

MASTER THESIS IN QUANTITATIVE FINANCE

Modeling the Credit Default

Swap Spread Curve

**A Gibbs Implementation of the Nelson-Siegel Model with
Time-Varying Factors Including Markov Switching**

2009

(Submitted October 6, 2009)

Christian Baer

Supervised by

Michel van der Wel (Erasmus Universiteit Rotterdam)

Regina Egorova and Joeri Potters (Cardano)

Modeling the Credit Default Swap Curve

A Gibbs Implementation of the Nelson-Siegel Model with
Time-Varying Factors Including Markov Switching

Abstract

The term structure of credit default swap spreads and interest rates shows similar stylized facts, especially in terms of persistency and cross correlations. We implement the Nelson-Siegel yield curve model to fit the term structure of credit default swap spreads. The factors in our model evolve according to a vector autoregressive process. We rewrite the model in state-space form and estimate the parameters in a Bayesian framework, using the multimove Gibbs sampling and extend it by adding Markov switching to the intercept and error covariance matrix of the transition equation. We employ the model to forecast credit default swap spreads, finding that the model outperforms the forecasts of the random walk approach and autoregressive models directly applied to the spreads, in terms of sum of squared prediction errors and the sign forecast performance.

Keywords:

Credit Default Swaps, Gibbs Sampling, Hamilton Filter, Kalman Filter, Nelson-Siegel Model, Spread Curve, State-Space Model, Markov Switching

TABLE OF CONTENTS

Acknowledgements.....	V
Tables and Figures.....	VI
1 Introduction.....	1
1.1 Preface.....	1
1.2 Literature Overview	4
1.3 Contribution and Outline	5
2 Data Description	8
3 The Nelson Siegel Model	10
3.1 Credit Default Swap Spreads and Vector-Autoregression.....	10
3.2 The Nelson Siegel Model.....	11
3.3 Least Square Estimation	14
3.4 Results from Least Square Estimation.....	16
4 State-Space Models with Gibbs Sampling.....	19
4.1 State-Space Representation.....	19
4.2 The Kalman Filter	20
4.3 The Gibbs Sampling Approach.....	23
4.3.1 Introduction to Bayesian Approaches.....	23
4.3.2 The Benefit of Gibbs Sampling	25
4.3.3 Prior Distribution for the Transition Matrix \mathbf{A}	27
4.3.4 Prior Distribution for the Covariance Matrix \mathbf{Q}	28
4.3.5 Prior Distribution for the Covariance Matrix \mathbf{R}	29
4.3.6 Generating the State Matrix $\boldsymbol{\beta}$	30
4.3.7 Generating the Parameters, Conditional on $\boldsymbol{\beta}$	32
4.4 Results from the Model without Regime Switching.....	34

5	Regime Switching.....	37
5.1	Markov-Switching Model and the Hamilton Filter	37
5.2	Estimation using Gibbs Sampling.....	40
5.2.1	Estimating $\mathbf{c}(r_t)$ and \mathbf{A}	40
5.2.2	Estimating the Covariance Matrix $\mathbf{Q}(r_t)$	42
5.2.3	Estimating the Transition Matrix \mathbf{K}	45
5.2.4	Generating the Regime Vector \mathbf{r}	46
6	Regime Switches in State-Space Models.....	47
6.1	Building the Algorithm for the Combined Model	47
6.2	Results from the Model including Regime Switching.....	47
7	Forecasting Performance Compared.....	52
8	Conclusions.....	56
8.1	Summary.....	56
8.2	Recommendations for Further Research.....	57
	Appendices.....	59
	References.....	VII

Acknowledgements

First, I would like to thank the management of Cardano for giving me the opportunity to make an internship and to write my thesis within their company. The thesis would not have been possible without the support of many people. Special thanks are due to my supervisors at Cardano, Regina Egorova and Joeri Potters, especially for giving me many suggestions for improvements, for checking and reading my thesis and for answering my tedious questions. I want to thank Bart Bos for many helpful suggestions, in particular specifying the topic and the model. I am further indebted to many other people at Cardano for their patience with answering questions concerning programming and econometrics, but also for the enabling a very nice atmosphere. I had a great time here in the last six and a half month.

Special thanks go also to my supervisor from the Erasmus University, Michel van der Wel and the second reader (co-reader) of this thesis, Lennart Hoogerheide for many great suggestions concerning my research and the presentation of my results. Thanks for carefully reading my thesis with respect to content and language many times and for helping me to arrange administrative things.

Needless to say, while I am extremely grateful for all the support I received during the time it took to write this thesis, I still claim sole responsibility for any errors that occur.

Tables and Figures

Figures:

Figure 1 – Commerzbank CDS Spread Term Structure over Time.....	8
Figure 2 – The “least square” Nelson-Siegel Factors	17
Figure 3 – Nelson-Siegel Loadings and Spread Curves	18
Figure 4 – A Gibbs Sampling Illustration.....	26
Figure 5 – The State-Space Gibbs Sampling Nelson-Siegel Factors.....	34
Figure 6 – Least Squares and Gibbs Sampling Nelson-Siegel Factor Correlations	36
Figure 7 – The State-Space Gibbs Sampling Nelson-Siegel Factors with two Regimes..	49
Figure 8 – The State-Space Gibbs Sampling Nelson-Siegel Factor Volatilities	50

Tables:

Table 1 – Cross and Auto Correlations of Commerzbank AG CDS Spreads.....	9
Table 2 – Sum of Squared Errors for Different Models	52
Table 3 – Sign Forecast Performance for Different Models.....	53
Table 4 – Diebold-Mariano Statistics for Comparing Predictive Accuracy	55

1 Introduction

1.1 Preface

A probable cause of the financial crisis from 2007 was the underrating of the risks involved in credit derivatives by investment banks and rating agencies. A credit derivative is used to transfer credit risk from a legal entity, called reference entity, to another party, the seller of the product. This way, the buyer of a credit derivative receives protection against the exposure to credit risk of a reference entity. With the start of the crisis, the default correlation of banks and other companies significantly increased. This behavior is also referred to as crash correlation. Due to high correlations, the diversification effect for many products decreased, resulting in a loss in value.

Since credit derivatives allow for a simple transfer of credit risk, the market has grown rapidly, together with the field of modeling credit risk. Due to the complexity of most of the credit derivatives, there is no common valuation model that rules out arbitrage opportunities for the majority of available products.

The most commonly traded products are credit default swaps (CDS) and collateralized debt obligations (CDO). In a credit default swap, the buyer agrees to pay a fixed coupon, dependent on the contractual credit default swap spread, to the seller of the product. Usually the coupon is paid quarterly. In return, the buyer receives a payment from the seller when a pre-specified credit event occurs at the reference entity. The buyer and the seller of the product define those credit events, the events which trigger the protection. Typical credit events are bankruptcy, default on payments (e.g. on obligations like bonds or loans) or debt restructuring. In contrast to insurance, credit default swaps can also be used for speculating, since the buyer does not necessarily have to be exposed to risk of the reference entity.

The contractual spread defines the price that the buyer pays for receiving the protection. In contrast to the contractual spread, the break-even spread implies a value of zero for the credit default swap. When the break-even spread is higher than the contractual spread, the credit default swap has a positive value for the buyer, since the fixed coupon that the

buyer pays is lower than the break even price. Assuming that other influencing factors like interest and recovery rate have not changed, increasing spreads go together with higher implied default probabilities of the reference entity.

The payment of the seller in case of a default depends on properties of the contract between the two parties, especially the notional and the recovery rate. Settlement and other characteristics are also defined in the contract. Since defaults are rare events, usually occurring unexpectedly and with unknown size of the losses, the payments of the sell-side are difficult to evaluate.

The probability that a credit event occurs (default probability) and the assumptions about the recovery rate play an important role in pricing credit default swaps. Most of the existing pricing models also rely on crucial assumptions about correlations between default probabilities, interest rates and recovery rates, such as the pricing models in Hull and White (2000) or Duffie (1999).

The payments of collateralized debt obligations are derived from a collection of assets which pay a fixed coupon, e.g. loans. The exposure to risk of this collection is sold in so called tranches, which are classified by their seniority. The lowest or junior tranche has the highest exposure to risk. In case one of the assets fails to pay the coupon, the lowest or junior tranche is the first to absorb the loss. In return for the higher risk, the tranche receives a higher coupon for the money invested in the collateralized debt obligation. In contrast, the senior tranches suffer from a much lower risk. Assuming a senior tranche absorbs only the top thirty percent of the losses, seventy percent of the assets need to default on their payments, before the senior tranche incurs any losses. Hence, for bearing a lower risk than the junior tranche, they receive a lower coupon for their investment.

High default correlations between the assets imply that a default of one payment is likely to come together with defaults of other payments. This increases the risk of the product and therefore investors require a higher compensation, respectively coupon on their investment.

In 1998, Paul Wilmott, a quantitative-finance consultant, already claimed that "the correlations between financial quantities are notoriously unstable". This fact was ignored

by rating agencies, banks and investors, resulting in highly overrated products. The widely used Gaussian copula model by Li (2000), which implies a fixed correlation structure, performed well, as long as the market was stable. Nevertheless, after the crash, the default correlation structure of the underlying products changed and the model failed. Especially in the mortgage market, this had devastating effects. After mortgage prices started falling, their correlation increased rapidly, leading to a decrease of collateralized debt obligation prices. Due to high exposures to collateralized debt obligations, many banks and other companies incurred huge losses and their credit ratings fell, implying higher default probabilities for the companies. Thus, the credit default swap spreads for those reference entities increased as well, resulting in an increase not only in credit default swap prices, but also in their volatility.

The importance of credit default swap spreads is not only characterized by the valuation of credit default swaps, but also reaches other financial fields. It plays an important role for accounting and disclosure of employee benefits, i.e. all forms of return given by a company in exchange for service rendered by employees, especially for defined benefit plans. According to the International Financial Reporting Standards (IFRS), the rate used to discount estimated cash flows should be determined by reference to market yields on high quality corporate bonds (IAS 19.78). Since highly rated corporate bonds are not well indexed (there is no index available for high quality corporate bonds), companies make use of the swap rate plus the spread of a credit default swap index (CDX) for highly rated financial institutions, e.g. the iTraxx Europe. The iTraxx Europe includes 125 credit default swap investment grade credits. Normally, the reference entities of the iTraxx Europe have very high credit ratings.

Furthermore, credit spreads are not quoted in the market, i.e. they have to be calculated from bond yields less a “risk-free” rate (usually Treasury bond yields are used) while credit default swap spreads are quoted. Due to their importance for discounting liabilities, the spreads are also important for risk-modeling.

Our focus is on modeling and forecasting credit default swap spread curves, dependent on their historical evolution. We use the Nelson and Siegel (1987) model to interpolate the

spread curve for each point in time and let the factors evolve according to a vector autoregressive process. To account for jumps in level, volatility and correlation of the factors, we allow for regime switches in the constant and the error covariance matrix of the vector autoregressive model.

1.2 Literature Overview

The credit default swap in its modern form was invented by a working group of JP Morgan Chase in 1997. Since the market grew rapidly, valuation of credit default swaps and modeling of credit default swap spreads gained popularity very fast. Currently two main approaches for pricing a credit default swap exist. Duffie (1999) and later Hull and White (2000) base their pricing model on an arbitrage-free market, which means there should be no risk-free possibility to make a profit without own investment. Both models within this approach make assumptions about correlations between recovery rates, interest rates and default probabilities, but also on transaction costs and other market frictions.

Another approach is based on discounted cash flows. The price of a credit default swap is given by the present values of all cash flows, weighted by the survival probability of the reference entity, the probability that no credit event occurs till the cash flow is paid. Schönbucher (2003) discusses the different pricing models for credit derivatives, amongst others credit default swaps and addresses various recovery rate, default probability and credit spread models.

As mentioned, credit default swap spreads gained more importance over the last decade. Nevertheless, there only exist few models to capture their dynamics, amongst others the model of Zhang, Zhou and Zhu (2006), which relates the spreads to equity, macroeconomic variables and firm characteristics. There exist a number of models for credit related spreads (for instance the Markov model of Jarrow, Lando and Turnbull (1997)) which try to capture the dynamics of corporate bond spreads. As discussed by Alexopoulou, Andersson and Georgescu (2009), credit spreads of corporate bonds are closely related to credit default swap spreads in the long-run, but show higher deviations

for spreads in the short-run. We need to take this fact into account with care. A good model for credit spreads is not necessarily a good model for credit default swap spreads and vice versa.

Credit default swap spread curves and yield curves show similar stylized facts, especially concerning the cross correlations and persistency of the rates. Since earlier models for fitting the term structure, amongst others Vasicek (1977), Cox, Ingersoll and Ross (1985) or Hull and White (1990), are not able to outperform the random walk, we focus on the Nelson and Siegel (1987) model, to fit the credit default swap spread curve. The factors of the model evolve according to a vector autoregressive process, as implemented in Diebold and Li (2006) for government bond yields. They show that the model performs significantly better than the random walk hypothesis for certain forecast horizons.

The model can be represented in state-space form and the Kalman filter is applicable to receive the latent factors. Based on this approach, a few extensions on the model have been developed. Diebold, Rudebusch and Aruoba (2006) add macroeconomic variables to the transition equation of the model to explain the dynamics of the factors. Koopman, Mallee and van der Wel (2007) allow for a time-varying convexity factor and investigate whether the overall volatility in the yields is constant over time, finding empirical evidence for an increase in in-sample goodness-of-fit. Based on Kim and Nelson (1999), who explain the estimation of state-space models with regime switching using classical and Bayesian approaches, Bernadell, Coche and Nyholm (2005) add Markov switching to the slope factor of the yield curve.

1.3 Contribution and Outline

The goal of this thesis is to develop a model that captures the behavior of credit default swap spreads. We can clearly observe a structural break in 2007 concerning level and volatility of the credit default swap spreads, but also in the shape of the term structure for almost all reference entities. Although we can identify where the structural break occurs, our goal is building a model to generate scenarios and forecast spreads.

We use the Nelson and Siegel (1987) model to fit the credit default swap curve and let the factors evolve according to a vector autoregressive process, as suggested in Diebold and Li (2006) for bond yields. To capture the structural break(s) we introduce regime switching for all factors of the model, i.e. for their constant and their covariance matrix in the transition equation. The focus of this thesis lies in the identification of the state vectors and the regime vector, as well as the estimation of the parameters in the model.

The estimation of many parameters is involved and the dataset for credit default swap spreads for different maturities is quite small. The large number of parameters and the small dataset make a classical approach for the estimation less attractive. Therefore we decide to estimate the model in a Bayesian framework. Using a Gibbs sampling approach, we have several advantages over the classical approach. It is not necessary to have a large dataset. We can have prior beliefs about the parameters and their distributions and update those prior beliefs using the available data. Another advantage occurs in the estimation itself. Gibbs sampling estimates the parameters by drawing from their distributions, conditional on the other parameters. Since we aim to add more flexibility to the model by introducing regime switching, we benefit from an easy implementation in this framework.

We build upon the approach of Kim and Nelson (1999). They estimate state-space models with regime switching using a classical approach and Gibbs sampling. Additionally, we propose to extend this framework in two ways. We allow for correlations in the state vector and we do not draw every element of the parameter vectors and matrices separately, but from their multivariate distributions, as done in Sugita (2008) for vector autoregressive models. We use first order Markov switching, leading to modifications in the estimation of the model. To generate draws of the state vectors and regime vector, we use multimove Gibbs sampling, originally motivated by Carter and Kohn (1994) and first implemented in a Markov switching framework by Kim and Nelson (1998). The decision against singlemove Gibbs sampling from Carlin, Polson and Stoffer (1992) is motivated by faster convergence of multimove sampling.

We test the forecast performance of the model against the same model without regime switching, autoregressive models on the spreads itself and the random walk using credit default swap spread data of Commerzbank AG. We estimate the performance by calculating the sum of squared forecast errors and the relative frequency of correctly predicted signs to measure the performance. We conclude that the Gibbs sampling estimated state space models perform better than the other models for all maturities.

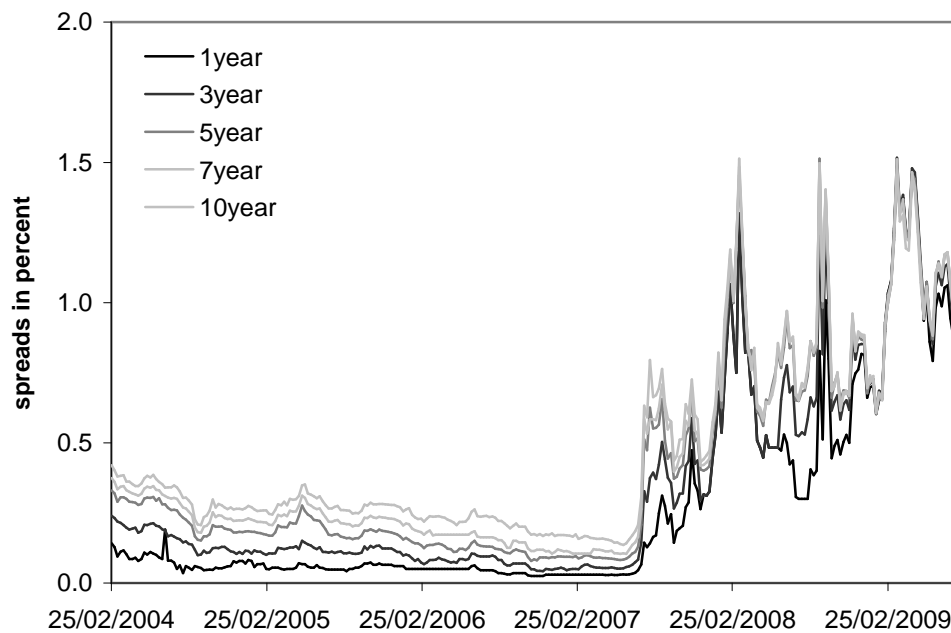
In the next section, we describe the data we use throughout the thesis to demonstrate the model. In Chapter 3, we introduce the Nelson and Siegel (1987) model and let the factors evolve according to a vector autoregressive process. Further we present a least square approach to estimate the model. In Chapters 3 and 4 we show the results of the estimation using the data of credit default swap spreads. In Chapter 4, we rewrite the model in state-space form and show how the Kalman filter can be used to obtain the factors, conditional on the hyperparameters. Using the Kalman filter, we illustrate a Gibbs sampling approach to estimate the hyperparameters. In Chapter 5, we add regime switching to the model. We point out the changes that have to be applied compared to the estimation steps in Chapter 4. Chapter 6 contains the estimation results of the Markov switching model, obtained for the Commerzbank dataset. In Chapter 7, we test the forecast performance of the presented models and compare them to the random walk and other autoregressive models on the spreads. In the final chapter, we summarize our results and give recommendations for further research.

