Autoregressive Model

$$x_t = \sum_{i=1}^{N} a_i x_{t-i} + \varepsilon_t$$

Selection of the right order of the autoregressive model is traditionally approach by studying the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the underlying time series.

Dynamic Nelson Siegel model features the following stochastic process, which serves as the baseline for the prediction.

6.1.1. AR(1)

$$\beta_{0,t} = c + \varphi \beta_{0,t-1} + \varepsilon_t$$

$$\beta_{1,t} = c + \varphi \beta_{1,t-1} + \varepsilon_t$$

$$\beta_{2,t} = c + \varphi \beta_{2,t-1} + \varepsilon_t$$

6.2. Vector Autoregressive Model

VAR or Vector Autoregression is an extension of autoregressive models to n-dimensional space. Created in 1980 by Sims, it has found its way in macroeconomic analysis, as it captures complicated dynamics between variables and can be used to visualize an impact of a shock to one variable to the rest. Each variable is modeled using its lags and lags of all the remaining $n-1$ variables in the model. Vector autoregression models differ not only in their dimensions, but also in their order – a VAR(p) model is of order p includes p lags of each of n variables.

A VAR(1) model with three variables can be written in matrix form as

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

where $y_1 \dots y_3$ are the modeled variables, $c_1 \dots c_3$ are the constant terms, $\varepsilon_1 \dots \varepsilon_3$ are the error terms and $A_{1,1} \dots A_{3,3}$ are the autoregression coefficients.

The use of vector autoregression can be divided in two parts. The first is the descriptive stage, where impulse responses and variance decomposition are used to describe co-movements in the data and even causal relationships. This is widely used to model effects of macroeconomic policy, as it can visualize the effect of shocks to one or more variables and the persistence of the shock throughout the time. It must be noted that one of the limitation of this approach is a case of high persistence in the underlying data in which case the results given by impulse responses may be misleading (Stock and Watson 2001). This can be partly solved by using bootstrap methods in the inference (Kilian 1999). The second stage is forecasting in which the methodology and interpretation of the results mirrors the single-variate autoregression approach. As with the choice of lags in the AR models, forecasting using VAR models must take into account the number of variables, as the number of unknown parameters increases in a non-linear way with an addition of each variable (Stock and Watson 2001).

6.2.1. VAR(1)

For the Vector Autoregression based forecast, there are two models. The first one with only β parameters comes from the Dynamic Nelson-Siegel model and will therefore serve as a benchmark.

$$\begin{bmatrix} \beta_{0,t} \\ \beta_{1,t} \\ \beta_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix} \begin{bmatrix} \beta_{0,t-1} \\ \beta_{1,t-1} \\ \beta_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}$$

6.2.2. VAR(1) – RV

The second model includes realized variance of each of the predicted latent factors, augmenting the standard Dynamic Nelson-Siegel model with volatility information. To focus solely on the added value of including the volatility information in the vector autoregressive model over the benchmark, the augmented model retains order of the original model. Autocorrelation functions show high integration in the data even after first differencing which can have implication for significance of autoregressive parameters as described by Toda and Yamamoto (1995).

$$\begin{bmatrix}\beta_{0,t}\\ \beta_{1,t}\\ \beta_{2,t}\\ \beta^{RV}_{0,t}\\ \beta^{RV}_{1,t}\\ \beta^{RV}_{2,t}\end{bmatrix}=\begin{bmatrix}c_1\\ c_2\\ c_3\\ c_4\\ c_5\\ c_6\end{bmatrix}+\begin{bmatrix}A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} & A_{1,5} & A_{1,6}\\ A_{2,1} & A_{2,2} & A_{2,3} & A_{2,4} & A_{2,5} & A_{2,6}\\ A_{3,1} & A_{3,2} & A_{3,3} & A_{3,4} & A_{3,5} & A_{3,6}\\ A_{4,1} & A_{4,2} & A_{4,3} & A_{4,4} & A_{4,5} & A_{4,6}\\ A_{5,1} & A_{5,2} & A_{5,3} & A_{5,4} & A_{5,5} & A_{5,6}\\ A_{6,1} & A_{6,2} & A_{6,3} & A_{6,4} & A_{6,5} & A_{6,6}\end{bmatrix}\begin{bmatrix}\beta_{0,t-1}\\ \beta_{1,t-1}\\ \beta_{2,t-1}\\ \beta^{RV}_{0,t-1}\\ \beta^{RV}_{1,t-1}\\ \beta^{RV}_{2,t-1}\end{bmatrix}+\begin{bmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\\ \varepsilon_{3,t}\\ \varepsilon_{4,t}\\ \varepsilon_{5,t}\\ \varepsilon_{6,t}\end{bmatrix}$$

6.3. Deep Neural Network

Based on the review of the current machine learning methods used for time series prediction in Chapter 4, the long short-memory neural network is selected as an alternative to the aforementioned autoregressive methods. Their architecture enabled inclusion of exogenous variables and can be therefore used to incorporate volatility information as a predictor.

6.3.1. LSTM Architecture for Nelson-Siegel Parameters

The long short-memory neural network is often used for classification of time sequences (in problems such as speech recognition or audio categorization), yet they can be used to predict time series in a similar manner to autoregressive models if the right architecture is selected. Selecting the right model of a deep neural network is not a straightforward process, as there are many more parameters than in autoregressive models, where the choice of variables, their transformation and the lag are usually the only parameters to decide for their estimation. Deep neural networks are of modular nature where the choice of batch size, activation function, number and size of layers, dropout, optimization algorithms may have serious implications over the accuracy of the forecast. There is no precise formula for the choice of these hyperparameters, but several strategies can be taken to find the optimal set, as described by Larochelle et al. (2009). The most important parameters for the Nelson-Siegel parameter prediction are listed below.

Batch size

In the LSTM architecture, it is important to choose the size of a batch of data on which the model is trained – for the direction of change forecast on daily, monthly, quarterly and annual data. For the batch size setting, Keskar et al. (2016) estimates that the usual choice ranges between 32 and 512 data points and warns that larger batch size may lead to significant degradation of the model performance. The reason for this type of behavior is that the model evaluates smaller number of very specific batches of data, which results in an inability to generalize and perform on out-of-sample data. Forecasting using daily, monthly, quarterly and annual data naturally suffers from a smaller number of observations and therefore the LSTM model in this forecast operates on batches of 32 observations – on the lower end of the spectrum specified by Keskar et al. (2016).

Number of epochs

Another important parameter of a neural network is the number of training epochs, a number specifying how many times the whole training sample will go through the model. The optimal number of epochs necessary to train the model depend on a speed of convergence to the result. If the model reaches equilibrium, it iterates on the data until the desired number of epochs is found. The choice of the number of epochs depends on the computing power as well, so it is often a trade-off between speed and accuracy. In order to balance the speed and the accuracy of the results, after evaluation of the convergence of the neural network prediction, 100 training epochs were selected for this network. The number of epochs does not depend only on the computational power, but also on the size of the data set. Early stopping at the 100th epoch is therefore used as a regularization technique to prevent overfitting. An example of the convergence can be found in the Appendix.

Hidden layers

The word ’deep’ in the deep learning discipline signifies how deep the system, in other words how many layers form the network. There is usually one input layer and one output/dense layer and the layers in between are not visible in classical prediction problems, hence they are called ’hidden’. Their choice depends on how complex the problem is – if the problem is a classification of a large picture, it is necessary to have more complex architecture with more layers and nodes in each layer, so the system can capture abstract relationships between different parts on the picture. For stochastic neural networks, such as LSTM, the depth and size of the network determines how complex intertemporal relationships can be captured by the network, but it does not mean that deeper and larger network will perform better, as there is a threat of overfitting. Overfitting can happen is cases where the network is so complex that it has enough memory to capture the whole training set and then performs worse on the testing set, as it fails to generalize.

There are techniques to reduce overfitting, such as dropout, therefore a complex system of three hidden layers with 200 nodes each was selected upon further testing on the Nelson-Siegel parameters.

Dropout

Dropout is a regularization technique used in neural networks in order to simplify the network and avoid overfitting. The idea behind dropout is randomly removing neurons in each layer of the neural network with given probability. This reduces the overall number of neurons in each layer and therefore simplifies complexity of the network. As different layers of a network have a different meaning, dropout parameter can vary between layers. For input layers, using dropout can result in an unnecessary loss of input data due to their random removal, therefore its use is not recommended.

