Unterschrift des Betreuers:

Diplomarbeit

Estimation and Analysis of Interest Rate Structure Models with Shadow Rates

Ausgeführt am Institut für

Institut für Stochastik und Wirtschaftsmathematik
der Technischen Universität Wien

unter Anleitung von

Ao.Univ.Prof. Dipl.-Ing. Mag.rer.nat. Dr.techn. Wolfgang Scherrer

durch

Lukas Fertl Bsc

[redacted]

Datum

Unterschrift (Student)

Acknowledgement:

I want to thank all people who supported me during my work on this Master's Thesis. Especially my family and girlfriend who encouraged and supported me during the process of writing. Special thanks to my mother who always encouraged me to give my best and supported me during my studies. I also want to mention and thank my friends and colleagues at the university.

Furthermore I want to thank my advisor Prof Dr Wolfgang Scherrer for his incredibly valuable advice and support. During the making of this thesis I learned a lot due to his ongoing supervision.

Contents

1 Introduction

Since the great financial crisis in 2008 the world entered a new normal. In this setting a lot of economic rules and principles had to be reconsidered. One of the cornerstones of modern economics is the interaction between interest rates and other economic variables. There is a great quantity of models that use the interest rate, set by the central bank, as an input. Before the crisis nearly nobody thought about negative interest rates, since it is natural to assume a lower bound of 0 (as long as cash with a nominal interest of 0 is available). It was also thought of as unlikely that this zero lower bound becomes relevant in modern developed economies with the exception of Japan. But since 2009 everything changed and the central banks of all major economies had to lower the benchmark interest rate to 0 or in some cases even into negative territory (i.e. the deposit rate of the European Central Bank ECB is currently -0.4 % [2]). The Federal Reserve System (thereafter Fed) lowered the effective federal funds rate (thereafter EFFR) to $0 - 0.25$ % on 16/12/2008 [1]. Since then the EFFR has been stuck at the zero lower bound and didn't display any meaningful variation that can be used to explain movements in other economic variables (see Figure 1). Furthermore unconventional monetary policy tools like forward guidance and quantitative easing (QE) were used. Quantitative easing are large scale asset purchases with newly created central bank money. These purchases of mostly long dated securities were intended to reduce long term interest rates. Forward guidance is a communication strategy by central bankers to explain their decision to the public. The goal is to reduce uncertainty about future monetary policy decisions. Both were used to stimulate the economy when it was stuck at the zero lower bound. Therefore a lot of effort has been invested to construct new models where the zero lower bound is modelled explicitly.

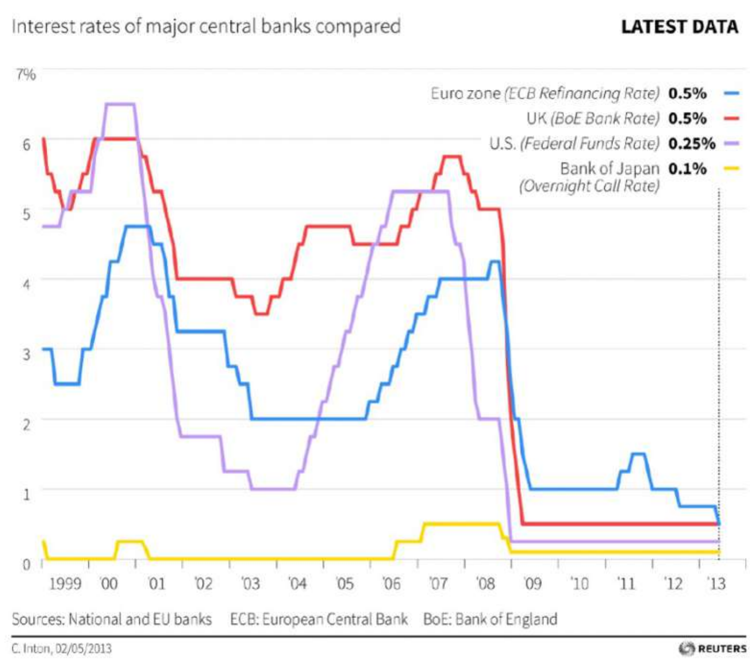

Figure 1: Benchmark interest rates of major central banks

One of the most popular, conventional models deployed in the literature to link the EFFR or the nominal short rate set by the central bank to the whole interest rate complex is the Gaussian Affine Term Structure Model (GATSM [3]). In this model the nominal interest rate can get negative due to the affine structure (i.e. the short rate is affine in the Gaussian factor, see also section Affine interest rate structure models). This structure allows easy computations and analytical solutions.
But in the real world nominal rates are bounded from below by 0 or more general a constant $\bar{r}$ ($\bar{r}$ can be thought of as small positive or negative constant). The extended model, first proposed by Black (1995) [7], is called shadow rate term structure model (SRTSM or shadow rate model). In this framework the lower bound of nominal interest rate is explicitly modelled. The shadow rate s_t is introduced and the nominal short rate is given by $r_t = \max(\bar{r}, s_t)$. There the shadow rate is affine in the Gaussian factors. This floor is a non-linearity and makes an analytic solution intractable, in fact only for one factor models the solution is known.

In my Master thesis I will take an in depth look at the model proposed by Jing Cynthia Wu and Fan Dora Xia [5]. They propose a framework (SRTSM) how to model the short rate in a world where the EFFR is at the zero lower bound and estimate the so called shadow rate. They use a simple analytical approximation for bond prices that makes the model tractable and allows easy implementation. The shadow rate can go deeply into negative territory and captures the effects of unconventional tools. This means that the shadow rate can be used instead of the EFFR in conventional models which relay on a short rate as input. In fact in normal times when the zero lower bound is not active both models overlap perfectly.

I begin with briefly introducing short rate models, especially affine ones in section 2. In section 3, I describe Kalman Filters for linear state space models and extended Kalman filters for non-linear systems since they are used for estimation. Section 4 explains the construction of the forward rates used in the SRTSM. Then the shadow rate and Gaussian affine term structure is introduced, estimated and analysed in section 5, 6, 7 and 8.

2 Affine interest rate structure models

Here I will formally introduce the short rate model and explain the special case of affine short rate models. This section is based on the lecture course "Interest rate theory" of Dr Paul Krühner TU Vienna [3]. The model is presented in continuous time since more mathematical tools and theorems are available in continuous time (i.e. Girsanov's Theorem). This section lays out the theoretical foundation and motivation of the GATSM and SRTSM mentioned above. The implementation of the models is in a discrete time setting, think of a discretely sampled continuous time model due to the fact that we want to estimate them by means of time series analysis tools. A little bit of basic knowledge about stochastic analysis (i.e. Brownian Motion, stochastic integration and pricing measures) is a prerequisite for this section. First we will see what a short rate model is and then extend it by a state variable.

2.1 Affine short rate model without state variable

In the whole section we assume a filtered probability space $(\Omega,\mathscr{A},\mathbb{F},P)$ where the Filtration $\mathbb{F} = (\mathscr{F}_t)_{t\geq 0}$ satisfies the usual conditions of completeness and right continuity (most of the time the natural filtration $\mathscr{F}_t = \sigma(W_s : s \leq t)$ is used). We denote by $(W_t)_{t\geq 0}$ a d dimensional standard Brownian Motion
A short rate model is a quadruple $(r_0, b, \sigma, \lambda)$ that satisfies:

- $r_0 \in \mathbb{R}$ or $\mathbb{R}_+$
- $\sigma, \lambda \in \mathscr{L}(W)$ with $\sigma \in \mathbb{R}^{1\times d}$ and $\lambda \in \mathbb{R}^d$
- $b \in \mathscr{L}(I)$ with $b \in \mathbb{R}$
- $\mathbb{E}(e^{\int_0^t |\lambda_s|^2 ds}) < \infty \; \forall t$

where $\mathscr{L}(W)$ stands for the set of progressive measurable processes ν that are integrable with respect to W_t (i.e. $\int_0^t |\nu_s|^2 ds < \infty$ P a.s $\forall t > 0$) and $\mathscr{L}(I)$ are the processes that are absolute integrable (i.e. $\int_0^t |\nu_s| ds < \infty$ P a.s $\forall t > 0$)[1].
Then we define[2]:

- $r_t := r_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s$ short rate
- $M_t := e^{\int_0^t r_s ds}$ bank account or numeraire
- $dQ := ZdP$ the pricing measure Q where $Z_t = \mathscr{E}(\int_0^t \lambda_s dW_s) > 0$

[1] Here $|\cdot|$ stands for a suitable norm.
[2] In the definition of r_t we see that for some choices of σ (i.e. deterministic ones) the stochastic integral part is normal distributed and nothing prevents it from becoming arbitrarily small (i.e. $\forall M < 0 : \mathbb{P}(\int_0^t \sigma_s dW_s < M) > 0$ so the short rate r_t can get negative, especially if for small $\tau > 0$, $r_{t-\tau}$ is close to zero the increment from $t - \tau$ to t can push r_t below zero.

Where $\mathscr{E}(X)_t = \exp(X_t - \frac{1}{2}[X]_t)$ is the stochastic exponential[3] of the process X and $[X]$ is the quadratic variation process of X. We call b drift, σ diffusion coefficient. The pricing measure Q can be identified with λ through Girsanov's theorem[4] and we have $dW_t^Q := -\lambda_t dt + dW_t$ is a Q standard Brownian Motion. Note that $dW_t^Q := -\lambda_t dt + dW_t$ is a short notation for $\int_0^t dW_s^Q = W_t^Q = W_t - \int_0^t \lambda_s ds$

Further note that due to Girsanov we have different representations for the density process of the measure change Z_t if we fix a finite time horizon $T^* < \infty$:

$$Z_t = \mathscr{E}(\int_0^t \lambda_s dW_s) = \mathbb{E}(Z_{T^*}|\mathscr{F}_t) = (\frac{dQ}{dP})_t = (\frac{dQ}{dP})|_{\mathscr{F}_t}$$

Next I show how the drift change of the Q Brownian motion (the density process is given by Z_t) works, using Ito's formula[5] for $f(x,y) = x \cdot y$. This is done to motivate the set-up used in GATSM and SRTSM. I will show that $(W^Q)_{t\geq 0}$ is a local Q martingale. This is equivalent

[3]Dolean's stochastic exponential $\mathscr{E}(X)_t$ of the process X is defined as the unique solution of the following stochastic differential equation:

$$dY = Y dX$$

This can be seen with Ito's formula for $f(x) = e^x$ $(\partial^2 f(x) = \partial f(x) = f(x))$ for the 1 dimensinal semimartingal (a process with local martingale and finite variation part) $\tilde{X}_t = X_t - \frac{1}{2}[X]_t$:

$$\mathscr{E}(X)_t = f(\tilde{X}_t) = f(X_0) + \int_0^t \partial f(\tilde{X}_s) d\tilde{X}_s + \frac{1}{2}\int_0^t \partial^2 f(\tilde{X}_s) d[\tilde{X}]_s = f(X_0) + \int_0^t \mathscr{E}(X)_s d(X_s - \frac{1}{2}[X]_s) + \frac{1}{2}\int_0^t \mathscr{E}(X)_s d[X - \frac{1}{2}[X]]_s$$

using the linearity of the integrator and that $[X]$ is a finite variation process, therefore $[[X]] = 0$

$$= f(X_0) + \int_0^t \mathscr{E}(X)_s dX_s - \frac{1}{2}\int_0^t \mathscr{E}(X)_s d[X]_s + \frac{1}{2}\int_0^t \mathscr{E}(X)_s d[X]_s = f(X_0) + \int_0^t \mathscr{E}(X)_s dX_s$$

Therefore we see that the stochastic exponential solves the sde $d(\mathscr{E}(X)_t) = \mathscr{E}(X)_t dX_t$ and is a local martingal if X is a local martingal (as a stochastic integral with respect to a local martingal is a local martingal). Together with Novikov's condition $\mathbb{E}(e^{\int_0^t |\lambda_s|^2 ds}) < \infty$ $\forall t$ we have that $\mathscr{E}(\int_0^t \lambda_s dW_s) > 0$ is a true martingale with constant expectation. If $X_0 = 0 \Rightarrow f(X_0) = 1$ and the martingale has constant expectation 1, since $\mathbb{E}(\mathscr{E}(X)_t) = f(X_0) + \mathbb{E}(\int_0^t \mathscr{E}(X)_s dX_s) = 1 + 0$, since the stochastic integral part is a true martingale and starts at 0 for $t = 0$. Otherwise we can normalize the expectation to 1 and use it as the density process of a measure change. Since it is positive wee see that the new measure is equivalent to P.

[4]Girsanov Theorem:
1) Let $Q \sim P$ then there exists $\lambda \in \mathscr{L}(W)$ such that $dW_t^Q := -\lambda_t dt + dW_t$ is a Q Brownian Motion.
2) If $\lambda \in \mathscr{L}(W)$ with $\mathbb{E}(e^{\int_0^t |\lambda_s|^2 ds}) < \infty$ $\forall t$ then there exists a random variable Z such that:

- $\mathbb{E}(Z) = 1$
- $\mathscr{E}(\int_0^t \lambda_s dW_s) = \mathbb{E}(Z|\mathscr{F}_t) = Z_t$
- $dQ := ZdP$ defines a probability measure $Q \sim P$ where $dW_t^Q := -\lambda_t dt + dW_t$ is a Q Brownian Motion with respect to the Filtration $\mathbb{F}$

[5]Ito's formula for a function $f \in C^2(\mathbb{R}^d, \mathbb{R})$ and a d dimensional semimartingal (a process with local martingale and finite variation part) X, see Rheinländer and Sexton [22]:

$$f(X_t) = f(X_0) + \sum_{j=1}^d \int_0^t \partial_j f(X_s) dX_s^j + \frac{1}{2}\sum_{j=1}^d \sum_{i=1}^d \int_0^t \partial_{i,j} f(X_s) d[X^j, X^i]_s$$

to $(W_t^Q \cdot Z_t)_{t\geq 0}$ being a local P martingale. Then I use levy's characterisation of a standard Brownian motion to show that W_t^Q is really a standard Brownian motion.

I start by showing that $W_t^Q = W_t - \int_0^t \lambda_s ds$ is a local Q martingale when using Z_t for the measure change. This is accomplished by showing that $W^Q \cdot Z$ can be written as two stochastic integrals with respect to local martingales. Slightly abusing notation we assume $(W_t)_{t\geq 0}$ to be 1 dimensional you can think of an arbitrary component of the d dimensional Brownian motion. Keep in mind that $dZ_t = Z_t d(\int_0^t \lambda_s dW_s) = Z_t \lambda_t dW_t$ and $dW_t^Q = dW_t - \lambda_t dt$ due to the chain rule:

$$W_t^Q Z_t = f(W_t^Q, Z_t) = f(W_0^Q, Z_0) + \int_0^t W_u^Q dZ_u + \int_0^t Z_u dW_u^Q + [W_t^Q, Z_t] \tag{1}$$

Since $\int_0^u \lambda_s ds$ has finite variation[6] (thus we have $[\int_0^u \lambda_s ds, Y] = 0$ for every continuous process Y) and the stochastic exponential Z_t solves a sde (stochastic differential equation see footnote 3) it follows ($[\cdot,\cdot]$ is bilinear):

$$[W_t^Q, Z_t] = [W_t - \int_0^t \lambda_s ds, Z_t] = [W_t, Z_t] - [\int_0^t \lambda_s ds, Z_t] = [\int_0^t dW_u, \int_0^t Z_u d(\int_0^u \lambda_s dW_s)] =$$

using $W_t = \int_0^t dW_u$, the chain rule $d(\int_0^u \lambda_s dW_s) = \lambda_u dW_u$ and $[\int A dX, \int B dY] = \int A \cdot B d[X,Y]$

$$= [\int_0^t dW_u, \int_0^t Z_u \lambda_u dW_u] = \int_0^t Z_u \lambda_u d[W]_u = \int_0^t Z_u \lambda_u du$$

Continuing from equation (1) it follows:

$$W_t^Q Z_t = \int_0^t W_u^Q dZ_u + \int_0^t Z_u dW_u^Q + \int_0^t Z_u \lambda_u du =$$

using the linearity of the stochastic integral with respect to the integrator and $dW_t^Q = dW_t - \lambda_t dt$:

$$= \int_0^t W_u^Q dZ_u + \int_0^t Z_u dW_u - \int_0^t Z_u \lambda_u du + \int_0^t Z_u \lambda_u du = \int_0^t W_u^Q dZ_u + \int_0^t Z_u dW_u$$

Here we see that both integrals on the right hand side are local P martingales as they are stochastic integrals with respect to a local martingale (Z,W are both local martingales. It is a well known result that W is a (local) P martingale[7] and Z solves the sde with respect to the local martingale $\int_0^t \lambda_s dW_s$, see footnote 3). Therefore we showed that $W^Q \cdot Z$ is a

---

An alternative representation of Ito's formula using the differential notation is:

$$df(X_t) = \sum_{j=1}^{d} \partial_j f(X_t) dX_t^j + \frac{1}{2} \sum_{j=1}^{d} \sum_{i=1}^{d} \partial_{i,j} f(X_t) d[X^j, X^i]_t$$

[6] $\int_0^u \lambda_s ds = \int_0^u (\lambda_s)^+ ds - \int_0^u (\lambda_s)^- ds$ where $(x)^+ := \max(0,x)$ and $(x)^- := |\min(0,x)|$ are the positive and negative part of x. Therefore every Lebesgue is of finite variation since it can be written as difference of two monotone processes. Alternativly we could argue that it is absolute continous and therefore of finite variation.

[7] If a process is a martingale then it is also a local martingale.

local P martingale. This is equivalent to: $W^Q = W - \int \lambda_s ds$ is a local Q martingale. Using levy's characterisation of the Brownian motion[8] gives that W^Q is a Q Brownian motion since $[W_t^Q] = [W_t - \int \lambda_s ds] = [W_t] = t$ (since $\int \lambda_s ds$ has finite variation).

Pricing of securities:

In a short rate model the pricing measure is given. We can price any security with pay-off $C(T)$ where T stands for the maturity. The price at time t is given by:

$$C_t = M_t \mathbb{E}^Q(\frac{C(T)}{M_T}|\mathscr{F}_t)$$

Alternatively we could define the stochastic discount factor[9] $\tilde{M}_t$ by:

$$\tilde{M}_t := \frac{1}{M_t}(\frac{dQ}{dP})_t = e^{-\int_0^t r_s ds}\mathscr{E}(\int_0^t \lambda_s dW_s) = e^{-\int_0^t r_s ds + \int_0^t \lambda_s dW_s - \frac{1}{2}\int_0^t \lambda_s' \lambda_s d_s} \tag{2}$$

using that for a stochastic integral the quadratic variation is given by $[\int_0^t \lambda_s dW_s] = \int_0^t \|\lambda\|^2 ds = \int_0^t \lambda_s' \lambda_s d_s$ and $[W]_t = tI$ for a Brownian motion ($\lambda_s'\lambda_s$ is needed because in general $\lambda_t, W_t \in \mathbb{R}^d$ and $[\int_0^t \lambda_s dW_s] = [\sum_{i=1}^d \int_0^t \lambda_s^i dW_s^i] = \sum_{i=1}^d \int_0^t (\lambda_s^i)^2 d[W]_s^i = \sum_{i=1}^d \int_0^t (\lambda_s^i)^2 ds = \int_0^t \|\lambda\|^2 ds = \int_0^t \lambda_s' \lambda_s d_s$ using the fact that for a standard Brownian motion the components are independent, see also the diagonal structure of $[W_t]$).

Therefore using the stochastic discount factor for pricing a security with pay-off C(T) under the real world measure P:[10]

$$C_t = \frac{1}{\tilde{M}_t}\mathbb{E}^P(C(T)\tilde{M}_T|\mathscr{F}_t) = \frac{M_t}{(\frac{dQ}{dP})_t}\mathbb{E}^P(C(T)\frac{1}{M_T}(\frac{dQ}{dP})_T|\mathscr{F}_t) = M_t\mathbb{E}^Q(C(T)\frac{1}{M_T}|\mathscr{F}_t)$$

using the fact that due to Girsanov's theorem $\mathscr{E}(\int_0^{\cdot} \lambda_s dW_s) = \frac{dQ}{dP}.$ is the density process between Q and P and Bayes Theorem for conditional expectation[11].

If we denote by $B(t,T)$ the price of a zero coupon bond at time t that matures at time T (i.e. $B(T,T) = 1$). We have [12]:

[8] It is equivalent for a 1 dimensional process on a filtered probability space with the natural filtration:

1. X is a Brownian motion with respect to the measure μ
2. X is a local martingale with respect to μ and $[X]_t = t$

[9] A stochastic discount factor is defined as the discount factor used to discount under the real world measure to price securities.

[10] Because we want the price of C(T) at any point $t < T$ and not just $t = 0$ we have to compound interest from 0 to t by multiplying with $1/\tilde{M}_t$, since multiplying with $\tilde{M}_T$ discounts from T to 0.

[11] Bayes Theorem (see Rheinländer and Sexton [22]): Let $X \in L^1(Q)$ be $\mathscr{F}_T$ measurable and integrable then:

$$\mathbb{E}^Q(X|\mathscr{F}_t) = \frac{1}{(\frac{dQ}{dP})_t}\mathbb{E}^P(X(\frac{dQ}{dP})_T|\mathscr{F}_t)$$

[12] M_t is $\mathscr{F}_t$ measurable.

$$B(t,T) = M_t \mathbb{E}^Q(\frac{B(T,T)}{M_T}|\mathscr{F}_t) = \mathbb{E}^Q(\frac{M_t}{M_T}|\mathscr{F}_t) = \mathbb{E}^Q(e^{-\int_t^T r_s ds}|\mathscr{F}_t) \quad (3)$$

The integral in (1) is replaced by the sum $\sum_{j=0}^{n-1} r_{t+j}$ (think about approximating the integral with n+1 ($n = T - t \in \mathbb{N}$ in a discrete setting) supporting points $\{t, 1, ..., t+n\}$ which gives n summands) in the discrete setting of the GATSM and SRTSM.

A short rate model has affine term structure (ATS) if there exist $C^1(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R})$ functions A, C with $A(T,T) = C(T,T) = 0$ such that:

$$B(t,T) = \exp(-A(t,T) + C(t,T)r_t)$$

A short rate model is affine (for dimension d = 1) if:

- $b_t = \beta_0(t) + \beta_1(t)r_t$
- $\sigma_t^2 = \alpha_0(t) + \alpha_1(t)r_t$

with $\alpha_0, \alpha_1, \beta_0, \beta_1 \in C(\mathbb{R}_+, \mathbb{R})$[13]. Under some technical assumptions, which are fulfilled in our setting, affine models have affine term structure. Most of the time no distinction is made between those two properties.