2 Data Description

We use credit default swap spread data on Commerzbank AG, one of the biggest banks in Germany. As many other banks, their spreads suffered from a huge, but compared to other not seriously different, structural break in 2007. Not only the level of the spread curve grew tremendously, but also its volatility grew and the shape of the curve changed several times. This behavior can be observed in many markets, when jumps occur. We use data from the 25th February 2004 to the 12th of August 2009 for one, three, five, seven and ten year maturities to capture the shape of the spread curve.

Figure 1 shows the evolvement of the spreads of Commerzbank AG over time. From February 2004 to June 2007, the spreads were decreasing, implying that the default probability of Commerzbank AG was evaluated very low. The one year spread reached its minimum of 2.5 bp on the 8th November 2006 and within 1.5 years reached a value above 130 bp. Regarding Figure 1, it is obvious, that the time series are not covariance stationary.

Figure 1 – Commerzbank CDS Spread Term Structure over Time



The figure shows the movement of the credit default swap spreads of Commerzbank for maturities from one to ten years in percent. The data goes from 25th February 2004 to the 12th August 2009.

We abandon the use of the iTraxx Europe, since the composition of the index changes every half a year. It is difficult to filter the bias that occurs due to the changes in the companies. Further, the price of the iTraxx changes not only due to the default risk of the new companies, but also due to their correlations with the other companies in the index. Since the default correlation between the companies is lower than one, we have a diversification effect. This leads to smaller jumps in the level and volatility of the index, compared to single name credit default swaps. Our goal is to show the impact of the structural break to the performance of the models. We chose to use spreads on one reference entity to clearly highlight the effect of a credit crisis.

Since we use the Nelson and Siegel (1987) model with time-varying factors to capture the evolution of the spread curve, we shortly emphasize, that credit default swap spreads and the yields show similar stylized facts. Regarding Figure 1, we observe that the credit default swap curve can take on various shapes, as the yield curve. From 2004 to 2007, the curve is slightly upward sloping. Immediately after the jump in July 2007, slope increases a lot. Further, the curve is downward sloping in January and February 2009. The auto correlations and cross correlations of credit default swap spreads are also very high. The same behavior also applies for yields. Table 1 shows the first order auto and cross correlations of the Commerzbank AG spreads.

Table 1 – Cross and Auto Correlations of Commerzbank AG CDS Spreads

	auto correlation	cross correlation				
Maturity	first order	1 year	3 year	5 year	7 year	10 year
1 year	0.9707	1.0000	0.9839	0.9596	0.9559	0.9505
3 year	0.9667	0.9839	1.0000	0.9882	0.9864	0.9829
5 year	0.9657	0.9596	0.9882	1.0000	0.9989	0.9969
7 year	0.9644	0.9559	0.9864	0.9989	1.0000	0.9983
10 year	0.9574	0.9505	0.9829	0.9969	0.9983	1.0000

The table contains the first order auto correlations and cross correlations of Commerzbank AG credit default swap spreads with five different maturities., using data from the 25th February 2004 to the 12th August 2009.

3 The Nelson Siegel Model

In this chapter we explain the use of a factor model to capture the term structure of credit default swap spreads. We focus on a three-factor Nelson-Siegel model. Furthermore, we assume that the factors follow a vector autoregressive process. In the end of this chapter we present one approach for the estimation of the hyperparameters, using linear least square methods.

3.1 Credit Default Swap Spreads and Vector-Autoregression

Credit default swap spreads for different maturities have high auto- and cross-correlations. We construct a vector-autoregressive (*VAR*) model to forecast the spreads and to build scenarios for their development over time. The *VAR* model is a linear model that tempts to capture the evolution of N different time series¹ with set of observed values $\mathbf{y}_t = (\mathbf{y}_t^1 \mathbf{y}_t^2 \dots \mathbf{y}_t^N)'$ over a sample period $t = 1, 2, 3 \dots T$, only dependent on their recent, respectively lagged evolution

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \dots + \mathbf{A}_k \mathbf{y}_{t-k} + \boldsymbol{\varepsilon}_t. \quad (2.1)$$

The error term $\boldsymbol{\varepsilon}_t = (\boldsymbol{\varepsilon}_t^1 \boldsymbol{\varepsilon}_t^2 \dots \boldsymbol{\varepsilon}_t^N)'$ denotes white noise with zero mean, covariance matrix $\boldsymbol{\Sigma}$ and no serial correlation in the single elements of $\boldsymbol{\varepsilon}_t$. The parameters of the *VAR* model are $N \times N$ matrices \mathbf{A}_l , a $N \times 1$ vector of constants \mathbf{c} and the covariance matrix of the error terms $\boldsymbol{\Sigma}$. The order k of a *VAR* model determines the number of lagged regressors.

The higher the order of the model, the better we can fit the data, since we add more degrees of freedom. Nonetheless we choose a vector-autoregressive model of order one. That means the spreads at time t only depend on their most recent values (observed at time $t - 1$). This way we reduce, but not eliminate, the risk of overfitting the model, which in many cases results in poor forecast performance.

¹ In special case of only one time series the model reduces to an autoregressive model (*AR*).

The model can be represented in different ways. We start with the general matrix notation to point out the different parameters of the model²

$$\mathbf{s}_t = \mathbf{c} + \mathbf{A}\mathbf{s}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (2.2)$$

where \mathbf{s}_t denotes a vector of spreads for N different maturities at time t and $\boldsymbol{\varepsilon}_t$ denotes white noise with covariance matrix $\boldsymbol{\Sigma}$. Under assumption of a multivariate normal distribution for the error terms, $\boldsymbol{\varepsilon}_t \sim MN(\mathbf{0}_N, \boldsymbol{\Sigma})$, the maximum likelihood estimation reduces to ordinary least squares (OLS) estimation. The unbiased estimate of $\boldsymbol{\Sigma}$ is the sample covariance matrix of the residuals.

Using a small number of spreads for different maturities, e.g. five (ten), we have to estimate 45 (165) parameters in total³. For a small number of maturities and a sufficiently large dataset, this might still be reasonable, but for many maturities or a small dataset, our model has too many degrees of freedom which restricts the predictive power. The following section describes how factor models can be used to limit the number of parameters. Our main focus is on the Nelson Siegel three-factor model.

3.2 The Nelson Siegel Model

We can relate the dynamics of the spread curve to a couple of factors, to reduce the number of parameters in our model. This allows including more dependent variables to the model with limited increase of the number of parameters.

We focus on the Nelson-Siegel model⁴. In Diebold and Li (2006), the model was applied to the yield curve and tested for its forecast performance, achieving sustained success and

² The concise matrix notation simplifies the equation. We use it in latter subsequent sections, since it allows for a more efficient way of presenting the formulas and results. We use the general notation here to emphasize the difference between the constant \mathbf{c} and the matrix \mathbf{A} .

³ That is 5 (10) parameters for \mathbf{c} , 25 (100) parameters for \mathbf{A} and 15 (55) parameters for $\boldsymbol{\Sigma}$. The formula to calculate the number of parameters is given by $N + N^2 + N(N + 1) / 2 = (3N / 2) (1 + N)$.

⁴ There are several other possible models to consider, amongst others the Vasicek model, the Cox-Ingersoll-Ross model and the Hull-White model.

outperforming the random walk approach. In practice it is widely used to fit the yield curve e.g. by the European and other Central Banks (some of them use the Svensson model, an extended version). As mentioned above, the spread and the yield curve show various similar stylized facts, especially concerning persistency and cross-correlations which puts forward good reasons to test the model on spreads.

The Nelson Siegel model is a factor model, which, when applied to yields or to spreads, approximates the term structure. In contrast to e.g. a principal component analysis, the Nelson Siegel model is continuous in the maturities, i.e. for each maturity it delivers the corresponding yield or spread hence additional interpolation is not necessary. In case of credit default swap spreads, this result is adjutant, since quotes are usually only available for certain maturities, mostly for maturities of 3, 5, 7 and 10 years. For discounting assets and liabilities and for traders of credit default swap spread related products, maturities in between are mandatory.

There are several representations of the Nelson Siegel model, depending on the characteristics of the factors. The most common form is

$$s_t(\tau) = \beta_{1t} + \beta_{2t} \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} \right) + \beta_{3t} \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) + v_t(\tau). \quad (2.3)$$

The spread s at time t with maturity τ only depends on the level β_{1t} , the slope β_{2t} , the curvature β_{3t} and the convexity λ . The estimation error $v_t(t)$ is normally distributed with mean zero and standard deviation σ_t . In the scope of this thesis we assume λ to be known, which simplifies the estimation⁵, because then the model is linear in β_{1t} , β_{2t} , β_{3t} .

The convexity λ is usually fixed to maximize the curvature loading at a chosen maturity (for yields this is usually a maturity between two and three years). We follow this practice and maximize the curvature loading at the three year maturity, resulting in a value of 0.598 for λ . Diebold and Li (2006) maximize the curvature loading at the thirty month maturity. Also note that they state τ in month, while we express it in years.

⁵ This assumption is quite common, although the identification of the model with a time-varying λ is possible. We refer to Koopman, Mallee and van der Wel (2007) for further details.

We again consider a VAR model of order one for the evolution of the factors, given by

$$\boldsymbol{\beta}_t = \mathbf{c} + \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{u}_t, \quad (2.4)$$

where $\boldsymbol{\beta}_t$ denotes the vector of the Nelson Siegel factors $(\beta_{1t}, \beta_{2t}, \beta_{3t})'$. The vector \mathbf{u}_t is multivariate normally distributed with $\mathbf{u}_t \sim MN(\mathbf{0}, \mathbf{Q})$. This assumption is not necessary, but allows for an ordinary least square estimation of the VAR model. Note that this assumption implies a positive probability for the factors to take negative values. Since Equation (2.3), which links the factors to the spreads, is linear in the factors, it is also possible for the spreads to take negative values. For more information on the Nelson-Siegel model and its forecasting performance on yield curves, we refer to the article of Diebold and Li (2006).

The identification of the hyperparameters \mathbf{c} , \mathbf{A} , \mathbf{Q} and \mathbf{R} in the model is not straight forward. If the factors $\boldsymbol{\beta}_t$ are observable variables, we could estimate the parameters in Equation (2.4) using multivariate least squares. However, the factors are unobserved. They themselves can be estimated based on available spread observations using Equation (2.3), but then contain an estimation error $v(t)$.

The following sections provide two solutions to this problem. In the first one, we estimate the Nelson-Siegel factors from Equation (2.3) for each point in time. We receive the estimated factors \mathbf{b}_t (estimate of $\boldsymbol{\beta}_t$) for all time points t up to T . Treating these factors as if they were observed, we can estimate the hyperparameters of Equation (2.4). The estimation error $v(t)$, therefore is ignored.

In this thesis, we pay more attention to the second solution, which does not require the assumption of observed factors. The solution provides us with a consistent estimation of the hyperparameters and the factors of the model. We make use of the specific characteristics of the model. We rewrite the model given by Equations (2.3) and (2.4) in state-space form. Due to the assumptions we made above, our model is linear in the factors. This allows us to use the basic Kalman filter.

The parameters can be estimated with several methods such as maximizing the likelihood. We apply Gibbs sampling, a Bayesian approach, to estimate the model. We aim to add more flexibility into the model to achieve a better fit to market data. We do this by adding regime switches to some of the hyperparameters in the model. The model becomes more complex and is non-linear in the Nelson-Siegel factors. Nevertheless, the Gibbs algorithm allows estimating the parameters and factors conditional on the regime.

3.3 Least Square Estimation

In this section we discuss the estimation of the model using ordinary and multivariate least squares. As mentioned above, the Nelson-Siegel model is linear in the parameters $\beta_{1t}, \beta_{2t}, \beta_{3t}$. Equation (2.3) can be identified for each point in time, given a set of spreads for different maturities τ_n for $n = 1, 2, 3 \dots N$. It provides us with an estimate of β_t for each t , denoted as \mathbf{b}_t . Combining the loadings in a matrix for all maturities of all factors, where the loading from the first factor is just given by 1, provides us with matrix \mathbf{H}

$$\mathbf{H} = \begin{pmatrix} 1 & \left(\frac{1 - e^{-\tau_1 \lambda}}{\tau_1 \lambda} \right) & \left(\frac{1 - e^{-\tau_1 \lambda}}{\tau_1 \lambda} - e^{-\tau_1 \lambda} \right) \\ 1 & \left(\frac{1 - e^{-\tau_2 \lambda}}{\tau_2 \lambda} \right) & \left(\frac{1 - e^{-\tau_2 \lambda}}{\tau_2 \lambda} - e^{-\tau_2 \lambda} \right) \\ \dots & \dots & \dots \\ 1 & \left(\frac{1 - e^{-\tau_N \lambda}}{\tau_N \lambda} \right) & \left(\frac{1 - e^{-\tau_N \lambda}}{\tau_N \lambda} - e^{-\tau_N \lambda} \right) \end{pmatrix}. \quad (2.5)$$

Note that all elements in the matrix are known upfront, since the convexity λ is fixed.

We estimate β_t by rewriting the model in concise matrix notation⁶ and applying linear least squares for each point in time.

⁶ Concise means that the constant is included into the vector of the other parameters, which requires that we add a vector of ones to the matrix of loadings, given by \mathbf{H} . For more information on general and concise matrix notation in VAR models, see Lütkepohl (2005).

The more spreads for different maturities we include into the regression, the better we can fit the spread curve for each t . The spread model and its unbiased estimate \mathbf{b}_t are given by

$$\mathbf{s}_t = \mathbf{H}\boldsymbol{\beta}_t + \mathbf{v}_t, \quad (2.6)$$

$$\mathbf{b}_t = (\mathbf{H}'\mathbf{H})^{-1} \mathbf{H}'\mathbf{s}_t, \quad (2.7)$$

where \mathbf{v}_t is a vector of error terms $(v_t(\tau_1), v_t(\tau_2) \dots v_t(\tau_N))'$. A common assumption is that \mathbf{v}_t follows a multivariate normal distribution. The error covariance matrix is commonly assumed to be diagonal (respectively uncorrelated). That reduces the number parameters to be estimated. Nevertheless, it is not necessary to apply this assumption here. We make use of it in subsequent chapters, when the parameters are estimated with Gibbs sampling.

As discussed above, the approximation with \mathbf{b}_t and hence contains an estimation error. If we do not take the error into account, or in other words, if we treat the factors as observed values (i.e. $\boldsymbol{\beta}_t = \mathbf{b}_t$), we can use them to estimate the parameters in the vector-autoregressive model, given by Equation (2.4).

As applied to the previous equation, we rewrite the VAR model (2.4) in concise matrix notation, including the information over all t . We can denote the collection of all factors $\boldsymbol{\beta}_t$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \dots \boldsymbol{\beta}_T)'$. We take the same step for the lagged set of factors $\boldsymbol{\beta}_c^L$, but connect it with a vector of ones, given by

$$\boldsymbol{\beta}_c^L = \begin{pmatrix} 1 & \beta_{1,1} & \beta_{2,1} & \beta_{2,1} \\ 1 & \beta_{1,2} & \beta_{2,2} & \beta_{3,2} \\ \dots & \dots & \dots & \dots \\ 1 & \beta_{1,T} & \beta_{2,T} & \beta_{3,T} \end{pmatrix}. \quad (2.8)$$

Respectively, the matrix \mathbf{A}_c is given by

$$\mathbf{A}_c = \begin{pmatrix} c_1 & A_{1,1} & A_{1,2} & A_{1,3} \\ c_2 & A_{2,1} & A_{2,2} & A_{2,3} \\ c_3 & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix}. \quad (2.9)$$

And the model can be rewritten as

$$\boldsymbol{\beta} = \boldsymbol{\beta}_c^L \mathbf{A}_c' + \mathbf{u}, \quad (2.10)$$

The error term \mathbf{u} is a $T \times N$ matrix of the error terms $\mathbf{u} = (\mathbf{u}_1 \mathbf{u}_2 \dots \mathbf{u}_T)'$. This estimator is equal to the maximum likelihood estimator⁷, applied to each equation separately, conditional on the others.