Using dropout for regularization is used in areas such as computer vision where the abundance of data leads to complicated neural networks which yield high variance of results.

According to Zaremba et al. (2014), regular dropout does not work well with recurrent neural networks and long short memory neural networks. Despite the criticisms of dropout use in recurrent neural networks, using recurrent dropout proved to be beneficial in Nelson-Siegel parameters estimation (as tested on the testing set), therefore a dropout of 10% is applied in each hidden layer.

Activation function

In neural networks, the activation function is a function assigned to each node of a layer, which decides on the output of the particular node given its input. It can be either binary, in which case the node either returns an output or not, or it can be a filter of the output, which amplifies or reduces the output based on its inputs. The input of the activation function may vary between the node output itself (it lets it through only if it reaches a particular value of the activation function), or it can be a combination of different events, such as the recent activity of the node itself.

As described by Bishop (2006), commonly used activation functions include

- linear function
- exponential function
- logistic function
- softmax activation function – a normalization of the exponential function

and others.

The choice of the appropriate activation function depends on the problem being solved. While softmax is used in problems with binary output, the linear function is more appropriate for continuous output, such as time series forecast in this thesis.

Loss function

Loss function is an integral part of a neural network, as it represents the price of inaccuracy of its predictions. It can be generally specified as

$$V : \mathbb{R} \times Y \to [0, +\infty)$$

It is convex and therefore a subject to optimization of the output of a neural network. The choice of loss function generally depends on the nature of the problem the neural network is solving.

As described by Rosasco et al. (2004), the commonly used loss functions in regression tasks are

- the square loss function, $V(\hat{y}, y) = (\hat{y} - y)^2$
- the absolute value loss function, $V(\hat{y}, y) = \left|\hat{y} - y\right|$
- the ε-insensitive loss function, $V(\hat{y}, y) =:= \left|\hat{y} - y\right|_{\varepsilon} = \max\left\{\left|\hat{y} - y\right| - \varepsilon, 0\right\}$

where y is the real value of the response variable, and $\hat{y} = f(\vec{x})$ is its estimate based on explanatory variables.

And for classification problems, the commonly used loss functions are

- the square loss function, $V(\hat{y}, y) = (\hat{y} - y)^2 = (1 - \hat{y}y)^2$
- the hinge loss function, $V(\hat{y}, y) =:= \left|\hat{y} - y\right|_{+} = \max\{1 - \hat{y}y, 0\}$

- the logistic loss function, $V(\hat{y}, y) = \frac{\ln\left(1+e^{-\hat{y}y}\right)}{\ln(2)}$
- the cross-entropy loss function, $V(\hat{y}, \tilde{y}) = -\tilde{y}\ln(\hat{y}) - (1-\tilde{y})\ln(1-\hat{y})$, where $\tilde{y} = \frac{1+y}{2}$ and $\tilde{y} \in (0,1)$

In the regression problem of forecasting the Nelson-Siegel parameters, the mean squared error was selected as the loss criteria.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Functional model vs Sequential model

In deep learning, it is important not only which types of layers are chosen, or how many are stacked on each other, but also in which way they are stacked. In the sequential architecture of a neural network, the system has one input layer at the beginning and one output layer at the end. Input data are inserted in the input layer and the performance of the network is weighted according to its accuracy in matching predictions on the output layer. In some applications, it can be beneficial to have multiple output layers in the network and assign weight to each output, to determine how significant each output prediction is in the training of the neural network. The added value of including intermediate output in a neural network can be described as a solution to a problem of predicting one time series using another. With the response variable present on the main output of the neural network and the secondary time series on the intermediate output, the system can focus on the main task of the response variable prediction, while being able to learn an appropriate representation of the secondary time series through running backpropagation from the intermediate output.

It is important to note that choosing the functional model results in more hyperparameters to tune, therefore sequential model could be more appropriate to reduce the complexity of the system and it is therefore selected as the architecture for the Nelson-Siegel parameter forecast.

Variable choice and transformation

The selection of a more complex system allows for capturing more complex relationships between stochastic variables, therefore all β_0, β_1, β_2 and their realized variances were selected as input variables for the network. To improve the accuracy over different parameters, each β parameter was estimated separately, resulting in a network with single node with the response variable on the output layer and the rest of the lagged variables on the input layer.

To improve the forecasting accuracy and to be able to better generalize over samples, the data was differenced and scaled on interval $(-1,1)$ and later transformed back into the original state to compare forecasts with other methods.

Initialization

There are various techniques to initialize neural network weights and the right choice can speed up the process of learning. In this particular problem, a random initialization is selected and through the epochs, the system converges to optima. A drawback of this approach is problematic reproducibility of the results, leading to results difficult to verify by other researchers. To account for this problem, a fixed state for the random process is selected by setting the seed of the process to the number ’2017’, therefore allowing reproducibility by others.

6.4. Model Evaluation

To compare the performance of proposed models with the models chosen as benchmarks, the root-mean-squared-error (RMSE) metric has been chosen, as it is scale independent (Hyndman and Koehler 2006) and can be therefore used across the forecasted variables. The comparison of error scores is an important rule of thumb on which model fits the testing data more closely. To compare the in-sample and out-of-sample forecasts, the data have been divided in a 80%-20% split, as recommended by Haykin (2004), where $4/5$ of data belong to the training sample and the last $1/5$ of data belong to the testing sample. Data from the 1991-2010 period belong in the training sample and data from the 2011-2015 period in the testing sample. While the train-test split is the same for all models, it is important to note the occurrence of the Dot-Com bubble and the financial crisis of 2007-2008 in the train sample, while no crisis of the same proportions is included in the testing sample. Their inclusion in the training sample may affect performance of some models, but as a compact window of observations between 1991-2015 is used, a possible narrowing of training and testing windows (in order to avoid crises in the data) is not viable due to a resulting selection bias.

6.5. Technical Aspects of High Frequency Time Series Analysis

To analyze data sampled in high frequencies in an efficient manner, it is necessary to choose the appropriate software stack and tools to handle large quantities of data. For the analysis of the term structure of government securities futures for years 1991-2015, the sample of each variable ranges between 50 million and 228 million observations and the timestamp of each sample is measured in milliseconds, which can result to up to more than 400 million different observations across the four available variables. Computer memory needed to match each observation across all four available data frames and even trivial re-sampling operating exceeds the memory of standard computers as of the time of writing this theses. The burden of computational overhead has been reflected in the choice of tools for the analysis and data management.

The majority of computations and re-sampling has been possible using the Metacentrum service, which allows to use a distributed network of computers for academic purposes.

Initial import and re-sampling of the data (originally sampled with the variable tick frequency) is done in Python and with the use of its libraries for data analysis. Estimation of the parameters of the Dynamic Nelson-Siegel model has been implemented in Python as well, rewriting the code from the R package **termstrc** into a new library, made available for public use. Analysis and forecast of Dynamic Nelson-Siegel parameters were both done in Python (especially the prediction using Neural Networks) and in R. The particular tools used in the analysis are listed and explained below, to enable reproducibility of the results and to review current capabilities in analysis of high frequency data.

R statistical programming language has been widely used in data science and econometrics for its extensibility. Most of the features are available through modules and that is the case with high frequency data analysis in R. Built in support for time series is provided with the **ts** function of the **stats** module, while externally available **xts** and **zoo** modules provide more advanced time series subsetting and manipulation. Tools for high frequency data can be found in the package **highfrequency** built with support for data aggregation, calculating liquidity measures, forecasting volatility and analyzing intraday periodicities (Cornelissen et al. 2017). While the development of neural network software has been more active on other platforms, neural networks can be developed in R as well. It is possible mainly through the **neuralnet** module by Günther and Fritsch (2010) and through interfaces to MXNet and Keras frameworks. For yield curve modelling, the estimation of the Nelson-Siegel and Nelson-Siegel-Svensson parameters is available in the **YieldCurve** package by Guirreri and Salvino (2013). This package was used in this thesis as an inspiration for the re-implementation of Nelson-Siegel estimates in Python.