2.2 Affine short rate models with state variable

Now we can easily extend the short rate model with a state variable $X_t \in \mathbb{R}^n$ that follows a stochastic differential equation:

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t \quad (4)$$

with $X_0 = x_0 \in \mathbb{R}^n$.[14] Here the markovian structure is emphasized through the explicit dependence of the drift $b(t, X_t) \in \mathbb{R}^n$ and diffusion $\sigma(t, X_t) \in \mathbb{R}^{n \times d}$ of only t and X_t. Now we define the short rate through:

$$r_t = \delta_0 + \delta_1' X_t$$

where $\delta_0 \in \mathbb{R}$ and $\delta_1 \in \mathbb{R}^n$. Since we could use Ito's Lemma to calculate the dynamics of r_t [15] which would be given through a stochastic differential equation with more explicit drift and diffusion than in the previous section. Therefore the extension with a state variable is complete analogous to a short rate model only that the state variable drives all the dynamics

[13] For the general d dimensional case $\beta_0, \beta_1 \in C(\mathbb{R}_+, \mathbb{R}^d)$ and σ is a $d \times d$ diagonal matrix with $\sqrt{\alpha_0^l(t) + \alpha_1^l(t)r_t}$ $l \in \{1, ..., d\}$ on the digonal

[14] Or to be more precise since dX_t is just a short notation for: $X_t = x_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s$

[15] $f(x) := \delta_0 + \delta_1' x$ then $\partial_x f = \delta_1'$ and $\partial_x^2 f = 0$:

$$r_t = f(X_t) = f(X_0) + \int_0^t \delta_1' dX_s = f(X_0) + \int_0^t \delta_1' b(s, X_s)ds + \int_0^t \delta_1' \sigma(s, X_s)dW_s$$

We now define $r_0 = f(X_0)$, $\tilde{b}_t = \delta_1' b(t, X_t)$ and $\tilde{\sigma}_t = \delta_1' \sigma(t, X_t)$ and end up with a short rate model with $\tilde{b}, \tilde{\sigma}$.

in the system.
Thus we call such a model affine or more accurate has the affine term structure property if there is $A \in C^1(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R})$ and $C \in C^1(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R}^n)$ with $A(T,T) = C(T,T) = 0$ such that:

$$B(t,T) = \mathbb{E}^Q(e^{-\int_t^T r_s ds}|\mathscr{F}_t) = \exp(-A(t,T) + C(t,T)'X_t)$$

Furthermore in a model with affine term structure the zero coupon yield (continuously compounded) is given by:

$$y(t,T) := -\frac{1}{T-t}\ln(B(t,T)) = \frac{A(t,T)}{T-t} - \frac{C(t,T)'}{T-t}X_t$$

We see that it is affine in the state variable X_t.

So if we start with a state model characterised by $(\delta_0, \delta_1, b, \sigma, \lambda)$ we get a short rate model $(r_0, \tilde{b}, \tilde{\sigma}, \lambda)$ (see footnote 15). When this short rate model is affine (has ATS) the zero coupon bond prices $B(t,T)$ are affine in the state variable. The Q dynamic of the state variable is given by (due to the linearity of the stochastic integral with respect to the integrator):

$$dX_t = b(t,X_t)dt + \sigma(t,X_t)dW_t = b(t,X_t)dt + \sigma(t,X_t)(dW_t + \lambda_t dt - \lambda_t dt)$$

$$= (b(t,X_t) - \sigma(t,X_t)\lambda_t)dt + \sigma(t,X_t)dW_t^Q \tag{5}$$

3 Construction of forward rates

In this section I will briefly summarize the construction of the forward rate data used in the GATSM and SRTSM. These rates are the observations $(y_t)_{t=1}^T$ in the Kalman filter problem. This section is based on the paper "The U.S. Treasury Yield Curve: 1961 to the Present, Refet S. Gürkaynak, Brian Sack, and Jonathan H. Wright" [4],[14].
The basic idea is to estimate a yield curve with a fixed functional form. The bond market is not perfectly homogeneous since the treasury does not issue a continuum (over maturity) of bills each day and every issue of securities has different specifications (i.e. maturities, coupon payments, liquidity, ...). We can observe only zero coupon prices for specific maturities $T_1, ..., T_p$ (i.e. 1 month, 3 month, 1 year, 5 years, 10 years and many more). Therefore we have to fit a yield curve through the yields we can observe at any given moment in the US treasury market. Our goal is to get a yield curve for every point in time (i.e. for every day or month we get a new yield curve) that maps the time to maturity to the yield.
We begin by deriving the link between yields and forward rates, after that we look at the estimation of the curve.

First we re-parametrize from the notation used in the short rate model section where we denote by $B(t,T)$ the price of a zero coupon bond at time t that matures at T $(t \leq T)$. These prices are observable in the market. The time to maturity is given by $n := T - t$.

Therefore we can write $B(t, t+n)$ and the continuously compounded yield of a zero coupon bond with duration n is given by (compare short rate section[16]):

$$y_t(n) = -\frac{\ln(B(t, t+n))}{n} \tag{6}$$

Forward rates:

To express the forward rate through yields we use the following investment strategy:

at t: buy one $n+m$ zero coupon bond for $B(t, t+n+m)$, sell $\frac{B(t,t+n+m)}{B(t,t+n)}$ n bonds for $B(t, t+n)$

at t+n: the n bond matures $B(t+n, t+n) = 1$ so the investor must pay $\frac{B(t,t+n+m)}{B(t,t+n)}$

at t+n+m: the n+m bond matures

The cash flows of the strategy are given by:

at t: $-1B(t, t+n+m) + \frac{B(t,t+n+m)}{B(t,t+n)}B(t, t+n) = 0$

at t+n: $-\frac{B(t,t+n+m)}{B(t,t+n)}$

at t+n+m: the investor gets $B(t+n+m, t+n+m) = 1$

So basically with this strategy the investor can lock in the cash-flows of a m year bond n years ahead. The continuous compound yield of this strategy also called the forward rate is by given by:

$$f_t(n, m) = -\frac{1}{m}\ln(\frac{B(t, t+n+m)}{B(t, t+n)}) = \frac{1}{m}(-\ln(B(t, t+n+m)) + \ln(B(t, t+n))$$

$$= \frac{1}{m}(-\frac{n+m}{n+m}\ln(B(t, t+n+m)) - -\frac{n}{n}\ln(B(t, t+n)) = \frac{1}{m}((m+n)y_t(n+m) - ny_t(n)) \tag{7}$$

Equation (7) will be used for constructing the forward rates out of the yield curve. The instantaneous forward rate is the limit as m goes to zero[17] (recall the definition of differential quotient).:

$$f_t(n, 0) = \lim_{m \to 0} f_t(n, m) = -\partial_2 \ln(B(t, t+n))$$

Integrating the last equation and using (6) we get:

$$y_t(n) = \frac{1}{n}\int_0^n f_t(x, 0)dx$$

[16] $y(t, T) = y(t, t+n) = -\frac{\ln(B(t,t+n))}{t+n-t} = -\frac{\ln(B(t,t+n))}{n}$

[17] A standard assumption in financial mathematics is that the price of a zero coupon bond $B(t, T)$ is continuous in the first argument and continuously differentiable in the second. ∂_2 denotes the derivative of a function with respect to the second argument.

and due to (6), it follows

$$B(t,t+n) = e^{-ny_t(n)} = e^{-\int_0^n f_t(x,0)dx} \tag{8}$$

Note that $f_t(0,0)$ (the instantaneous interest rate at time t) corresponds to the short rate introduced in the short rate section[18].

Yield curve:

For estimating the yield curve we need a functional form to fit to the observed data. The function used by Gürkaynak, Sack and Wright is an extended Nelson-Siegel function, so the instantaneous forward curve at time t is given through[19]:

$$f_t(n,0) = \beta_{0,t} + \beta_{1,t} e^{-(\frac{n}{\tau_{1,t}})} + \beta_{2,t}(\frac{n}{\tau_{1,t}})e^{-(\frac{n}{\tau_{1,t}})} + \beta_{3,t}(\frac{n}{\tau_{2,t}})e^{-(\frac{n}{\tau_{2,t}})}$$

using the formula from above and integrating we get the formula for the yield curve:

$$y_t(n) = \beta_{0,t} + \beta_{1,t}\frac{1-e^{-(\frac{n}{\tau_{1,t}})}}{(\frac{n}{\tau_{1,t}})} + \beta_{2,t}(\frac{1-e^{-(\frac{n}{\tau_{1,t}})}}{(\frac{n}{\tau_{1,t}})} - e^{-(\frac{n}{\tau_{1,t}})}) + \beta_{3,t}(\frac{1-e^{-(\frac{n}{\tau_{2,t}})}}{(\frac{n}{\tau_{2,t}})} - e^{-(\frac{n}{\tau_{2,t}})}) \tag{9}$$

Gürkaynak, Sack and Wright fit the theoretical bond prices given by (8) and (9) to the observed bond prices of a given day. This is done by minimizing the squared difference between the theoretical and observed prices with respect to $(\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \tau_{1,t}, \tau_{2,t})$. For a fixed t we observe $B^o(t,T_1), ..., B^o(t,T_p)$ in the US treasury market and solve the following minimization problem $(n_s = T_s - t)$:

$$\min_{\beta_{1,t},\beta_{2,t},\beta_{3,t},\tau_{1,t},\tau_{2,t}} \sum_{s=1}^{p} (B^o(t,t+n_s) - e^{-n_s y_t(n_s)})^2$$

where $y_t(n)$ is given by (9). Gürkaynak, Sack and Wright publish the resulting parameters $\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \tau_{1,t}, \tau_{2,t}$ for every day. This means we do not need to fit but just use the published parameters. For the GATSM and SRTSM it is sufficient to use a monthly time period, therefore we pick out the parameters for every month end (we use a monthly grid for t). Out of these parameters the forward rates are constructed using equation (9) and (7).

4 Gaussian affine term structure model, GATSM

4.1 The GATSM model

Next I present the standard Gaussian affine term structure model GATSM. The GATSM is a discrete time model and you can think about it as discretely sampled continuous time model. It is the workhorse model of affine term structure model due to its easy implementation and

[18] We have different pricing formulas: $B(t,T) = \mathbb{E}^Q(e^{-\int_t^T r_s ds}|\mathscr{F}_t) = \mathbb{E}^Q(e^{-\int_t^T f_s(0,0)ds}|\mathscr{F}_t) = e^{-\int_0^{T-t} f_t(x,0)dx}$

[19] if we set $\beta_{3,t} = 0$ we end up at the normal Nelson-Siegel yield curve.

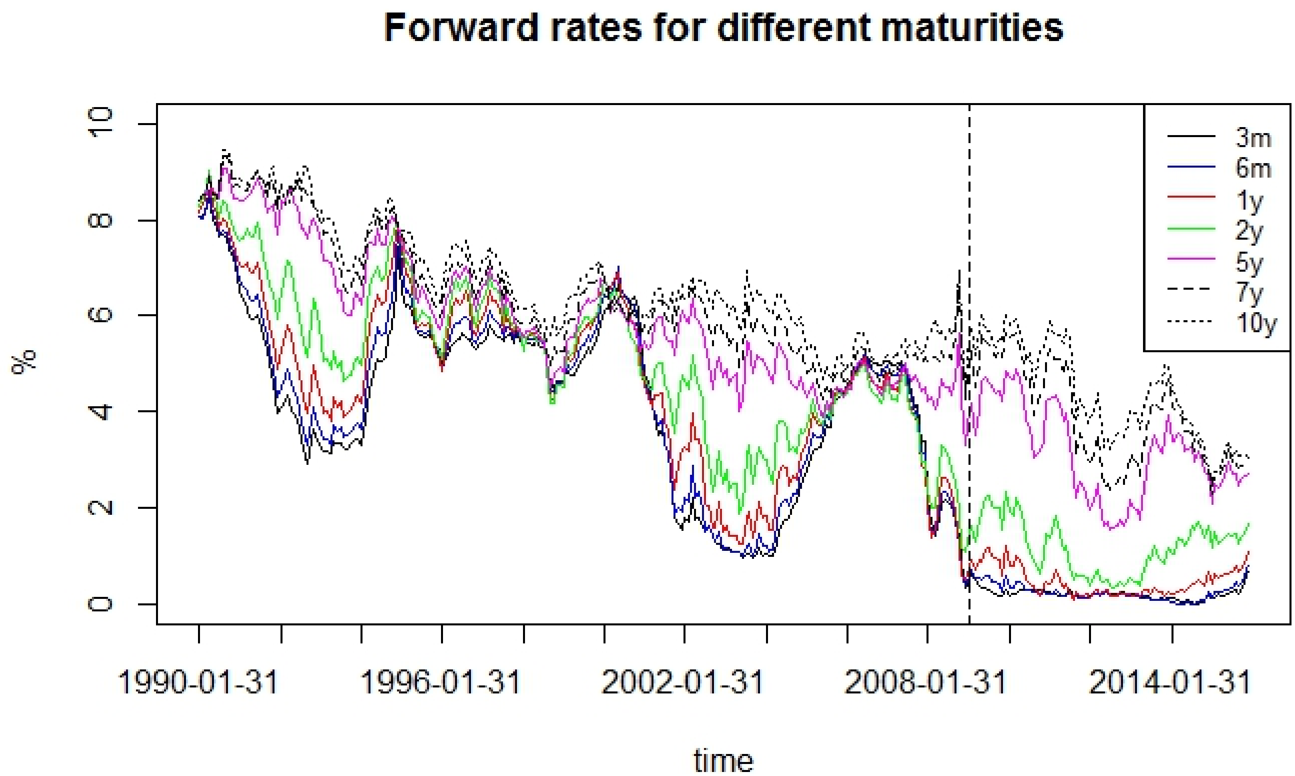

Figure 2: Forward rates created from Gurkaynak, Sack and Wright data set.

structure. In the next section we will extend it and derive the shadow rate term structure model. For the GATSM and SRTSM we use a monthly time grid. This means if t goes to t + 1 a step of one month is made. The GATSM is given by:

$$r_t = \delta_0 + \delta_1' X_t \tag{10}$$

with $\delta_0 \in \mathbb{R}, \delta_1 \in \mathbb{R}^n$. r_t is the short rate (compare with r_t in the Affine interest rate structure model section). The dynamics of the 3 dimensional state variable $X_t \in \mathbb{R}^3$, under the real world measure P, are given by the following VAR(1) (compare with (4)):

$$X_{t+1} = \mu + \rho X_t + \Sigma \epsilon_{t+1} \tag{11}$$

with $\mu \in \mathbb{R}^3$, $\rho \in \mathbb{R}^{3\times3}$, $\Sigma \in \mathbb{R}^{3\times3}$ a lower triangular and $\epsilon_t \sim N(0, I)$. It is common knowledge in the financial literature that 3 state variables are sufficient and extra dimensions for the state do not add much explanatory power to the model (see Wu and Xia [5]). Analogous to the continuous time short rate models the Q dynamics are characterised by the stochastic discount factor (compare with equation (2))[20]:

[20] Suppose you want to discount from $T = t + n$ to t, where $n \in \mathbb{N}$. Then we have to use:

$$\prod_{i=t}^{T} \tilde{M}_i = e^{-\sum_{j=0}^{n-1} r_{t+j} + \sum_{j=0}^{n-1} \lambda_{t+j}' \epsilon_{t+j+1} - \frac{1}{2} \sum_{j=0}^{n-1} \lambda_{t+j}' \lambda_{t+j}}$$

$$\ln \tilde{M}_{t+1} = -r_t + \lambda_t' \epsilon_{t+1} - \frac{1}{2}\lambda_t' \lambda_t$$

with λ_t given by:

$$\lambda_t = \lambda_0 + \lambda_1 X_t$$

with $\lambda_0 \in \mathbb{R}^3$ and $\lambda_1 \in \mathbb{R}^{3\times 3}$. According to Wu and Xia [5] the Q dynamics are given by (compare with Duffee (2002)[17], [6] and (5)):

$$X_{t+1} = \mu + \rho X_t + \Sigma(\epsilon_{t+1} + \lambda_t - \lambda_t) = \mu^Q + \rho^Q X_t + \Sigma \epsilon_{t+1}^Q \tag{12}$$

with $\mu^Q = \mu - \Sigma\lambda_0$, $\rho^Q = \rho - \Sigma\lambda_1$ and $\epsilon_t^Q = \epsilon_t + \lambda_{t-1}$

4.2 Derivation of forward prices in the GATSM:

Next we derive a formula how the short rate or equivalently the state X_t determines the forward rates. This formula will be used as observation equation when estimating the model with the Kalman filter. First we need to calculate the conditional moments of the short rate before we can derive the final formula for the forward rate.
So we solve the r_t equation recursively to get a handy representation for the calculations of conditional moments. For convenience I will slightly abuse notation and write $\epsilon_t \hat{=} \Sigma\epsilon_t \sim N(0, \Sigma\Sigma')$:

$$r_{t+n} = \delta_0 + \delta_1' X_{t+n} = \delta_0 + \delta_1'(\mu\rho X_{t+n-1} + \epsilon_{t+n}) = \delta_0 + \delta_1'\mu\rho(\mu + \rho X_{t+n-2} + \epsilon_{t+n-1}) + \delta_1'\epsilon_{t+n} =$$
$$= ... = \delta_0 + \delta_1' \sum_{j=0}^{n-1} \rho^j \mu + \delta_1' \rho^n X_t + \delta_1' \sum_{j=0}^{n-1} \rho^j \epsilon_{t+n-j}$$

or analogue with the Q dynamics of the state variable:

$$r_{t+n} = \delta_0 + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \mu^Q + \delta_1'(\rho^Q)^n X_t + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \epsilon_{t+n-j}^Q \tag{13}$$

Next we define the following quantities:

$$\bar{a}_n := \delta_0 + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \mu^Q$$

$$a_n := \bar{a}_n - \frac{1}{2}\delta_1' (\sum_{j=0}^{n-1} (\rho^Q)^j) \Sigma\Sigma' (\sum_{j=0}^{n-1} (\rho^Q)^j)' \delta_1$$

$$b_n' := \delta_1' (\rho^Q)^n$$

This allows us to write $r_{t+n} = \bar{a}_n + b_n' X_t + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \epsilon_{t+n-j}^Q$. Next we compute the conditional moments of the shadow rate, for convenience we will write for the conditional

as dicount factor since $\tilde{M}_t$ discounts just one period. This can be interpreted as discretely sampled discount factor from the continous case.

expectation and variance, $\mathbb{E}_t^Q(X) := \mathbb{E}^Q(X|\mathscr{F}_t)$ and $\mathbb{V}ar_t^Q(X) := \mathbb{V}ar^Q(X|\mathscr{F}_t)$:

$$\mathbb{E}_t^Q(r_{t+n}) = \mathbb{E}_t^Q(\bar{a}_n + b_n' X_t + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \epsilon_{t+n-j}^Q) = \bar{a}_n + b_n' X_t + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \mathbb{E}_t^Q(\epsilon_{t+n-j}^Q)$$

where we used that $\bar{a}_n + b_n' X_t$ is $\mathscr{F}_t = \sigma(\epsilon_s : \forall s \in \mathbb{N}, s \leq t)$ measurable with respect to the natural Filtration generated by the ϵ_t. Furthermore since ϵ_t is i.i.d. and normal distributed it follows $\mathbb{E}_t^Q(\epsilon_{t+s}) = \mathbb{E}^Q(\epsilon_{t+s}) = 0 \; \forall s \in \mathbb{N} : s > 0$ due to the independence. Therefore we get:

$$\mathbb{E}_t^Q(r_{t+n}) = \bar{a}_n + b_n' X_t \tag{14}$$

Next we compute the conditional variance:

$$\mathbb{V}ar_t^Q(r_{t+n}) = \mathbb{V}ar_t^Q(\delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \epsilon_{t+n-j}^Q) = \delta_1' \mathbb{V}ar_t^Q(\sum_{j=0}^{n-1} (\rho^Q)^j \epsilon_{t+n-j}^Q) \delta_1 =$$

due to the independence of the ϵ_t we get:

$$= \delta_1' \sum_{j=0}^{n-1} \mathbb{V}ar_t^Q((\rho^Q)^j \epsilon_{t+n-j}^Q) \delta_1 = \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \Sigma\Sigma' ((\rho^Q)')^j \delta_1 =: (\sigma_n^Q)^2$$

Now we want to show the following equality:

$$\frac{1}{2}(\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) - \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j})) = \bar{a}_n - a_n \tag{15}$$

where the right hand side is given by $\bar{a}_n - a_n = \frac{1}{2}\delta_1' (\sum_{j=0}^{n-1} (\rho^Q)^j) \Sigma\Sigma' (\sum_{j=0}^{n-1} (\rho^Q)^j)' \delta_1$.