We estimate matrix \mathbf{A}_c using multivariate least squares, as follows

$$\hat{\mathbf{A}}_c = \boldsymbol{\beta}_c' \boldsymbol{\beta}_{c-1} (\boldsymbol{\beta}_{c-1}' \boldsymbol{\beta}_{c-1})^{-1}. \quad (2.11)$$

3.4 Results from Least Square Estimation

The results for \mathbf{A} and \mathbf{c} for the Commerzbank dataset of credit default swap spreads (the one, three, five, seven and the ten year spreads), starting at the 21st of March 2005 and ending at the 11th of August 2009, are given by

$$\hat{\mathbf{A}} = \begin{pmatrix} 0.925 & 0.040 & 0.040 \\ 0.037 & 0.903 & -0.020 \\ 0.132 & -0.063 & 0.825 \end{pmatrix}, \quad \hat{\mathbf{c}} = \begin{pmatrix} 0.067 \\ -0.058 \\ -0.134 \end{pmatrix}, \quad \hat{\mathbf{Q}} = \begin{pmatrix} 0.011 & -0.005 & -0.002 \\ -0.005 & 0.008 & -0.001 \\ -0.002 & -0.001 & 0.017 \end{pmatrix}.$$

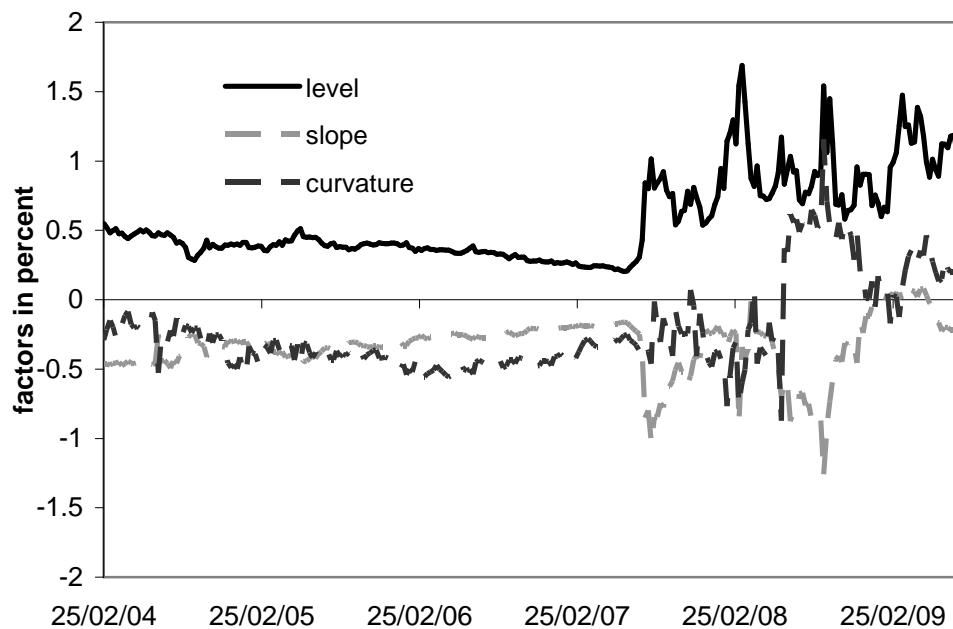
Figure 2 shows the estimated \mathbf{b}_t from the first least square estimation. We remark that many statistic tests, e.g. T-tests for significance of the parameters in the VAR model are not reliable in our case, because we ignored the estimation error in \mathbf{b}_t .

Note that a negative value for the slope factor does not imply a negative slope for the spread curve. Given time t , the factor is the same for each maturity. Crucial for the slope of the spread curve is the loading of the slope factor. To see the relation of the slope factor to the spreads, we multiply it with the loading of the slope factor. Figure 3 shows

⁷ See also Hamilton (1994).

the loading of slope and curvature for different maturities with λ equal to 0.59. It further provides fitted spread curves for different dates. Due to the low available cross sectional dimension, it fits the observed spreads quite well. We observe that the slope loading is monotonically decreasing. Given a negative slope, that implies an upward sloping spread curve. Only in the beginning of year 2009, we observe a downward sloping spread curve. We also see that the curvature loading reaches its maximum for a maturity of three years, which follows from the predefinition of λ .

Figure 2 – The “least square” Nelson-Siegel Factors

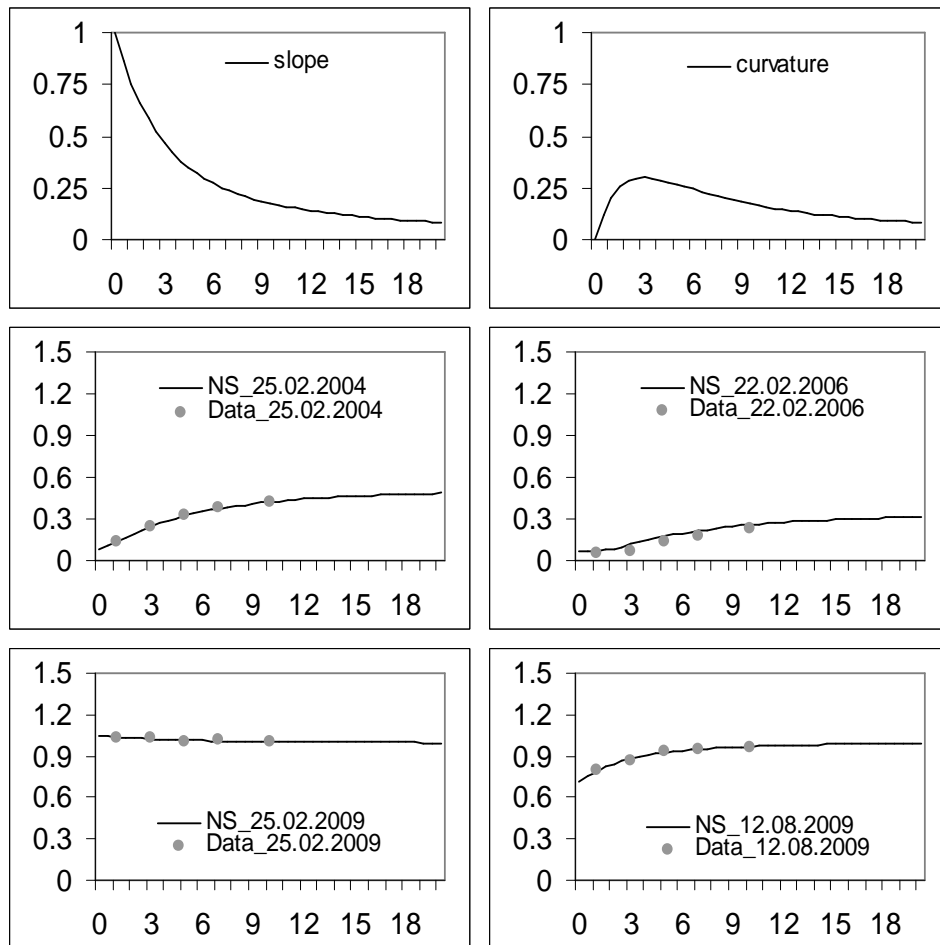


The figure shows the results from estimating the Nelson-Siegel factors using ordinary least squares. We used weekly data of the credit default swap spread for maturities one, three, five, seven and ten years, starting at the 25th February 2004 to the 12th of August 2009. Note that the scale of the figure is in percentage points, not in basis points.

Although the least square results are simple to obtain, their forecasting performance differs strongly from the results attained with the other methods that are presented further in this thesis, as we will see later. It does not outperform the random walk or autoregressive models applied directly to the spreads. See Chapter 7 for more detailed information.

We remark that the described procedure is not as accurate as the procedure described in the subsequent chapters, since more assumptions are needed. The assumption of observed factors is not necessary since methods already exist to estimate models with unobserved factors. Especially for models that are linear in the factors we can obtain results without substantial complexity. In the next chapter we present the model, given by Equations (2.3) and (2.4) in state-space form and implement the Kalman filter, our key instrument to obtain the unobserved factors and estimate the model.

Figure 3 – Nelson-Siegel Loadings and Spread Curves



The figure shows the Nelson-Siegel loadings for slope (top-left) and curvature (top-right). The loading for the level factor is constantly one. The horizontal axes show the time to maturity. Below, there are four fitted spread curves at different dates. The horizontal axes show the time to maturity and the vertical axes show the spread in percent. The grey points represent the data, from which the curve was build.

4 State-Space Models with Gibbs Sampling

In this chapter we estimate the model from Chapter 3 using the Kalman filter and Gibbs sampling. In the following section, we rewrite the model in state-space form. We present the Kalman filter, which we use to generate the factors in Section 4.2. We give a short introduction to Bayesian approaches in Section 4.3, focusing on a Gibbs sampling algorithm. The main part of the section deals with the estimation of the state vectors and the hyperparameters in the model. In the last section we identify the model using the dataset of Commerzbank AG. In the following chapter we add Markov Switches to the model. The combination of Bayesian statistics with autoregressive time series and Markov switches in mean and variance was considered by Albert and Chib (1993).

4.1 State-Space Representation

In economics, state-space models typically deal with dynamic time series that involve unobserved variables, as in our case $\boldsymbol{\beta}_t$. The basic tool to estimate this kind of models is the Kalman Filter. A state-space model can be represented in the following way

$$\mathbf{s}_t = \mathbf{d} + \mathbf{H}\boldsymbol{\beta}_t + \mathbf{v}_t, \quad (3.1)$$

$$\boldsymbol{\beta}_t = \mathbf{c} + \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{u}_t \text{ with } \mathbf{c} = (\mathbf{I}_M - \mathbf{A})\boldsymbol{\mu}, \quad (3.2)$$

where Equation (3.1) is called the *measurement* equation with $\boldsymbol{\beta}_t = (\boldsymbol{\beta}_t^1 \boldsymbol{\beta}_t^2 \dots \boldsymbol{\beta}_t^F)'$ being the state vector of the system with F different factors. In the remaining subsections, we assume that the error terms \mathbf{v}_t follow a multivariate normal distribution, which is given by $\mathbf{v}_t \sim MN(\mathbf{0}_N, \mathbf{R})$ with diagonal covariance matrix \mathbf{R} .

The *transition* equation is given by Equation (3.2). We assume that \mathbf{u}_t follows a multivariate normal distribution $MN(\mathbf{0}_N, \mathbf{Q})$, but we do not restrict \mathbf{Q} to be diagonal. The vector $\boldsymbol{\mu}$ denotes the unconditional mean of $\boldsymbol{\beta}_t$ and the matrix \mathbf{I}_M denotes an identity matrix of size M .

To generate the state vector of the model, it is necessary to assume that $E(\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0$ and $\text{Var}(\boldsymbol{\beta}_0) = \mathbf{P}_0$, with \mathbf{P}_t the covariance matrix of $\boldsymbol{\beta}_t$, i.e. the initial state and covariance

matrix of the state are equal to their expectations. We present the optimal initialization when we introduce the Kalman filter.

We further assume that the error terms \mathbf{u}_t and \mathbf{v}_t are uncorrelated. Furthermore there is no serial correlation in the single elements of the error term vectors which means that the serial covariance matrices $\text{Cov}(\mathbf{u}_t, \mathbf{u}_{t-i})$ and $\text{Cov}(\mathbf{v}_t, \mathbf{v}_{t-i})$ for all $i > 0$ are diagonal (zero elements on the off-diagonal). The latter two restrictions can be relaxed, but it would lead to modifications in the Kalman filter, which is beyond the scope of this thesis.

We already derived the state-space equations of the Nelson Siegel model and the vector autoregressive model for the factors in the previous section – they are given by Equations (2.4) and (2.6). Here is the complete state-space model:

$$\begin{aligned} \textit{Measurement equation:} \quad & \mathbf{s}_t = \mathbf{H}\boldsymbol{\beta}_t + \mathbf{v}_t, \text{ with } \mathbf{v}_t \sim MN(\mathbf{0}_n, \mathbf{R}) \\ \textit{Transition equation:} \quad & \boldsymbol{\beta}_t = \mathbf{c} + \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{u}_t, \text{ with } \mathbf{u}_t \sim MN(\mathbf{0}_3, \mathbf{Q}) \end{aligned}$$

Given a state-space representation together with the assumptions made above, the state vector can be generated by applying the Kalman filter with appropriate initial values for the factors $\boldsymbol{\beta}_0$ and its covariance matrix \mathbf{P}_0 , given the hyperparameters \mathbf{c} , \mathbf{A} , \mathbf{Q} and \mathbf{R} . The matrix \mathbf{H} is known upfront, assuming that we fix λ . Note that the system parameters are time invariant here, i.e. they have the same values for each point in time. We relax this assumption as soon as regime switches are embedded into the model.

4.2 The Kalman Filter

The state vector of a state-space model can be obtained using the Kalman filter, conditional on the hyperparameters. The Kalman filter is used to calculate the state of linear dynamic systems, starting at $t = 1$ and working recursively till the last state vector at time T is reached.

Nowadays there exist several versions of the Kalman filter. The “plain” Kalman filter, as presented in this section, can only be used if the *measurement* equation is linear in the

state vector⁸. There exist several extended versions of the filter, which allow for non-linearities, i.e. the extended Kalman filter, the unscented Kalman filter and for continuous time the Kalman-Bucy filter. For a full derivation of the Kalman filter and the correct initialization, we refer to Hamilton (1994) or Harvey (1993).

To provide a better overview of the functionality of the Kalman filter, we first describe the filter in words and then provide a flowchart with the relevant equations. The steps of the linear filter can be divided into a *prediction* and an *updating* part.

In the prediction step, we use the information from the previous time $t - 1$ to make inference about t using the *transition* equation. Based on the information up to $t - 1$, we predict the state vector for time t , assuming that we do not have the information for the next point in time, t , yet.

In a next step, we assume that the information of the current time step t arrives (in our case, the new information are the spreads for different times to maturity). We use the new information to update the predicted state vector or, more precise, we calculate the residual between the forecasts and the new information and correct for it using the residuals itself and their error covariance matrix.

The likelihood or *log*-likelihood of the parameters can be calculated within the recursion. This is necessary when the estimation of the hyperparameters is done via maximizing the likelihood.

Given a set of values for the hyperparameters, the state vector for each t , starting with time $t = 1$, can be calculated recursively now. The following equations show, in which order the steps have to be taken. The sign $\mathbf{X} \otimes \mathbf{Y}$ denotes the Kronecker product of the matrix \mathbf{X} with matrix \mathbf{Y} and $\text{vec}(\mathbf{X})$ denotes the vectorized version of matrix \mathbf{X} , meaning

⁸ The assumption for a linear model in the factors arises from the derivation of the Kalman filter.

we write each column of the matrix \mathbf{X} below each other. The function $l(\boldsymbol{\theta})$ denotes the \log -likelihood function, dependent on the actual set of parameters. By $|\mathbf{X}|$ we denote the determinant of matrix \mathbf{X} by \mathbf{X}^{-1} the inverse of \mathbf{X} , as is customary. We present the filter in notation of the spread model, given in Section 4.1:

Initialization:

$$\boldsymbol{\beta}_0 = \boldsymbol{\mu} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{c}$$

$$\text{vec}(\mathbf{P}_0) = (\mathbf{I} - \mathbf{A} \otimes \mathbf{A})^{-1}\text{vec}(\mathbf{Q})$$

Prediction:

$$\boldsymbol{\beta}_{t|t-1} = \mathbf{c} + \mathbf{A}\boldsymbol{\beta}_{t-1|t-1} \quad (3.3)$$

$$\mathbf{P}_{t|t-1} = \mathbf{A}\mathbf{P}_{t-1|t-1}\mathbf{A}' + \mathbf{Q} \quad (3.4)$$

Updating:

$$\boldsymbol{\eta}_{t|t-1} = \mathbf{s}_t - \mathbf{s}_{t|t-1} = \mathbf{s}_t - \mathbf{H}\boldsymbol{\beta}_{t|t-1}$$

$$\mathbf{F}_{t|t-1} = \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R}$$

$$\boldsymbol{\beta}_{t|t} = \boldsymbol{\beta}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}'\mathbf{F}_{t|t-1}^{-1}\boldsymbol{\eta}_{t|t-1} \quad (3.5)$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}'\mathbf{F}_{t|t-1}^{-1}\mathbf{H}\mathbf{P}_{t|t-1} \quad (3.6)$$

Likelihood:

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2}\ln((2\pi)^n |\mathbf{F}_{t|t-1}|) - \frac{1}{2}\boldsymbol{\eta}'_{t|t-1}\mathbf{F}_{t|t-1}^{-1}\boldsymbol{\eta}_{t|t-1} \quad (3.7)$$

We start at $t = 1$ and repeat the steps until we reach T , setting $t = t + 1$ after each recursion. We attain a full set of state vectors and the \log -likelihood of the parameters, given the set of spreads and values for the hyperparameters.

The optimal estimates can be obtained by different methods. One approach is the maximization of the likelihood function using an appropriate optimizer. Durbin and Koopman (2001) suggest several methods to solve such optimization problems, amongst others the expectation maximization algorithm. Diebold, Rudebusch and Aruoba (2006)

use the Marquart and Berndt–Hall–Hall–Hausman algorithms. We focus on a Bayesian approach to estimate the model. The next chapter describes how to estimate a state-space model using Gibbs sampling.

4.3 The Gibbs Sampling Approach

4.3.1 Introduction to Bayesian Approaches

The classical and the Bayesian approach are different in many ways. Given a dataset, the focus of the classical approach is on estimates of parameters that have the highest probability to be correct. In our case, we usually have only one dataset to analyze⁹. We make use of maximum likelihood methods to estimate the parameters. In Bayesian inference the parameters are stochastic variables. They are estimated using Bayes theorem.

The concept of probability is a different one in the two approaches. In the classical framework, probability is defined as relative frequency of occurrences, when we repeat a process infinitely often, whereas in Bayesian inference, probability can be seen as the confidence that certain events occur.

In contrast to the classical approach, the Bayesian estimation does not attempt to find the best fit estimates to the dataset. Instead we take the prior knowledge we have about the parameters and update this information using the data we observe. This especially has advantages when we suffer from small datasets, where we are afraid that observed data might not be representative enough.

Suppose, for example, we have two chess players, playing against each other four times in a row. Let us assume our prior belief is that they play equally well. Using a Bernoulli distribution with q the probability that player one wins, we can estimate q . If player one wins four times, the classical approach produces an estimate of q equal to one, whereas the Bayesian approach produces a probability dependent on the prior and therefore lower

⁹ In a laboratory, you can repeat tests, always getting new datasets. You can estimate their parameters for each sample and finally get a distribution for the parameters. In economics that is not the case.

than one. For small datasets it might therefore be useful to implement Bayesian estimation instead of using a classical approach.