Python is an open source programming language in development since 1991. In recent years, it has successfully gained ground in statistical and data analysis, disciplines historically dominated by the aforementioned R language, but also Matlab, Stata or recently Julia. The gap of statistical packages for Python has been closing with the availability of powerful mathematical libraries, such as NumPy or SciPy and also due to the rising popularity of neural networks. Neural network library interfaces for Python – Theano, Tensorflow, Caffe, MXNet or Keras – have been supported by companies such as Google, Facebook, Microsoft, Amazon or Intel which accelerated development of other statistical libraries. Python (in versions 3.4 and 3.6) and its data manipulation, statistical and neural network libraries (listed below) we selected as the main environment for most of computations in the thesis for its flexibility and flat learning curve.

NumPy is a standard mathematical library for Python. It is built on a logic of n-dimensional arrays, providing an easy access to vector calculations and transformations of time series. Its use ranges from scalar, vector and matrix operation to supervised problem creation for neural networks. Many other scientific libraries and neural network frameworks maintain interface compatibility with NumPy, allowing for a unified style of mathematical expression and array manipulations across multiple platforms. Besides the capabilities of a mathematical engine, it can be utilized as a n-dimensional data storage. Having been partly implemented in the Fortran programming language, its speed and memory efficiency predetermines its use in high frequency data manipulation, such as yield calculation or Nelson-Siegel parameter inference in this thesis. Having been created in the 1990s, it is currently being used in both the academic world and industries, with use cases in gaming and space exploration (Walt et al. 2011). It is available under the open source BSD license, which allows it to be used with little restrictions.

Pandas is a data structure library for Python, initially developed for quantitative finance applications (McKinney et al. 2010). It expands the capabilities of NumPy arrays by allowing the use of multiple data formats in the same data-set and by using unique labels on data (such as labeled columns). Its out-of-the-box support for time/date indexes and time selection functions makes it the tool of choice for time series storage and manipulation in Python. Time series oriented functions in Pandas are comparable to R libraries **xts** or **ts**, but its native database-like data joining and merging capabilities make it a more powerful tool for time series manipulation. Additional features include native support for data summaries, plotting and database access. Support for the HDF database format (details below) allows to analyze large data-sets efficiently. It is available under the open source BSD license (McKinney 2014).

HDF (or the Hierarchical Data Format) is a data storage technology, designed to store enormous and complex data collections. It is a storage of choice for storing big data, especially extremely large time series, such as stock data. It can be interfaced from a variety of programming and statistical languages, such as Python, R or Matlab. The original version was created in 1987 at the National Center for Supercomputing Applications at the University of Illinois and became widely used after being chosen by NASA as a data storage tool for their projects. The current version (HDF5) was released in 2002 and allows to store multiple data-sets in a single database file with a possibility to compress data using various compression libraries. While there are storage and analysis frameworks more fitting for the purpose of big data analyses, such as Hadoop with its parallel computational capabilities, the hierarchical data format is sufficient for high frequency data for its speed and focus on time series storage and analysis. HDF5 was used as the main data storage for analysis in this thesis, with the storage compressed with the bzip2 compression library.

Tensorflow is an interface to express machine learning algorithms. Its logic is based on the concept of tensors – typed multidimensional arrays. Tensors are directly tied to mathematical concepts of scalars and vectors – scalar can be described as a tensor of zero rank, vector as a first rank tensor and a matrix as a second rank tensor (of Cambridge (2015)). It can be therefore seen as an extension of vectors into n-dimensional space or a description of relationships between vectors.

Keras (Chollet et al. 2015) is a neural network abstraction library for Python, built on an interface which uses different backends for the neural network training and prediction. It currently supports Tensorflow and Theano as the underlying computation libraries and creates a unified interface which results in a one-design multiple deployments approach, where one model can be trained using either library. It is one of the most used neural network libraries for its simplicity, with which complicated deep learning network can be assembled with a few lines of code. Its architecture is based on two approaches to layer stacking – the first, sequential, is a layer stacking design with a single input and single output, with layers in between. The second, functional, approach is more complicated to assemble, but provides the possibility of creating multiple inputs and outputs and even re-using whole neural networks as layers in different models. Initially released in 2015, it is available under the open source MIT license. Keras was used as an abstraction over Tensorflow in the LSTM model architecture for this thesis. The functional model approach was used for the ability to include exogenous variables in the model.

StatsModels (Perktold et al. 2006) is a Python library for statistical models estimation and testing. It features autoregressive models such as AR or VAR.

PyFlux (Taylor 2016) is a Python library for time series modelling and inference. It is built upon the aforementioned libraries such as Numpy, Scipy and Pandas and was created in 2016 by Ross Taylor, a former economist of the HM Treasury of Great Britain. It is available under the Apache 2.0 license and, among others, it implements time series models from ARIMA, VAR and GARCH families. It has been used in the thesis to model Dynamic Nelson-Siegel parameters using ARIMA, ARIMAX (ARIMA with external variables) and VAR models. Its capabilities also include an implementation of the Metropolis-Hastings algorithm, which is a basis of multiple inference steps in the DNS – Realized Volatility model for interval estimation for Dynamic Nelson-Siegel parameters using high frequency data.

Chapter 7

Results and Discussion

Modelling Nelson-Siegel parameters for U.S. government bonds using high frequency data offers possibilities to extract and use intraday information from the tick-sampled data and use it in the prediction models. Additionally, the availability of a vast quantity of data suggests recent techniques in machine learning, such as neural networks can be applied alongside or instead of traditional predicting methods.

In order to allow for broader generalization of the analysis, the results are divided into a theoretical part and an empirical part. The theoretical part studies causal relationships between Nelson-Siegel parameters and their realized variance using Granger causality. The empirical part applies aforementioned techniques on the same data set, to propose a method to incorporate realized variance as a predictor for Nelson-Siegel parameters.

7.1. Causal Relationships

To study the viability of using daily realized variance of Nelson-Siegel parameters, the Granger causality test is used to test for the causal relationship between the realized variance and the parameters. Granger causality test uses the F-test to compare two models. The first is the traditional AR(1) model proposed by Diebold in the Dynamic Nelson-Siegel model and the second one is an AR(1) model, augmented with the lagged realized variance of each parameter.

7.1.1. Granger causality

As the Granger causality test assumes stationarity of the data, both variables (the β parameter and its daily realized variance) are first-differenced

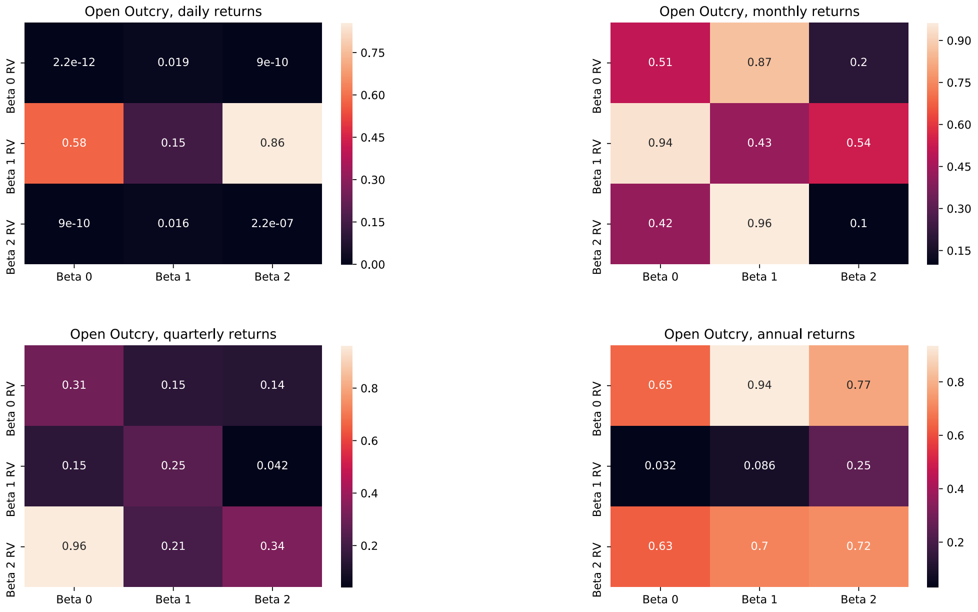

Figure 7.1.: Granger causality for returns

$$\beta_t' = \beta_t - \beta_{t-1}$$

$$\beta_t^{RV\prime} = \beta_t^{RV} - \beta_{t-1}^{RV}$$

and the two models are specified with the lagged variables

$$\beta_t' = c + \varphi\beta_{t-1}' + \varepsilon_t$$

$$\beta_t' = c + \varphi\beta_{t-1}' + \gamma\beta_{t-1}^{RV\prime} + \varepsilon_t$$

Under the null hypothesis of the Granger test, the $\beta^{RV\prime}$ does not Granger-cause β'.