Proof of equation (15):

First we observe that $\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) = \mathbb{V}ar_t^Q(\sum_{j=1}^{n} \delta_1' \sum_{i=0}^{j-1} (\rho^Q)^i \epsilon_{t+j-i}^Q)$ due to equation (13). The $\mathscr{F}_t$ measurable terms $\bar{a}_j + b_j' X_t$ drop out since they are like constants.
Furthermore we can write:

$$\sum_{j=1}^{n} \delta_1' \sum_{i=0}^{j-1} (\rho^Q)^i \epsilon_{t+j-i}^Q$$

as the sum of the entries of the following matrix:
define $A_j := \sum_{i=0}^{j-1} (\rho^Q)^i$ and collecting terms in the diagonals where the ϵ have the same index, we get:

$$\sum_{j=1}^{n} \delta_1' \sum_{i=0}^{j-1} (\rho^Q)^i \epsilon_{t+j-i}^Q = \delta_1' \sum_{j=1}^{n} A_j \epsilon_{t+n+1-j}^Q$$

Using the equation above and the independence of ϵ_t^Q with $\epsilon_s^Q \; \forall s \neq t$ we get:

$$\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) = \mathbb{V}ar_t^Q(\delta_1' \sum_{j=1}^{n} A_j \epsilon_{t+n+1-j}^Q) = \delta_1' \sum_{j=1}^{n} A_j \Sigma\Sigma' A_j' \delta_1$$

$$\begin{array}{llllllllll}
 & \text{i = 0} & & \text{i = 1} & & \text{i = 2} & & \dots & & \text{i = n-1} \\
\text{j = 1} & I\epsilon_{t+1}^Q & & & & & & & & \\
 & + & & & & & & & & \\
\text{j = 2} & I\epsilon_{t+2}^Q & + & \rho^Q\epsilon_{t+1}^Q & & & & & & \\
 & + & & & & & & & & \\
\text{j = 3} & I\epsilon_{t+3}^Q & + & \rho^Q\epsilon_{t+2}^Q & + & (\rho^Q)^2\epsilon_{t+1}^Q & & & & \\
 & + & & + & & + & & & & \\
\vdots & \vdots & & \vdots & & & & \ddots & & \\
 & + & & + & & & & & & \\
\text{j = n} & I\epsilon_{t+n}^Q & + & \rho^Q\epsilon_{t+n-1}^Q & + & \dots & + & (\rho^Q)^{n-2}\epsilon_{t+2}^Q & + & (\rho^Q)^{n-1}\epsilon_{t+1}^Q
\end{array}$$

Inserting the equation from above into the left hand side of (15) yields:

$$\frac{1}{2}(\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) - \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j})) = \frac{1}{2}\delta_1' A_n \Sigma\Sigma' A_n' \delta_1 = \frac{1}{2}\delta_1'(\sum_{j=0}^{n-1}(\rho^Q)^j)\Sigma\Sigma'(\sum_{j=0}^{n-1}(\rho^Q)^j)'\delta_1 = \bar{a}_n - a_n$$

This is the equation (15) we wanted to show. □
Now we have everything in place to derive the formula for the forward rates in the GATSM.

Derivation of forward formula in GATSM:

Note that in a discrete setting we have (see (3) and (6)):

$$n \cdot y_t(n) = \frac{n}{n} - \ln(B(t, t+n)) = -\ln(\mathbb{E}_t^Q(e^{-\int_t^T r_s ds})) = -\ln(\mathbb{E}_t^Q(e^{-\int_0^n r_{t+s} ds})) \hat{=} -\ln(\mathbb{E}_t^Q(e^{-\sum_{j=0}^{n-1} r_{t+j} 0}))$$

The forward rate between t+n and t+n+1 is given by (see (3), (6) and (7)):

$$f_{t,n,n+1} = f_t(n, n+1) = (n+1)y_t(n+1) - n y_t(n) = -\ln(\mathbb{E}_t^Q(e^{-\sum_{j=0}^{n} r_{t+j}})) + \ln(\mathbb{E}_t^Q(e^{-\sum_{j=0}^{n-1} r_{t+j}})) =$$

using the approximation $\ln(\mathbb{E}(e^X)) \approx \mathbb{E}(X) + \frac{1}{2}\mathbb{V}ar(X)$[21] which is accurate for normal distributed random variables since the moment generating function of a normal variable is $\mathbb{E}(e^{tX}) = e^{\mathbb{E}(X)t + \frac{1}{2}\mathbb{V}ar(X)t^2}$

$$= -\ln(e^{-r_t}\mathbb{E}_t^Q(e^{-\sum_{j=1}^{n} r_{t+j}})) + \ln(e^{-r_t}\mathbb{E}_t^Q(e^{-\sum_{j=1}^{n-1} r_{t+j}})) = -\ln(\mathbb{E}_t^Q(e^{-\sum_{j=1}^{n} r_{t+j}})) + \ln(\mathbb{E}_t^Q(e^{-\sum_{j=1}^{n-1} r_{t+j}})) =$$

[21] To see this define the cumulant generating function of X by $\kappa_X(t) = \ln(\mathbb{E}(e^{tX})) = \ln(M_x(t))$ where $M_X(t)$ denotes the moment generating function. We observe that $\partial_t \kappa_X(t)|_{t=0} = \frac{\partial_t M_X(t)}{M_X(t)}|_{t=0} = \mathbb{E}(X)$ and $\partial_t^2 \kappa_X(t)|_{t=0} = \frac{\partial_t^2 M_X(t) M_X(t) - (\partial_t M_X(t))^2}{M_X(t)^2}|_{t=0} = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{V}ar(X)$ where we used that $\mathbb{E}(X^n) = \partial_t^n M_X(t)|_{t=0}$. Next we approximate the kumulant with a Taylor expansion of second order around the support point $t_0 = 0$ and get:

$$\kappa_X(t) \approx \kappa_X(0) + \partial_t \kappa_X(0) t + \frac{\partial_t^2 \kappa_X(0)}{2} t^2 = \mathbb{E}(X)t + \frac{\mathbb{V}ar(X)}{2} t^2$$

In the end it follows:

$$\ln(\mathbb{E}(e^X)) = \kappa_X(1) \approx \mathbb{E}(X) + \frac{\mathbb{V}ar(X)}{2}$$

$$= -\mathbb{E}_t^Q(-\sum_{j=1}^{n} r_{t+j}) - \frac{1}{2}\mathbb{V}ar_t^Q(-\sum_{j=1}^{n} r_{t+j}) + \mathbb{E}_t^Q(-\sum_{j=1}^{n-1} r_{t+j}) + \frac{1}{2}\mathbb{V}ar_t^Q(-\sum_{j=1}^{n-1} r_{t+j}) =$$

using $\mathbb{E}_t^Q(r_{t+n}) = \bar{a}_n + b_n' X_t$ (14) and $\frac{1}{2}(\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) - \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j})) = \bar{a}_n - a_n$ (15) gives:

$$= \mathbb{E}_t^Q(r_{t+n}) - \frac{1}{2}(\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) - \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j})) = \bar{a}_n + b_n' X_t - (\bar{a}_n - a_n) \tag{16}$$

Therefore we get the following formula for the forward rates in the GATSM:

$$f_{t,n,n+1} = a_n + b_n' X_t \tag{17}$$

4.3 Summary of GATSM

In the GATSM we have the equations (11) and (17), the others where just used to derive the forward rate formula. I present them like they are used in the Kalman filter problem. The transition equation is given by the dynamics of the state X_t under the real world measure P (see (11)):

Transition equation:

$$X_{t+1} = \mu + \rho X_t + \Sigma \epsilon_{t+1} \tag{18}$$

with $\epsilon_t \sim N(0, \Sigma\Sigma')$. (17) is used to derive the observation equation. We use seven different maturities (3months, 6m, 1year, 2y, 5y, 7y, 10y) = $(n_1, n_2, ..., n_7)$. Due to market inefficiencies and pricing errors we have to modify (17) by a stochastic error term ν_t. The observed forward rates are generated by the procedure described in the section 3. Putting everything together the observation equation is given through:

Measurement/ observation equation:

$$f_t^o = a + b' X_t + \nu_t \tag{19}$$

where $f_t^o = (f_{t,n_1,n_1+1}^o, ..., f_{t,n_7,n_7+1}^o)' \in \mathbb{R}^7$ is the stacked vector with the forward rates for the seven maturities and the 7 dimensional measurement error $\nu_t \sim N(0, \omega I_7)$, $\omega \in \mathbb{R}_+$, $a = (a_{n_1}, ..., a_{n_7})' \in \mathbb{R}^7$ and $b' = (b_{n_1}, ..., b_{n_7})' \in \mathbb{R}^{7\times 3}$. As a starting point I use $x_0 = 0$ and $P_0 = diag(100, 100, \frac{100}{144})$ like Wu and Xia [5].
a_n, $\bar{a}_n$ and b_n are given by:

$$\bar{a}_n := \delta_0 + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \mu^Q$$

$$a_n := \bar{a}_n - \frac{1}{2}\delta_1' (\sum_{j=0}^{n-1} (\rho^Q)^j) \Sigma\Sigma' (\sum_{j=0}^{n-1} (\rho^Q)^j)' \delta_1$$

$$b'_n := \delta'_1(\rho^Q)^n$$

5 Shadow rate term structure model, SRTSM

5.1 The SRTSM model

The shadow rate term structure model is given by the following equations. Everything is completely analogous except for the nominal short rate that is given by:

$$r_t = \max(\bar{r}, s_t) \tag{20}$$

where $\bar{r}$ is the lower bound (i.e. a small positive constant, for the estimation I use $\bar{r} = 0.25$). The shadow rate is defined like the short rate in the GATSM

$$s_t = \delta_0 + \delta'_1 X_t \tag{21}$$

with $\delta_0 \in \mathbb{R}, \delta_1 \in \mathbb{R}^n$. It follows that we have the formulas (13), (14) and (15) for the shadow rate s_t since everything works out completely analogously. The dynamics of the state variable X_t is given through the following VAR(1):

$$X_{t+1} = \mu + \rho X_t + \Sigma\epsilon_{t+1} \tag{22}$$

with $\epsilon_t \sim N(0, I)$. Analogue to the continuous time short rate models and the GATSM the Q dynamics are given by:

$$X_{t+1} = \mu + \rho X_t + \Sigma(\epsilon_{t+1} + \lambda_t - \lambda_t) = \mu^Q + \rho^Q X_t + \Sigma\epsilon^Q_{t+1} \tag{23}$$

with $\mu^Q = \mu - \Sigma\lambda_0$, $\rho^Q = \rho - \Sigma\lambda_1$ and $\epsilon^Q_t = \epsilon_t + \lambda_{t-1}$

5.2 Derivation of the forward rate formula

We notice that we have (16) for the forward rate between t+n and t+n+1 in the SRTSM too. Since we can follow the same reasoning we used to obtain (16), only this time the approximation is not exact. In the SRTSM the nominal short rate r_t (and the sum of it) is not normal distributed due to the maximum with $\bar{r}$. Therefore we have:

$$f^o_{t,n,n+1} \approx \mathbb{E}^Q_t(r_{t+n}) - \frac{1}{2}(\mathbb{V}ar^Q_t(\sum_{j=1}^{n} r_{t+j}) - \mathbb{V}ar^Q_t(\sum_{j=1}^{n-1} r_{t+j})) \tag{24}$$

Next we calculate $\mathbb{E}^Q_t(r_{t+n})$ and approximate the term with the variances. It holds that $r_{t+n} = \max(\bar{r}, s_{t+n})$ with $s_{t+n}|\mathscr{F}_t \sim N(\bar{a}_n + b'_n X_t, (\sigma^Q_n)^2)$ (under the measure Q), for convenience I write $\mu_n = \bar{a}_n + b'_n X_t$ and define:

$$\alpha_n := \frac{\bar{r} - \mu_n}{\sigma^Q_n}$$

$$\mathbb{E}_t^Q(r_{t+n}) = \bar{r}\mathbb{E}_t^Q(1_{\{s_{t+n}\le\bar{r}\}}) + \mathbb{E}_t^Q(s_{t+n}1_{\{s_{t+n}>\bar{r}\}}) = \bar{r}\Phi(\alpha_n) + \mathbb{E}_t^Q(s_{t+n}1_{\{s_{t+n}>\bar{r}\}}) \tag{25}$$

due to $\mathbb{E}_t^Q(1_{\{s_{t+n}\le\bar{r}\}}) = Q_t(\{s_{t+n}\le\bar{r}\}) = Q_t(\{\frac{s_{t+n}-\mu_n}{\sigma_n^Q}\le\frac{\bar{r}-\mu_n}{\sigma_n^Q}\}) = \Phi(\alpha_n)$ where $\Phi(x)$ denotes the distribution function of a standard normal random variable. We write the second term in (25) as integral and using the transformation $y = \frac{x-\mu_n}{\sigma_n^Q}$ with $dx = \sigma_n^Q dy$ it follows:

$$\mathbb{E}_t^Q(s_{t+n}1_{\{s_{t+n}>\bar{r}\}}) = \frac{1}{\sqrt{2\pi}\sigma_n^Q}\int_{\bar{r}}^{\infty} xe^{-(\frac{x-\mu_n}{\sqrt{2}\sigma_n^Q})^2}dx = \frac{1}{\sqrt{2\pi}}(\sigma_n^Q\int_{\alpha_n}^{\infty} ye^{-\frac{y^2}{2}}dy + \mu_n\int_{\alpha_n}^{\infty} e^{-\frac{y^2}{2}}dy) =$$

We denote the density function of a standard normal distribution by $\phi(x) := \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ and observe $\partial_x\phi(x) = -x\phi(x)$. Using this we get:

$$= \sigma_n^Q\int_{\alpha_n}^{\infty} -\partial_y\phi(y)dy + \mu_n(1-\Phi(\alpha_n)) = \sigma_n^Q(-\phi(\infty)+\phi(\alpha_n)+\mu_n\Phi(-\alpha_n) = \sigma_n^Q\phi(-\alpha_n)+\mu_n\Phi(-\alpha_n)$$

where we used the symmetry of the standard normal distribution $1-\Phi(\alpha_n) = \Phi(-\alpha_n)$ and $\phi(\alpha_n) = \phi(-\alpha_n)$. Therefore we continue from (25):

$$\mathbb{E}_t^Q(r_{t+n}) = \bar{r}(1-\Phi(-\alpha_n)) + \sigma_n^Q\phi(-\alpha_n) + \mu_n\Phi(-\alpha_n) = \bar{r} + \sigma_n^Q(\frac{\mu_n-\bar{r}}{\sigma_n^Q}\Phi(-\alpha_n) + \phi(-\alpha_n)) =$$
$$= \bar{r} + \sigma_n^Q(-\alpha_n\Phi(-\alpha_n) + \phi(-\alpha_n)) = \bar{r} + \sigma_n^Q g(-\alpha_n) \tag{26}$$

with the function $g(x) := x\Phi(x) + \phi(x)$.
For the second term in equation (24) we use the following approximations:

$$\mathbb{V}ar_t^Q(r_{t+n}) \approx Q_t(s_{t+n}\ge\bar{r})\mathbb{V}ar_t^Q(s_{t+n})$$
$$\mathbb{C}ov_t^Q(r_{t+j}, r_{t+n}) \approx Q_t(s_{t+j}\ge\bar{r}, s_{t+n}\ge\bar{r})\mathbb{C}ov_t^Q(s_{t+j}, s_{t+n}) =$$

using $Q_t(s_{t+j}\ge\bar{r}|s_{t+n}\ge\bar{r}) \approx 1$ for $j \in \{1, ..., n-1\}$ gives:

$$= Q_t(s_{t+j}\ge\bar{r}|s_{t+n}\ge\bar{r})Q_t(s_{t+n}\ge\bar{r})\mathbb{C}ov_t^Q(s_{t+j}, s_{t+n}) \approx Q_t(s_{t+n}\ge\bar{r})\mathbb{C}ov_t^Q(s_{t+j}, s_{t+n})$$

Wu and Xia (2016) [5] showed through simulation studies and analytic analysis that the approximation error for the SRTSM is just a few basis points and therefore negligible.
We also need the following relation:

$$\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) = \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j}+r_{t+n}) = \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j})+2\mathbb{C}ov_t^Q(\sum_{j=1}^{n-1} r_{t+j}, r_{t+n})+\mathbb{V}ar_t^Q(r_{t+n}) \tag{27}$$

using equation (27) from above, it follows:

$$\frac{1}{2}(\mathbb{V}ar_t^Q(\sum_{j=1}^{n} r_{t+j}) - \mathbb{V}ar_t^Q(\sum_{j=1}^{n-1} r_{t+j})) = \frac{1}{2}(2\sum_{j=1}^{n-1}\mathbb{C}ov_t^Q(r_{t+j}, r_{t+n}) + \mathbb{V}ar_t^Q(r_{t+n})) \approx$$

where we used equation (27) and (15) for the shadow rate s_t for the last equality below

$$\approx Q_t(s_{t+n}\ge\bar{r})\frac{1}{2}(2\sum_{j=1}^{n-1}\mathbb{C}ov_t^Q(s_{t+j}, s_{t+n}) + \mathbb{V}ar_t^Q(s_{t+n})) = \Phi(-\alpha_n)(\bar{a}_n - a_n) \tag{28}$$

since $Q_t(s_{t+n} \geq \bar{r}) = Q_t(\frac{s_{t+n}-\mu_n}{\sigma_n^Q} \geq \frac{\bar{r}-\mu_n}{\sigma_n^Q}) = 1 - \Phi(\frac{\bar{r}-\mu_n}{\sigma_n^Q}) = \Phi(-\alpha_n)$ with $\alpha_n := \frac{\bar{r}-\mu_n}{\sigma_n^Q}$ defined as above.

Furthermore we observe that $\partial_x g(x) = \Phi(x)$[22]. Now everything is in place to derive the observation equation of the SRTSM. Continuing from equation (24) using (26) and (28):[23]

$$f_{t,n,n+1} = \bar{r} + \sigma_n^Q g(-\alpha_n) - \Phi(-\alpha_n)(\bar{a}_n - a_n) = \bar{r} + \sigma_n^Q g(\frac{\mu_n - \bar{r}}{\sigma_n^Q}) + \Phi(\frac{\mu_n - \bar{r}}{\sigma_n^Q})(a_n - \bar{a}_n) =$$

$$= \bar{r} + \sigma_n^Q g(\frac{\bar{a}_n + b_n' X_t - \bar{r}}{\sigma_n^Q}) + \sigma_n^Q \frac{\partial g(\frac{\bar{a}_n + b_n' X_t - \bar{r}}{\sigma_n^Q})}{\partial \bar{a}_n}(a_n - \bar{a}_n) \approx \bar{r} + \sigma_n^Q g(\frac{a_n + b_n' X_t - \bar{r}}{\sigma_n^Q}) \quad (29)$$

□

Equation (29) is the final formula for the forward rates in the SRTSM. For a short comparison of the GATSM and SRTSM we plot the function $g(x)$ in Figure 3:

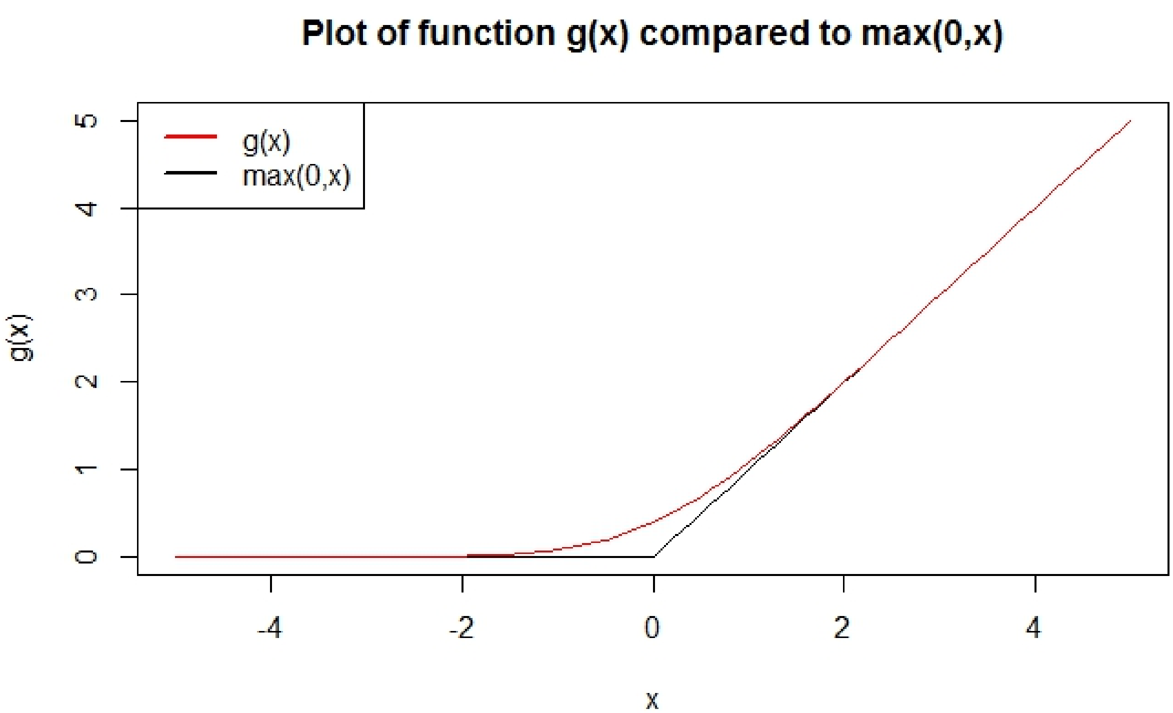

Figure 3: Plot of $g(x) = x \cdot \Phi(x) + \phi(x)$

In Figure 3 we note that for $x \geq 2$ $g(x) \approx x$. Therefore if $\frac{a_n + b_n' X_t - \bar{r}}{\sigma_n^Q} \geq 2$ $\forall t, n$ the SRTSM

[22] $\partial_x g(x) = \partial_x(x\Phi(x) + \phi(x)) = \Phi(x) + x\phi(x) - x\phi(x) = \Phi(x)$ using $\partial_x \phi(x) = -x\phi(x)$ once more

[23] Using $\sigma_n^Q \frac{\partial g(\frac{\bar{a}_n + b_n' X_t - \bar{r}}{\sigma_n^Q})}{\partial \bar{a}_n} = \sigma_n^Q \Phi(\frac{\bar{a}_n + b_n' X_t - \bar{r}}{\sigma_n^Q}) \frac{1}{\sigma_n^Q} = \Phi(\frac{\mu_n - \bar{r}}{\sigma_n^Q})$ and the first order Taylor approximation around $\bar{a}_n$ of $h(a_n) := \sigma_n^Q g(\frac{a_n + b_n' X_t - \bar{r}}{\sigma_n^Q}) \approx \sigma_n^Q g(\frac{\bar{a}_n + b_n' X_t - \bar{r}}{\sigma_n^Q}) + \frac{\partial \sigma_n^Q g(\frac{\bar{a}_n + b_n' X_t - \bar{r}}{\sigma_n^Q})}{\partial \bar{a}_n}(a_n - \bar{a}_n)$

equals the GATSM[24]. We will see empirically that this corresponds to times where the effective federal funds rate is not stuck at the lower bound.