Bayesian statistics gained more and more popularity over the recent decades. There are two favored arguments, why classical approaches are preferred over Bayesian approaches. The first argument affects the choice of a prior (subjective) in the Bayesian approach, although one can choose a non-informative prior, which means that the prior sets no preference for any value of the parameters.

The second argument is based on the derivations of (complex) posterior densities (the concept will be explained throughout this chapter). Since they can also be computed numerically and the development of computers and algorithms for numerical computations improved tremendously, this problem is not significant anymore (for most of the cases). This as well explains the gaining popularity of Bayesian statistics nowadays.

We assume that the parameters in our model are stochastic variables with probability distribution $p(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the set of parameters. Let \mathbf{y} denote the observable data. In the remaining sections, \mathbf{y} is the credit default swap spread for different maturities.

Let us first recall the Bayes' theorem:

$$p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y} | \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \mathbf{y})p(\mathbf{y})$$

and therefore

$$p(\boldsymbol{\theta} | \mathbf{y}) = \frac{p(\mathbf{y} | \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \text{ or } p(\boldsymbol{\theta} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta})p(\boldsymbol{\theta}). \quad (3.8)$$

Since $p(\mathbf{y})$ has no operational significance (it does not depend on $\boldsymbol{\theta}$), it will not change the posterior density for different sets of parameters. That is the reason why we can write the proportional relationship in Equation (3.8). Note also that $p(\mathbf{y} | \boldsymbol{\theta})$ is nothing else than the likelihood function $L(\boldsymbol{\theta} / \mathbf{y})$ of the parameters $\boldsymbol{\theta}$ given the observations \mathbf{y} . Defining a prior belief via the distribution of the parameters $p(\boldsymbol{\theta})$, we are able to calculate a posterior

distribution, given the set of observations. Note that different priors $p(\boldsymbol{\theta})$ lead to different posterior distributions.

4.3.2 The Benefit of Gibbs Sampling

Implementing the approach described above for models with more than one unknown parameter, especially when they have different priors, can lead to problems. We need to derive the marginal and joint prior distribution first to get the joint posterior distribution of the parameters.

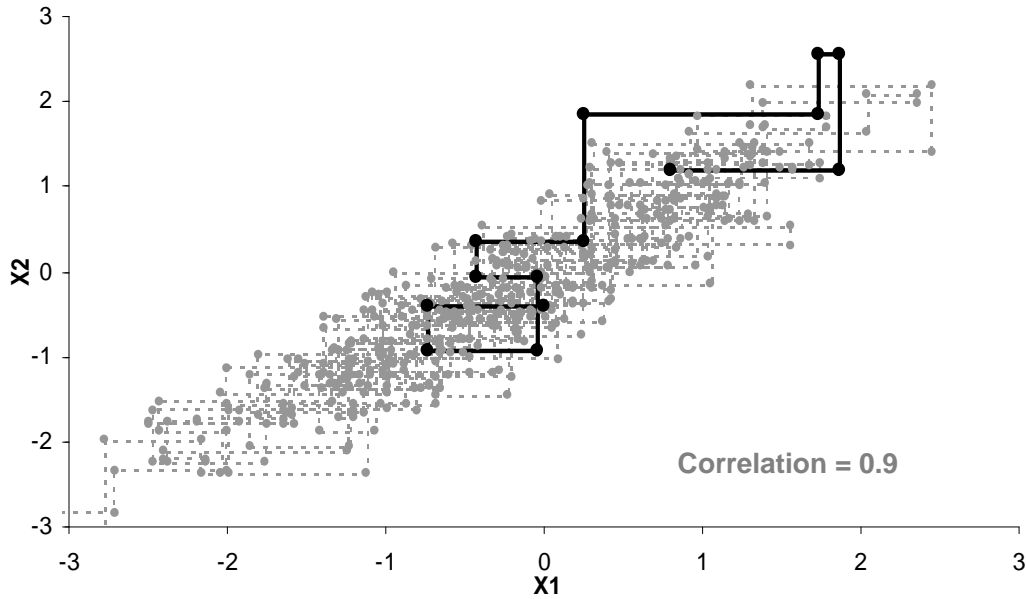
After calculating the joint posterior density of the unknowns using Equation (3.8), we still need the marginal posterior distributions of the parameters to make inference on them. The straightforward solution involves integration of the joint distribution of the parameters, where in many cases analytical solutions do not exist. Gibbs Sampling is a Bayesian approach used to approximate joint and marginal distributions from their conditional distributions. That way we avoid the mentioned integration problem. In a first step we choose priors for the parameters. We follow the ideas of Kim and Nelson (1999) and Sugita (2008) and choose natural conjugate priors for each of the parameters, conditional on all other parameters. By natural conjugate prior we understand a prior that, multiplied with the likelihood function, has the same posterior distribution as the prior (conditional on the other parameters). The conditional distributions are used to obtain the marginal distributions. The Gibbs sampling algorithm can be divided in five steps:

1. Initialize the set of parameters $\boldsymbol{\theta}$, with M the number of parameters.
2. Draw the first parameter θ_1 from its posterior distribution, conditional on the other parameters $\theta_2, \theta_3 \dots \theta_M$ and save the value.
3. Generate a draw of the second parameter θ_2 conditional on the draw of θ_1 and the other parameters $\theta_3, \theta_4 \dots \theta_M$.
4. Do this for all parameters, save the values and repeat the steps 1 to 3 J times.
5. Discard the first draws, since they might depend strongly on the initialization of the first step. Take the average over the remaining draws to get an estimate.

To illustrate how this approach captures the marginal distributions, we consider a simple example. Figure 4 shows the results from Gibbs sampling on two normally distributed

random variables with high correlation. The generated marginal and posterior distributions converge at an exponential rate to the desired distributions. For a proof, see Geman and Geman (1984). Given the conditional distributions, we can approximate the marginal distributions of the random variables without integrating the joint distribution over one of the parameters.

Figure 4 – A Gibbs Sampling Illustration



Result from 250 Gibbs sampling rounds on two normally distributed random variables X_1 and X_2 with correlation 0.9. The dark line highlights the first ten draws, initializing X_1 at 0.

In the state-space model, we analyze throughout this thesis, we deal with much more complex situations. We consider a multivariate normal distributions and a multivariate representation of the gamma distribution, a Wishart distribution. In the following sections we describe which prior distributions we need to choose and how we approximate the marginal distributions. The mean of the generated values represents the estimators for the parameters.

Another issue will be the generation of the unobserved state vector using Gibbs sampling, since we have to generate a $3 \times T$ matrix there. There are several ways to generate the state vector. We use multimove Gibbs sampling, as suggested by Carter and Kohn (1994). The latter will be described Section 4.3.6.

4.3.3 Prior Distribution for the Transition Matrix \mathbf{A}

The assumption of independence between the error terms in the *measurement* and the *transition* equations and the conditional drawing of the parameters allow us to treat them separately. We start by estimating the *transition* equation. To derive the posterior distribution of \mathbf{A} , we first need to rewrite the *transition* equation in concise matrix notation as discussed in Section 3.3. We express Equation (3.2) for all time points as¹⁰

$$\boldsymbol{\beta}_c = \boldsymbol{\beta}_c^L \mathbf{A}_c' + \mathbf{u}. \quad (3.9)$$

The likelihood function of \mathbf{A}_c , conditional on $\boldsymbol{\beta}$ and \mathbf{Q} in Equation (3.9) is proportional to

$$l(\mathbf{A}_c, \mathbf{Q} | \boldsymbol{\beta}) \propto |\mathbf{Q}|^{-T/2} \exp \left[-\frac{1}{2} \text{tr} \{ \mathbf{Q}^{-1} (\boldsymbol{\beta}_c - \boldsymbol{\beta}_c^L \mathbf{A}_c')' (\boldsymbol{\beta}_c - \boldsymbol{\beta}_c^L \mathbf{A}_c') \} \right], \quad (3.10)$$

A non informative prior for \mathbf{A}_c is given by the multivariate normal distribution, where the conditional posterior density of \mathbf{A}_c can be written as

$$\text{vec}(\mathbf{A}_c) | \mathbf{Q}, \boldsymbol{\beta} \sim MN(\text{vec}(\mathbf{A}_{c1}), \boldsymbol{\Phi}_1) \quad (3.11)$$

and where $\boldsymbol{\Phi}_1$ denotes the covariance matrix of $\text{vec}(\mathbf{A}_c)$. We can derive \mathbf{A}_{c1} and $\boldsymbol{\Phi}_1$ using the results from section 4.3.1. We insert the multivariate normal prior and the likelihood, given by Equation (3.10), into Equation (3.8) and rearrange terms. \mathbf{A}_{c1} and $\boldsymbol{\Phi}_1$ are then given by

$$\boldsymbol{\Phi}_1 = [\boldsymbol{\Phi}_0^{-1} + \mathbf{Q}^{-1} \otimes (\boldsymbol{\beta}_c^{L'} \boldsymbol{\beta}_c^L)]^{-1}, \quad (3.12)$$

$$\text{and } \text{vec}(\mathbf{A}_{c1}) = \boldsymbol{\Phi}_1 [\boldsymbol{\Phi}_0^{-1} \text{vec}(\mathbf{A}_{c0}) + (\mathbf{Q} \otimes \mathbf{I})^{-1} \text{vec}(\boldsymbol{\beta}_c^{L'} \boldsymbol{\beta}_c)]. \quad (3.13)$$

The parameters \mathbf{A}_{c0} and $\boldsymbol{\Phi}_0$ have to be set before we draw the parameters \mathbf{A}_c and $\boldsymbol{\Phi}_1$. They are prior beliefs about the parameters and their covariance matrix.¹¹

For a derivation of the results in Equations (3.12) and (3.13), see Sugita (2008).

¹⁰ Note that the result of the following steps will deliver \mathbf{A}_c' and not \mathbf{A} . Before running the Kalman Filter again, we need to transpose the matrix and split it into \mathbf{A} and \mathbf{c} again.

¹¹ The initial matrix $\boldsymbol{\Phi}_0$ should have high values on the diagonal, if we are unsure about our guess for \mathbf{A}_{c0} . This ensures that we are able to capture the joint distribution of the elements in \mathbf{A}_c .

Conditional on \mathbf{A}_c and $\boldsymbol{\beta}$, our next step is the derivation of \mathbf{Q} . Here we face another problem, since the elements on the diagonal of \mathbf{Q} are restricted to values larger than zero.

4.3.4 Prior Distribution for the Covariance Matrix \mathbf{Q}

The error covariance matrix \mathbf{Q} in our model is a 3×3 matrix. Since it is a covariance matrix, we should choose the natural conjugate prior (and therefore posterior distribution) in such a way, that the elements on the diagonal are larger or equal to zero. To achieve this goal, we pick an inverse Wishart distribution as prior for the conditional distribution of \mathbf{Q} . The inverse Wishart distribution is the multivariate case of an inverse gamma distribution. It produces a symmetric matrix with positive values on the diagonal. Its density function can be written as

$$p(\mathbf{Q} | \boldsymbol{\Psi}, m) = \frac{|\boldsymbol{\Psi}|^{m/2} |\mathbf{Q}|^{-(m+p+1)/2} \exp(-tr(\boldsymbol{\Psi}\mathbf{Q}^{-1})/2)}{2^{mp/2} \Gamma_p(m/2)}. \quad (3.14)$$

The $p \times p$ scale matrix $\boldsymbol{\Psi}$ has the same size as \mathbf{Q} , respectively $p = 3$. The parameter m denotes the degrees of freedom and Γ denotes the multivariate gamma function. Since the density in Equation (3.14) is a natural conjugate prior, the posterior density of \mathbf{Q} , conditional on all other parameters, follows the inverse Wishart distribution¹²

$$\mathbf{Q} | \mathbf{A}_c, \boldsymbol{\beta} \sim IW(\boldsymbol{\Psi}_1, m_1). \quad (3.15)$$

Deriving the conditional posterior distribution of \mathbf{Q} , we get the following changes in the scale matrix and the degrees of freedom¹³

$$\boldsymbol{\Psi}_1 = (\boldsymbol{\beta}_c - \boldsymbol{\beta}_c^L \mathbf{A}_c)' (\boldsymbol{\beta}_c - \boldsymbol{\beta}_c^L \mathbf{A}_c) + \boldsymbol{\Psi}_0, \quad (3.16)$$

$$\text{and } m_1 = m_0 + T. \quad (3.17)$$

We choose a factor model in order to reduce the number of parameters to be estimated. This worked perfectly for \mathbf{c} , \mathbf{A} and \mathbf{Q} . We have 18 unknown parameters, independent of

¹² Most programming languages do not offer the possibility to draw from an inverse Wishart distribution. Appendix A shows how we solve this issue using standard normal and chi-square random numbers.

¹³ For a detailed derivation of (3.16) and (3.17) see again Sugita (2008).

the number of maturities N we include in our analysis. This number changes with the error covariance matrix \mathbf{R} , dependent on the number of maturities N .

4.3.5 Prior Distribution for the Covariance Matrix \mathbf{R}

The previous sections described the way to generate draws of the parameters in the *transition* equation. Now we focus on the *measurement* equation. Since the matrix of loadings, \mathbf{H} , given λ , is already known and our measurement equation does not contain a constant, we only need to estimate the covariance matrix of the errors, \mathbf{R} . Since the number of elements in \mathbf{R} depends on the size of \mathbf{s}_t , the number of parameters increases significantly with the number of spreads we include for different maturities. For five different maturities in \mathbf{s}_t , we have a symmetric 5×5 matrix with 15 different parameters. In the style of Diebold, Rudebusch and Aruoba (2006), we assume that \mathbf{R} is diagonal. This can reduce the number of parameters significantly¹⁴. The matrix \mathbf{R} is given by

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_N \end{pmatrix}. \quad (3.18)$$

The elements of \mathbf{R} , as before the elements of the diagonal of \mathbf{Q} , should be strictly positive, since they denote variances of the error terms \mathbf{v}_t . The natural conjugate prior for each variance σ_n , conditional on $\boldsymbol{\beta}$ that assures positive elements for the variances is an inverse gamma distribution with density function

$$p(\sigma_i | \kappa, \omega) = \frac{\omega^\kappa}{\Gamma(\kappa)} (1/\sigma_n)^{\kappa+1} \exp(-\omega/\sigma_n). \quad (3.19)$$

Note that there are different representations of the gamma distribution and the inverse gamma distribution. It can be parameterized in terms of a scale parameter or an inverse

¹⁴ In case we would use ten different maturities, the number of additional parameters reduces from 55 to 10. Nevertheless, if we do not make this assumption, the generation of draws of \mathbf{R} will not be different from the steps taken in the previous section for \mathbf{Q} . The natural conjugate prior would again be an inverse Wishart distribution.

scale parameter, also referred to as rate parameter. The shape parameter does not change for the two representations. We make use of the scale parameter notation.

The conditional posterior density and the parameters of the posterior density are given by

$$\sigma_n | \boldsymbol{\beta} \sim IG\left(\frac{\kappa_1}{2}, \frac{\omega_1}{2}\right), \quad (3.20)$$

with the two parameters

$$\text{and } \omega_1 = (\mathbf{S}_n - \boldsymbol{\beta}\mathbf{H}'_n)'(\mathbf{S}_n - \boldsymbol{\beta}\mathbf{H}'_n) + \omega_0, \quad (3.21)$$

$$\kappa_1 = \kappa_0 + T \quad (3.22)$$

where κ denotes the shape of the gamma distribution, ω denotes the scale of the gamma distribution and \mathbf{S} is the collection of spreads for all points in time $(s_0 s_1 \dots s_T)'$. The index n in \mathbf{S}_n and \mathbf{H}_n is a column index, implying that we only take the n -th column of the matrices. For further information about the derivation of the univariate results, see Kim and Nelson (1999, § 7.4). Note also that this result is just a special case of the inverse Wishart distribution, where the dimension of \mathbf{Q} equals a 1×1 matrix.

4.3.6 Generating the State Matrix $\boldsymbol{\beta}$

We know how to generate draws of the hyperparameters of our model. But the distributions of the hyperparameters are also conditional on $\boldsymbol{\beta}$, the unobserved state vector of the model. We consider the state vectors, as the hyperparameters before, to be random variables whose natural conjugate prior given by a multivariate normal distribution with covariance matrix \mathbf{P} .

This state-space model is a linear Gaussian model given the observations \mathbf{S} . Multimove Gibbs sampling, developed by Carter and Kohn (1994), uses the following result, which can be derived from the Markov property of $\boldsymbol{\beta}_t$

$$p(\boldsymbol{\beta} | \mathbf{S}) = p(\boldsymbol{\beta}_T | \mathbf{S}) \prod_{t=1}^{T-1} p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t+1}, \mathbf{S}). \quad (3.23)$$

That means we can create the whole state vector recursively by first generating $\boldsymbol{\beta}_T$ conditional on \mathbf{S} . Given $\boldsymbol{\beta}_{t+1}$ and \mathbf{S} , we know that the density of $\boldsymbol{\beta}_t$ is again normally

distributed. Going back to the notation we used for the state vector in the Kalman filter for predicted variables $\boldsymbol{\beta}_{t|t-1}$ and $\boldsymbol{\beta}_{t|t}$, we can write the densities for $\boldsymbol{\beta}_T | \mathbf{S}$ and $\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t+1}, \mathbf{S}$ as

$$\boldsymbol{\beta}_T | \mathbf{S} \sim N(\boldsymbol{\beta}_{T|T}, \mathbf{P}_{T|T}) \quad (3.24)$$

$$\text{and } \boldsymbol{\beta}_t | \boldsymbol{\beta}_{t+1}, \mathbf{S} \sim N(\boldsymbol{\beta}_{t|t, \boldsymbol{\beta}_{t+1}}, \mathbf{P}_{t|t, \boldsymbol{\beta}_{t+1}}). \quad (3.25)$$

We need to derive the mean and the variance of distribution (3.25) in the next step. Receiving a draw for $\boldsymbol{\beta}_T$ is straight forward. We just need to run the Kalman filter once to provide $\boldsymbol{\beta}_{T|T}$ and $\mathbf{P}_{T|T}$. We also need to save the path of $\boldsymbol{\beta}_{t|t}$ and $\mathbf{P}_{t|t}$ for all t , since these values are necessary to draw $\boldsymbol{\beta}_t$ for $t < T$.