The results in Figure 7.1 are based on the Open Outcry dataset from years 1991-2015 on daily, monthly, quarterly and annual basis and show p-values of the Granger test on the aforementioned models. For example, if the value in the $\left[\beta_0^{RV}, \beta_1\right]$ cell is smaller than a chosen significance level, the null hypothesis of no Granger causality of β_0^{RV} on β_1 can be rejected and concluded that β_0^{RV} could be a viable predictor for β_1. On the other hand, if the p-value is larger, one cannot come to the same conclusion on the chosen significance level. It does not imply no Granger causality, but it simply fails to reject the null hypothesis of the Granger causality test.

In the daily returns data set, the null hypothesis of no Granger causality is strongly rejected at the $\alpha = 0.01$ significance level for the effect of the realized variance of β_0 and β_2 on the β_0 parameter. The null hypothesis of no Granger causality is also rejected for the effect of the realized variance of β_0 and β_2 on the β_2 parameter, suggesting a causal relationship between the volatility of the level and curvature parameters of the Nelson-Siegel yield curve and the parameters themselves.

On the $\alpha = 0.05$ level, the null hypothesis of no Granger causality is rejected for the stochastic effect of the realized variance of the β_0 and β_2 parameters on the β_1 parameter, suggesting a causal relationship between the volatility of the level and curvature of the Nelson-Siegel yield curve and the slope of the curve.

Results of the monthly returns dataset lead to a failure to reject the null hypothesis on any relevant significance level for any causal relationships. The only relationships close to a relevant significance level (of $\alpha = 0.1$) is the causal relationship between the realized variance of the β_2 parameter and the parameter itself. This suggests a longer temporal relationship between the volatility of the curvature of the Nelson-Siegel yield curve and the curvature itself.

In the quarterly returns dataset, the null hypothesis of no Granger causality is rejected on the $\alpha = 0.05$ significance level for the relationship between the realized variance of β_1 and the change in the β_2 parameter. This suggests even longer temporal relationship between the volatility of the slope of the Nelson-Siegel yield curve and the curvature of the curve.

In the annual returns dataset, the null hypothesis of no Granger causality is rejected on the $\alpha = 0.05$ significance level for the relationship between the realized variance of β_1 and the change in theβ_2 parameter, and on the $\alpha = 0.1$ level for the relationship between the same realized variance (of β_1) and the change in the β_1 parameter itself. Even though this suggests a longer temporal significance of the volatility of the slope of the Nelson-Siegel yield curve in prediction of the curve itself, it is important to note the small sample size, as the Granger causality test is applied on a sample of 25 years, with one year being truncated in order to difference the series and account for non-stationarity.

For day-to-day forecasts, using the daily realized variance of the level and curvature of the Nelson-Siegel yield curve in the term structure forecasts could be a viable alternative to the current autoregressive models.

For longer temporal periods, such as quarters and years, the average daily realized variance of the slope of the Nelson-Siegel yield curve could be used as an augmenta-

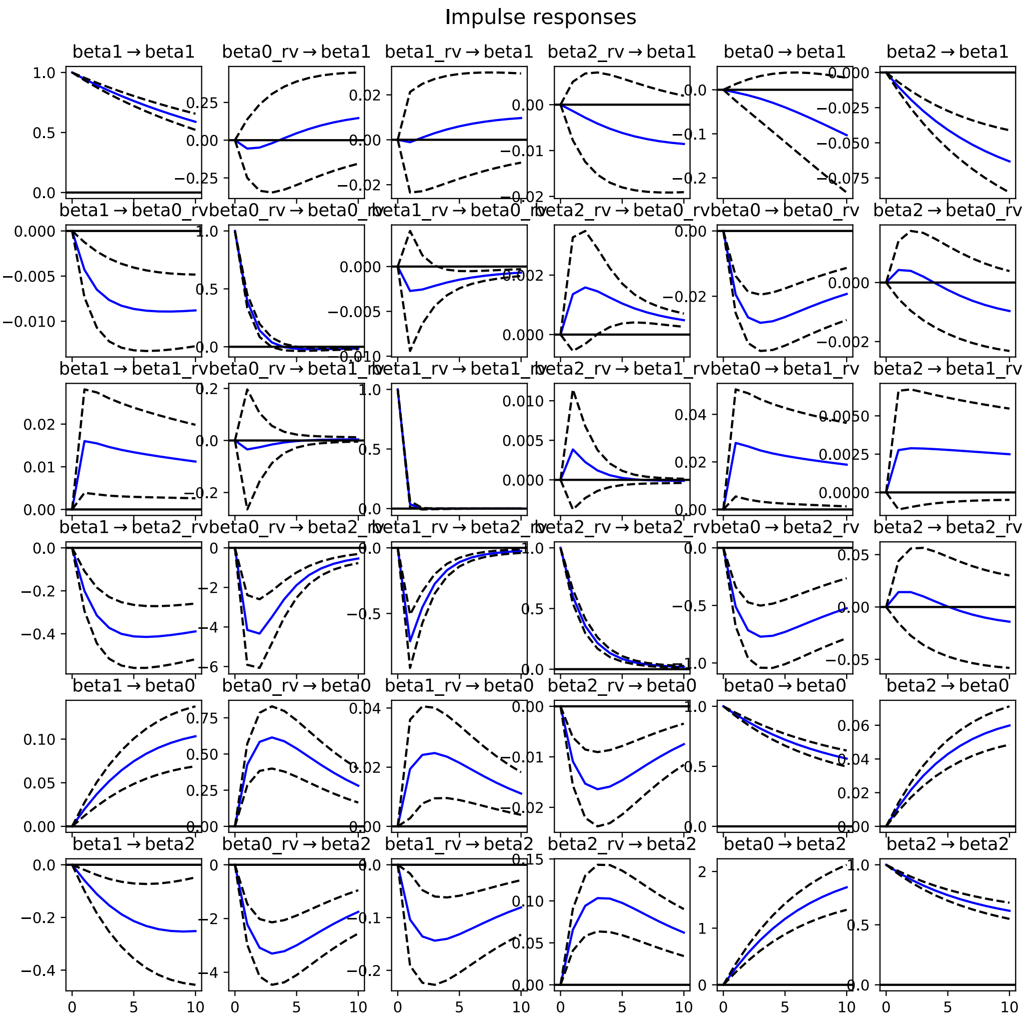

Figure 7.2.: Impulse responses on monthly data

tion of current autoregressive models for the term structure prediction. As the failure to reject the null hypothesis of no Granger causality between the rest of the variables does not imply no Granger causality, their use in the prediction could be evaluated as well.

7.1.2. Impulse Responses

Impulse responses estimated by the augmented VAR model can show not only the causal effect, described by Granger causality, but also the direction of the change and the dynamics of the effect.

Impulse responses of the VAR model variables can be seen on Figure 7.2. This particular model was estimated on monthly data and the rest of the frequencies can be found in the Appendix.

Impulse response functions from the VAR system show that a positive shock to the volatility of the β_0 variable has a short term negative effect on the β_0 factor. When the interpretations of the Nelson-Siegel latent factors are taken into account, it can be seen as an effect of higher volatility of the level of yield over maturities is negative in the short run to the level itself. The volatility feedback theory cannot be applied here without abstraction, as the latent factors are a decomposition of the yield curve, but as the first latent factor is the level of the yield across maturities, it could be seen as an implication of the volatility feedback.

The effect of increased volatility of the first latent factor has positive effect on both the slope and the curvature of the yield curve, with lasting effect over 10 periods. Higher volatility of the level therefore contributes to a rise in slope and curvature of the Nelson-Siegel yield curve.