5.3 Summary of SRTSM

Here I present the shadow rate term structure model equations. The equations (22) and (29) are used for the implementation. The other equations especially the one with the Q dynamics where just used to derive the forward rate formula. Like in the GATSM summary I present the SRTSM in the form it is used for the Kalman filter. Like in the GATSM the forward rates for the seven different maturities are used as observation. Furthermore due to market inefficiencies, pricing errors and the approximations used in the derivation, the forward rate formula has to be extended with a random error term ν_t. The transition equation is again given by the P dynamics of X_t:

Transition equation:

$$X_{t+1} = \mu + \rho X_t + \Sigma \epsilon_{t+1} \tag{30}$$

Observation/ measurement equation:

$$f_t^o = \bar{r}\iota_7 + G(X_t) + \nu_t \tag{31}$$

where $f_t^o = (f^o_{t,n_1,n_1+1}, ..., f^o_{t,n_7,n_7+1})' \in \mathbb{R}^7$ is the stacked vector with the forward rates for the seven maturities and the 7 dimensional measurement error $\nu_t \sim N(0, \omega I_7)$, $\omega \in \mathbb{R}_+$, $G(X_t) := (\sigma^Q_{n_1} g(\frac{a_{n_1}+b'_{n_1}X_t-\bar{r}}{\sigma^Q_{n_1}}), ..., \sigma^Q_{n_7} g(\frac{a_{n_7}+b'_{n_7}X_t-\bar{r}}{\sigma^Q_{n_7}}))' \in \mathbb{R}^7$ and $\iota_7 = (1, ..., 1)' \in \mathbb{R}^7$. As a starting point I use $x_0 = 0$ and $P_0 = diag(100, 100, \frac{100}{144})$ like Wu and Xia [5].
a_n, $\bar{a}_n$, b_n and $g(x)$ are given by:

$$\bar{a}_n := \delta_0 + \delta_1' \sum_{j=0}^{n-1} (\rho^Q)^j \mu^Q$$

$$a_n := \bar{a}_n - \frac{1}{2}\delta_1' (\sum_{j=0}^{n-1} (\rho^Q)^j) \Sigma\Sigma' (\sum_{j=0}^{n-1} (\rho^Q)^j)' \delta_1$$

$$b_n' := \delta_1' (\rho^Q)^n$$

$$g(x) := x \cdot \Phi(x) + \phi(x)$$

where $\Phi(x)$ is the distribution function and $\phi(x)$ the density function of a standard normal random variable. Like Wu and Xia [5] I use $\bar{r} = 0.25$ as a lower bound for the nominal short rate in the SRTSM.

[24] Due to $\bar{r} + \sigma^Q_n g(\frac{a_n+b'_nX_t-\bar{r}}{\sigma^Q_n}) \approx \bar{r} + \sigma^Q_n \frac{a_n+b'_nX_t-\bar{r}}{\sigma^Q_n} = a_n + b'_n X_t$ if the argument of the function g is larger then 2.

6 Kalman Filters and extension

In this section I will summarize the theory behind Kalman filters. They are very popular in the literature due to their easy and intuitive implementation. The section is based on the books of Dan Simon, Optimal State Estimation [20] and Robert Shumway and David Stoffer, Time series analysis and its applications [21].
First I will present the linear filter problem and then the derivation of the simple Kalman filter as minimal variance solution to the given problem. The simple Kalman filter is presented in a discrete time setting as it is used for estimating the Gaussian affine term structure model (GATSM). After that I introduce the extended Kalman filter for non-linear filter problems like it appears in the shadow rate term structure model (SRTSM).

6.1 The filter problem

Here I will present the filter problem to which the Kalman filter is the optimal solution. The objective is to estimate a non-observable n dimensional state variable $x_t \in \mathbb{R}^n$ where we can only observe noisy d dimensional measurements $y_t \in \mathbb{R}^d$ that are linear transformations of the state. The basic set-up is that we have a measurement and a transition equation. The transition equation gives the dynamics of the state variable and the measurement equation gives the link between observation and state.
The transitional dynamic of the state is given through the following equation:

$$x_{t+1} = \mu + F_t x_t + w_t \tag{32}$$

where $\mu \in \mathbb{R}^n$, $F_t \in \mathbb{R}^{n \times n}$ $\forall t$ describes the influence of the last state on the next one and $w_t \sim N(0, Q_t)$ is the stochastic noise term. The intercept μ is only relevant for the a priori and a posteriori predictions since for all covariance terms the constant μ drops out.

The measurement equation is given by:

$$y_t = c + H_t x_t + v_t \tag{33}$$

where $c \in \mathbb{R}^d$, $H_t \in \mathbb{R}^{d \times n}$ $\forall t$ and the noise term $v_t \sim N(0, R_t)$ is normal distributed. Furthermore we can demean the measurements $y_t^* = y_t - c$ since the observation intercept c is known in the filter problem. If we have a prediction for y_t^* we can easily get the prediction for y_t by adding c. Therefore without loss of generality we can set $c = 0$. Furthermore we assume that w_t and v_t are independent and therefore uncorrelated[25]. Out of this assumption it follows that (v_t, w_t) is joint normal distributed. Another assumption we have to make is that w_t and w_s are uncorrelated for $t \neq s$ and the same for the measurement noise. Furthermore we have to know the initial distribution x_0 of the state (since it is normal distributed it is suffice to know $\hat{x}_{0|0} := \mathbb{E}(x_0)$ and $\hat{P}_{0|0} := \mathbb{E}(x_0 x_0') - \hat{x}_{0|0}\hat{x}_{0|0}'$) and the initial state is assumed to be uncorrelated with the noise terms w_t, v_t. To summarize this assumption the following holds:

$$\mathbb{E}(w_t w_s') = Q_t \delta_{t-s} \tag{34}$$

$$\mathbb{E}(v_t v_s') = R_t \delta_{t-s} \tag{35}$$

[25] If v_t and w_t follow a joint normal distribution, independence is equivalent to being uncorrelated.

$$\mathbb{E}(w_t v_s') = 0 \qquad \forall s,t \in \mathbb{N} \tag{36}$$

$$\mathbb{C}ov(x_0 v_s') = \mathbb{C}ov(x_0 w_s') = 0 \qquad \forall s,t \in \mathbb{N} \tag{37}$$

where δ_{t-s} stands for the Kronecker delta ($\delta_{t-s} = 1$ for $t = s$ 0 otherwise). The goal is to get the best estimates for the normal distributed state x_t given only realisations of the measurements $y_1, ..., y_t$.

So first we analyse the Kalman filter in a setting where we assume that the parameters $(\mu, F_t, c, H_t, R_t, Q_t)$ are known then we discuss how to estimate them if they are not known. Furthermore for the GATSM and SRTSM we have the simplification $F_t = F$ and $H_t = H$ (the second is valid only for the GATSM). The solution of the given filter problem is a recursive algorithm that consists of a prediction step where x_t is predicted with all information up to time t $(y_1, ..., y_{t-1})$ and an update step where we update the prediction when the new measurement y_t comes in.

6.2 Derivation of the Kalman filter

The goal of the Kalman filter is to compute the conditional distribution of x_t given the information up to time t-1 or t, i.e. it is given the Information set $\mathscr{Y}_s := (y_1, ..., y_s)$ for s = t-1 or t respectively.
Due to the linear setup and the Gaussian assumption on the initial state and the noise sources, it follows that the random variables $(x_1, ..., x_t, y_1, ..., y_t)$ are jointly normal distributed. This implies that the conditional distribution $x_t|\mathscr{Y}_s$ is a normal distribution with expectation $\mathbb{E}(x_t|\mathscr{Y}_s)$ and variance $\mathbb{V}ar(x_t|\mathscr{Y}_s)$[26].
Furthermore the conditional expectation is linear and thus given by the projection of x_t on the space span$\{1, y_1, ..., y_s\}$. The conditional variance $\mathbb{V}ar(x_t|\mathscr{Y}_s)$ is constant and equal to the unconditional variance of the projection error $x_t - \mathbb{E}(x_t|\mathscr{Y}_s)$ (see Appendix).
For convenience we will use the following notation:

- $\hat{x}_{t|t} = \mathbb{E}(x_t|y_1, ..., y_t)$ a posteriori estimate of x_t that uses all information up to t including t, updated state
- $\hat{x}_{t|t-1} = \mathbb{E}(x_t|y_1, ..., y_{t-1})$ a priori estimate of x_t that uses all information up to t, predicted state
- $\tilde{x}_{t|k} = x_t - \hat{x}_{t|k}$ estimation error
- $P_{t|t} = \mathbb{V}ar_t(x_t - \hat{x}_{t|t}) = \mathbb{V}ar(x_t - \hat{x}_{t|t}) = \mathbb{E}(\tilde{x}_{t|t}\tilde{x}_{t|t}')$ a posteriori error covariance matrix or error variance of updated state estimation
- $P_{t|t-1} = \mathbb{V}ar_{t-1}(x_t - \hat{x}_{t|t-1}) = \mathbb{V}ar(x_t - \hat{x}_{t|t-1}) = \mathbb{E}(\tilde{x}_{t|t-1}\tilde{x}_{t|t-1}')$ a priori error covariance matrix or error variance of predicted state estimation
- $\hat{y}_{t|t-1} = H_t\hat{x}_{t|t-1}$ predicted measurement or if we incorporate the intercept $\hat{y}_{t|t-1} = c + H_t\hat{x}_{t|t-1}$

[26]For the normal distribution it holds that the whole distribution is determined by the expectation and variance.

- $\tilde{y}_t = y_t - \hat{y}_{t|t-1} = y_t - H_t\hat{x}_{t|t-1}$ prediction or forecast error (residuals), if we incorporate c we have to subtract it too

- $S_t = \mathbb{V}ar(\tilde{y}_t)$ variance matrix of the prediction errors/ residuals

Furthermore we notice that due to the Gaussian assumption (for joint normal random variables it holds that independence is equivalent to being uncorrelated) we have $\mathbb{E}(\tilde{x}_{t|k}\tilde{x}'_{t|k}|y_1, ..., y_k) = \mathbb{E}(\tilde{x}_{t|k}\tilde{x}'_{t|k})$ since the projection theorem (see Appendix and [24]) states that $\tilde{x}_{t|k}$ is orthogonal[27] (uncorrelated) to the space $span\{1, y_1, ..., y_k\}$ (see Theorem 1 in Appendix) we are projecting on. Due to the normality we get the independence and can drop the condition.

To get the a priori estimate under the assumption that the a posteriori estimate of the previous step is known we take the conditional expectation of the transition equation (31):

$$\hat{x}_{t|t-1} = \mathbb{E}(x_t|\mathscr{Y}_{t-1}) = \mu + F_{t-1}\mathbb{E}(x_{t-1}|\mathscr{Y}_{t-1}) + \mathbb{E}(w_{t-1}|\mathscr{Y}_{t-1}) = \mu + F_{t-1}\hat{x}_{t-1|t-1} \tag{38}$$

where we used $\mathbb{E}(w_{t-1}|\mathscr{Y}_{t-1}) = \mathbb{E}(w_{t-1}) = 0$ for the last equation. This holds due to the independence of w_t to v_t and itself for all lags $\neq 0$. Since into $\mathscr{Y}_{t-1}$ enter only the v_t and the w_s with $s < t-1$. Completely analogously we get:

$$\hat{y}_{t|t-1} = \mathbb{E}(y_t|\mathscr{Y}_{t-1}) = c + H_t\mathbb{E}(x_t|\mathscr{Y}_{t-1}) + \mathbb{E}(v_t|\mathscr{Y}_{t-1}) = c + H_t\hat{x}_{t|t-1}$$

Therefore it follows $\tilde{y}_t = y_t - \hat{y}_{t|t-1} = y_t - H_t\hat{x}_{t|t-1} - c$. Next we want to calculate the a priori error covariance matrix.

$$(x_t - \hat{x}_{t|t-1}) = (F_{t-1}x_{t-1} + w_{t-1} - F_{t-1}\hat{x}_{t-1|t-1}) = (F_{t-1}(x_{t-1} - \hat{x}_{t-1|t-1}) + w_{t-1})$$

Therefore we get the following for the error variance:

$$P_{t|t-1} = \mathbb{C}ov(x_t - \hat{x}_{t|t-1}) = \mathbb{C}ov(F_{t-1}(x_{t-1} - \hat{x}_{t-1|t-1}) + w_{t-1}) = \mathbb{C}ov(F_{t-1}(x_{t-1} - \hat{x}_{t-1|t-1})) + \mathbb{C}ov(w_{t-1}) =$$

the last equality holds because $(x_{t-1} - \hat{x}_{t-1|t-1})$ is uncorrelated with w_{t-1}. In the end we get:

$$P_{t|t-1} = F_{t-1}\mathbb{C}ov(x_{t-1} - \hat{x}_{t-1|t-1})F'_{t-1} + Q_t = F_{t-1}P_{t-1|t-1}F'_{t-1} + Q_t \tag{39}$$

Next we calculate the variance matrix S_t of the residuals $\tilde{y}_t$:

$$S_t = \mathbb{V}ar(\tilde{y}_t) = \mathbb{V}ar(y_t - H_t\hat{x}_{t|t-1}) = \mathbb{V}ar(H_t x_t + v_t - H_t\hat{x}_{t|t-1}) =$$

$(x_t - \hat{x}_{t|t-1})$ is uncorrelated with v_t, therefore:

$$= \mathbb{V}ar(H_t(x_t - \hat{x}_{t|t-1}) + v_t) = \mathbb{V}ar(H_t(x_t - \hat{x}_{t|t-1})) + \mathbb{V}ar(v_t) = H_t\mathbb{V}ar(x_t - \hat{x}_{t|t-1})H'_t + R_t = H_tP_{t|t-1}H'_t + R_t$$

The covariance between the residuals and the state are given by:

$$\mathbb{C}ov(x_t, \tilde{y}_t) = \mathbb{C}ov((x_t - \hat{x}_{t|t-1}) + \hat{x}_{t|t-1}, y_t - H_t\hat{x}_{t|t-1}) = \mathbb{C}ov((x_t - \hat{x}_{t|t-1}) + \hat{x}_{t|t-1}, H_t x_t - H_t\hat{x}_{t|t-1} + v_t) =$$
$$= \mathbb{C}ov((x_t - \hat{x}_{t|t-1}) + \hat{x}_{t|t-1}, H_t(x_t - \hat{x}_{t|t-1}) + v_t) = \mathbb{C}ov(x_t - \hat{x}_{t|t-1}, H_t(x_t - \hat{x}_{t|t-1})) = P_{t|t-1}H'$$

[27] With respect to $< X, Y > := \mathbb{E}(XY)$ on L^2.

where the second equality in the last equation is due to the fact that $(x_t - \hat{x}_{t|t-1})$ and v_t are uncorrelated due to the assumed correlation structure and $(x_t - \hat{x}_{t|t-1})$ is uncorrelated to $\hat{x}_{t|t-1}$ due to the projection theorem. Now we calculate the posteriori expectation $\hat{x}_{t|t}$ using theorem 1 (Appendix):

$$\hat{x}_{t|t} = \mathbb{E}(x_t|y_1, ..., y_t) = \mathrm{Pr}_{span\{1,y_1,...,y_t\}} x_t = \mathrm{Pr}_{span\{1,y_1,...,y_{t-1}\}} x_t + \mathrm{Pr}_{span\{\tilde{y}_t\}} x_t =$$

The last equality is due to the fact that $span\{1, y_1, ..., y_t\} = span\{1, y_1, ..., y_{t-1}\} \oplus span\{\tilde{y}_t\}$[28] since due to the projection theorem $\tilde{y}_t$ is orthogonal to $span\{1, y_1, ..., y_{t-1}\}$, therefore we get the direct sum. Using that $(x_t, \tilde{y}_t)$ is joint normal we can use Theorem 2 for the second term $\mathrm{Pr}_{span\{\tilde{y}_t\}} x_t = \mathbb{E}(x_t|\tilde{y}_t)$.

$$= \hat{x}_{t|t-1} + K_t \tilde{y}_t$$

where $K_t = \mathbb{C}ov(x_t, \tilde{y}_t)\mathbb{V}ar(\tilde{y}_t)^{-1} = P_{t|t-1} H' S_t^{-1}$. The posteriori variance can also be computed using Theorem 2 or using $x_t = x_t - \hat{x}_{t|t} + \hat{x}_{t|t} = (x_t - \hat{x}_{t|t}) + \hat{x}_{t|t-1} + K_t \tilde{y}_t \Leftrightarrow \tilde{x}_{t|t-1} = x_t - \hat{x}_{t|t-1} = (x_t - \hat{x}_{t|t}) + K_t \tilde{y}_t$

$$P_{t|t-1} = \mathbb{C}ov(\tilde{x}_{t|t-1}) = \mathbb{C}ov(x - \hat{x}_{t|t}) + K_t \mathbb{C}ov(\tilde{y}_t) K_t' = P_{t|t} + K_t S_t K_t'$$

where we used that due to the projection theorem $x - \hat{x}_{t|t}$ is orthogonal to $\tilde{y}_t$. Inserting the expression obtained for K_t and rearranging the terms gives:

$$P_{t|t} = P_{t|t-1} - K_t S_t K_t' = P_{t|t-1} - K_t H_t' P_{t|t-1} = (I - K_t H_t') P_{t|t-1}$$

6.3 Summary of simple Kalman filter

Here I will quickly summarize what we have shown so far and describe the Kalman filter algorithm step by step:
Given filter problem:

- $x_{t+1} = \mu + F_t x_t + w_t$
- $y_t = c + H_t x_t + v_t$
- $v_t \sim N(0, R_t)$ and $w_t \sim N(0, Q_t)$,
- $\mathbb{E}(w_t w_s') = Q_t \delta_{t-s}$ $\mathbb{E}(v_t v_s') = R_t \delta_{t-s}$ and $\mathbb{E}(w_t v_s') = 0 \qquad \forall s, t \in \mathbb{N}$
- $x_0 \sim N(\hat{x}_0, P_0)$
- $\mathbb{E}(x_0 w_t') = \mathbb{E}(x_0 w_t') = 0\ \forall t$

We initialise the algorithm with given $\hat{x}_{0|0} := \hat{x}_0, P_{0|0} := P_0$ then we recursively calculate:

[28] It is clear that $span\{1, y_1, ..., y_{t-1}\} \oplus span\{\tilde{y}_t\} \subseteq span\{1, y_1, ..., y_t\}$ since $\tilde{y}_t = y_t - \hat{y}_{t|t-1}$ where $\hat{y}_{t|t-1} \in span\{1, y_1, ..., y_{t-1}\}$. The other inclusion follows if we pick an element of $span\{1, y_1, ..., y_t\}$ we can subtract and ad $\hat{y}_{t|t-1}$ and then collect the terms in the right way.

1) $\hat{x}_{t|t-1} = \mu + F_{t-1}\hat{x}_{t-1|t-1}$

2) $P_{t|t-1} = F_{t-1}P_{t-1|t-1}F'_{t-1} + Q_t$

3) $S_t = H_t P_{t|t-1} H'_t + R_t$

4) $K_t = P_{t|t-1} H'_t S_t^{-1}$

5) $\tilde{y}_t = y_t - H_t \hat{x}_{t|t-1} - c$

6) $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t \tilde{y}_t$

7) $P_{t|t} = (I - K_t H_t) P_{t|t-1}$

This algorithm is easy to understand and to implement. It is called the simple Kalman filter. In practice the filter is often initialized with $x_0 \sim N(0, rI)$ with a suitable constant $r > 0$, which corresponds to the uncertainty of the initial guess (the more uncertain the bigger r).

6.4 Extended Kalman Filter

Now we want to solve a non-linear filter problem with an analogue algorithm. The idea is pretty straight forward namely to linearise all non-linear functions, but in this setting we do not know any more if the Kalman filter solution is the best (i.e. minimal variance) filter. First I will present the non-linear filter problem and then the solution:

Extended filter problem:

- $x_{t+1} = g(x_t) + w_t$
- $y_t = f(x_t) + v_t$
- $v_t \sim N(0, R_t)$ and $w_t \sim N(0, Q_t)$,
- $\mathbb{E}(w_t w_s') = Q_t \delta_{t-s}$ $\mathbb{E}(v_t v_s') = R_t \delta_{t-s}$ and $\mathbb{E}(w_t v_s') = 0 \qquad \forall s, t \in \mathbb{N}$
- $x_0 \sim N(\hat{x}_0, P_0)$

where $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and $f \in C^1(\mathbb{R}^n, \mathbb{R}^d)$. If g and f are linear functions the extended Kalman filter collapses to the simple one. The solution concept is to linearise the non-linear functions with a first order taylor series approximation around the previous state. We initialize the algorithm with $\hat{x}_0 =: \hat{x}_{0|0}, P_0 =: P_{0|0}$. Then for given $x_{t-1|t-1}, P_{t-1|t-1}$ we calculate:

Extended Kalman filter algorithm

1) $\hat{x}_{t|t-1} = g(\hat{x}_{t-1|t-1})$
2) $P_{t|t-1} = F_t P_{t-1|t-1} F_t' + Q_t$
3) $S_t = H_t P_{t|t-1} H_t' + R_t$
4) $K_t = P_{t|t-1} H_t' S_t^{-1}$
5) $\tilde{y}_t = y_t - f(\hat{x}_{t|t-1})$
6) $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t \tilde{y}_t$
7) $P_{t|t} = (I - K_t H_t) P_{t|t-1}$

with the matrices $F_t = \frac{\partial g(x)}{\partial x}|_{x=\hat{x}_{t-1|t-1}}$ and $H_t = \frac{\partial f(x)}{\partial x}|_{x=\hat{x}_{t|t-1}}$

6.5 Parameter estimation of Kalman filter

So far we have assumed that all the parameters of the filter problem are known. But in our applications specifically the Gaussian affine term structure model and the shadow rate term structure model we face the problem that we only have the measurements $(y_t)_{t=1}^T$ at hand. So we have to estimate all the parameters of the filter problem. After we have estimated them we want to use the state variable which is filtered with the estimated parameters to calculate the short rate (GATSM) or the shadow rate (SRTSM). This estimation is done by maximizing a likelihood function.
First I present the problem once again and then discuss the solution concept:

Given filter problem for GATSM:

- $x_{t+1} = \mu + Fx_t + w_t$
- $y_t = c + Hx_t + v_t$
- $v_t \sim N(0, R)$ and $w_t \sim N(0, Q)$,
- $\mathbb{E}(w_t w_s') = Q_t \delta_{t-s}$ $\mathbb{E}(v_t v_s') = R_t \delta_{t-s}$ and $\mathbb{E}(w_t v_s') = 0 \qquad \forall s, t \in \mathbb{N}$
- $x_0 \sim N(\hat{x}_0, P_0)$

where the parameters are unknown and we only have $(y_t)_{t=1}^T$ (i.e. for the GATSM and SRTSM we use the forward rates from 01/01/1990 to 01/12/2013 as measurements). Note that in this filter problem we have $R_t = R$, $Q_t = Q$, $F_t = F$ and $H_t = H$ $\forall t \in \{1, ..., T\}$ For convenience we will use the following notation:

$$h(\theta) := (\mu, F, c, H, R, Q) \tag{40}$$

where theta is the stacked vector with all unknown parameters (actually the initial state x_0 and P_0 are parameters too, but for the GATSM we do not incorporate them into $h(\theta)$ and use $x_0 = 0$ and $P_0 = rI$). For a description of θ and the map $h(\cdot)$ see the next section 6.6. We also write:

$$\mathscr{Y}_t = (y_1, ..., y_t)$$

where $\mathscr{Y}_t$ is the information available up to time t inclusive t.