We first draw $\boldsymbol{\beta}_T$ from (3.24). The next step is drawing $\boldsymbol{\beta}_{T-1}$, conditional on the draw of $\boldsymbol{\beta}_T$ and \mathbf{S} . Hence, we need to derive $\boldsymbol{\beta}_{T-1|T-1, \boldsymbol{\beta}_T}$ and $\mathbf{P}_{T-1|T-1, \boldsymbol{\beta}_T}$, the mean and covariance matrix of the distribution. They can be derived by rewriting the updating part in the Kalman filter. The final expressions are given by

$$\boldsymbol{\beta}_{T-1|T-1, \boldsymbol{\beta}_T} = \boldsymbol{\beta}_{T|T} + \mathbf{P}_{T|T} \mathbf{A}' (\mathbf{A} \mathbf{P}_{T|T} \mathbf{A}' + \mathbf{Q})^{-1} (\boldsymbol{\beta}_T - \boldsymbol{\mu} - \mathbf{A} \boldsymbol{\beta}_{T|T}) \quad (3.26)$$

$$\text{and } \mathbf{P}_{T-1|T-1, \boldsymbol{\beta}_T} = \mathbf{P}_{T|T} - \mathbf{P}_{T|T} \mathbf{A}' (\mathbf{A} \mathbf{P}_{T|T} \mathbf{A}' + \mathbf{Q})^{-1} \mathbf{P}_{T|T}. \quad (3.27)$$

For more information, see Kim and Nelson (1999, § 8.1). We now have the instruments to draw $\boldsymbol{\beta}_{T-1}$, using (3.25) with (3.26) and (3.27). Setting $T = T - 1$, we can proceed. We repeat the steps until we obtain the whole set of factors $\boldsymbol{\beta}$. Note that for generating $\boldsymbol{\beta}$, we need to know the hyperparameters. Their initialization might be important, since the Kalman filter does not run for all values. For instance the initialization of the filter does no work if $\mathbf{I} - \mathbf{A}$ is not invertible.

Finally we have all building blocks to run the Gibbs sampling algorithm. Note that the order in which we start the algorithm does not affect the results, at least not after a sufficient number of runs. Nevertheless it makes sense to initialize the hyperparameters first and start sampling the state vector, because initializing the state vector for each point in time needs more effort (since it contains much more values). In most cases, it is very difficult to come up with a good guess of the unobserved variables. The next part describes in short, how we build the algorithm.

4.3.7 Generating the Parameters, Conditional on β

We start generating the collection of state vectors β as described in the previous section. Afterwards we can carry on generating the hyperparameters, step by step. We start with the transition matrix \mathbf{A} . We need to set \mathbf{A}_0 and the covariance matrix Φ_0 of $\text{vec}(\mathbf{A})$. As mentioned before, we should choose high values for the elements of the covariance matrix, the more we are unsure about our guess of \mathbf{A}_0 , especially for the diagonal. We should also take care that we chose the matrix in such a way, that it is invertible, since this is necessary to calculate Φ_1 .

Using a VAR model, we imply that our process is covariance stationary, which means that the first and the second moments $E(\beta_t)$ and $\text{Cov}(\beta_t)$ of the factors are independent of the date t . This restriction may not be fulfilled for all draws of the parameters. This may lead to problems in other building blocks of the algorithm, especially in the generation of the state vector β_t , e.g. we receive covariance matrices which are not Hermitian or have negative eigenvalues (problems may arise drawing from a multivariate normal distribution, since we need a square root of the covariance matrix). Discarding these draws, we also prevent a bias in the result when averaging over all draws¹⁵.

Concerning the choice of \mathbf{A}_0 , \mathbf{c}_0 , Φ_0 , Ψ_0 , m_0 , κ_0 and ω_0 , there exist several approaches.¹⁶ The choice can significantly influence the estimation of the models. Litterman (1986) suggests picking the priors such that the time series is impossible to forecast. He chose the prior covariance matrix to diagonal, where in addition Zellner (1971) suggests a solution, where this assumption is not necessary. Jorion (1986) also suggests an approach, based on the underlying dataset.

The choice of the prior is beyond the scope of our research. In this thesis, we use a trivial initialization, setting the values to zeros and ones where possible and high values for Ψ_0 , such that we make up for the trivial initialization of \mathbf{A}_0 and \mathbf{c}_0 .

¹⁵ To test for stationarity, we can apply a unit root test. See also Hamilton (1994) for the conditions of covariance stationarity in vector autoregressive models.

¹⁶ Since the values are (with some restrictions) free to choose, it is not necessary to make use of those approaches.

In a next step, we generate \mathbf{Q} . Here we need to set the scale matrix Ψ_0 . As described in **Appendix A**, we need a square root of Ψ_1 (e.g. Cholesky decomposition). Therefore we need to make sure that Ψ_1 will be positive definite (Hermitian for Cholesky decomposition). The initial degrees of freedom m_0 need to be set as well. We generate \mathbf{Q} from the inverse Wishart distribution.

The last matrix to generate is \mathbf{R} , the covariance matrix of the error terms in the *measurement* equation. Here we need to initialize the shape κ_0 and scale ω_0 . Then we draw each σ_n for $n = 1, 2 \dots N$.

We repeat the generation of the state vectors and hyperparameters and save the values for the hyperparameters until we consider our sample to be large enough, starting at $j = 0$

1. Generate a draw of the state vectors β_j :

$$\beta_{j+1} \leftarrow \mathbf{c}_j, \mathbf{A}_j, \mathbf{Q}_j, \mathbf{R}_j \quad \text{Equations (3.24) - (3.27)}$$

▼
2. Generate a draw of the parameters \mathbf{A} and \mathbf{c} :

$$\mathbf{c}_{j+1}, \mathbf{A}_{j+1} \leftarrow \beta_{j+1}, \mathbf{Q}_j, \mathbf{R}_j \quad \text{Equations (3.11) - (3.13)}$$

▼
3. Generate a draw of the error covariance matrix \mathbf{Q} :

$$\mathbf{Q}_{j+1} \leftarrow \beta_{j+1}, \mathbf{c}_{j+1}, \mathbf{A}_{j+1}, \mathbf{R}_j \quad \text{Equations (3.15) - (3.17)}$$

▼
4. Generate a draw of the error covariance matrix \mathbf{R} :

$$\mathbf{R}_{j+1} \leftarrow \beta_{j+1}, \mathbf{c}_{j+1}, \mathbf{A}_{j+1}, \mathbf{Q}_{j+1} \quad \text{Equations (3.20) - (3.22)}$$

▼
5. Repeat steps 1 to 4 until we have a sufficient number of draws:

$$j = j + 1.$$

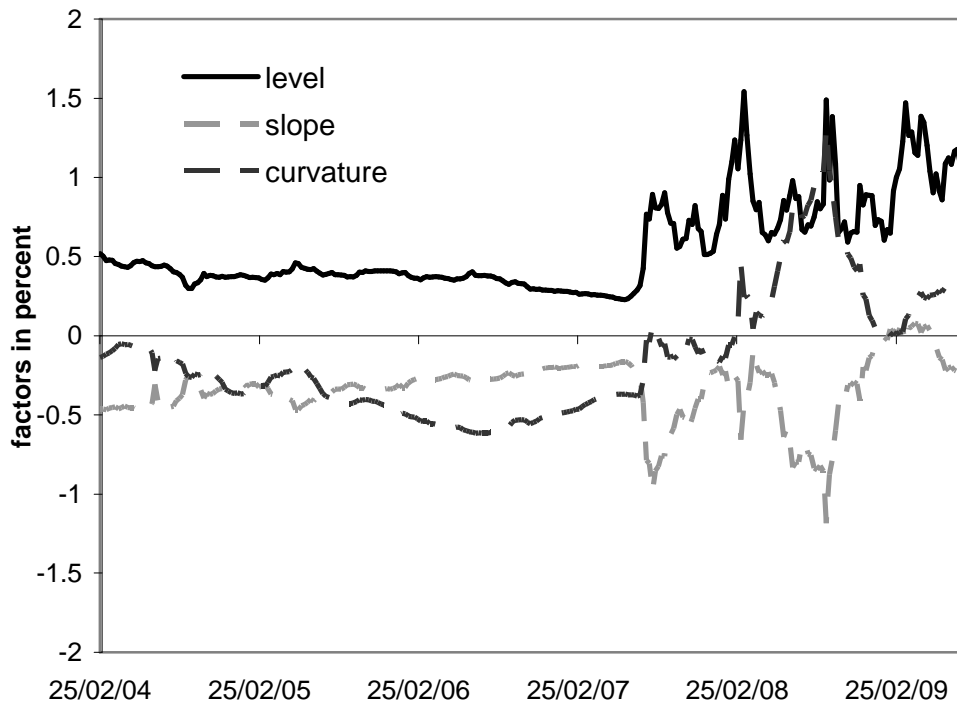
Again we discard the first draws, since they strongly depend on the initialization of the hyperparameters. From the remaining draws we take the average to get estimates for the hyperparameters.

4.4 Results from the Model without Regime Switching

We implement the algorithm for the Commerzbank dataset, as discussed in Section 4.3.7. We trivially choose the scale matrix Ψ_0 to be an identity matrix and the degrees of freedom κ_0 and m_0 to be zero. The elements of \mathbf{A}_0 and \mathbf{c}_0 where also set to zero and the covariance matrix of $\text{vec}(\mathbf{A}_c)$ to a diagonal matrix with 10^3 on the diagonal.

We run the filter, discard the first 2.000 draws and take the average over the next 10.000 draws. The results are shown in Figure 5.

Figure 5 – The State-Space Gibbs Sampling Nelson-Siegel Factors



The figure shows the results from estimating the Nelson-Siegel factors using the Kalman Filter and Gibbs Sampling for the estimation, again using the Commerzbank dataset. They were obtained after running the algorithm 12.000 times, averaging over the last 10.000 outputs.

Comparing Figure 2 and Figure 5, we observe small differences for the time series of the level factor. However, the factors slope and curvature differ strongly from the results in Section 3.4, not only concerning their values, but also their correlations. Figure 6 shows the correlation for the two estimation methods, calculated using a moving window with data for twelve weeks. We take the state vectors β_t for $t = 1, 2 \dots 12$ and calculate their correlation. The next correlations are calculated using the state vectors β_t for $t = 2, 3 \dots 13$ and so on.

As mentioned before, the correlation between the factors for credit default swap spreads takes very high values (positive and negative), compared to the correlation of the factors for yields. Especially the correlation between level and slope is constantly close to minus one, except for a few dates. We observe that the correlation between level and slope does not differ strongly between least square estimation and the estimation via Gibbs sampling, but for the other correlations, especially the correlation between level and curvature, the difference is huge.

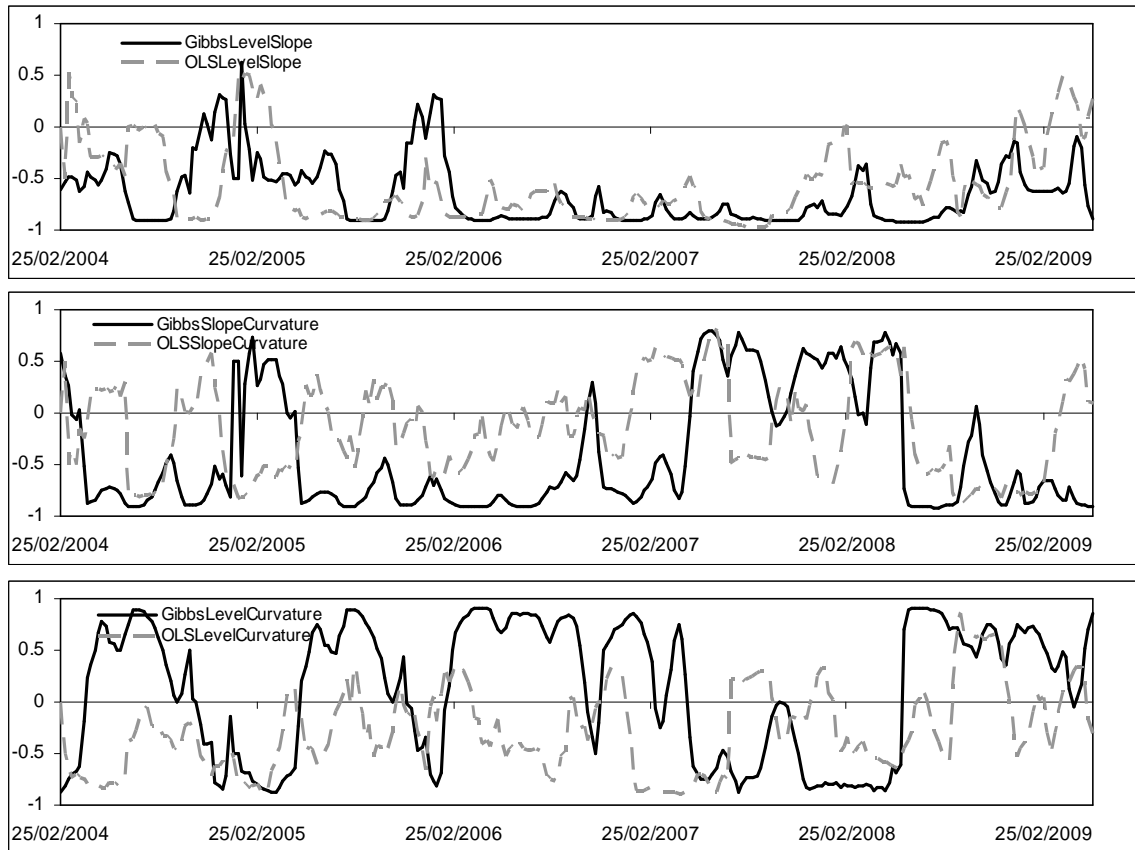
The parameters, resulting from the Gibbs sampling algorithm, averaged over the last 10.000 iterations are given by

$$\mathbf{A} = \begin{pmatrix} 0.850 & 0.137 & -0.099 \\ 0.025 & 0.921 & -0.017 \\ -0.001 & 0.049 & 0.990 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 0.142 \\ -0.043 \\ 0.015 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0.009 & -0.004 & 0.002 \\ -0.004 & 0.005 & -0.004 \\ 0.002 & -0.004 & 0.006 \end{pmatrix},$$

$$\mathbf{R}_{diag} = (0.00033, 0.00179, 0.00015, 0.00012, 0.00009).$$

At this point we should mention again, that the results obtained from Gibbs sampling depend on the prior distributions and the values we choose for \mathbf{A}_0 , \mathbf{c}_0 , Φ_0 , Ψ_0 , m_0 , κ_0 and ω_0 . This is likely to be also (next to the fact that we used the Kalman filter to back out the state vector) a reason for deviations from the results obtained in Section 3.4.

Figure 6 – Least Squares and Gibbs Sampling Nelson-Siegel Factor Correlations



The figure shows the cross-correlations of the Nelson Siegel factors of the Commerzbank dataset. The correlation is calculated using a moving window over twelve weeks, which explains the very high correlations when a jump occurs in two of the factors at the same time.

Regarding Figure 2 or Figure 5 again, we observe that the credit default swap spreads and therefore the Nelson-Siegel factors are not covariance-stationary over time. We do not only observe a jump in the level of the spreads, but also their slope and curvature change significantly. Furthermore level, slope and curvature get more volatile after the jump. The accuracy of our linear model suffers from such jumps, since we assume covariance stationarity for the $VAR(1)$ model. We therefore introduce regime switches as a possible solution to this issue in the following chapter.

5 Regime Switching

Observing the movement of credit default swap spreads for different reference entities over the last couple of years, we see that a low level of the spread comes together with a low volatility of the spread. At the same time, we observe periods with high and highly volatile spreads. This behavior is typical for many markets, e.g. jumps in the equity market usually come together with a huge increase in their volatilities.

We want to account for this additional information in the state-space model by allowing for more flexibility. This results in an increase of the parameters in the model and also in more computational effort.

In this chapter we add regime switching to our model by allowing for shifts in the level and error covariance matrix of the factors. We consider two different regimes for credit default swap spreads to allow for low levels with low volatility and vice versa. We adjust the estimation steps from the previous chapter to use regime switching.

In Section 5.1, we present the Hamilton filter, which provides us with the probabilities to be in a certain regime. In Section 5.2 we show how we generate the hyperparameters and the regime vector. For the latter we make use of multimove Gibbs sampling. Finally, we generate the transition matrix, which contains the probabilities to move from one regime to another.

5.1 Markov-Switching Model and the Hamilton Filter

There are several models applicable, e.g. independent switching models, which are relatively simple to implement. We consider a more complex switching model, i.e. a first order Markov switching model with two different regimes. This means that the probability to be located in one regime only depends on the previous regime and the observations of the spreads. The regime at time t is denoted with r_t . It is equal to zero, if we are in regime zero at time t and equal to one, when we are in regime one.

Allowing for more regimes is possible, but increases the number of parameters. Also the parameters, for which we allow for switches, influence the number of parameters to estimate. We decide to allow for switches in the level \mathbf{c} and the error covariance matrix \mathbf{Q} of the factors, from now denoted as $\mathbf{c}(r_t)$ and $\mathbf{Q}(r_t)$. The other hyperparameters do not change in their notation.