The effect of increased volatility of the third latent factor has a strong, but diminishing effect on the level of the Nelson-Siegel curve, slightly negative lasting effect on the slope of the curve and stronger lasting effect on the curvature.

Despite the non-significance of the volatility of the β_1 factor in Granger causality test, we can observe a similar effect of a positive shock to volatility of the slope of the curve on the slope itself. The negative effect of higher volatility of a factor on the factor itself is a finding which can prove useful in forecasting the Nelson-Siegel factors.

7.2. Prediction

As an extension to the theoretical findings of the Granger causality test, several models are selected to evaluate a concrete proposal to incorporate the realized variance of Nelson-Siegel parameter as predictors in the parameter prediction.

Four models are estimated on four frequencies of the Open Outcry dataset from years 1991-2015. The basis of the estimated Nelson-Siegel parameters are closing prices of the U.S. government bond futures in the Open Outcry system, therefore the daily data are based on the closing prices at 14:00 Central Time on a working day. Monthly, quarterly and annual data are based on the closing prices of the last day of the month, quarter or year, respectively. Realized variance of the aforementioned parameters is calculated from intraday samples with 5-minute frequency in the opening time of the market (7:20-14:00 Central Time) using formulas described in the Methodology chapter.

The first model is the standard AR(1) model from the Dynamic Nelson-Siegel model by Diebold and acts as a baseline for other forecasts.

The second model is an extension to the AR(1) model, capturing the dynamics of the yield curve by using a VAR(1) model with β_0, β_1 and β_2 variables. It serves as a baseline for the vector-based models, but unlike the AR(1) model from the Dynamic Nelson-Siegel model, it uses first differences to account for the presence of unit root.

The third model is an extension to the VAR(1) model, augmenting the β_0, β_1 and β_2 variables with their daily realized variances. Like the VAR(1) model above, it uses first differences to account for the presence of unit root.

The fourth model is a Long Short-Term Memory (LSTM) model using the β_0, β_1 and β_2 variables and their daily realized variances. Its architecture and hyperparameters are described in the previous chapter.

All four models are estimated/trained on the first 80% of the data and tested on the remaining 20%. The Root-mean-squared-error (RMSE) metric is used to compare the accuracy of the respective models. Finally, a Random Walk model, which uses the previous value of the series as the prediction, was calculated to put the forecast into perspective.

The respective RMSE accuracy metrics are replicated in the Table 7.1 and for each dataset/variable the model with the highest accuracy is highlighted.

The daily and monthly prediction results suggest that LSTM neural networks work better with larger data samples, while they are outperfomed by standard autoregressive methods in the periods with smaller training samples.

LSTM yielded higher prediction accuracy than other methods in 2/3 of daily and monthly data forecast, in both cases for the level and curvature parameters of the Nelson-Siegel yield curve. Surprisingly, it surpassed the other prediction models in annual forecasting accuracy of the level parameter of the Nelson-Siegel yield curve, by a small margin. It is important to note that the sample size for the annual data is too low to be able to generalize the result to longer series and unit root is still present in differenced series, which hurts the interpretation of autoregressive models.

In forecasting the year-to-year change in the curvature parameter of the Nelson-Siegel yield curve and in forecasting the level parametr on month-to-month time series, the baseline AR(1) model from the Dynamic Nelson-Siegel modelling approach

	RW	AR	VAR	VAR-RV	LSTM-RV
β_0^{d}	0.00473	0.00518	0.00452	0.00448	**0.00438**
β_1^{d}	0.00728	**0.00663**	0.00680	0.00680	0.00684
β_2^{d}	0.02658	0.02502	0.02533	0.02519	**0.02487**
β_0^{m}	0.01041	0.01508	0.01026	0.01028	**0.00979**
β_1^{m}	**0.01482**	0.01807	0.01561	0.01555	0.01679
β_2^{m}	0.06106	0.06203	0.05885	0.05862	**0.05491**
β_0^{q}	0.01852	0.02419	**0.01427**	0.01450	0.01808
β_1^{q}	**0.02628**	0.03172	0.02990	0.02927	0.03098
β_2^{q}	0.10415	0.11055	0.08757	**0.08496**	0.11801
β_0^{a}	0.02440	0.03729	0.02598	0.02877	**0.02402**
β_1^{a}	**0.02738**	0.09069	0.03784	0.03605	0.03849
β_2^{a}	**0.11998**	0.33761	0.135311	0.18695	0.17852

Table 7.1.: Open Outcry RMSE

and the rest of the models were surpassed by the non-augmented VAR(1) model. It should be noted that the AR(1) model is based on forecasting the level of the parameter, while the VAR(1) model uses differenced data to account for unit-root presence and therefore the autoregressive model can suffer from the presence of unit-root.

The inclusion of realized variance in the model led to higher precision in 50% of all prediction cases, most remarkably in cases where the neural network failed to provide higher accuracy. 1/3 of all cases was dominated by the random walk which points to the complicated nature of the testing set compared to the training set. While the limited nature of the underlying data should be noted, it can be viewed as a proposal for including intraday features of high frequency data in forecasting the term structure of government bond yields.

7.3. Findings

A subset of realized variance features from high frequency data was shown to have a statistically significant Granger causality relationship with the Nelson-Siegel model parameters and can be therefore useful in their prediction. The impulse response functions estimated on the vector autoregressive model show the effect of shocks to the factors of the yield curve and their dynamics on different horizons.

To allow for comparison with dynamic yield curve modelling approaches, a single LSTM architecture was used in a similar manner to using an autoregressive model of one order for all yield curve parameters in the Dynamic Nelson-Siegel model. Further improvements of the LSTM approach could be possibly attained by creating a personalized set of hyperparameters for each frequency and data set. Choosing a larger training set instead of the 80/20 sample split could also help in improving the accuracy of predictions. The same approach could be applied in the choice of lags for autoregressive models and the neural network, but it could complicate comparisons to the baseline model.

Finally, the comparison of four forecast models show that an inclusion of the realized variance features from high frequency data results in higher accuracy of forecasting the yield curve on the testing dataset. Furthermore, combining the long short-term memory neural network with the aforementioned features leads to higher accuracy on daily and monthly data.

Chapter 8

Conclusion

This thesis focuses on evaluation of yield curve forecasting techniques using high frequency data. While the availability of high frequency data creates new opportunities to forecast intraday changes, the macroeconomic implications of the yield curve create demand to improve forecasts in longer periods. The thesis discusses different approaches to enrich these forecasts by high frequency data and proposes a new method of incorporating daily realized variances of the Nelson-Siegel parameters as proxies for intraday activity in the high frequency domain.

The proposed method improves the prediction of the yield curve itself, which can be useful in forecasting recessions or GDP growth. The abstraction from particular yields allows application independently on the number of bonds, as long as a Nelson-Siegel curve can be estimated.

Furthermore, the thesis proposes a use of multivariate neural networks in government bond yield modelling as an alternative to autoregressive models. With the increasing availability of data, stochastic networks such as Long Short-Term Memory (LSTM) neural network can be used to predict the yield curve parameters on daily, monthly, quarterly and annual basis. Results from estimating the autoregressive model, vector autoregressive model, augmented vector autoregressive model and LSTM with parameters from high frequency data show that including the intraday trading information in the yield curve prediction yields better results in 50% of cases. While using neural networks on smaller samples yields lower forecast accuracy than parametric methods, the 1-day ahead and 1-month ahead forecasts result in higher accuracy with LSTM in 2/3 of the cases.

Further research can be done in utilizing recent machine learning methods, such as attention or sequence-to-sequence architectures, in combination with exploring additional intraday features of high frequency data, including intraday trading structure.

Bibliography

Yen-Ming Chiang, Li-Chiu Chang, and Fi-John Chang. Comparison of static-feedforward and dynamic-feedback neural networks for rainfall–runoff modeling. *Journal of hydrology*, 290(3):297–311, 2004.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 2016.

Martin Martens. Measuring and forecasting S&P 500 index-futures volatility using high-frequency data. *Journal of Futures Markets*, 22(6):497–518, 2002.

Bevan J Blair, Ser-Huang Poon, and Stephen J Taylor. Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. In *Handbook of Quantitative Finance and Risk Management*, pages 1333–1344. Springer, 2010.