Now we observe that y_t conditional on $\mathscr{Y}_{t-1}$ is normal distributed[29]:

$$y_t|\mathscr{Y}_{t-1} \sim N(c + H'\hat{x}_{t|t-1}, HP_{t|t-1}H' + R) = N(c + H'\hat{x}_{t|t-1}, S_t) \tag{41}$$

Therefore the density function is given by:

$$f_{y_t|\mathscr{Y}_{t-1}}(z|\mathscr{Y}_{t-1}, \theta) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(S_t)^{\frac{1}{2}}} \exp(-\frac{1}{2}(z - c - H'\hat{x}_{t|t-1})' S_t^{-1}(z - c - H'\hat{x}_{t|t-1}))$$

[29]To get the conditional moments either look to the derivation of the Kalman filter or calculate them again: $y_t = c + Hx_t + v_t$ with $\mathbb{E}(y_t|(\mathscr{Y}_{t-1})) = c + H\mathbb{E}(x_t|(\mathscr{Y}_{t-1})) + \mathbb{E}(v_t|(\mathscr{Y}_{t-1})) = cu_t + H\hat{x}_{t|t-1} + \mathbb{E}(v_t) = c + H\hat{x}_{t|t-1}$ and $\mathbb{V}ar(y_t|(\mathscr{Y}_{t-1})) = H\mathbb{V}ar(x_t|(\mathscr{Y}_{t-1}))H' + \mathbb{V}ar(v_t|(\mathscr{Y}_{t-1})) = HP_{t|t-1}H' + R$ due to the assumed correlation structure and the gaussian assumption. For the extended Kalman filter the mean is replace by $f(\hat{x}_{t|t-1})$.

Now we construct the likelihood function:

$$L(\theta) = \prod_{t=1}^{T} f_{y_t|(\mathscr{Y}_{t-1})}(y_t|\mathscr{Y}_{t-1}, \theta)$$

Next we take the logarithm to get the log likelihood:

$$\ln(L(\theta)) = \sum_{t=1}^{T} \ln(f_{y_t|(\mathscr{Y}_{t-1})}(y_t|\mathscr{Y}_{t-1}, \theta))$$

for convenience we drop the explicit dependence of the right hand side on θ

$$\ln(L(\theta)) = \sum_{t=1}^{T} -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(S_t)) - \frac{1}{2}(y_t - c - H'\hat{x}_{t|t-1})'S_t^{-1}(y_t - c - H'\hat{x}_{t|t-1})$$

Now we observe that $\tilde{y}_t = y_t - c - H'\hat{x}_{t|t-1}$

$$\ln(L(\theta)) = -\frac{dT}{2}\ln(2\pi) - \frac{1}{2}(\sum_{t=1}^{T}\ln(\det(S_t(\theta))) + \tilde{y}_t(\theta)'S_t^{-1}(\theta)\tilde{y}_t(\theta)) \tag{42}$$

Now we have everything in place to estimate θ with:

$$\hat{\theta} = \arg\max_{\theta} \ln(L(\theta))$$

So for given θ we can evaluate the log likelihood function through running the Kalman filter algorithm and constructing $\ln(L(\theta))$ with the residuals $\tilde{y}_t$ and the covariance matrices S_t. To finish this section I will present the estimation of the parameters θ with pseudo-code ($\epsilon > 0$ given tolerance):

1 guess an arbitrary $\theta^{(0)}$

2 run the Kalman filter algorithm for the filter problem given by $(y_t)_{t=1}^T$ and $\theta^{(i)}$ and get $(\tilde{y}_t(\theta^{(i)}), S_t(\theta^{(i)})_{t=1}^T$

3 construct the log likelihood $\ln(L(\theta^{(i)}))$

4 numerically update $\theta^{(i+1)}$ with a suitable optimization algorithm

5 as long as $\ln(L(\theta^{(i+1)})) - \ln(L(\theta^{(i)})) > \epsilon$ go to step 2

Note that the log likelihood can have multiple global maxima (this is the generic case). To prevent this we have to make certain identification assumptions (i.e. restrictions) for θ to prevent rotation, translation and rescaling of the parameters (see GATSM and SRTSM section which identification assumptions are made for our specific models).

6.6 The GATSM and SRTSM as filter problem

In the next paragraph I describe the connection between the GATSM model parameters and the parameters used in the Kalman filter problem. So I summarize how the Kalman filter is parametrized using $\theta = (\mu, \mu^Q, \rho, \rho^Q, \Sigma, \omega, \delta_0, \delta_1)$ (Σ is a lower triangular matrix).
The GATSM:

- $X_{t+1} = \mu + FX_t + w_t$ corresponds to $X_{t+1} = \mu(\theta) + \rho(\theta)X_t + \Sigma(\theta)\epsilon_{t+1}$
- $y_t = c + Hx_t + v_t$ corresponds to $f_t^o = a(\theta) + b(\theta)'X_t + \nu_t$
- $\epsilon_t \sim N(0, I)$ and $\nu_t \sim N(0, w(\theta)I)$,
- $\mathbb{E}(\epsilon_t\epsilon_s') = I\delta_{t-s}$ $\mathbb{E}(\nu_t\nu_s') = w(\theta)I\delta_{t-s}$ and $\mathbb{E}(\epsilon_t\nu_s') = \mathbb{E}(x_0 w_t') = \mathbb{E}(x_0 v_t') = 0 \quad \forall s, t \in \mathbb{N}$
- $x_0 \sim N(0, 100I)$

where $f_t^o = (f_{t,n_1,n_1+1}^o, ..., f_{t,n_7,n_7+1}^o)' \in \mathbb{R}^7$ is the stacked vector with the forward rates for the seven maturities and the 7 dimensional measurement error $\nu_t \sim N(0, \omega(\theta)I_7)$, $\omega \in \mathbb{R}_+$, $a = (a_{n_1}, ..., a_{n_7})' \in \mathbb{R}^7$ and $b' = (b_{n_1}, ..., b_{n_7})' \in \mathbb{R}^{7\times 3}$
Where $a_n(\theta)$, $\bar{a}_n(\theta)$ and $b_n(\theta)$ are given by:

- $\bar{a}_n(\theta) := \delta_0 + \delta_1' \sum_{j=0}^{n-1}(\rho^Q)^j \mu^Q$
- $a_n(\theta) := \bar{a}_n(\theta) - \frac{1}{2}\delta_1'(\sum_{j=0}^{n-1}(\rho^Q)^j)\Sigma\Sigma'(\sum_{j=0}^{n-1}(\rho^Q)^j)'\delta_1$
- $b_n(\theta)' := \delta_1'(\rho^Q)^n$
- $\rho(\theta) = \rho$
- $\mu(\theta) = \mu$
- $\Sigma(\theta) = \Sigma$
- $\omega(\theta) = \omega$

Next we repeat the task for the shadow rate model and show the parametrization of the extended Kalman filter through θ.
The SRTSM:

- $X_{t+1} = \mu + FX_t + w_t$ corresponds to $X_{t+1} = \mu(\theta) + \rho(\theta)X_t + \Sigma(\theta)\epsilon_{t+1}$
- $y_t = G(x_t) + v_t$ corresponds to $f_t^o = \bar{r}\iota_7 + G(X_t) + \nu_t$
- $\epsilon_t \sim N(0, I)$ and $\nu_t \sim N(0, w(\theta)I)$,
- $\mathbb{E}(\epsilon_t\epsilon_s') = I\delta_{t-s}$ $\mathbb{E}(\nu_t\nu_s') = w(\theta)I\delta_{t-s}$ and $\mathbb{E}(\epsilon_t\nu_s') = \mathbb{E}(x_0 w_t') = \mathbb{E}(x_0 w_t') = 0 \quad \forall s, t \in \mathbb{N}$
- $x_0 \sim N(0, 100I)$

where $G(X_t) := (\sigma_{n_1}^Q g(\frac{a_{n_1} + b_{n_1}' X_t - \bar{r}}{\sigma_{n_1}^Q}), ..., \sigma_{n_7}^Q g(\frac{a_{n_7} + b_{n_7}' X_t - \bar{r}}{\sigma_{n_7}^Q}))' \in \mathbb{R}^7$ and $\iota_7 = (1, ..., 1) \in \mathbb{R}^7$. The rest is like for the GATSM.

7 Estimation of the shadow rate and short rate

7.1 GATSM short rate estimation

The estimation is done through maximizing the likelihood function described in the Kalman filter section. The transition equation is given by (18) and the observation equation is given by (19).
First we have to make identification assumptions such that the parameters $\theta = (\mu, \mu^Q, \rho, \rho^Q, \Sigma, \omega, \delta_0, \delta_1)$ become identifiable. We will restrict the parameters under the risk neutral or pricing measure Q similar to Joslin, Singleton, and Zhu (2011)[9] and Hamilton and Wu (2014)[10]. For this we assume that X_t follows the VAR(1) given by equation (18) under the real world measure P and also under Q. We assume fixed parameters θ. When we apply the affine transformation $X_t^* = C + D^{-1}X_t$[30] with $C \in \mathbb{R}^n$ and an invertible matrix $D \in \mathbb{R}^{n \times n}$ it follows:

$$X_t^* = C + D^{-1}X_t = C + D^{-1}(\mu^Q + \rho^Q X_{t-1} + \epsilon_t^Q) = (C + D^{-1}\mu^Q) + D^{-1}\rho^Q DD^{-1}X_{t-1} + D^{-1}\Sigma\epsilon_t^Q =$$

using $X_{t-1}^* - C = D^{-1}X_{t-1}$

$$= (C + D^{-1}\mu^Q) + D^{-1}\rho^Q D(X_{t-1}^* - C) + D^{-1}\Sigma\epsilon_t^Q = (I - D^{-1}\rho^Q D)C + D^{-1}\mu^Q + \tilde{\rho}^Q X_{t-1}^* + \tilde{\epsilon}_t^Q$$

with $\tilde{\rho}^Q := D^{-1}\rho^Q D$, $\tilde{\epsilon}_t^Q \sim N(0, B)$ and $B := D^{-1}\Sigma\Sigma'(D^{-1})'$. B is a symmetric covariance matrix (positive definiteness follows from the assumption that D is regular and $\Sigma\Sigma'$ is positive definite[31]), therefore we can take the cholesky decomposition of $B = \tilde{\Sigma}\tilde{\Sigma}'$. We define $\tilde{\mu}^Q := (I - \tilde{\rho}^Q)C + D^{-1}\mu^Q$ and get the following VAR for the transformed state:

$$X_t^* = \tilde{\mu}^Q + \tilde{\rho}^Q X_{t-1}^* + \tilde{\Sigma}\epsilon_t^Q \tag{43}$$

The short rate equation is transformed too:

$$r_t = \delta_0 + \delta_1' X_t = \delta_0 + \delta_1' DD^{-1}X_t = \delta_0 - \delta_1' DC + \delta_1' DX_t^* = \tilde{\delta_0} + \tilde{\delta_1}' X_t * \tag{44}$$

with $\tilde{\delta_0} := \delta_0 - \delta_1' DC$ and $\tilde{\delta_1}' := \delta_1' D$. The observation equation is transformed too:

$$f_t^o = a + b' X_t + \nu_t = a - b' DC + b' DX_t^* + \nu_t = \tilde{a} + \tilde{b}' X_t^* + \nu_t \tag{45}$$

Here we see that the model given by (43), (44) and (45) is observational equivalent to the original one described in the summary of GATSM section. Therefore given only the observation f_t^o we can not distinguish between those models, since X_t and X_t^* can both produce the same observations. It follows that the parameters are not identifiable (i.e. multiple global maxima of the likelihood function). To solve this problem we set restrictions for the Q parameters such that the transformation with C, D is unique. We set D such that $\tilde{\rho}^Q$ is in real Jordan form (you can think of diagonalising the given matrix ρ^Q). Creal and Wu (2015)[13] showed that empirical results indicate, for our set-up when estimating term structure models with

[30] D^{-1} instead of D is used to make the argument why under the identification restrictions the transformation is uniquely determined easier to see.

[31] $\forall x \in \mathbb{R}^n/\{0\}$ $x'D^{-1}\Sigma\Sigma'(D^{-1})'x = ((D^{-1})'x)'\Sigma\Sigma((D^{-1})'x) > 0 \leftrightarrow (D^{-1})'x \neq 0$ which is true if D^{-1} is invertible therefore injective (only the zero vector is projected on the zero vector) and $x'\Sigma\Sigma x > 0$ $\forall x \in \mathbb{R}^n/\{0\}$.

three state variable $X_t \in \mathbb{R}^3$, that there is a repeated eigenvalue with geometric multiplicity 1. To identify the intercept under the measure Q we set:

$$0 = \tilde{\mu}^Q = (I - D\rho^Q D^{-1})C + D\mu \qquad \Leftrightarrow \qquad C = -(I - D\rho^Q D^{-1})^{-1} D\mu$$

The inversion of $(I - D^{-1}\rho^Q D)$ is possible if we assume that $\tilde{\rho}^Q := D^{-1}\rho^Q D$ has eigenvalues ρ_1^Q and ρ_2^Q smaller than 1[32]. This corresponds to the stationarity assumption of VAR theory and is a classic assumption in time series analysis. With this identification restriction C of the transformation is uniquely determined.
Furthermore we set the restriction $\tilde{\delta_1}' = (1,1,0)$. Together with the restriction that $\tilde{\rho}^Q$ is in real Jordan normal form, the matrix D is then uniquely determined. To see this, lets do a little bit of linear algebra (For convenience I drop the suffix Q):

Given a matrix $\rho \in \mathbb{R}^{3\times 3}$ with eigenvalues ρ_1, ρ_2 where the second is a repeated one with geometric multiplicity 1 and $\delta_1 \in \mathbb{R}^3/\{0\}$ (if it is zero we would have a deterministic short rate and therefore deterministic forward rates, furthermore we assume that δ_1 is not orthogonal to v_1 and v_2) I show that the equations $D^{-1}\rho D = J$ and $\delta_1' D = (1,1,0)$ uniquely determines the matrix D. J is the block diagonal matrix given by the Jordan normal form[33]:
Due to the Jordan normal form we have (see Havlicek, Lineare Algebra für Technische Mathematiker Theorem 8.7.7 and 8.7.10 [23]): $ker\{\rho - \rho_1 I\} = span\{v_1\}$, $ker\{\rho - \rho_2 I\} = span\{v_2\}$ and $ker\{(\rho - \rho_2 I)^2\} = span\{v_2, v_3\}$ where v_1, v_2 are arbitrary eigenvectors for ρ_1, ρ_2 and v_3 is an arbitrary generalized eigenvector for ρ_2 (i.e. $(\rho - \rho_2 I)^2 v_3 = 0$ and $(\rho - \rho_2 I)v_3 \neq 0$ see [23] 8.7.3). First we observe:

$$0 = (\rho - \rho_2 I)^2 v_3 = (\rho - \rho_2 I)(\rho - \rho_2 I)v_3$$

Therefore $0 \neq x := (\rho - \rho_2 I)v_3 \in ker\{\rho - \rho_2 I\} = span\{v_2\}$. That means $\exists! z \in \mathbb{R}/\{0\}$: $x = zv_2$ where z is uniquely determined by v_3 and not free any more. Therefore we have $\rho v_3 = \rho_2 v_3 + z v_2$.
Due to the theorem about the normal form we have the existence of at least one base (where the basis vectors are out of the three kernels mentioned above) that transforms the matrix ρ into Jordan normal form. We can therefore write the matrix D a priori as $D = (c_1 v_1, c_2 v_2, c_3 v_2 + c_4 v_3)$ with $c_i \neq 0$ for $i \in \{1,2,4\}$(otherwise the matrix D would not be invertible). We want to have $\rho D = DJ$ (then we can multiply with D^{-1} from the left) which gives the following equation:

$$\rho D = \rho(c_1 v_1, c_2 v_2, c_3 v_2 + c_4 v_3) = (c_1 \rho v_1, c_2 \rho v_2, c_3 \rho v_2 + c_4 \rho v_3) =$$
$$= (c_1 \rho_1 v_1, c_2 \rho_2 v_2, c_3 \rho_2 v_2 + c_4(\rho_2 v_3 + z v_2)) = (c_1 \rho_1 v_1, c_2 \rho_2 v_2, (c_3 \rho_2 + c_4 z) v_2 + c_4 \rho_2 v_3) =$$

[32] Note that $(I - \tilde{\rho}^Q)$ is an upper triangular matrix, therefore $\det(I - \tilde{\rho}^Q) = (1 - \rho_1^Q) \cdot (1 - \rho_2^Q)^2 \neq 0$ if $|\rho_1^Q| < 1$ and $|\rho_2^Q| < 1$.
[33]

$$J = \begin{pmatrix} \rho_1^Q & 0 & 0 \\ 0 & \rho_2^Q & 1 \\ 0 & 0 & \rho_2^Q \end{pmatrix}$$

$$\stackrel{!}{=} (c_1v_1, c_2v_2, c_3v_2 + c_4v_3) \begin{pmatrix} \rho_1 & 0 & 0 \\ 0 & \rho_2 & 1 \\ 0 & 0 & \rho_2 \end{pmatrix} = (c_1\rho_1v_1, c_2\rho_2v_2, (\rho_2c_3 + c_2)v_2 + \rho_2c_4v_3)$$

The last equality is a straight forward calculation. We see that the last column of the matrices on the right side of the two equations above must match for the equality denoted with ! to hold. Because $v_2 \neq 0$ it follows:

$$c_3\rho_2 + c_4z = \rho_2c_3 + c_2 \Leftrightarrow c_4 = \frac{c_2}{z}$$

Here we see that three degrees of freedom c_i $i \in \{1, 2, 3\}$ are left for the matrix D and it can be written as:

$$D = (c_1v_1, c_2v_2, c_3v_2 + \frac{c_2}{z}v_3) = (v_1, v_2, v_3) \begin{pmatrix} c_1 & 0 & 0 \\ 0 & c_2 & c_3 \\ 0 & 0 & \frac{c_2}{z} \end{pmatrix} = \tilde{D}P$$

where P is the matrix containing the c_i and $\tilde{D} = (v_1, v_2, v_3)$. The last restriction $\delta_1'D = (1, 1, 0)$ determines the values uniquely if we assume $\delta_1 \neq 0$ (else the short rate in the original model would be constant) and δ_1 is not orthogonal to v_1 and v_2 since then $\delta_1'\tilde{D} \neq 0$ and the first two entries are non zero:

$$\delta_1'D = \delta_1'\tilde{D}P = (1, 1, 0)$$

Here we see three equations for the three parameters c_i $i \in \{1, 2, 3\}$ that determine the three parameters uniquely under the assumptions made.

Therefore we impose the following restrictions on the Q parameters $\mu^Q = 0$, Σ is a lower triangular matrix, $\delta_1' = (1, 1, 0)$ and ρ^Q is in Jordan form to prevent an affine transformation of X_t (with these restrictions the transformation is uniquely determined).
Furthermore the specific choice $\delta_1' = (1, 1, 0)$ can be explained a little bit more. In some models the specific choice of δ_1 and ρ^Q allows the state variable to be interpreted. That means in these models we can interpret X_t as the time varying coefficients $\beta_{i,t}$ of the Nelson Siegel yield curve. Therefore using this interpretation $X_{1,t}$ stands for the level, $X_{2,t}$ the slope and $X_{3,t}$ the curvature of the yield curve. Nevertheless the specific restriction $\delta_1 = (1, 1, 0)$ is somewhat arbitrary[34].
Compare with Jens H. E. Christensen, Francis X. Diebold, Glenn D. Rudebusch, [8] (page 2 and 3, especially Proposition 1), Scott Joslin, Kenneth J. Singleton, Haoxiang Zhu [9], James D. Hamilton and Jing Cynthia Wu [10] and Jens H. E. Christensen [11].

Summarizing the identification restrictions we have:

$$\rho^Q = \begin{pmatrix} \rho_1^Q & 0 & 0 \\ 0 & \rho_2^Q & 1 \\ 0 & 0 & \rho_2^Q \end{pmatrix}$$

$\mu^Q = 0$, Σ a lower triangular matrix and $\delta_1' = (1, 1, 0)$. So all in all we have to estimate 22 parameters $\theta = (\mu, \rho, \rho^Q, \Sigma, \omega, \delta_0)$, 3 for $\mu \in \mathbb{R}^3$, 9 for $\rho \in \mathbb{R}^{3\times3}$, 2 for $\rho^Q \in \mathbb{R}^{3\times3}$, 6 for

[34] Also this interpretation found in the literature is for the normal Nelson-Siegel function and not the extended one used here.

$\Sigma \in \mathbb{R}^{3\times 3}$, 1 for $\omega \in \mathbb{R}_+$ and 1 for $\delta_0 \in \mathbb{R}$.

7.2 SRTSM shadow rate estimation

The estimation of the shadow rate in the SRTSM is realized through a non-linear filter problem (see section extended Kalman filter). The transition equation is like in the GATSM model given by (30) and the observation equation is given by (31):

$$f_t^o = \bar{r} + G(X_t) + \nu_t \tag{46}$$

where $f_t^o = (f_{t,n_1,n_1+1}^o, ..., f_{t,n_7,n_7+1}^o)' \in \mathbb{R}^7$ is the stacked vector with the forward rates for the seven maturities and the 7 dimensional measurement error $\nu_t \sim N(0, \omega^2 I_7)$, $\omega \in \mathbb{R}$, $G(X_t) := (\sigma_{n_1}^Q g(\frac{a_{n_1}+b'_{n_1}X_t-\bar{r}}{\sigma_{n_1}^Q}), ..., \sigma_{n_7}^Q g(\frac{a_{n_7}+b'_{n_7}X_t-\bar{r}}{\sigma_{n_7}^Q}))' \in \mathbb{R}^7$.
Like Wu and Xia [5] I use $\bar{r} = 0.25$ as a lower bound for the nominal short rate in the SRTSM.