We have to use a filter method to obtain the whole regime vector $\mathbf{r} = (r_1, r_2 \dots r_T)'$, which contains the regimes for all t from 1 to T . The transition matrix \mathbf{K} contains the probabilities of shifts from one point in time to the next. K_{ij} is the probability to move from regime i to regime j . Note that the rows of the matrix have to sum up to one, which means that the matrix does not contain four, but only two unknown probabilities, that we have to estimate. The matrix can be written as

$$\mathbf{K} = \begin{pmatrix} K_{0|0} & K_{0|1} \\ K_{1|0} & K_{1|1} \end{pmatrix} \text{ or } \mathbf{K} = \begin{pmatrix} K_{0|0} & 1 - K_{0|0} \\ 1 - K_{1|1} & K_{1|1} \end{pmatrix}. \quad (4.1)$$

The filter method that we use, does not deliver the regime itself, but the probability to be in a certain regime at time t , given the data up to time t and the transition probability matrix \mathbf{K} . Note that the matrix \mathbf{K} does not depend on the time or on the observations.

The following chart shows the steps of the Hamilton (1989) basic filter, starting with time t equal to 1. To get the filter started, we first need to initialize the steady state probability $p(r_0)$, the probability to be in regime zero or regime one at $t = 0$. Then we move from probabilities that are conditional on the set of observations and previous regime to probabilities that are only conditional on the set of observations. Finally we obtain the joint probability of r_t and r_{t-1} , conditional on the observations up to that point in time.

Next to the initialization of the filter and the steps to get from the conditional probability $p(r_t, r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t-1})$ to $p(r_t | \boldsymbol{\beta}_{0 \rightarrow t})$, the probability of r_t and r_{t-1} , conditional on all information up to time t , the flowchart contains the likelihood function (4.5) of the parameters.

Note that we do not make use of the likelihood, since we use Gibbs sampling to estimate the model. For maximum likelihood estimation techniques, the function is crucial. It only depends on the cumulative probability of the data $\boldsymbol{\beta}_t$ at time t , conditional on the data of

the previously available data $\boldsymbol{\beta}_{0 \rightarrow t-1}$. $\boldsymbol{\beta}_{0 \rightarrow t-1}$ denotes the data (the state vectors) from time zero to time t .

Initialization:
$$p(r_0) = \left(\begin{pmatrix} \mathbf{I}_2 - \mathbf{K} \\ \mathbf{1}_2 \end{pmatrix}' \begin{pmatrix} \mathbf{I}_2 - \mathbf{K} \\ \mathbf{1}_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{I}_2 - \mathbf{K} \\ \mathbf{1}_2 \end{pmatrix}' \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (4.2)$$

Filter Steps:
$$p(r_t, r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t-1}) = p(r_t | r_{t-1}) p(r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t-1}) \quad (4.3)$$

$$\boldsymbol{\varphi}(r_t) = \boldsymbol{\beta}_t - \mathbf{c}(r_t) - \mathbf{A}\boldsymbol{\beta}_{t-1}$$

$$p(\boldsymbol{\beta}_t | r_t, r_{t-1}, \boldsymbol{\beta}_{0 \rightarrow t-1}) = \frac{1}{(2\pi)^{N/2} |\mathbf{Q}(r_t)|} \exp\left(-\frac{1}{2} \boldsymbol{\varphi}(r_t)' \mathbf{Q}(r_t)^{-1} \boldsymbol{\varphi}(r_t)\right)$$

$$p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{0 \rightarrow t-1}) = \sum_{r_t} \sum_{r_{t-1}} p(\boldsymbol{\beta}_t | r_t, r_{t-1}, \boldsymbol{\beta}_{0 \rightarrow t-1}) p(r_t, r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t-1})$$

$$p(r_t, r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t}) = \frac{p(\boldsymbol{\beta}_t | r_t, r_{t-1}, \boldsymbol{\beta}_{0 \rightarrow t-1}) p(r_t, r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t-1})}{p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{0 \rightarrow t-1})}$$

$$p(r_t | \boldsymbol{\beta}_{0 \rightarrow t}) = \sum_{r_{t-1}} p(r_t, r_{t-1} | \boldsymbol{\beta}_{0 \rightarrow t}) \quad (4.4)$$

Likelihood:
$$l(\theta) = l(\theta) + \ln(p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{0 \rightarrow t-1})) \quad (4.5)$$

N is the number of factors. In our case N equals three. A derivation for the univariate case of these results can be found in Kim and Nelson (1999, § 4.2). The only difference to the multivariate case is the probability distribution function of the normal distribution.

Replacing the simple density function by the multivariate version, the steps to be taken are equal.

Note the similarity between the steps for the estimation of the state-space model and now the regime switching model. We first run a filter method to obtain the unobserved regime vector and then, conditional on these variables we start drawing the parameters, including the transition matrix \mathbf{K} . The following steps describe the building blocks for a successful estimation in a VAR model with regime switching using Gibbs sampling.

5.2 Estimation using Gibbs Sampling

As in the previous sections for the estimation of the state-space model, we explain how to obtain and draw from the posterior densities of the parameters, and afterwards how to generate the regime vector and the transition matrix \mathbf{K} . We consider the case for switches in the vector of constants \mathbf{c} and the covariance matrix \mathbf{Q} .

The switches only occur in the transition equation. Since the error terms of the measurement and the transition equation are independent, we can treat the transition equation separately. Hence the building blocks of the regime switching estimation are shown in a vector autoregressive model of order one, the same given by Equation (3.2), assuming that the state vectors are observed. Later we draw the regime vector conditional on the state vector; therefore this assumption does not affect our building blocks.

We first focus on the estimation of $\mathbf{c}(r_t)$ and \mathbf{A} . Both can be treated together. Further we take a look at the estimation of the covariance matrix \mathbf{Q} for different regimes in Section 5.2.2. In Section 5.2.3 we show a way to generate the transition matrix \mathbf{K} . In the last section of this chapter, we generate the regime vector \mathbf{r} using multimove Gibbs sampling, as implemented for generating the state vector in the previous chapter.

5.2.1 Estimating $\mathbf{c}(r_t)$ and \mathbf{A}

The estimation of \mathbf{c} and \mathbf{A} in Section 4.3.3 cannot be applied here, since we allow for shifts in the vector \mathbf{c} . We again choose the multivariate normal prior distribution. To

simplify the estimation, we rewrite Equation (3.2) such that it includes the dummy variable¹⁷ r_t

$$\mathbf{c}(r_t) = \mathbf{c}_0 + r_t \mathbf{c}_1, \quad (4.6)$$

$$\text{and therefore } \boldsymbol{\beta}_t = \mathbf{c}_0 + r_t \mathbf{c}_1 + \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{u}_t. \quad (4.7)$$

We need to get rid of the regime dependence in the covariance matrix $\mathbf{Q}(r_t)$ of the error term $\mathbf{u}_t \sim WN(\mathbf{0}, \mathbf{Q}(r_t))$, otherwise the selected prior cannot be used. For the univariate case, there is a simple trick that can be applied, since matrix \mathbf{Q} reduces to a scalar. We can divide the whole equation by the corresponding standard deviation of the error term for each point in time, conditional on the regime. We then get an adjusted regressor and regressand with a variance of one independent of the regime.

In a multivariate scenario, we need to pre-multiply the error term with the inverse of a root¹⁸ of the error covariance matrix $\mathbf{Q}(r_t)$, denoted as $\mathbf{L}^{-1}(\mathbf{Q}(r_t))$. This results in a new error term, which has an identity matrix as covariance matrix. Note that for very high correlations in the matrix, a root might not exist, because the matrix is nearly singular. Since we are drawing the matrix, this event can occur. A solution to this problem is to redraw the covariance matrix and discard the previous draw.

Rewriting the model in matrix notation, we get

$$\mathbf{L}^{-1}[\mathbf{Q}(r_t)]\boldsymbol{\beta}_t = \mathbf{L}^{-1}[\mathbf{Q}(r_t)]\mathbf{c}_0 + \mathbf{L}^{-1}[\mathbf{Q}(r_t)]r_t\mathbf{c}_1 + \mathbf{L}^{-1}[\mathbf{Q}(r_t)]\mathbf{A}\boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t \quad (4.8)$$

where $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{I})$. To simplify the estimation, we rearrange terms. The unknown matrix of parameters is placed between the matrix $\mathbf{L}^{-1}[\mathbf{Q}(r_t)]$ and the lagged factors $\boldsymbol{\beta}_{t-1}$. We

¹⁷ Note that $\mathbf{c}(0) = \mathbf{c}_0$ but $\mathbf{c}(1) = \mathbf{c}_0 + \mathbf{c}_1$. The vector \mathbf{c}_1 only denotes the change in \mathbf{c} , but not the absolute value of \mathbf{c} for the second regime.

¹⁸ We can use a Cholesky decomposition of the matrix \mathbf{Q} , since it is Hermitian. Also other methods can be used, e.g. an eigenvalue decomposition, where the matrix of eigenvectors, multiplied with the diagonal matrix of the eigenvalues gives another root of the matrix.

rewrite the equation in such a way, that the known regressor is on the left hand side of the unknown parameters. The result is given by

$$\text{vec}(\mathbf{L}^{-1}[\mathbf{Q}(r_t)]\boldsymbol{\beta}_t) = (1 \quad r_t \quad \boldsymbol{\beta}_{t-1}') \otimes \mathbf{L}^{-1}[\mathbf{Q}(r_t)] \text{vec}(\mathbf{c}_0 \quad \mathbf{c}_1 \quad \mathbf{A}) + \boldsymbol{\varepsilon}_t, \quad (4.9)$$

or written again in concise notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\eta} \quad \text{with } \boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I}_{3T}), \quad (4.10)$$

where the regressand and the regressor are given by

$$\mathbf{X} = \begin{pmatrix} 1 & r_0 & \boldsymbol{\beta}_0' \\ \dots & \dots & \dots \\ 1 & r_t & \boldsymbol{\beta}_{t-1}' \end{pmatrix} \otimes \mathbf{L}^{-1}[\mathbf{Q}(r_t)], \quad \boldsymbol{\theta} = \text{vec}(\mathbf{c}_0 \quad \mathbf{c}_1 \quad \mathbf{A}) \quad \text{and} \quad \mathbf{Y} = \text{vec}(\mathbf{L}^{-1}[\mathbf{Q}(r_t)]\boldsymbol{\beta}').$$

The procedure to generate draws from the posterior distribution is the same as given in section 4.3.3. For the set of parameters and variables under consideration, this is

$$\boldsymbol{\theta} \mid \mathbf{Q}(r_t), \boldsymbol{\beta}, \mathbf{x} \sim MN(\boldsymbol{\theta}_1, \boldsymbol{\Phi}_1) \quad (4.11)$$

with

$$\boldsymbol{\Phi}_1 = [\boldsymbol{\Phi}_0^{-1} + \mathbf{I} \otimes \mathbf{X}'\mathbf{X}]^{-1}, \quad (4.12)$$

$$\text{and } \boldsymbol{\theta}_1 = \boldsymbol{\Phi}_1[\boldsymbol{\Phi}_0^{-1}\boldsymbol{\theta}_0 + \text{vec}(\mathbf{X}'\mathbf{Y})]. \quad (4.13)$$

In the next section, we derive the posterior distribution of the covariance matrix $\mathbf{Q}(r_t)$. The idea is based on the steps in this section. We rewrite the regression model in such a way that the covariance matrix of the error term \mathbf{Q} does not depend on the regime.

5.2.2 Estimating the Covariance Matrix $\mathbf{Q}(r_t)$

The estimation of the covariance matrices $\mathbf{Q}(0)$ and $\mathbf{Q}(1)$ in the multivariate case differs from the univariate derivation¹⁹. We rewrite the model in such a way that the estimation of the covariance matrix for the two regimes can be separated. Then we estimate one of the covariance matrices, conditional on the other one.

¹⁹ See Kim and Nelson (1999, § 9.1.4) for the univariate case.

We first rewrite the covariance matrix as

$$\mathbf{Q}(r_t) = \mathbf{Q}(0)(\mathbf{I} - r_t\mathbf{I}) + r_t\mathbf{Q}(1) = \mathbf{Q}(0)(\mathbf{I} + r_t(\mathbf{\Lambda} - \mathbf{I})), \quad (4.14)$$

with $\mathbf{\Lambda}$ as a term that contains information about the covariance matrix of the first regime and the covariance matrix of the second regime. In order to estimate the covariance matrices, we need to set the values for $\mathbf{\Lambda}_0$. Note again that the initialization influences the relation between the covariance matrices. We can express $\mathbf{Q}(0)$ and respectively $\mathbf{\Lambda}$ as

$$\mathbf{Q}(0) = \mathbf{Q}(1)\mathbf{\Lambda}^{-1} \text{ or } \mathbf{\Lambda} = \mathbf{Q}(0)^{-1}\mathbf{Q}(1). \quad (4.15)$$

After setting the covariance matrix $\mathbf{Q}(1)$ we can pre-multiply Equation (4.7) with the inverse root of $\mathbf{Q}(r_t)$ and then pre-multiply the equation with the root of $\mathbf{Q}(0)$. The whole pre-multiplication term \mathbf{F} is given by

$$\mathbf{F} = \mathbf{L}(\mathbf{Q}(1)\mathbf{\Lambda}^{-1})\mathbf{L}^{-1}(\mathbf{Q}(r_t)), \quad (4.16)$$

so that the error covariance matrix is given by $\mathbf{Q}(0)$. The model can now be written as

$$\mathbf{F}\boldsymbol{\beta}_t = \mathbf{F}[\mathbf{c}_0 + r_t\mathbf{c}_1 + \mathbf{A}\boldsymbol{\beta}_{t-1}] + \boldsymbol{\varepsilon}_t, \quad (4.17)$$

where $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{Q}(0))$. The covariance matrix $\mathbf{Q}(0)$ again follows an inverted Wishart distribution. It is not necessary to rearrange terms as in the previous section, since we generate $\mathbf{Q}(0)$ conditional on the parameters. The remaining steps are the same as in section 4.3.4, calculating the residuals for the scale matrix using Equation (4.17). The posterior distribution can be written as

$$\mathbf{Q}(0) | \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x} \sim IW(\boldsymbol{\Psi}_1, m_1). \quad (4.18)$$

In contrast to the estimation of the covariance matrix without regime switching, we needed to initialize one of the covariance matrices. Additionally, we need to set the matrix $\mathbf{\Lambda}_0$. Deriving the conditional posterior distribution of $\mathbf{Q}(0)$, the same way the posterior distribution for \mathbf{Q} without regime switching has been derived, we obtain the values for the parameters $\boldsymbol{\Psi}_1$ and m_1 of the distribution. They are given by

$$\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_0 + \mathbf{E}'\mathbf{E}, \quad (4.19)$$

$$m_1 = m_0 + T, \quad (4.20)$$

where \mathbf{E} is the residual of the new equation,

$$\mathbf{E} = \mathbf{F} \left[\boldsymbol{\beta} - \begin{pmatrix} \mathbf{1} & \mathbf{x} & \boldsymbol{\beta}^L \end{pmatrix} \begin{pmatrix} \mathbf{c}_0' \\ \mathbf{c}_1' \\ \mathbf{A}' \end{pmatrix} \right]. \quad (4.21)$$

Drawing from distribution (4.18), we obtain an estimate for $\mathbf{Q}(0)$. We now need to estimate the covariance matrix for the second regime, $\mathbf{Q}(1)$. Instead of doing so, we could also estimate $\boldsymbol{\Lambda}$ and use the relationship between the matrices to transform it to $\mathbf{Q}(1)$. After rearranging again, the non-conjugate prior of matrix $\boldsymbol{\Lambda}$ is an inverse Wishart distribution. We pre-multiply Equation (4.7) with

$$\mathbf{F}^* = \mathbf{L}^{-1}(\boldsymbol{\Lambda}_0). \quad (4.22)$$

Note that the steps are similar to the previous steps; we only change the pre-multiplication term. The likelihood function $\boldsymbol{\Lambda}$ now only depends on the values in regime one, but not on the values from regime zero. This leads to the following posterior density

$$\boldsymbol{\Lambda} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x} \sim IW(\boldsymbol{\Psi}_1^*, m_1^*). \quad (4.23)$$

Deriving the conditional posterior distribution of $\boldsymbol{\Lambda}$, we get the following set of parameters for the distribution

$$\boldsymbol{\Psi}_1^* = \boldsymbol{\Psi}_0^* + \mathbf{E}^{*'} \mathbf{E}^*, \quad (4.24)$$

$$m_1^* = m_0^* + T, \quad (4.25)$$

$$\text{with } \mathbf{E}^* = \mathbf{F}^* \left[\boldsymbol{\beta} - \begin{pmatrix} \mathbf{1} & \mathbf{r} & \boldsymbol{\beta}^L \end{pmatrix} \begin{pmatrix} \mathbf{c}_0' \\ \mathbf{c}_1' \\ \mathbf{A}' \end{pmatrix} \right]. \quad (4.26)$$

Note again that the initialization of the matrix $\boldsymbol{\Lambda}_0$ significantly influences the discrepancy between the covariance matrices. Given a draw of $\boldsymbol{\Lambda}$ from Equation (4.23), we can calculate $\mathbf{Q}(1)$ using Equation (4.15). In contrast to the approach in the last section, we need a two step procedure to estimate the covariance matrices. We first draw the covariance matrix for the first regime and then we generate the covariance matrix for the second regime, conditional on this matrix.

5.2.3 Estimating the Transition Matrix \mathbf{K}

As seen before, we need a transition matrix \mathbf{K} to make any inference about the probability to be in either one or the other regime. This matrix needs to be generated as well, using Gibbs sampling. The elements in the matrix depend on the transitions that we have from the first regime to the first regime, the first to the second regime, the second to the first regime and the second to the second regime. Since we generate the transition matrix conditional on the regime vector, we can just count the number of transitions z_{ij} , the transitions from regime i to regime j . We furthermore know that the elements on the off-diagonal of \mathbf{K} can be calculated from the elements on the diagonal, since every column has to sum up to one.