Francis X Diebold, Canlin Li, and Vivian Z Yue. Global yield curve dynamics and interactions: a dynamic Nelson–Siegel approach. *Journal of Econometrics*, 146 (2):351–363, 2008.

Yacine Aït-Sahalia, Per A Mykland, and Lan Zhang. Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, 160(1): 160–175, 2011.

Richard Heckinger, David Mengle, Robert Steigerwald, Ivana Ruffini, and Kirstin Wells. Understanding Derivatives: Markets and Infrastructure. *Federal Reserve Bank of Chicago*, 2013.

Nicholas Johnson, John Kerpel, and Jonathan Kronstein. Understanding treasury futures. 2017.

Inc. CME Group. U.S. Treasury Futures and Options. 2013. URL `http://www.cmegroup.com/trading/interest-rates/files/IR-179_USTreasury_FO_Fact_Card.pdf`.

Inc. CME Group. Timeline of CME Achievements. 2017. URL `http://www.cmegroup.com/company/history/timeline-of-achievements.html`.

Charles M Jones, Gautam Kaul, and Marc L Lipson. Information, trading, and volatility. *Journal of Financial Economics*, 36(1):127–154, 1994.

Jean-Marie Dufour, René Garcia, and Abderrahim Taamouti. Measuring high-frequency causality between returns, realized volatility, and implied volatility. *Journal of Financial Econometrics*, 10(1):124–163, 2012.

Jianfeng Yu and Yu Yuan. Investor sentiment and the mean–variance relation. *Journal of Financial Economics*, 100(2):367–381, 2011.

Yaw-Huei Wang, Aneel Keswani, and Stephen J Taylor. The relationships between sentiment, returns and volatility. *International Journal of Forecasting*, 22(1):109–123, 2006.

Hui Guo and Robert Savickas. Understanding stock return predictability. 2006.

Peter F Christoffersen and Francis X Diebold. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287, 2006.

John M Maheu and Thomas H McCurdy. Do high-frequency measures of volatility improve forecasts of return distributions? *Journal of Econometrics*, 160(1):69–76, 2011.

Stanislav Anatolyev and Jozef Barunik. A simple model for forecasting conditional return distributions. *arXiv preprint arXiv:1711.05681*, 2017.

Minchul Shin and Molin Zhong. Does realized volatility help bond yield density prediction? *International Journal of Forecasting*, 33(2):373–389, 2017.

Anat R Admati and Paul Pfleiderer. A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, 1(1):3–40, 1988.

Paul F Cwik. The inverted yield curve and the economic downturn. *New Perspectives on Political Economy*, 1(1):1–37, 2005.

Frank J Fabozzi and Steven V Mann. *The handbook of fixed income securities*. McGraw Hill Professional, 2012.

Dimitri Vayanos and Jean-Luc Vila. A preferred-habitat model of the term structure of interest rates. Technical report, National Bureau of Economic Research, 2009.

John Y Campbell. Some lessons from the yield curve. Technical report, National Bureau of Economic Research, 1995.

Pasquale Della Corte, Lucio Sarno, and Daniel L Thornton. The expectation hypothesis of the term structure of very short-term rates: Statistical tests and economic value. *Journal of Financial Economics*, 89(1):158–174, 2008.

Rajna Gibson, Francois Lhabitant, and Denis Talay. Modeling the term structure of interest rates: A review of the literature. *Foundations and Trends in Finance*, 5 (1-2), 2010.

Arturo Estrella and Frederic S Mishkin. Predicting US recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1):45–61, 1998.

Jonathan H Wright. The yield curve and predicting recessions. 2006.

Arturo Estrella and Frederic S Mishkin. The yield curve as a predictor of US recessions. *Current Issues in Economics and Finance*, 2(7), 1996.

Michael J Dueker. Strengthening the Case for the Yield Curve as a Predictor of US Recessions. *Review*, 79, 1997.

Marcelle Chauvet and Simon Potter. Predicting a recession: evidence from the yield curve in the presence of structural breaks. *Economics Letters*, 77(2):245–253, 2002.

Joseph G Haubrich and Ann M Dombrosky. Predicting real growth using the yield curve. *Economic Review-Federal Reserve Bank of Cleveland*, 32(1):26, 1996.

Andrew Ang, Monika Piazzesi, and Min Wei. What does the yield curve tell us about GDP growth? *Journal of Econometrics*, 131(1):359–403, 2006.

Raffaella Giacomini and Barbara Rossi. How stable is the forecasting performance of the yield curve for output growth? *Oxford Bulletin of Economics and Statistics*, 68(s1):783–795, 2006.

Yu-chin Chen and Kwok Ping Tsang. What does the yield curve tell us about exchange rate predictability? *Review of Economics and Statistics*, 95(1):185–205, 2013.

Jan Marc Berk. The information content of the yield curve for monetary policy: A survey. *De Economist*, 146(2):303–320, 1998.

Torben G Andersen and Tim Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of empirical finance*, 4(2):115–158, 1997a.

Torben G Andersen and Tim Bollerslev. Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns. *The journal of Finance*, 52(3):975–1005, 1997b.

Daniel Klerfors and Terry L Huston. Artificial neural networks. *St. Louis University, St. Louis, Mo*, 1998.

David Lowe and D Broomhead. Multivariable functional interpolation and adaptive networks. *Complex systems*, 2(3):321–355, 1988.

Zhong-Yuan Zhang. Stochastic Neural Network. *Encyclopedia of Systems Biology*, pages 1998–1998, 2013.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH*, pages 3214–3218, 2015.

Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

Daniel S Clouse, C Lee Giles, Bill G Horne, and Garrison W Cottrell. Time-delay neural networks: Representation and induction of finite-state machines. *IEEE Transactions on Neural Networks*, 8(5):1065–1070, 1997.

Jozef Baruník and Barbora Malinska. Forecasting the term structure of crude oil futures prices with neural networks. *Applied Energy*, 164:366–379, 2016.

Eurípedes P dos Santos and Fernando J Von Zuben. Efficient second-order learning algorithms for discrete-time recurrent neural networks. *RECURRENT NEURAL NETWORKS*, 2001.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

LR Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5, 2001.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997a.

Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research*, 3 (Aug):115–143, 2002.

Sepp Hochreiter and Jürgen Schmidhuber. LSTM can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997b.

Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. LSTM network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75, 2017.

Jack V Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231, 1996.

Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1):135–150, 2002.

Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

Sabine Toulson. Forecasting level and volatility of high frequency exchange rate data. Technical report, Working Paper in Financial Economics, 1996.

Lily Y Liu, Andrew J Patton, and Kevin Sheppard. Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1):293–311, 2015.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

J Huston McCulloch. Measuring the term structure of interest rates. *The Journal of Business*, 44(1):19–31, 1971.

Oldrich A Vasicek and H Gifford Fong. Term structure modeling using exponential splines. *The Journal of Finance*, 37(2):339–348, 1982.

Oldrich Vasicek. An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188, 1977.

John C Cox, Jonathan E Ingersoll Jr, and Stephen A Ross. An intertemporal general equilibrium model of asset prices. *Econometrica: Journal of the Econometric Society*, pages 363–384, 1985.

Charles R Nelson and Andrew F Siegel. Parsimonious modeling of yield curves. *Journal of business*, pages 473–489, 1987.

Inc. CME Group. U.S. Treasury Futures Conversion Factor Look-Up Tables. 2011. URL `http://www.cmegroup.com/trading/interest-rates/us-treasury-futures-conversion-factor-lookup-tables.html`.

Francis X Diebold, Glenn D Rudebusch, and S Boragan Aruoba. The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of econometrics*, 131(1):309–338, 2006.

Siem Jan Koopman, Max IP Mallee, and Michel Van der Wel. Analyzing the term structure of interest rates using the dynamic Nelson–Siegel model with time-varying parameters. *Journal of Business & Economic Statistics*, 28(3):329–343, 2010.

Christopher A Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48, 1980.

James H Stock and Mark W Watson. Vector autoregressions. *The Journal of Economic Perspectives*, 15(4):101–115, 2001.

Lutz Kilian. Finite-sample properties of percentile and percentile-t bootstrap confidence intervals for impulse responses. *The Review of Economics and Statistics*, 81(4):652–660, 1999.

Hiro Y Toda and Taku Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1):225–250, 1995.

Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 10(Jan):1–40, 2009.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Simon Haykin. A comprehensive foundation. *Neural Networks*, 2(2004):41, 2004.

Jonathan Cornelissen, Kris Boudt, Scott Payseur, and Maintainer Jonathan Cornelissen. Package ‘highfrequency’. 2017.

Frauke Günther and Stefan Fritsch. neuralnet: Training of neural networks. *The R journal*, 2(1):30–38, 2010.

Guirreri and Sergio Salvino. Package ‘YieldCurve’. 2013.

Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. SciPy Austin, TX, 2010.

W McKinney. Pandas, Python Data Analysis Library. 2015. *Reference Source*, 2014.

University of Cambridge. What is tensor, 2015. URL `https://www.doitpoms.ac.uk/tlplib/tensors/what_is_tensor.php`.

François Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k*, 2015.

J Perktold, S Seabold, and J Taylor. StatsModels: Statistics in python, 2006.

Ross Taylor. PyFlux, 2016. URL `http://www.pyflux.com/`.

Appendix A

Data and prediction characteristics

```
        Augmented Dickey-Fuller Test
data:  dataset$beta0
Dickey-Fuller = -3.9195, Lag order = 18, p-value = 0.01297
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta1
Dickey-Fuller = -3.6004, Lag order = 18, p-value = 0.03256
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta2
Dickey-Fuller = -4.224, Lag order = 18, p-value = 0.01
alternative hypothesis: stationary

Differenced data

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta0))
Dickey-Fuller = -24.986, Lag order = 18, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta1))
Dickey-Fuller = -22.419, Lag order = 18, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta2))
Dickey-Fuller = -22.976, Lag order = 18, p-value = 0.01
alternative hypothesis: stationary
```

Figure A.1.: Daily level and differenced data

```
        Augmented Dickey-Fuller Test
data:  dataset$beta0
Dickey-Fuller = -0.56525, Lag order = 6, p-value = 0.9785
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta1
Dickey-Fuller = -2.886, Lag order = 6, p-value = 0.2028
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta2
Dickey-Fuller = -1.5418, Lag order = 6, p-value = 0.7694
alternative hypothesis: stationary

Differenced data

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta0))
Dickey-Fuller = -7.5174, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta1))
Dickey-Fuller = -8.7912, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta2))
Dickey-Fuller = -8.9087, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Figure A.2.: Monthly level and differenced data

```
        Augmented Dickey-Fuller Test
data:  dataset$beta0
Dickey-Fuller = -0.47194, Lag order = 4, p-value = 0.9815
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta1
Dickey-Fuller = -2.5328, Lag order = 4, p-value = 0.3559
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta2
Dickey-Fuller = -1.3437, Lag order = 4, p-value = 0.8485
alternative hypothesis: stationary

Differenced data

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta0))
Dickey-Fuller = -4.9665, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta1))
Dickey-Fuller = -4.353, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta2))
Dickey-Fuller = -4.3166, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figure A.3.: Quarterly level and differenced data

```
        Augmented Dickey-Fuller Test
data:  dataset$beta0
Dickey-Fuller = 0.19078, Lag order = 2, p-value = 0.99
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta1
Dickey-Fuller = -2.0801, Lag order = 2, p-value = 0.5419
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  dataset$beta2
Dickey-Fuller = -1.2291, Lag order = 2, p-value = 0.866
alternative hypothesis: stationary

Differenced data

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta0))
Dickey-Fuller = -2.2561, Lag order = 2, p-value = 0.4748
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta1))
Dickey-Fuller = -2.9986, Lag order = 2, p-value = 0.192
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
data:  na.omit(diff(dataset$beta2))
Dickey-Fuller = 0.63106, Lag order = 2, p-value = 0.99
alternative hypothesis: stationary
```