7.3 Results of estimation

I have implemented the estimation procedure in R (using version 3.3.2 and the packages vars, fields and Matrix). Basically I have implemented the Kalman filter like described in the section Kalman Filters and extensions. For given parameters θ the log likelihood function can be evaluated. Then I use the built into R optim() function to estimate the parameters. The forward rate data is constructed like it is described in the section Construction of forward rates using the data set published by Gürkaynak, Sack, and Wright[35]. The forward rates used are plotted in Figure 2. I use the same seven maturities like Wu and Xia (2016) [5], namely 3 months, 6 months, 1 year, 2 years, 5 years, 7 years and 10 years. The first goal was to replicate the shadow rate estimated by Wu and Xia [5], for this I used the exact same time interval from January 1990 to December 2013. I report the results in Table 1 and they are almost identical to the Wu and Xia estimates[36]. Also the log likelihood values are almost identical with the SRTSM giving a better fit with 855.8 compared to 755.7, the value for the GATSM. This is due to the period at the zero lower bound where the SRTSM fits the data better.

In Figure 4 I plot the estimated shadow rate of the SRTSM and the short rate from the GATSM model. Furthermore I plot the effective federal funds rate.
In Figure 4 we see that in normal times, when the effective federal funds rate (i.e. effr) set by the central bank is not at the zero lower bound (from December 2008 $\hat{=}$ dashed line to the end), the GATSM and SRTSM match each other almost perfectly. They are also nearly identical to the effr. This is exactly what one would expect. Also the short rate

[35] The data set can be found here: https://www.federalreserve.gov/pubs/feds/2006/200628/200628abs.html
[36] Since I use annualized percentage forward rates I rescale f_t^o with 1200. Therefore $\mu, \delta_0, \Sigma, \sqrt{\omega}$ have to be rescaled too. For $a_n = \bar{a}_n - \frac{1}{2}\delta_1'(\sum_{j=0}^{n-1}(\rho^Q)^j)\Sigma\Sigma'(\sum_{j=0}^{n-1}(\rho^Q)^j)'\delta_1\frac{1}{1200}$ we get the extra $\frac{1}{1200}$ factor since both Σ matrices have to be rescaled. This is also indicated by Wu and Xia in Table 1 where they report their estimated parameters.

		SRTSM			GATSM	
δ_0	13.372			11.675		
	0.964	-0.0026	0.345	0.968	-0.0043	0.488
ρ	-0.023	0.942	1.015	-0.0231	0.933	1.0135
	0.0033	0.0028	0.886	0.0029	0.0028	0.893
eig(ρ)	0.983	0.964	0.845	0.987	0.963	0.845
ρ^Q	0.998	0	0	0.997	0	0
	0	0.950	1	0	0.950	1
	0	0	0.950	0	0	0.950
μ	-0.304	-0.238	0.0253	-0.229	-0.207	0.0186
	0.416	0	0	0.474	0	0
Σ	-0.399	0.244	0	-0.459	0.218	0
	0.011	0.003	0.039	-0.017	0.0013	0.036
$\sqrt{\omega}$	0.089			0.093		
loglikelihood	855.7951			755.678		

Table 1: Estimated parameters for SRTSM and GATSM. ML estimation using data from 01/01/1990 to 31/12/2013. Sample size: $288 = 12 \cdot 24$.

looks nearly identical to the one estimated by Wu and Xia [5] and displayed in their Figure 4.

In Figure 5 I plot the short rate calculated with the parameters from Table 1 for the time horizon January 1990 to January 2016. To see how the shadow rate behaves in a time when there was hope for a monetary policy normalization. On 16/12/2015 the fed increased the federal funds rate by 0.25 basis points for the first time since the great recession.

In Figure 5 we see that the shadow rate (red) behaves like one would expect if a rate increase is on the horizon. On the last observation date 01/01/2016 the first rate hike can be seen. Before, the shadow rate rushes to the zero lower bound from below. Furthermore the period from 12/2013 onwards is not part of the dataset we used to estimate the parameters and the model behaves, on this separate "test-set", like expected. This demonstrates that the shadow rate term structure model captures the dynamics of monetary policy very well. It underscores the usefulness and plausibility of the model. Furthermore we can see that the short rate estimated by the GATSM turns also negative on the "test-set" before coming positive again with the first rate hike.

In the next plot (Figure 6) the parameters θ used to calculate the shadow rate are estimated by using the full data set from 01/01/1990 to 01/01/2016. This is done to compare

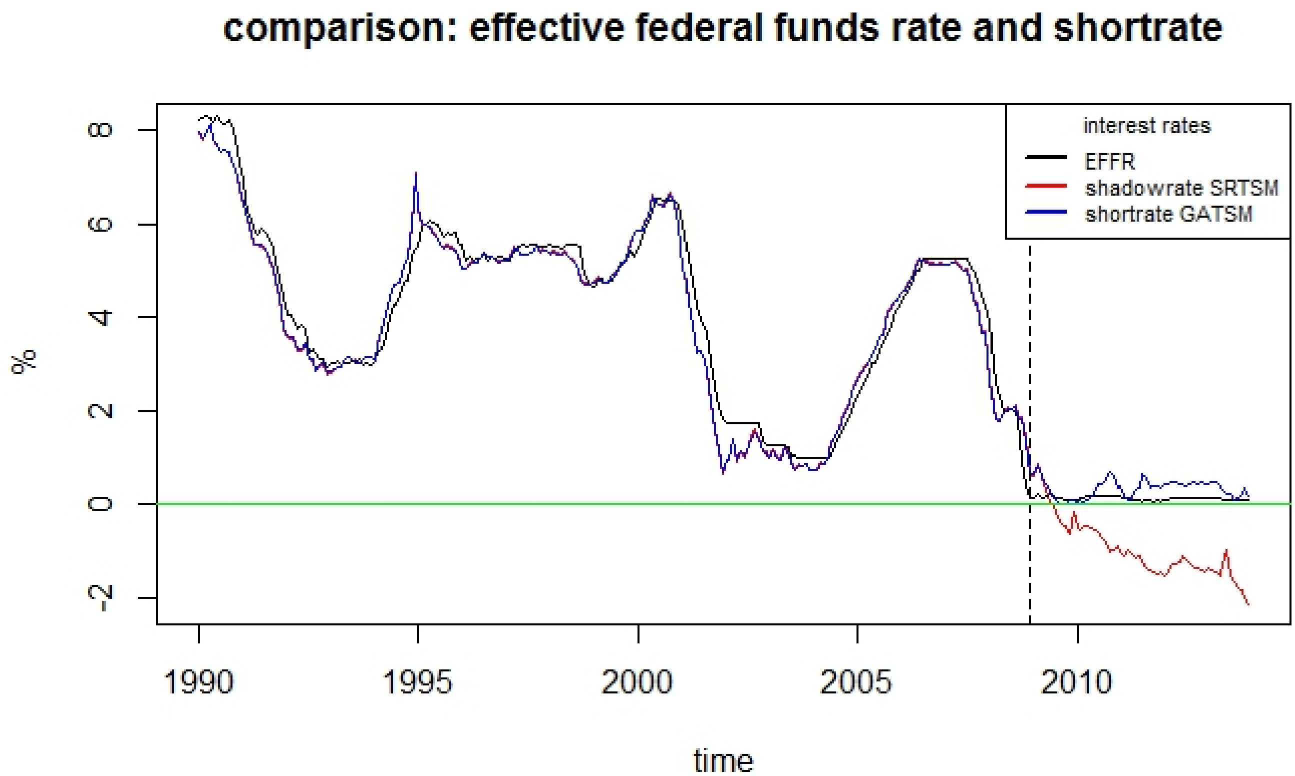

Figure 4: SRTSM, GATSM short rate and effr, dashed vertical line corresponds to December 2008.

the different shadow rates. We would expect that they are similar.
In Figure 6 we compare the shadow rate of the SRTSM with different estimation data sets. The red line is the "old" shadow rate (see Figure 5) and the blue dashed line is the shadow rate when the full data set is used as estimation period. Interestingly the log likelihood value drops off to 823.5609 from 855.8. The old parameters θ_{old} estimated from the sub sample are a good starting point for the optimization problem and the algorithm converges after a few iterations. The log likelihood value over the whole data set, is $\ln(L_{1990-2016}(\theta_{old})) = 678.8085$ lower than the previous value of $855.7951 = \ln(L_{1990-2014}(\theta_{old}))$[37]. Furthermore the new shadow rate (blue dashed line) displays much more extreme movements around 2015 than the old one. The new rate goes below -4% before shooting back up to 0. All in all the two shadow rates are similar. Nevertheless this is a disadvantage of the SRTSM since I expected a priori that the two rates differ not so much. It indicates that the model does not work as good in the years 2014 till 2016 as expected.

To investigate this issue further we repeat this exercise using the data period 01/01/1990 to 01/01/2017. Figure 7 shows the new shadow rate, the old one (using 1990-2014) and

[37]When we use the average log likelihood by dividing through T, we get the following picture: The value when using the old parameters and 1990 to 2014 is $2.97 = 855.79/(24 \cdot 12)$. For the old θ and 1990 to 2016 we get $2.17 = 678/(26 \cdot 12)$, for the new θ and 1990 to 2016 the average loh likelihood is $2.63 = 823/26 \cdot 12$. So the picture does not change when using the average log likelihood function.

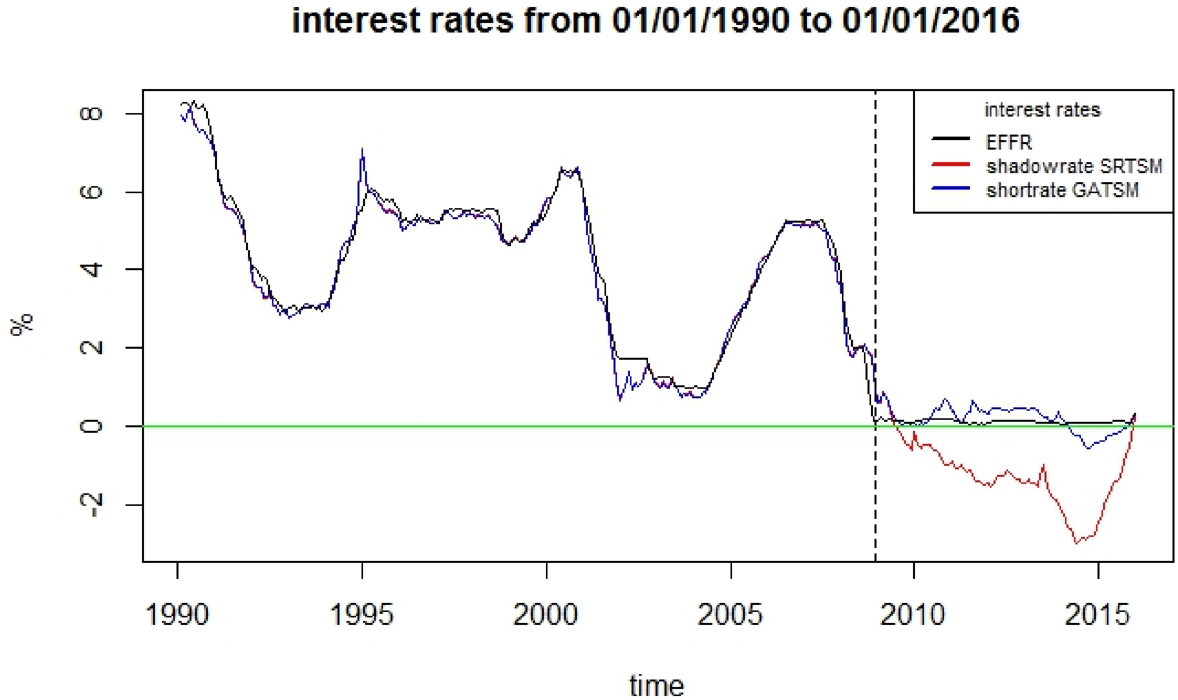

Figure 5: SRTSM, GATSM short rate and effr.

the one using 1190-2016. Again around 2015 the new rate experiences significant volatility. Both rates estimated from 1990-2016 and 1990-2017 look very similar but have this strange behaviour around 2015. Furthermore we see that as the economy is exiting the zero lower bound the shadow rate matches the effective federal funds rate almost perfectly. The log likelihood value increases slightly to 837.4809 when using the 1990-2017 data set.

All in all the shadow rate calculated with the parameters estimated from 1990 to 2014 looks the most plausible to me as it does not display such extreme behaviour around 2015. Therefore I will use it for further analysis.

7.4 Robustness and diagnostics of SRTSM

In this section I test the robustness of the model and whether the model assumptions are met. Since we assumed that the random shocks driving the system are i.i.d. sequences it is important to check if this assumption can be verified after the estimation is performed. By using the Kalman filter we also have implicitly assumed the assumptions of the Kalman filter problem (see section Kalman filter and extensions). We also note that the financial crisis caused a great shock to the system and caused a structural break (see plots below and structural break test section). This break complicates the analysis of the GATSM and SRTSM model since it is unclear how to decide whether the problems we detected are caused by model inherent failures or are just caused by the structural break. To decide this issue we compare the period 1990 to 2009 before the crisis with the period 2009 to 2016 (see Figure

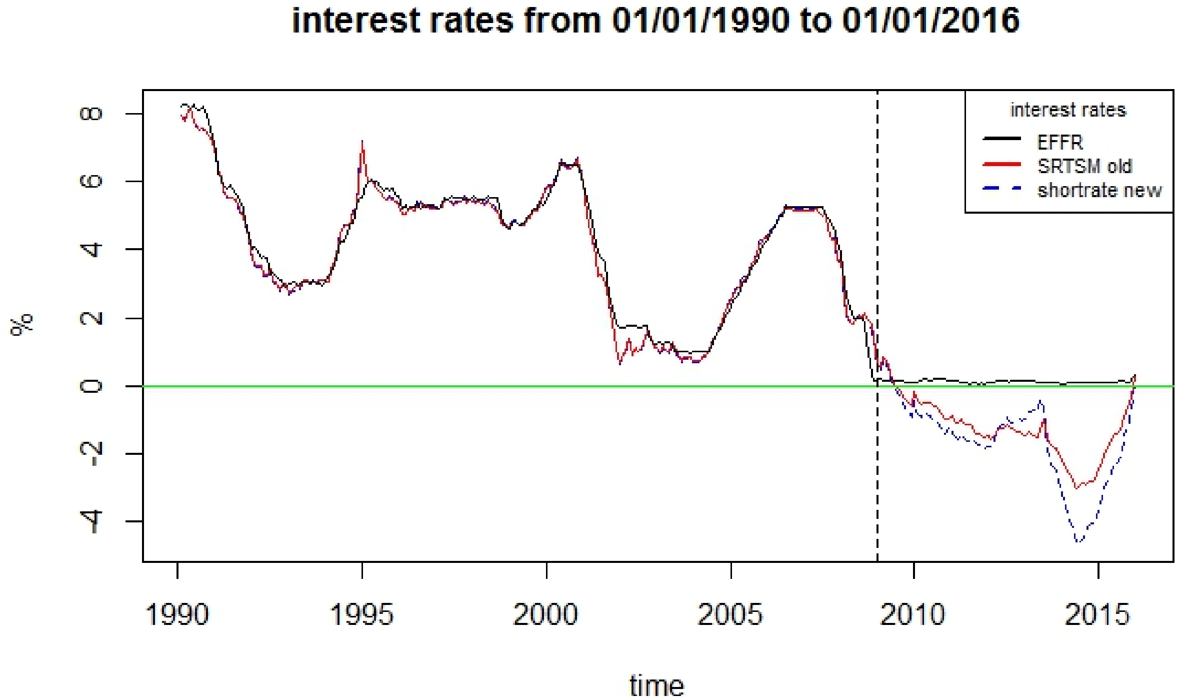

Figure 6: SRTSM, GATSM short rate and effr, red shadow rate (old) calculated with parameters estimated from period 1990 to 2014, blue dashed (new) using 1990-2016 as estimation period.

13, 14).
First I start by considering the robustness of the shadow rate with respect to different lower bounds $\bar{r}$. For this I now consider the lower bound as an additional parameter, that means for the parameter estimation I use $\tilde{\theta} := (\theta, \bar{r}) \in \mathbb{R}^{23}$ as new argument for the log likelihood function $\ln(L(\tilde{\theta}))$[38]. If the model is robust with respect to the lower bound the new shadow rate calculated with the new estimated parameters $\hat{\tilde{\theta}} := \arg\max \ln(L(\tilde{\theta}))$ should not differ to much from the old one. The estimation is performed exactly like before using the optim() function built into R. The log likelihood value increases to 865 from 855. In Figure 8 we see the result. The two rates differ when the economy is at the zero lower bound but overall they are quite similar. The estimated lower floor is 0.186 and the rest of the parameters are very similar to the old ones. To summarize, this exercise validated the shadow rate once more.

Next we will test for the assumptions of the Kalman filter. In particular we will test whether the standardized residuals (i.e. prediction errors) are uncorrelated. Under the assumptions made for the Kalman filter it follows that the residuals $\tilde{y}_t$ are uncorrelated, see Simon (2006) [20] and the variance matrix is given by S_t. For the whole section I use standardized residuals

[38]I also analysed how the model performs when we consider the starting values of the Kalman filter as parameters. Doing this the fit improves a little bit (the loglikelihood increases to 880) but overall the picture stays the same and the shadow rate is very similar to the shadow rate estimated by using $\tilde{\theta}$. Therefore I do not report them.

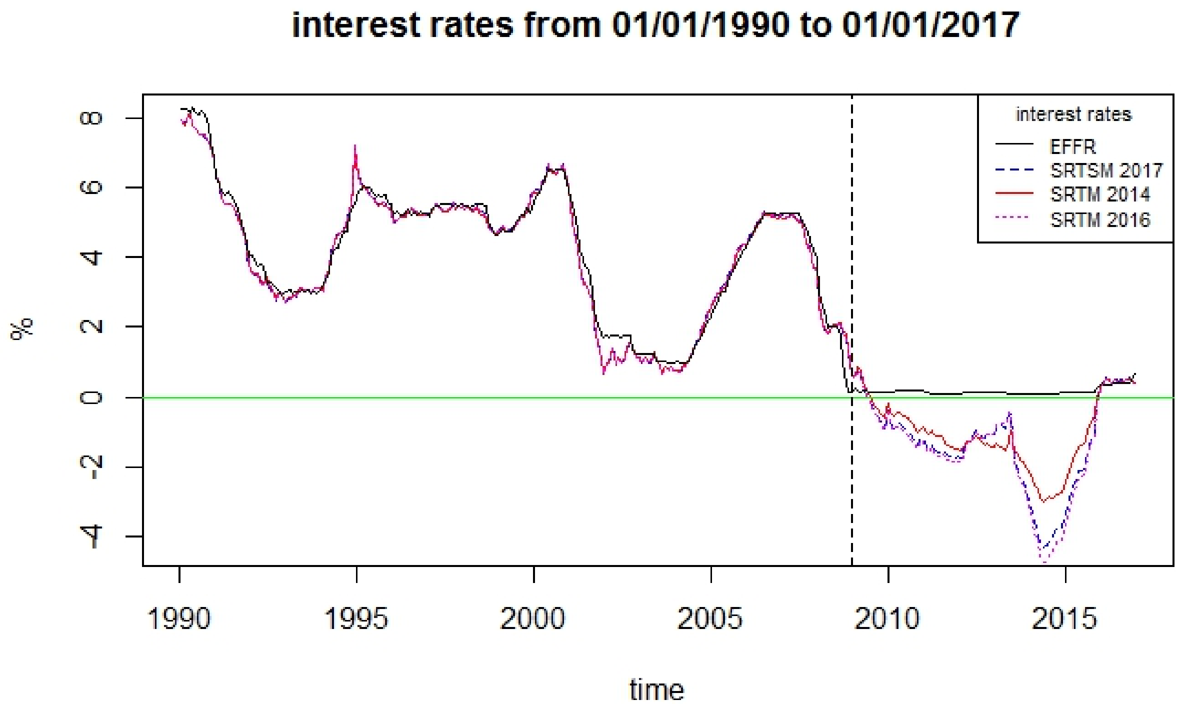

Figure 7: SRTSM, GATSM short rate and effr, red shadow rate (old) calculated with parameters estimated from period 1990 to 2014, blue dashed (new) using 1990-2017 as estimation period, red dashed line using 1990-2016.

where I rescale the residuals to unit variance by dividing through the square root of the diagonal entries (i.e. $(\tilde{y}_t^i)^* = \frac{\tilde{y}_t^i}{\sqrt{S_t^{i,i}}}$). The standardized residuals should be white noise under the Kalman filter assumptions since they have constant variance and zero autocorrelation. So we will use the Ljung–Box test[39]. This is a well known test statistic and is implemented in R. The basic idea is that we test whether the autocorrelation function ($\gamma_X(k) := \mathbb{C}or(X_{t+k}, X_t)$) is zero up to a certain lag l. The $H_0 : \gamma(1) = ... = \gamma(l) = 0$ vs that at least one is different from zero. The p values are reported in Table 2.

In Table 2 we see that almost all p values are below any conventional significance level, therefore we would reject the null hypothesis that the residuals are uncorrelated. Furthermore we note that the p values for the first components are basically zero. The other components have a little bit higher values especially for lower lags. This corresponds to the residual plots below where the first three components have the highest autocorrelation.
In Figure 9 I present a residual plot of the standardized residuals to check whether it can be plausibly assumed that they are uncorrelated. For the whole section I omit component seven of the residuals from the plots due to spacing problems. This causes no loss of information

[39] The Box tests are often used when testing VAR residuals since then under the H_0 the true distribution (χ_l^2 if l stands for the lags) of the test statistic is known. Such a result is not known for Kalman filter residuals and maybe, we would have to adjust the degrees of freedom. Nevertheless they offer an additional piece of information together with the residual plots to decide if the residuals are uncorrelated. I use the univariate test for each of the seven components of the residuals to get a better picture.