The natural conjugate prior of the diagonal elements in the transition matrix is a beta distribution (see Kim and Nelson (1999, § 9.1.2)). It is a continuous distribution defined on the interval between zero and one. The shape parameters of the distribution are given by α and β , which need to be strictly positive. The posterior distribution can be written as

$$K_{00} | \mathbf{r} \sim \text{beta}(u_{00} + z_{00}, u_{01} + z_{01}) \quad (4.27)$$

$$\text{and } K_{11} | \mathbf{r} \sim \text{beta}(u_{11} + z_{11}, u_{10} + z_{10}). \quad (4.28)$$

The parameters u_{ij} denote a prior belief about the transitions and should therefore be integers. By selecting u_{ij} we can change the number of transitions, given by z_{ij} . This implies a different view on the expected number of transitions. Note that u_{ij} should be larger than zero, since the beta distribution is only defined for parameters larger than zero. Otherwise we cannot draw the transition probabilities.

Given the regime vector, we can generate draws of the transition matrix. The regime vector is generated conditional on the transition matrix in the next step. We use the Hamilton filter, described in Section 5.1 to achieve our goal.

5.2.4 Generating the Regime Vector \mathbf{r}

We generate the regime vector, conditional on $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, \mathbf{Q} and \mathbf{K} using multimove Gibbs sampling again. We first run the Hamilton filter and save the probabilities. We can rewrite the joint density of the regime vector such that

$$p(\mathbf{r} | \boldsymbol{\beta}) = p(r_T | \boldsymbol{\beta}) \prod_{t=1}^{T-1} p(r_t | r_{t+1}, \boldsymbol{\beta}). \quad (4.29)$$

For a complete derivation of the multimove Gibbs sampling in Equation (4.29), see Kim and Nelson (1999, § 9.1.1) or Carter and Kohn (1994). We can now start to generate the regime vector. In a first step we take the last iteration of the Hamilton filter, which provides us with $p(r_t | \boldsymbol{\beta})$. We only consider the probability for $r_T = 1$, given by $p(1 | \boldsymbol{\beta})$. We obtain the regime corresponding to this probability in a very simple manner. We draw a random number from a uniform distribution. If the number is higher than the probability $p(1 | \boldsymbol{\beta})$, we are in regime zero; else we are in regime one. After generating the regime for time T , we need to generate the regime for all the other t , conditional on the collection of state vectors $\boldsymbol{\beta}$ and r_{t+1} by employing

$$p(r_t | r_{t+1}, \boldsymbol{\beta}_{0 \rightarrow t}) = \frac{p(r_{t+1} | r_t) p(r_t | \boldsymbol{\beta}_{0 \rightarrow t})}{p(r_{t+1} | \boldsymbol{\beta}_{0 \rightarrow t})}. \quad (4.30)$$

Once this probability is calculated, we can generate draws in the same way that we used for drawing r_T . Following this procedure till $t = 1$, we have a draw of the whole regime vector.

We now have all building blocks to estimate a model that combines the state-space form with first order Markov chain regime switching with two regimes. The order and the number of regimes can easily be increased; however, for our purpose this model is satisfying. In the next section we introduce the combined procedure, i.e. how the building blocks can be put together and how the model performs in comparison to the model without regime switching.

6 Regime Switches in State-Space Models

6.1 Building the Algorithm for the Combined Model

Given the building blocks, we implement the Gibbs sampling method to estimate the parameters. We first must decide, how we set the parameters \mathbf{A}_0 , \mathbf{c}_0 , Φ_0 , Ψ_0 , Λ_0 , m_0 , κ_0 , u_{ij} , and in which order we implement the building blocks. We suggest drawing first the state vectors, collected in β , since it contains many elements and is difficult to initialize. Hence, the other parameters and the regime vector need to be initialized. A trivial initialization for the regime vector would be to assume that it only consists of zeros or ones. The fit of the initial values for the parameters does not influence the results very much, if we make sufficiently many draws. However, the choice of \mathbf{A}_0 , \mathbf{c}_0 , Φ_0 , Ψ_0 , Λ_0 , m_0 , κ_0 , u_{ij} , influences the result.

After the initialization, we generate the transition matrix \mathbf{K} and then the regime vector \mathbf{r} . The hyperparameters are generated as the last (starting with $\mathbf{c}(r_t)$ and \mathbf{A} and afterwards $\mathbf{Q}(r_t)$), using the results from Sections 5.2.1 and 5.2.2. Finally, the covariance matrix \mathbf{R} is generated. The described procedure is repeated until convergence is achieved. The first draws should be discarded due to the bias of the initialization. Finally we take the average of the generated parameters.

6.2 Results from the Model including Regime Switching

The initialization is basically the same as in the case without regime switches. Additionally we set values for the elements in the transition matrix \mathbf{K} , given by u_{ij} , which are necessary for generating the transition probabilities. We set all of them equal to one, giving no preference for any transition (A value of zero is also possible, but we have to keep in mind that the shape parameters of the beta distribution have to be larger than one. Unfortunately, for some draws we are not allowed to have a switch from one specific regime to another, resulting in a shape parameter of zero). Furthermore we need to set the matrix Λ to generate the covariance matrices for the two regimes. We set the elements on

the diagonal equal to two. Finally, we need to set $\mathbf{c}_{1,0}$, the initial guess for the change in the constant. Note that the more we are unsure about the value, the higher we should set the initial covariance matrix of the parameters. We initialize the elements of $\mathbf{c}_{1,0}$ equal to zero and the diagonal of its covariance part equal to 10^3 , repeating the initialization from the previous Section 4.4.

We run the algorithm with 24.000 runs, discard the first 4.000 estimates and take the average over the last 20.000 draws. Since the results from the Kalman filter depend on the regime vector and vice versa, we also take the average over them, leaving us with the averaged factors level slope and curvature factors and the average of the regime that we were in each point in time for 20.000 draws.

Figure 7 shows clearly, that a regime shift occurs between 2007 and 2008, as expected. The difference in the constant is quite large. The shift in the error covariance matrix \mathbf{Q} is quite small in an absolute sense, but relatively seen, they change a lot. The estimates are given by

$$\mathbf{A} = \begin{pmatrix} 0.758 & 0.106 & 0.084 \\ 0.129 & 0.922 & -0.042 \\ -0.102 & 0.046 & 1.053 \end{pmatrix}, \quad \mathbf{c}_0 = \begin{pmatrix} 0.153 \\ -0.087 \\ 0.046 \end{pmatrix}, \quad \mathbf{c}_0 + \mathbf{c}_1 = \begin{pmatrix} 0.245 \\ -0.143 \\ 0.122 \end{pmatrix},$$

$$\mathbf{Q}_0 = \begin{pmatrix} 0.0029 & -0.0019 & 0.0024 \\ -0.0019 & 0.0023 & -0.0020 \\ 0.0024 & -0.0020 & 0.0050 \end{pmatrix}, \quad \mathbf{Q}_1 = \begin{pmatrix} 0.0057 & -0.0025 & 0.0060 \\ -0.0025 & 0.0031 & -0.0049 \\ 0.0060 & -0.0049 & 0.0116 \end{pmatrix},$$

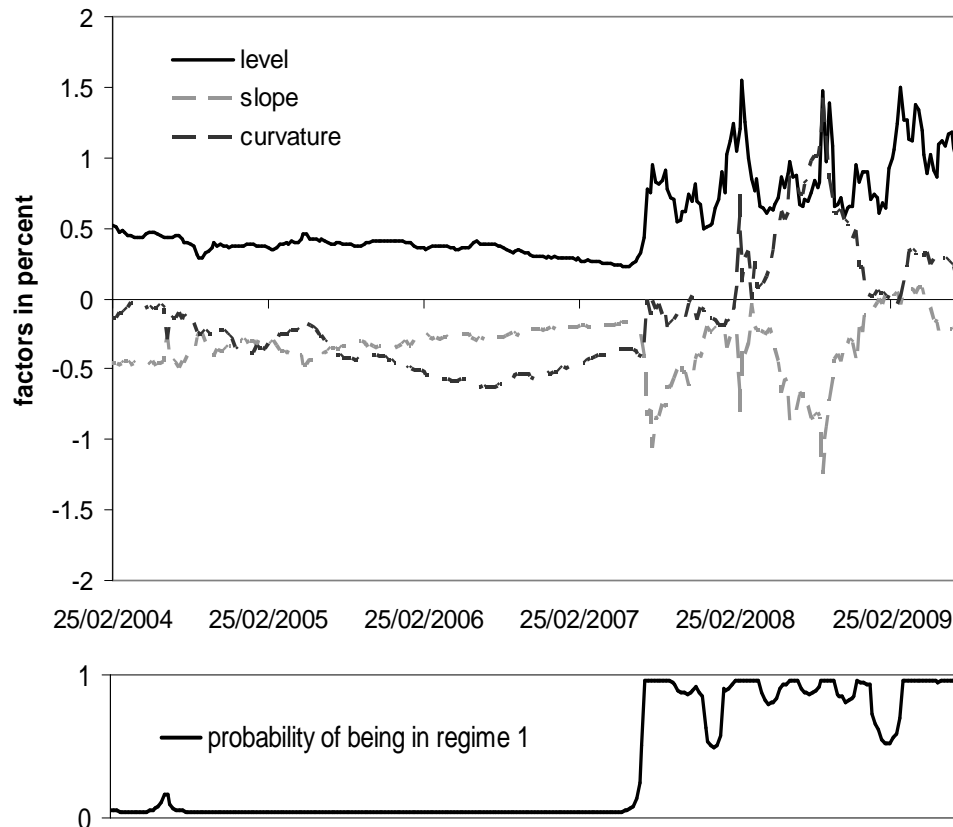
$$\mathbf{R}_{diag} = (0.00131, 0.00193, 0.00035, 0.00025, 0.00032).$$

Comparing the results to previous outcomes, we can see that the first element of matrix \mathbf{A} is much lower than before. Compared to the model without regime switching, the error covariance matrix \mathbf{R} shows much lower elements for the one and for the 10 year spread.

We further calculated the unconditional mean of the factors, using the relation given in (3.2) to show how level, slope and curvature differ for the two regimes

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 0.372 \\ -0.316 \\ -0.349 \end{pmatrix}, \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 0.895 \\ -0.525 \\ 0.325 \end{pmatrix}.$$

Figure 7 – The State-Space Gibbs Sampling Nelson-Siegel Factors with two Regimes



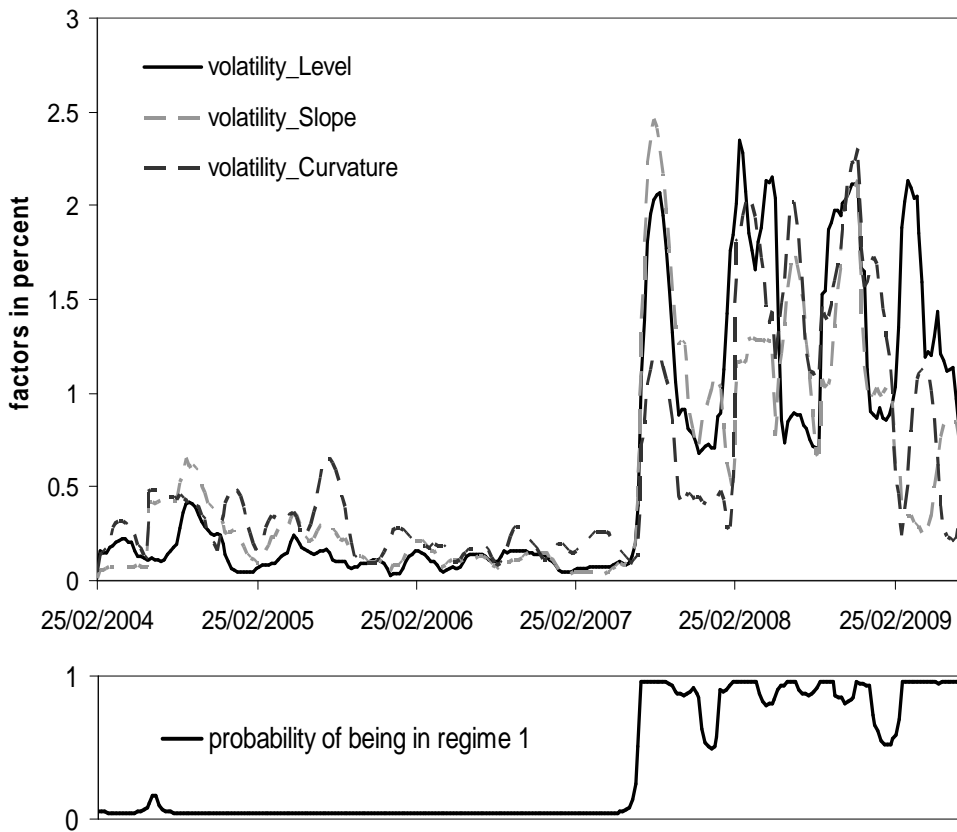
The figure shows the results from estimating the Nelson-Siegel factors using the Kalman Filter and Gibbs Sampling for the estimation, again with the Commerzbank dataset. This time, regime switches are included and the number of runs is increased to 24.000, averaging over the last 20.000 draws. The factors and the regime vector are the averaged values over the runs.

The jump in the probability of being in regime one occurs on the same date, where the jump in level, slope and curvature occur, that is on the 25th July 2007. When the factors revert to the level they had in regime zero, the probability of being in regime one decreases. We clearly observe this behavior after the jump from December 2007 to

January 2008 and from January 2009 to February 2009. In both cases, the probability goes down to nearly fifty percent. Both cases are followed by a jump in the factors.

The variance of the factors also increases at the same time when the regimes change. Including a complete shift in the variance would therefore fortify the gap between the regime probabilities. Figure 8 shows the variance of the factors (calculated on a twelve week basis using a moving window).

Figure 8 – The State-Space Gibbs Sampling Nelson-Siegel Factor Volatilities



The figure shows the yearly volatilities (standard deviations) of the Nelson Siegel factors of the Commerzbank dataset. The volatility is calculated using a moving window over 12 weeks, which explains the very high volatilities when a jump occurs in the factors.

We remark that in order to include a complete shift in the variance of the factors, we need to allow for shifts in the whole transition matrix \mathbf{A} , resulting in additional nine parameters to estimate. Furthermore, switches in \mathbf{A} , affect the probability to be either one

or the other regime very much. Especially in the volatile area, the regime we are in depends strongly on the direction of the change in the spread.

In the next chapter we present the forecast performance of different models with and without regime switching, using the Commerzbank AG dataset. We measure the performance using the sum of squared errors of the forecasts and check the performance of sign forecasts.

7 Forecasting Performance Compared

We know how to estimate the parameters of the models; we can test their forecast performance now. We divide the Commerzbank dataset in an in-sample and an out-of-sample period. The in-sample period starts at 21st March 2005 and goes to the 26th of December 2007. We forecast the remaining 85 spreads for each maturity (that means the out-of-sample period starts at the 2nd January 2008 and goes to the 12th August 2009).

We compare the random walk approach, where the forecast of the spread next week is equal to today's value, to an autoregressive model of order one on the spreads, to a vector autoregressive model of order one on the spreads and to the Nelson Siegel latent factor models with fixed convexity, set to 0.59 (which maximizes the curvature loading for the three year maturity). The factor models are estimated with least squares, as in Section 3.3, and with Gibbs sampling for no regime switches, regimes for the constant and for the covariance matrix of the transition equation separately and for both together. We re-estimate the parameters every week, using a moving window for the in-sample period. To measure the performance of the models, we calculate the sum of squared errors (SSQE) and the sign forecast performance, the relative frequency of correctly predicted spread return signs. Table 2 and Table 3 show the SSQE and the percentage of correctly predicted signs for the mentioned models.

Table 2 – Sum of Squared Errors for Different Models

		Least Squares			Gibbs Sampling			
					Nelson Siegel Factor Model			
SSQE	RW	AR(1)	VAR(1)	OLS	No Switch	Switch c, Q	Switch c	Switch Q
1 year	2.004	2.134	2.301	2.305	2.028	1.973	2.002	2.012
3 year	2.323	2.435	2.524	2.494	2.178	2.064	2.107	2.155
5 year	2.387	2.474	2.591	2.746	2.324	2.240	2.217	2.304
7 year	2.288	2.369	2.496	2.585	2.205	2.151	2.137	2.199
10 year	2.415	2.490	2.544	2.643	2.232	2.189	2.193	2.241
Sum	11.418	11.901	12.456	12.772	10.967	10.617	10.655	10.911

The table contains the sum of squared errors for the forecasts. They were calculated for different maturities and different models, using the Commerzbank dataset. The out-of-sample period starts at 2nd January 2008 and goes to the 12th August 2009. The bold values denote the lowest realizations.

Regarding the SSQE, the Nelson Siegel models, estimated using Gibbs sampling perform better than the other models. The AR(1) and the VAR(1) model, as well as the Nelson Siegel factor model, estimated with least squares perform worse than the random walk, while the Gibbs sampling models are constantly better for this dataset.

We observe the best performance for the models that include a regime switch for the constant in the transition equation, although the difference between the models, estimated with Gibbs sampling, is still small, with or without regime switching.

Table 3 – Sign Forecast Performance for Different Models

		Least Squares			Gibbs Sampling			
					Nelson Siegel Factor Model			
SIGN	Buy and Hold	AR(1)	VAR(1)	OLS	No Switch	Switch c, Q	Switch c	Switch Q
1 year	0.559	0.447	0.412	0.400	0.541	0.576	0.565	0.565
3 year	0.571	0.424	0.459	0.471	0.600	0.588	0.588	0.600
5 year	0.559	0.447	0.435	0.447	0.529	0.506	0.553	0.494
7 year	0.559	0.459	0.494	0.412	0.576	0.529	0.541	0.624
10 year	0.559	0.459	0.494	0.494	0.576	0.576	0.565	0.576
Average	0.561	0.447	0.459	0.445	0.565	0.555	0.562	0.572

The table contains the relative value of correctly predicted signs. They were calculated for different maturities and different models, using the Commerzbank dataset. The out-of-sample period starts at 2nd January 2008 and goes to the 12th August 2009. The bold values denote the highest realizations.