Figure A.4.: Annual level and differenced data

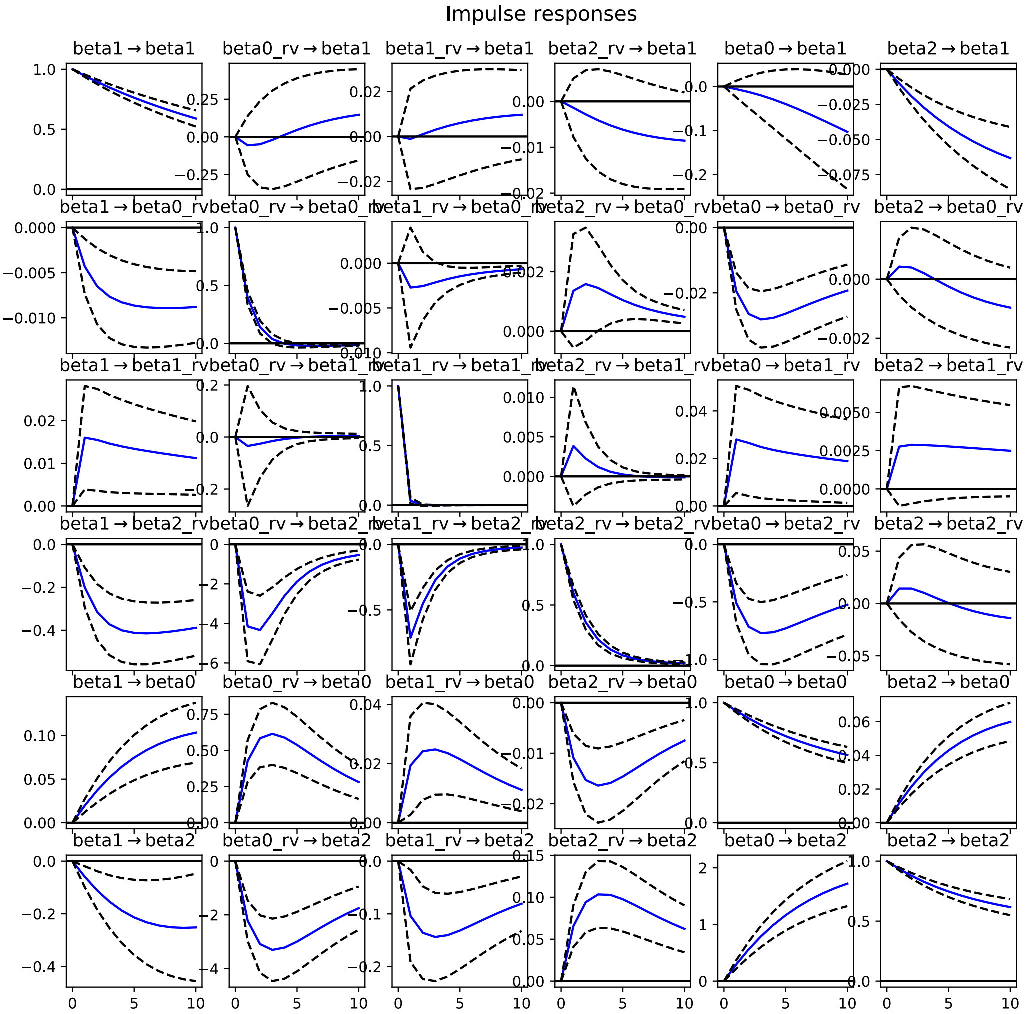

Figure A.5.: Impulse responses on daily data

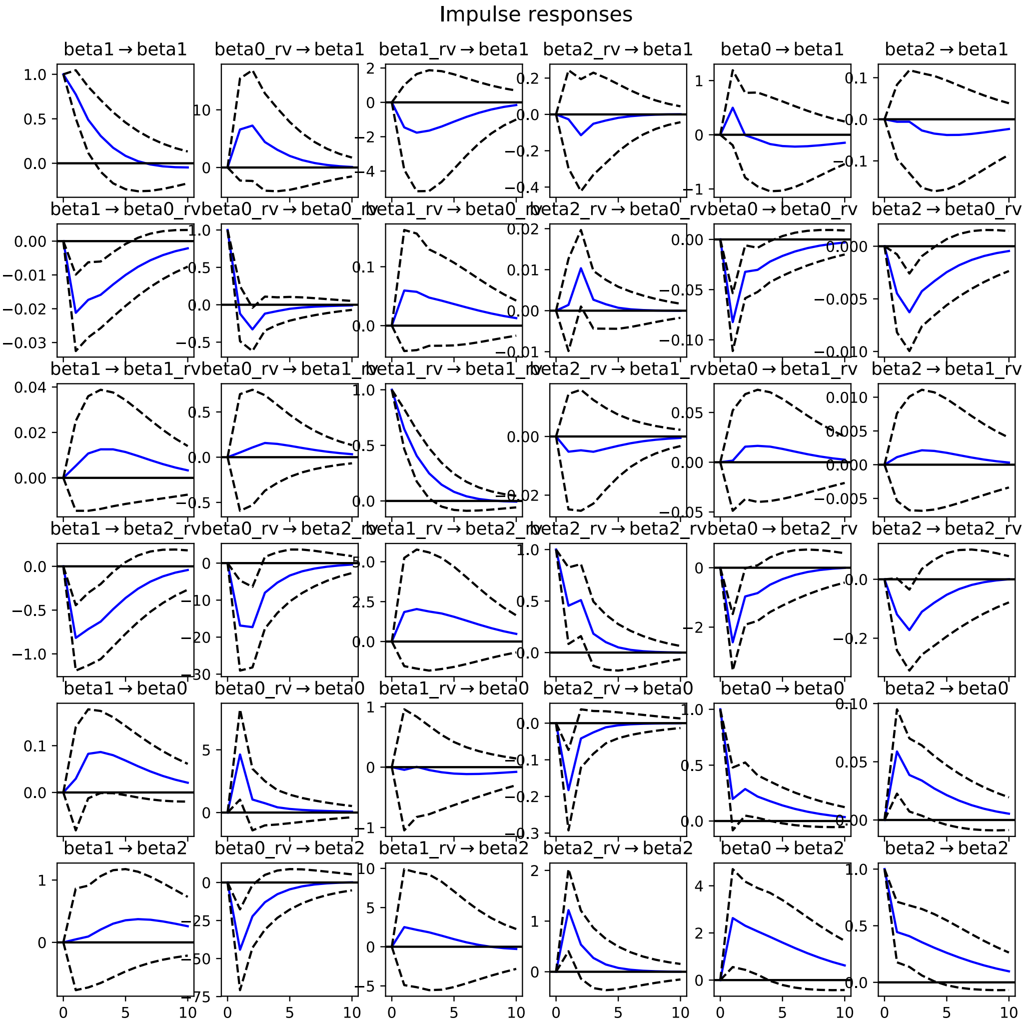

Figure A.6.: Impulse responses on quarterly data

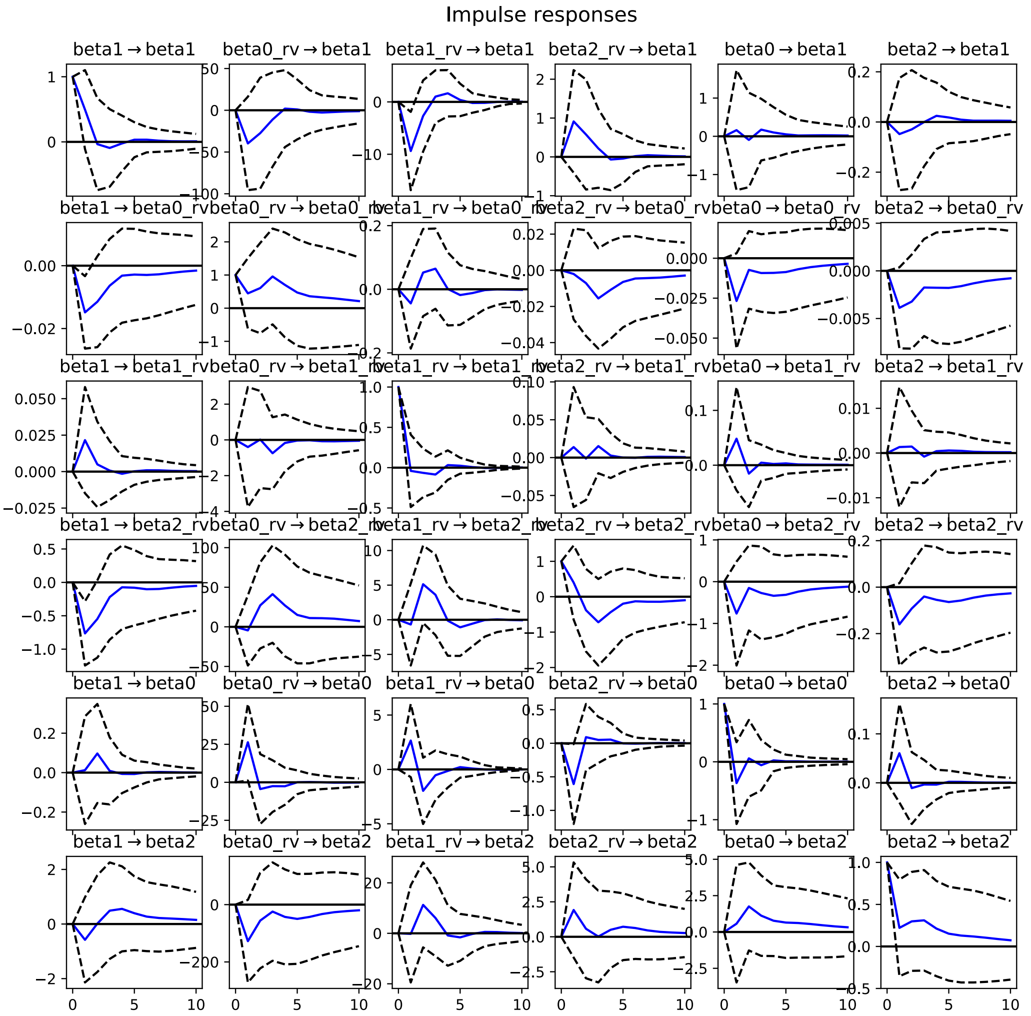

Figure A.7.: Impulse responses on annual data

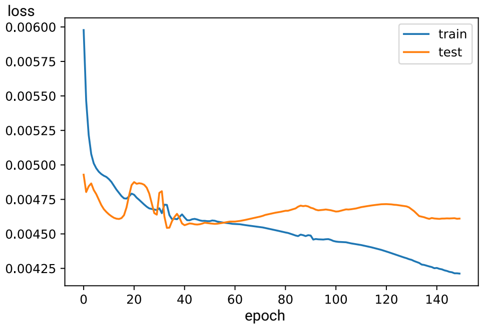

Figure A.8.: LSTM convergence on daily β_0

	RW	AR	VAR	VAR-RV	LSTM-RV
β_0^{d}	0.00473	0.00518	0.00962	0.01007	**0.00438**
β_1^{d}	0.00728	**0.00663**	0.00825	0.00827	0.00684
β_2^{d}	0.02658	0.02502	0.04539	0.04782	**0.02487**
β_0^{m}	0.01041	0.01508	0.02041	0.02163	**0.00979**
β_1^{m}	**0.01482**	0.01807	0.01630	0.01678	0.01679
β_2^{m}	0.06106	0.06203	0.11232	0.11417	**0.05491**
β_0^{q}	0.01852	0.02419	0.02492	0.02792	**0.01808**
β_1^{q}	0.02628	0.03172	**0.02545**	0.02959	0.03098
β_2^{q}	**0.10415**	0.11055	0.13002	0.13676	0.11801
β_0^{a}	0.02440	0.03729	0.02895	0.02910	**0.02402**
β_1^{a}	0.02738	0.09069	**0.02186**	0.02290	0.03849
β_2^{a}	**0.11998**	0.33761	0.12106	0.14989	0.17852

Table A.1.: Open Outcry RMSE (non-differenced VAR)

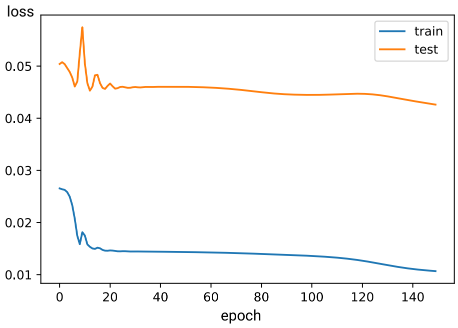

Figure A.9.: LSTM convergence on monthly β_0