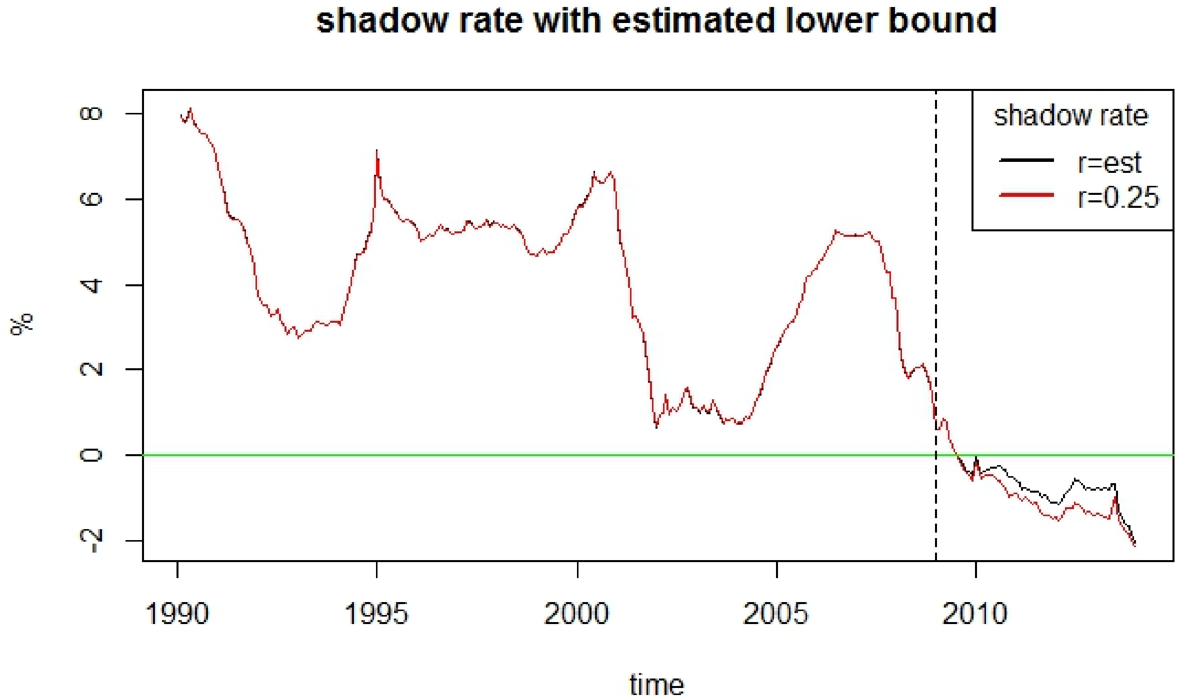

Figure 8: red old shadow rate with $\bar{r} = 0.25$, blue shadow rate where $\bar{r}$ is estimated too.

since this component looks the most white from all (compare with Box tests).

In Figure 9 we see that the standardized residuals of component four and higher look pretty white. In the first three pictures some kind of pattern is visible especially after the crisis. To decide the question whether the standardized residuals are uncorrelated, we take a look at the empirical autocorrelation function. The acfs of the components are plotted in Figure 10. In Figure 10 we see that the autocorrelation functions are mostly not significantly different from zero except for component 1, 2 and 3 (for 4 it is hard to decide).

The plots and Box tests both show the tendency that the lower components have a high autocorrelation and from component 4 upwards the correlation declines. Next I asses the puzzle seen in the previous section, namely that the SRTSM does not perform as good in the period 2014 to 2016 as hoped. In Figure 11 we see the standardized residuals for the period 1990 to 2016.

In Figure 11 we see the same picture as before. The higher the component the better the whiteness performance. For the first two some kind of pattern around the year 2015 (the last tick on the x-axis) is visible. So lets zoom into the period after the financial crisis. This is done in Figure 12.

In Figure 12 it can be seen that the first two components seem to be biased to the downside, too smooth and do not fluctuate evenly around zero, therefore they do not look white. The others look reasonable. It seems that the SRTSM model has some problems fitting the

component	lag 1	lag 2	lag 3	lag 4	lag5
1	7.598861e-05	2.117552e-05	7.749173e-05	2.322880e-04	4.623978e-04
2	1.090389e-07	2.244133e-09	1.843051e-09	4.633161e-09	5.984615e-09
3	3.528866e-06	5.264108e-06	1.310856e-05	3.456039e-05	9.283425e-05
4	0.001047801	0.004595805	0.01140794	0.01200308	5.575401e-03
5	0.0902211	0.01134887	0.002870362	0.002774384	4.840084e-03
6	0.4910506	0.0006260142	0.0001278134	9.979575e-05	2.127457e-04
7	0.09884328	0.008566805	0.001600062	0.003160873	5.165244e-03

Table 2: p values for Ljung Box test for different lags and the seven components of the residuals from 1990 to 2016.

period around 2015, this corresponds to what we have seen in the previous section. As the log likelihood value (as a measure for the goodness of fit) using the period 1990 to 2016 is lower than the value when using 1990 to 2014[40]. As a last check I plot the autocorrelation function of the sub period 2009 to 2016 in Figure 13. There the problem with the first two components is clearly visible for the other components it is hard to decide but I would rate them as reasonably uncorrelated.

Next I compare Figure 13 with the estimated acfs of the period 1990 to 2009 (see Figure 14). The acfs in Figure 14 look very similar to the ones using the whole 1990 to 2016 period. The picture is again the same, the first components are correlated and for the others it is not perfectly clear but I would rate them as uncorrelated.
As a last check whether the autocorrelations of the residuals are caused by model inherent assumption failures or the shock caused by the crisis I repeat the Box test for the sub period 1990 to 2009. The p values are reported in Table 3.

component	lag 1	lag 2	lag 3	lag 4	lag5
1	7.025634e-04	3.490061e-04	1.156762e-03	3.066124e-03	5.888361e-03
2	2.067576e-06	1.629812e-07	2.154518e-07	5.666318e-07	9.463505e-07
3	9.848393e-06	2.009416e-05	6.190157e-05	1.589858e-04	3.945461e-04
4	7.233482e-04	3.278539e-03	9.507216e-03	0.01152262	9.513611e-03
5	0.2225635	0.04829925	0.0649086	0.105199	0.1764296
6	0.664045	7.006002e-03	0.02832123	9.979575e-05	0.05228086
7	0.3055402	0.3928057	0.5993025	0.6233401	0.6891777

Table 3: p values for Ljung Box test for different lags and the seven components of the residuals from 1990 to 2009.

In Table 3 we see that the p values of component one to four are below all conventional significance levels therefore the H_0 is rejected even when using the sub period 1990 to 2009. For the remaining components we get a different picture than for the whole period. The p values for the components four and higher are above the 1% and 5% significance level and most are also above the 10% level. This corresponds well to the acf and residual plots where these components seem to have a low autocorrelation. We can conclude that the picture improved when looking at the sub period 1990 to 2009. Nevertheless the model assumptions

[40]The picture does not change much if I use the parameters estimated from the period 1990 to 2016.

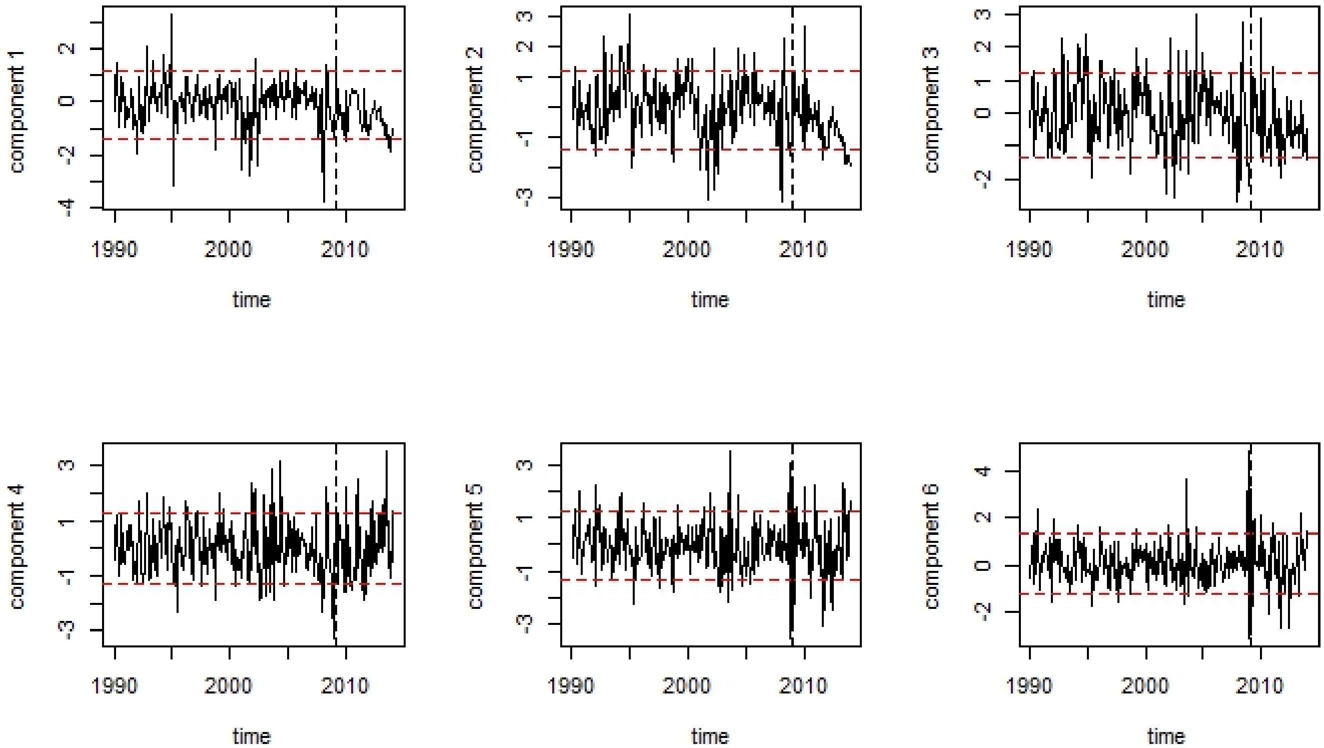

Figure 9: Residual plot of the first 6 components. Red dashed line are naive confidence intervals using the mean + $\Phi^{-1}(0.9)$ and mean - $\Phi^{-1}(0.1)$ where Φ stands for the standard normal distribution function. The vertical dashed line represents the financial crisis (01/01/2009). The parameters for the SRTSM are estimated from the 1990 to 2014 are used.

seem to be violated and it can not reasonably be assumed that the noise processes w_t and v_t of the state space models are i.i.d processes. So one part of the correlation can be attributed to the shock caused by the crisis the other part is model inherent. This means that the estimation of the SRTSM with the Kalman filter is based on false assumptions that are not met in reality. This could be a new direction for further research. Especially how to incorporate the observed autocorrelation in the residuals into the filter and find an estimation procedure where the assumptions are not in conflict with reality.

Further I present the residuals of the GATSM model for the period 1990 to 2016 in Figure 15. There we see a similar picture as before. The first 3 components show some kind of pattern and the others look reasonably uncorrelated. Further we note that the volatility of the standardized residuals (especially of the first two components) of the GATSM after the crisis seem to be lower and fluctuate more evenly around zero than the one from the SRTSM. This would point to a better fit for the GATSM in this period. Nevertheless we have to keep in mind that the fit (log likelihood) for the GATSM is lower than the fit for the SRTSM (see also mean squared error analysis below). One part of this puzzle can be explained when looking at the standard deviations. In the GATSM the variance of the residual components is constant over the whole period whereas the variance for the SRTSM drops off after the

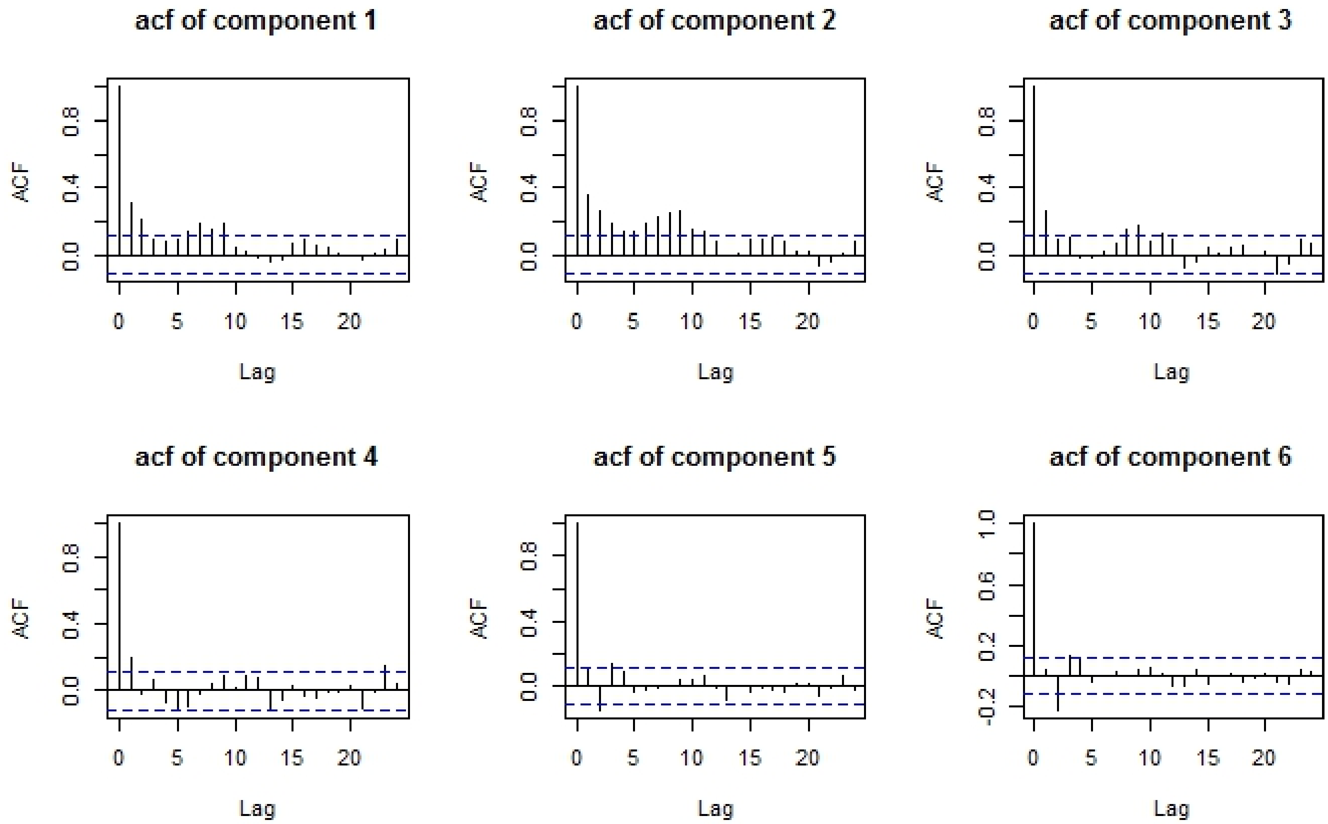

Figure 10: Plot of the autocorrelation function of the first 6 components of the residuals. Data from 1990 to 2016 and the parameters for the SRTSM estimated from the 1990 to 2014 period are used.

crisis[41]. For the GATSM, the period after the crisis, gets standardized by a variance that is to big. This explains one part of the low volatility of the GATSM residuals after the crisis. The different variances are plotted in Figure 16[42]. When I analysed the GATSM residuals with the Box test the picture matches the one for the SRTSM, so I do not report them. For the whole period the p values are across the board below all significance levels. When using the sub period 1990 to 2009 the first three components have low p values and the others behave like in the SRTSM above.

Last I analyse the SRTSM when using 1990 to 2009 as estimation and 2009 to 2016 as validation period. Interestingly when I consider the lower bound $\bar{r}$ as parameter for the optimization the SRTSM estimated from 1990 to 2009 matches the GATSM estimated from 1990 to 2014 almost perfectly even on the validation set 2014 to 2016. In the left plot of Figure 17 we see that the shadow rate produced by the SRTSM estimated from 1990 to 2009 matches the short rate of the GATSM estimated from 1990 to 2014. This effect is due to the fact that in this period a lower floor of $\bar{r} = -7.7$ is estimated for the SRTSM and

[41] The mean variance of the SRTSM equals the constant variance of the GATSM.

[42] In Figure 16 the first entry for $t = 1$ is not plotted since it would distort the picture due to the large entries caused by the initialization.

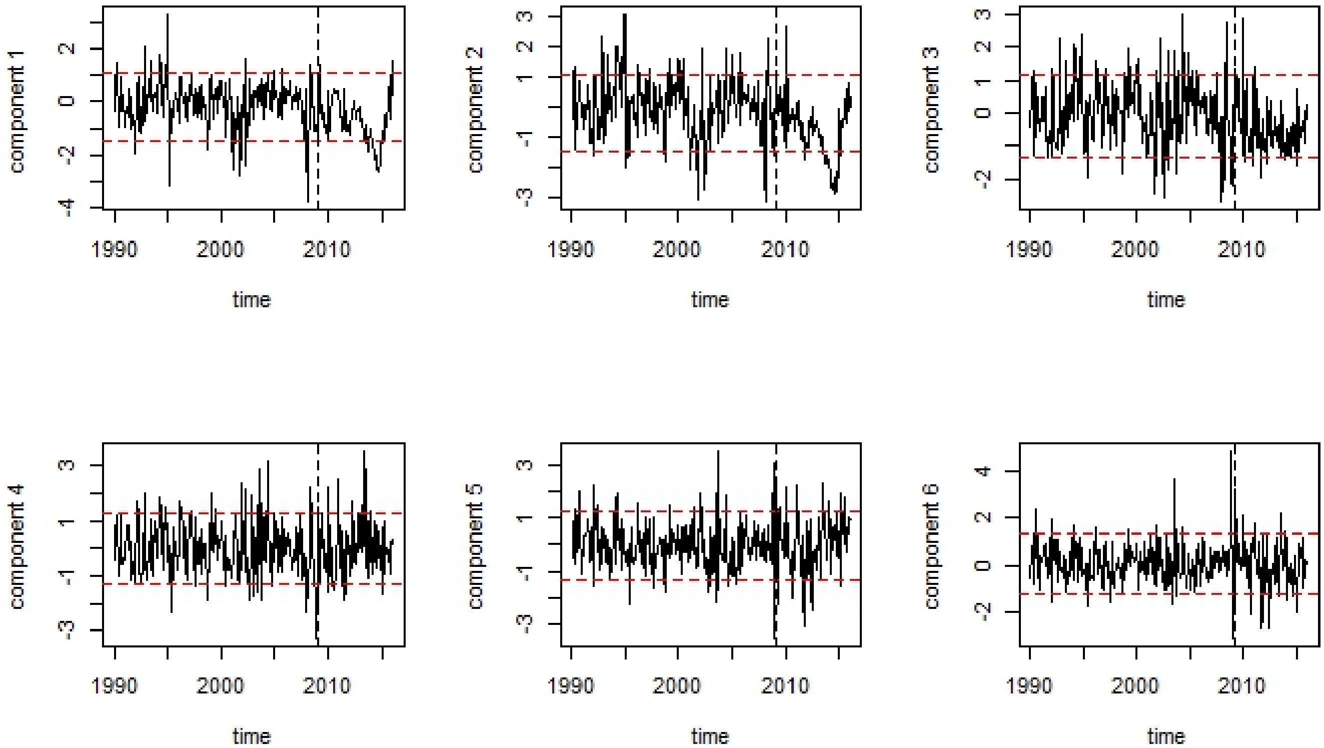

Figure 11: Residual plot of the first 6 components. Here the period 1990 to 2016 is used in conjunction with the parameters estimated from 1990 to 2014.

the other parameters are very similar to the GATSM parameters estimated from the 1990 to 2014 period. When I estimate the SRTSM from 1990 to 2009 with a fixed lower floor of 0.25 the shadow rate and parameters are very similar to the shadow rate of the SRTSM estimated from 1990 to 2014. This can be seen in Figure 17 in the right plot. There we see that the two shadow rates are very similar. Furthermore the standardized residual plots and the acfs of the two SRTSM models display the same picture and are very similar. The standardized residuals of the lower components (1 ,2, 3) have a high autocorrelation and the others have a lower autocorrelation. Therefore I do not report them since they contain no new information.

Furthermore I compare the baseline GATSM and SRTSM, both with the parameters estimated from the period 1990 to 2014, using the mean squared error (i.e. the mean of the raw, non standardized squared residuals $MSE^i = \frac{1}{t_1-t_0}\sum_{t=t_0}^{t_1}(\tilde{y}_t^i)^2$, thereafter MSE). The results are reported in Table 4. In Table 4 we see that the SRTSM performs a bit better than the GATSM over all periods (interestingly for the first component the GATSM performs better), but the differences are not dramatic. The picture matches the one obtained by the log likelihood value where the SRTSM performs also better than the GATSM. Note that in the period 1990 to 2009 the GATSM and SRTSM match each other, this corresponds to what we have seen so far.

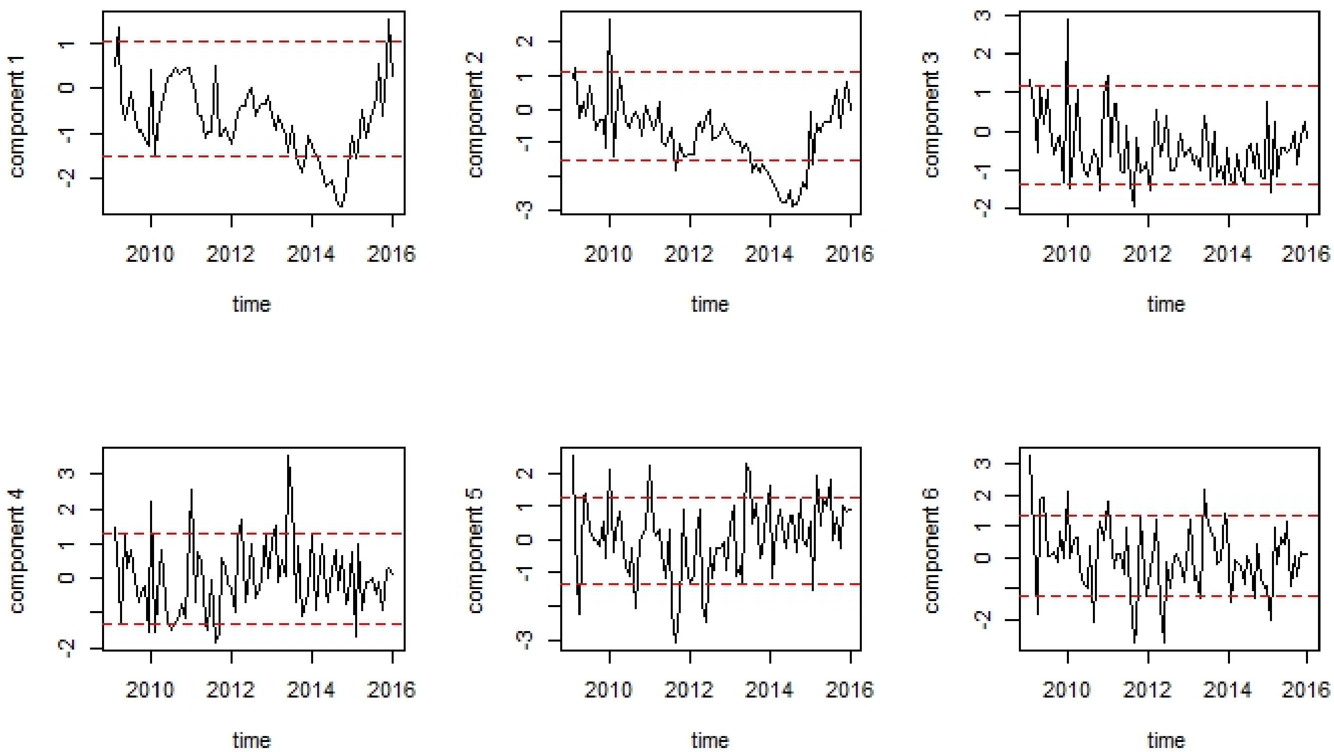

Figure 12: Residual plot of the first 6 components. Here the period 2009 to 2016 is plotted with the parameters estimated from 1990 to 2014.