The sign forecasts fortify the better performance over the random walk, the AR(1), the VAR(1) and the least square estimated Nelson Siegel model. Nevertheless, the percentage of correctly predicted signs does not differ significantly from the performance of predicting always a positive sign, i.e. we are not significantly outperforming a simple buy-and-hold strategy²⁰.

We further test, whether each of the models performs significantly different from the random walk and the regime switching model in terms of their squared prediction error.

²⁰ Note that we did not compare the magnitude of the prediction errors as before in the SSQE, but only the correct prediction of the signs.

We apply the Diebold and Mariano (1995) statistic for comparing predictive accuracy with the following null hypotheses

$$H_0 : E[SQE_{t+1|t}^{Model(1)}] = E[SQE_{t+1|t}^{Model(2)}],$$

against:

$$H_1 : E[SQE_{t+1|t}^{Model(1)}] \neq E[SQE_{t+1|t}^{Model(2)}].$$

$SQE_{t+1|t}$ denotes the squared prediction error at time $t + 1$, forecasted with information up to time t . Since we forecast only one-step-ahead, the implementation of the test is simple. We calculate the average of the difference d in the squared prediction errors of the two models, given by

$$d_{t+1|t} = SQE_{t+1|t}^{Model(1)} - SQE_{t+1|t}^{Model(2)},$$

$$\bar{d} = \frac{1}{T} \sum_{t=1}^T d_{t+1|t}.$$

The standard deviation $std()$ of the error term is given by the standard deviation of the differences $d_{t+1|t}$, divided by the square root of T

$$std(\bar{d}) = std(d_{t+1|t}) / \sqrt{T}.^{21}$$

We use the OLS estimator to calculate the standard deviation of $d_{t+1|t}$. Dividing the average of d by its standard deviation we get the Diebold and Mariano (1995) statistic. It is asymptotically normally distributed with mean 0 and variance 1.

We observe that the least square estimated Nelson Siegel model performs significantly different from (worse than) the random walk at the 5% level. None of the other differences in the squared prediction errors is significant at the 5% level.

Table 4 shows Diebold and Mariano (1995) statistic of the forecasts. Positive values indicate superiority of the models against the reference model (the random walk in the upper part of the table, respectively the Gibbs sampling switching model in the lower part). The test indicates, that the random walk is superior to the models that are estimated

²¹ The standard deviation term gets more complicated, when we want to forecast more than one step ahead, since the difference term $d_{t+h|t}$ gets serially correlated. The integer h denotes the steps ahead.

without Gibbs sampling. The switching model with switches in the constant and the error covariance matrix seems to be superior to the other models. The test further indicates that the switching model with only switches in the constant performs better than all other models for five and seven year maturities.

At this point we mention again that the available dataset is very small, which results in a relatively small out-of-sample period. The total number of forecasts for the different maturities is only 85. The smaller the number of forecasts we include in the test, the higher the probability that we do not find evidence for statistically significant differences.

Table 4 – Diebold-Mariano Statistics for Comparing Predictive Accuracy

	Least Squares			Gibbs Sampling			
RW				Nelson Siegel Factor model with convexity fixed to 0.59			
DM - Test	AR(1)	VAR(1)	OLS	No Switch	Switch c, Q	Switch c	Switch Q
1 year	-1.247	-1.578	-2.556	-0.132	0.137	0.010	-0.046
3 year	-0.430	0.082	-0.631	0.500	0.806	0.700	0.586
5 year	-1.241	-0.494	-2.438	0.315	0.422	0.516	0.422
7 year	-1.187	-1.721	-2.426	0.491	0.444	0.521	0.537
10 year	-1.042	-0.473	-1.349	1.009	0.724	0.762	0.968
	Least Squares			Gibbs Sampling			
Switch c, Q				Nelson Siegel Factor model with convexity fixed to 0.59			
DM - Test	RW	AR(1)	VAR(1)	OLS	No Switch	Switch c	Switch Q
1 year	-0.137	-0.586	-1.696	-1.209	-0.544	-0.337	-0.369
3 year	-0.806	-1.033	-0.839	-1.202	-0.581	-0.422	-0.454
5 year	-0.422	-0.629	-0.808	-1.538	-0.487	0.236	-0.367
7 year	-0.444	-0.664	-1.034	-1.134	-0.321	0.142	-0.282
10 year	-0.724	-0.935	-0.750	-1.014	-0.274	-0.041	-0.329

The table contains the test statistics of the Diebold-Mariano test. We test, whether the expectation of squared forecasting errors are equal for different models. In the upper part, we compare the random walk with the other models. In the second part we compare the regime switching model with switches in the constant and the covariance matrix with the other models. Bold values denote significance at the 5% level.

8 Conclusions

8.1 Summary

In this thesis we introduced and estimated a new credit default swap spread curve model. The model is based on the Nelson and Siegel (1987) model for yield curves, where the factors evolve according to a vector autoregressive process as in Diebold and Li (2006).

We first presented the estimation steps under the assumption that the Nelson-Siegel factors level, slope and curvature do not contain an estimation error by applying least square methods. We fixed the convexity factor, such that the transition equation of the state space model is linear in the factors level, slope and curvature. The observable variables were given by spreads for different maturities. Given the factors for each point in time, we built a vector autoregressive model and estimated the parameters.

Further we rewrote the model in state-space form, as suggested by Diebold, Rudebusch and Aruoba (2006). We illustrated the estimation of the state space model using Gibbs sampling, which provided two main advantages for the estimation. First, we suffered from small datasets for credit default swap spreads, not only in the cross-sectional dimension, but also in the serial dimension. Together with the large number of parameters, this would have made a classical estimation less accurate. Secondly, by implementing a Bayesian approach, we allowed for prior beliefs in the estimation. We used multimove Gibbs sampling as suggested by Carter and Kohn (1994) to generate the state vector. We observed that the factors of the least square model differ from the factors in the state space model, especially in the sense of their cross correlations.

We added more flexibility to the model by allowing for regime switching. We introduced switches in the constant and the error covariance matrix of the transition equation in the state-space model to capture the evolvement of the credit default swap spread curve and show how the parameters are estimated using Gibbs sampling. We generated the regime vector by applying multimove Gibbs sampling.

We divided a dataset of Commerzbank AG spreads into an in-sample and out-of-sample period and created one-step-ahead forecasts. We showed that the state-space model, estimated with Gibbs sampling outperforms the random walk, autoregressive models on the spreads and the least square estimated model in terms of their sign forecast performance and the sum of squared prediction errors. The superiority of the model is statistically not significant. The available out-of-sample period is very short, which makes it difficult to prove statistical significance.

8.2 Recommendations for Further Research

The estimation of parameters in a Bayesian framework allows for wide flexibility. Not only the chosen distributions and the prior beliefs about mean and covariance influence the results. The applied estimation techniques also play an important role.

We decided in favor of multimove Gibbs sampling because of its computational efficiency, but one could also consider using singlemove Gibbs sampling, suggested by Carlin, Polson and Stoffer (1992). Another procedure to simulate from the posterior densities of state is the simulation smoother by De Jong and Shephard (1995). It is an alternative multimove Gibbs sampler, but instead of drawing the random variables, it simulates their disturbance terms. The random variables can be constructed from the disturbances as required. Computationally, this approach is more efficient than the multimove Gibbs sampler by Carter and Kohn (1994).

Further, the whole model can be estimated in a classical framework. Estimation methods have already been implemented for state space models (see Durbin and Koopman (2001)), but also for state space models with regime switching in Kim and Nelson (1999). Nowadays, the estimation suffers also from a short time series, but due to the high trading volume of credit default swaps, it is likely that sufficient data is available in future, as the time series grows. Nevertheless, it is unclear whether the cross-sectional dimension increases, which restricts the quality of a classical estimation.

In the very beginning of this thesis, we assumed that the convexity factor of the model is fixed. The loadings of the Nelson and Siegel (1987) model only depend on this single factor. A fixed convexity assumption is very common, but the factor can also be time varying. Although this adds extra complexity to the estimation (the model is not linear in the factors anymore), Koopman, Mallee and van der Wel (2007) show that the extension leads to a significant improvement in model fit for the yield curve.

As suggested by Diebold, Rudebusch and Aruoba (2006) for yield curves, some macroeconomic variables like inflation, but also equity returns or interest rates might have influence on credit default swap spreads. They can be added to the transition equation of the model. This improves the fit of the model to the data, but might also add more predictive power to the model.

Appendices

Appendix A. Drawing from an inverse Wishart distribution

We follow the procedure from Smith and Hocking (1972) to draw from a Wishart distribution. We can then convert the results to an inverted Wishart distribution, since when the matrix \mathbf{A}^{-1} follows a Wishart distribution with $\mathbf{A}^{-1} \sim W(\mathbf{X}^{-1}, df)$, then the inverse of \mathbf{A} is inverse Wishart distributed with $\mathbf{A} \sim IW(\mathbf{X}, df)$.

To generate draws from a Wishart distribution, we need to generate random variables from the standard normal and the chi-square distribution. When \mathbf{A}^{-1} is of size $p \times p$, we need to build a lower triangular matrix \mathbf{T} with independent samples from the standard normal distribution on the off-diagonal. The elements on the diagonal are square roots of a chi-square distribution with $df - i + 1$ degrees of freedom, where i denotes the position on the diagonal in the matrix. In the end matrix should look like this

$$\mathbf{T} = \begin{pmatrix} \sqrt{\chi_{df}^2} & 0 & \dots & 0 \\ z \sim N(0,1) & \sqrt{\chi_{df-1}^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ z \sim N(0,1) & z \sim N(0,1) & \dots & \sqrt{\chi_{df-p+1}^2} \end{pmatrix}.$$

The next step is to compute a Cholesky decomposition of \mathbf{X}^{-1} , given by $\mathbf{X}^{-1} = \mathbf{L}\mathbf{L}'$, where \mathbf{L} denotes the lower triangular matrix. Building the following product $\mathbf{A}^{-1} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}'$, \mathbf{A} is Wishart distributed with $\mathbf{A}^{-1} \sim W(\mathbf{X}^{-1}, df)$, and therefore the inverse of \mathbf{A}^{-1} is inverse Wishart distributed with $\mathbf{A} \sim IW(\mathbf{X}, df)$.

Appendix B. Deriving regressand and regressor in the multivariate model

We start with Equation (4.8), the regression model pre-multiplied with the inverse of the Cholesky factorization of the error covariance matrix. For three factors, this can be written as

$$\mathbf{L}^{-1}[\mathbf{Q}(x_t)]\boldsymbol{\beta}_t = \begin{pmatrix} \omega_{11} & 0 & 0 \\ \omega_{21} & \omega_{22} & 0 \\ \omega_{31} & \omega_{32} & \omega_{33} \end{pmatrix} \begin{pmatrix} c_{0,1} & c_{1,1} & A_{11} & A_{12} & A_{13} \\ c_{0,2} & c_{1,2} & A_{21} & A_{22} & A_{23} \\ c_{0,3} & c_{1,3} & A_{31} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} 1 \\ x_t \\ \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \end{pmatrix} + \boldsymbol{\varepsilon}_t.$$

Multiplying the equation and rearranging terms, we get the results in Equation (4.9). The regressor \mathbf{X}_t for time t is given by

$$\begin{pmatrix} \omega_{11} & 0 & 0 & x_t\omega_{11} & 0 & 0 & \beta_{1,t}\omega_{11} & 0 & 0 & \beta_{2,t}\omega_{11} & 0 & \dots \\ \omega_{12} & \omega_{22} & 0 & x_t\omega_{12} & x_t\omega_{22} & 0 & \beta_{1,t}\omega_{12} & \beta_{1,t}\omega_{22} & 0 & \beta_{2,t}\omega_{12} & \beta_{2,t}\omega_{22} & \dots \\ \omega_{13} & \omega_{23} & \omega_{33} & x_t\omega_{13} & x_t\omega_{23} & x_t\omega_{33} & \beta_{1,t}\omega_{13} & \beta_{1,t}\omega_{23} & \beta_{1,t}\omega_{33} & \beta_{2,t}\omega_{13} & \beta_{2,t}\omega_{23} & \dots \end{pmatrix}.$$

This can be rewritten as in Equation (4.9) using the Kronecker product. The vector of parameters $\boldsymbol{\theta}$ equivalently is

$$(c_{0,1} \ c_{0,2} \ c_{0,3} \ c_{1,1} \ c_{1,2} \ c_{1,3} \ A_{11} \ A_{21} \ A_{31} \ A_{12} \ A_{22} \ A_{32} \ A_{13} \ A_{23} \ A_{33})'.$$

Note that the regressand changes to its vectorized version.

References

- Albert, J. H. & Chib, S. (1993): Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts. *Journal of Business and Economic Statistics* 11(1), 1 – 15.
- Alexopoulou, I. & Andersson, M. & Georgescu, O. M. (2009): An Empirical Study on the Decoupling between Corporate Bonds and CDS Spreads. *ECB Working Paper Series*, No. 1085.
- Bernadell, C. & Coche, J. & Nyholm, K. (2005): Yield Curve Prediction for the Strategic Investor. *ECB Working Paper Series*, No. 472.
- Carlin, B. P. & Polson, N. G. & Stoffer, D. S. (1992): A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association* 87(418), Theory and Methods, 493 – 500.
- Carter, C. K. & Kohn, P. (1994): On Gibbs Sampling for State Space Models. *Biometrika* 81, 541 – 553.
- Cox, J. C. & Ingersoll, J. E. & Ross, S. A. (1985): A Theory of the Term Structure of Interest Rates. *Econometrica* 53, 385 – 407.
- De Jong, P. & Shephard, N. (1995): The simulation smoother for time series models. *Biometrika* 82(2), 339 – 350.
- Diebold, F.X. & Li, C. (2006): Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics* 130, 337-364.
- Diebold, F.X. & Mariano, R. (1995): Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13, 253 – 265.
- Diebold, F. X., Rudebusch, G. D. & Aruoba, B. (2006): The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach. *Journal of Econometrics* 131, 309 – 338.
- Duffie, D. (1999): Credit Swap Valuation. *Financial Analysts Journal* Jan/Feb 1999, 73 – 87.

- Durbin, J. & Koopman S. J. (2001): Time Series Analysis by State Space Methods. Third Edition. Oxford: Oxford University Press.
- Geman, S. & Geman, D. (1984): “Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721 – 741.
- Hamilton, J. D. (1989): A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* 57(2), 357 - 384.
- Hamilton, J. D. (1994): Time Series Analysis. New Jersey: Princeton University Press.
- Harvey, A. C. (1993): Time Series Models. Second Edition. Cambridge: The MIT Press.
- Hull, J. & White, A. (1990): Pricing Interest Rate Derivative Securities. *Review of Financial Studies* 3, 573 – 592.
- Hull, J. & White, A. (2000): Valuing Credit Default Swaps I: No Counterparty Default Risk. *Journal of Derivatives* 8(1), 29 – 40.
- Jarrow, R., Lando, D. & Turnbull, S. M. (1997): A Markov Model for the Term Structure of Credit Spreads. *The Review of Financial Studies* 10, 481 – 523.
- Jorion, P. (1986): Bayes-Stein Estimation for Portfolio Analysis. *Journal of Financial and Quantitative Analysis* 21, 279 – 292.
- Kim, C.-J. & Nelson, C. R. (1998): Business Cycle Turning Points, a New Coincident Index, and Tests of duration Dependence Based on a Dynamic Factor Model with Regime Switching. *The Review of Economics and Statistics* 80(2), 188 – 201.
- Kim, C.-J. & Nelson, C. R. (1999): State-Space Models with Regime Switching – Classical and Gibbs-Sampling Approaches with Applications. Cambridge: MIT Press.
- Koopman, S. M., Mallee, M. I. P. & Wel, M. van der (2007): Analyzing the Term Structure of Interest Rates using the Dynamic Nelson-Siegel Model with Time-Varying Parameter. *Tinbergen Institute Discussion Papers*, 07 – 095 / 4.
- Landschoot, A. (2004): Determinants of Euro Term Structure of Credit Spreads. *ECB Working Paper Series*, No. 397.
- Li, D. X. (2000): On Default Correlation: A Copula Function Approach. *Journal of Fixed Income* 9(4), 43 – 54.

- Litterman, R. (1986): Forecasting with Bayesian Vector Autoregression – Five Years of Experience. *Journal of Business and Economic Statistics* 4, 25 – 38.
- Lütkepohl, H. (2005): New Introduction to Multiple Time Series Analysis. Second Edition. Berlin: Springer.
- Nelson, C. & Siegel, A. (1987): Parsimonious Modeling of Yield Curves. *Journal of Business* 60(4), 473 – 489.
- Smith, W. B & Hocking, R. R. (1972): Algorithm AS 53: Wishart Variate Generator. *Journal of the Royal Statistical Society Series C* 21(3), 341 – 345.
- Schönbucher, P. J. (2003): Credit Derivatives Pricing Models: Model, Pricing and Implementation. New York: Wiley.
- Sugita, K. (2008): Bayesian analysis of a vector autoregressive model with multiple structural breaks. *Economics Bulletin* 3(22), 1 – 7.
- Vasicek, O. (1977): An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics* 5, 177 – 188.
- Zellner, A. (1971): An Introduction to Bayesian Inference in Econometrics. New York: Wiley.
- Zhang, B.Y., Zhou, H. & Zhu, H. (2006): Explaining Credit Default Swap Spreads with the Equity, Volatility and Jump Risks of individual Firms. *BIS Working Paper*, No. 181.