To summarize the diagnostics section I conclude that in all models and periods the lower components corresponding to short maturities display an autocorrelation that is significantly different from zero. The other components can reasonably be assumed to be uncorrelated. So the model assumptions are violated and the question how to incorporate a non zero correlation structure into the model and estimation is an interesting question for further investigation.

8 Analysis of shadow rate. Factor augmented vector auto regression FAVAR

Now we have estimated the so-called shadow rate using the SRTSM. Next we want to analyse if this rate can replace the effective federal funds rate in various models. Specifically we want to know whether the interaction of the shadow rate with other macroeconomic variables is consistent with the link between effr and the variables in normal times. We have seen so far that in normal times when the economy is not stuck at the zero lower bound the shadow rate mirrors the effr almost perfectly, therefore the interactions between the rate and economic variables in normal times is like one would expect. The remaining question is whether there is meaningful information in the shadow rate when it is below zero and does not match the effr. This question is analysed using a factor augmented vector auto regression (FAVAR).

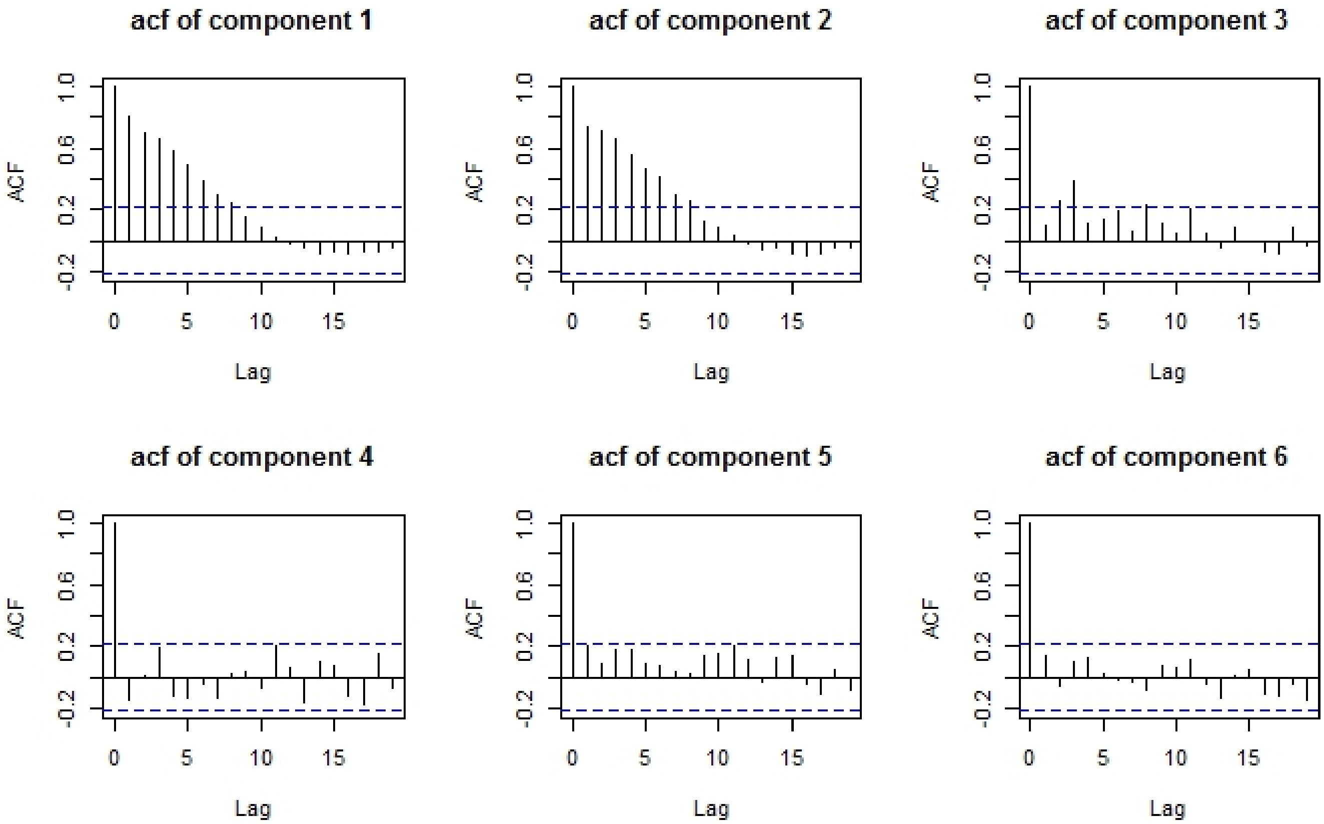

Figure 13: Plot of the autocorrelation function of the first 6 components of the residuals between 2009 and 2016. Using the parameters for the SRTSM estimated from the period 1990 to 2014.

First I will briefly summarize the FAVAR and then present the results. The FAVAR analysis section is mostly based on Wu and Xia [5] and Ben S. Bernanke, Jean Boivin, Piotr Eliasz (2005) [16] who proposed the FAVAR approach to measure the interactions between monetary policy and macroeconomic variables.

8.1 Summary of factor FAVAR

Here I will recap the FAVAR framework used to analyse whether the variation in the shadow rate contains meaningful information when the economy is stuck at zero lower bound. The FAVAR is very similar to normal VAR analysis. The basic idea is that in a normal vector auto regression set-up we focus on a few important macroeconomic variables together with the interest rate. Then we estimate the model and make certain identification assumptions to identify the monetary policy shocks (i.e. recursiveness assumption, that policy shocks only have a contemporaneous impact on the interest rate). When the shocks are identified the impulse response functions are used to assess the link between monetary policy and the economy.
The criticism with this approach is that due to the focus on a few important macroeconomic variables the model is far too narrow to capture what is going on in the economy.

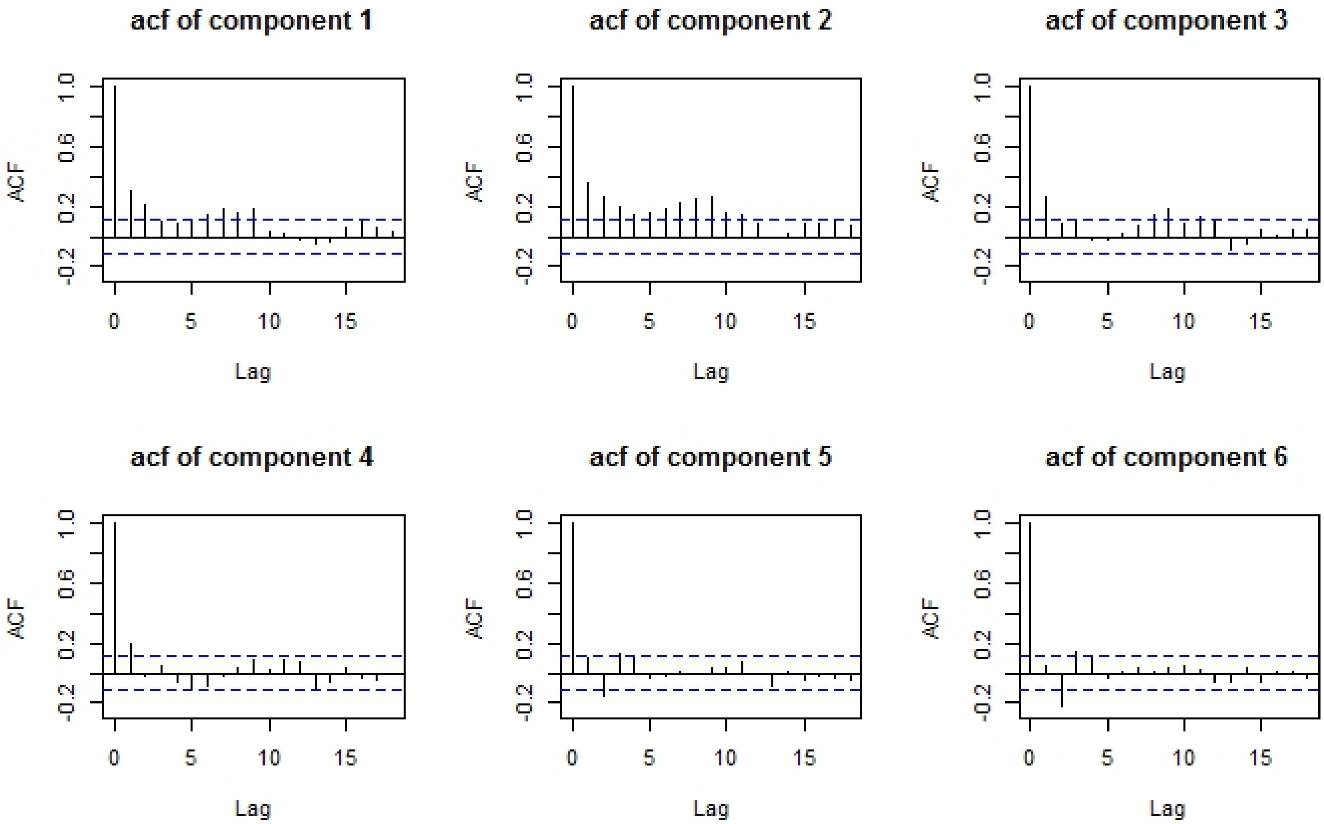

Figure 14: Plot of the autocorrelation function of the first 6 components between 1990 and 2009. Using the parameters for the SRTSM estimated from the period 1990 to 2014.

Therefore the factor augmented VAR is used. The idea is to take a large set of economic variables, extract the first three principal components (i.e. factors) of the data set and use these 3 factors together with the shadow rate in a VAR. Before estimating the VAR model the influence of the interest rate on the factors is removed using a simple OLS estimation technique. Furthermore we determine the loadings of the factors on the macroeconomic variables. This is done to gain insight into the impact of the shadow rate on each macroeconomic variable even though the variable is not incorporated in the VAR model. So we use a kind of dimensionality reduction to summarize the information contained in a large data set Y_t by a low dimensional vector x_t. We denote the shadow rate or policy rate by s_t. For the whole FAVAR analysis a time horizon of 1960 to 2014 is used. Therefore the rate s_t is constructed using the effr from 01/01/1960 to 01/01/2009 and afterwards the shadow rate, estimated in the previous section is used[43].
The FAVAR is summarized by three equations. The VAR uses 13 lags due to the monthly observation frequency[44] and is rewritten into a regression model, given by:

[43]This is done since we have estimated the shadow rate only from 1990 onwards but since the economy was far away from the zero lower bound between 1960 and 1990 the shadow rate would match the effr in this period.

[44]With the AIC information criterion we would select 14 lags using the VARselect function in R.

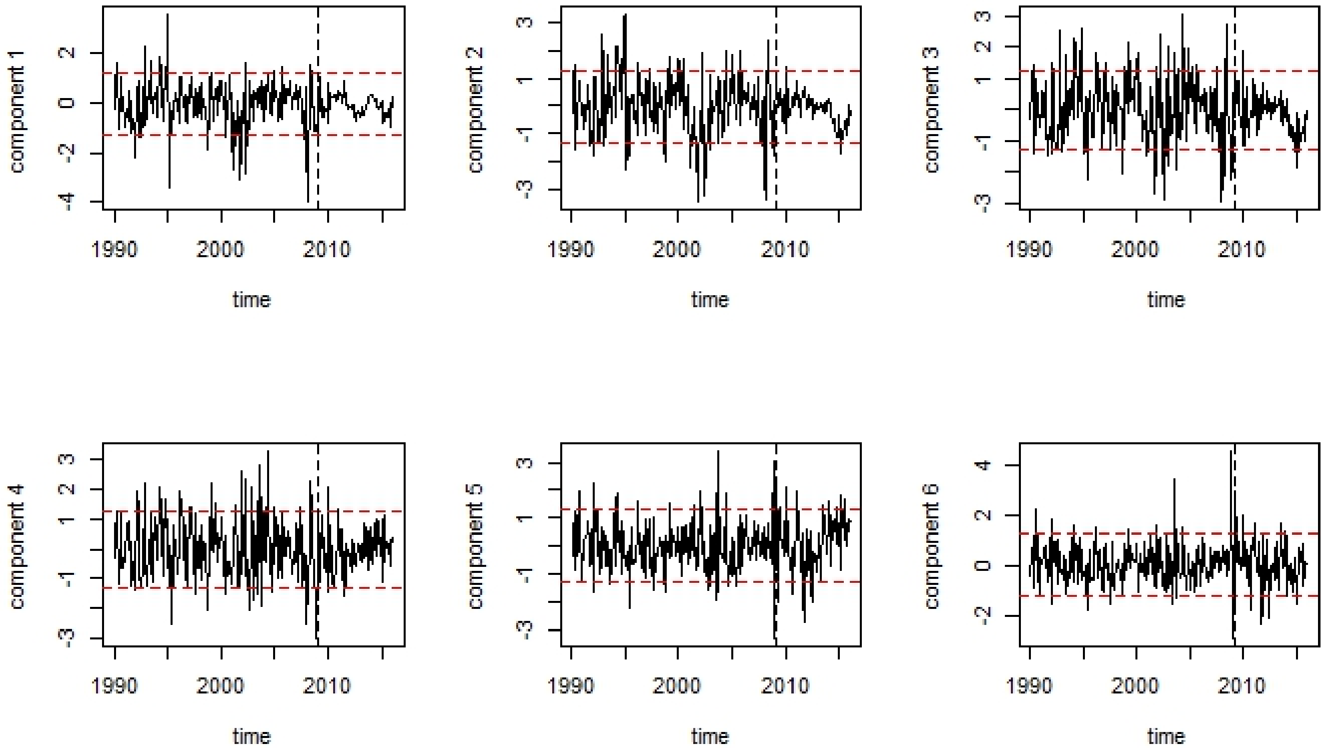

Figure 15: Residual plot of GATSM. First 6 components of the residuals between 1990 and 2016. Using the parameters for the GATSM estimated from 1990 to 2014.

$$\begin{bmatrix} x_t \\ s_t \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_s \end{bmatrix} + \rho^m \begin{bmatrix} X_{t-1} \\ S_{t-1} \end{bmatrix} + \Sigma^m \begin{bmatrix} \epsilon_t^m \\ \epsilon_t^{MP} \end{bmatrix}, \begin{bmatrix} \epsilon_t^m \\ \epsilon_t^{MP} \end{bmatrix} \sim N(0, I) \tag{47}$$

where $x_t \in \mathbb{R}^3$ are the macro factors, $s_t \in \mathbb{R}$ the shadow rate, $\mu_x \in \mathbb{R}^3$, $\mu_s \in \mathbb{R}$ the intercepts, $X_t := (x_t', ..., x_{t-12}')' \in \mathbb{R}^{13 \cdot 3}$ the lagged factors, $S_t := (s_t, ..., s_{t-12})' \in \mathbb{R}^{13}$ the lagged rates, $\Sigma^m \in \mathbb{R}^{4\times 4}$ the lower triangular matrix and $\rho^m \in \mathbb{R}^{4\times 13\cdot 4}$ (the subscript m just stands for macro). For the identification of the monetary policy shocks ϵ_t^{MP} the recursiveness assumption is made meaning that we take Σ^m as lower triangular. This means that only the shadow rate s_t reacts immediately on a monetary policy shock ϵ_t^{MP}.

The macro factors are created by estimating the first three principal components[45] of the

[45] PCA (see Johnson and Wichern (2013)[25]): Given a dataset $Y := (Y_t^m)_{t=1}^T \in \mathbb{R}^{n\times T}$. We assume that $y \in \mathbb{R}^n \sim N(\mu, B)$ and that the dataset consists of T realisation of this random variable. First we demean all the variables involved, so without loss of generality $y \sim N(0, B)$. We want to approximate y by a linear combination of k factors (i.e. principal components pc, $k << n$). This means we approximate y by W'pc, were pc is a k dimensional random vector and $W \in \mathbb{R}^{k\times n}$. Without loss of generality we may assume that $WW' = I_k$. The goodness of this approximation is measured by the mean squared error. Thus we want to minimize $\mathbb{E}((y - W'pc)'(y - W'pc))$. For given W the optimal chocie for the factors pc is $pc = (WW')^{-1}Wy = Wy$ and $\hat{y} = W'pc = W'Wy$ (compare with OLS estimation). The MSE for this choice is given by:

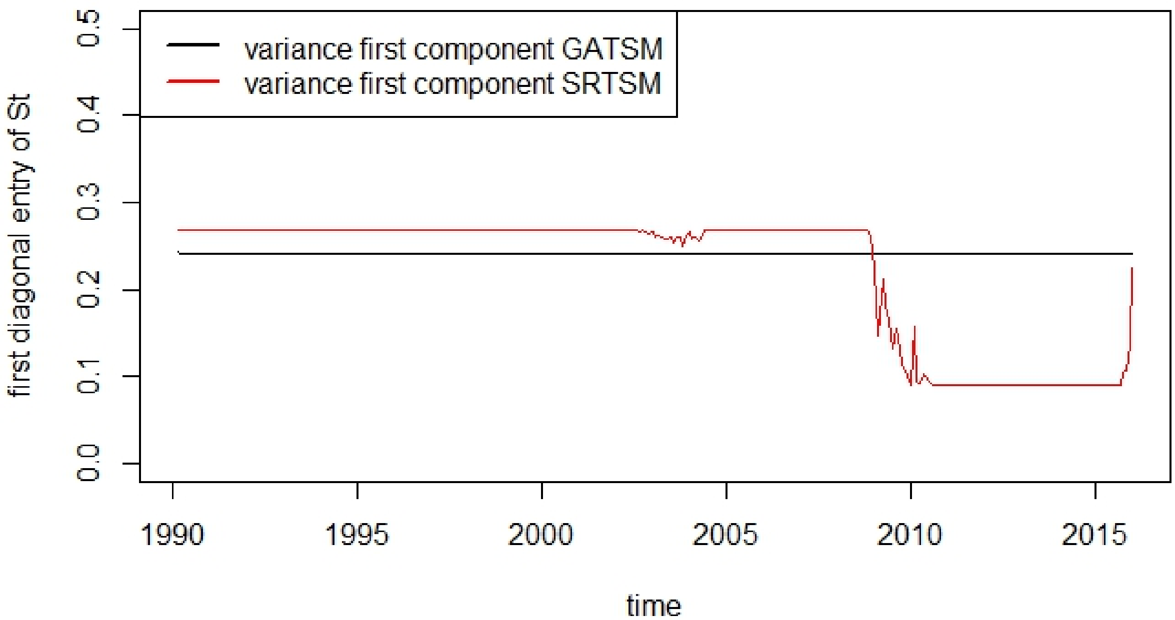

Figure 16: Plot of $S_t^{1,1}$ of GATSM and SRTSM.

high dimensional macro data set $Y := (Y_t^m)_{t=1960}^{2014}$ with $Y_t^m \in \mathbb{R}^{97}$, that means we use a bal-

---

$$\mathbb{E}(\|y - W'pc\|^2) = \mathbb{E}((y - W'Wy)'(y - W'Wy)) = \mathbb{E}(y'y - 2y'W'Wy + y'W'WW'Wy) =$$

using the constraint $WW' = I$, it follows

$$= \mathbb{E}(y'y) - \mathbb{E}(y'W'Wy)$$

Here we see that minimizing the reconstruction error is equivalent to maximizing the sum of the variance of the components of $Wy = pc$ since the first term in the last equation does not depend on W and is therefore a constant in the optimization with respect to W. For simplicity we assume $k = 1$ ($Wy \in \mathbb{R}$) and derive the first principal component. Using the Lagrangian function $L(W, \lambda) = \mathbb{E}((Wy)^2) - \lambda(WW' - 1)$ it follows:

$$\frac{\partial L(W, \lambda)}{\partial W} = 2\mathbb{E}(Wyy') - 2\lambda W' = 0 \in \mathbb{R}^n$$

using $(Wy)y = y(Wy) = yy'W' \in \mathbb{R}^n$ since Wy is a scalar for which $a = a'$ holds.

$$\mathbb{E}(yy')W' = \lambda W'$$

There we see an eigenvalue problem of the variance matrix of y (since y is demeaned) and using the saddle-point characterisation of the Lagrangian function we have to maximize with respect to λ since we started by minimizing the MSE. Therefore we see that for k = 1 W is the eigenvector to the highest eigenvalue of the variance matrix of y. For general dimensions it can be shown that the rows of W are the first k eigenvectors to the k highest eigenvalues [25]. To calculate them the unknown true variance matrix $B = \mathbb{E}(yy')$ is replaced by the empirical variance matrix $\hat{B} = \frac{1}{T}YY'$. After calculating the first k eigenvectors of $\hat{B}$ and putting them together in W, the principal components are given by $pc_t = WY_t$. Often either the principal components pc are normalized to unit variance or the whole data set Y is demeaned and normalized ($\hat{B}$ is then the correlation matrix) before doing the pca.

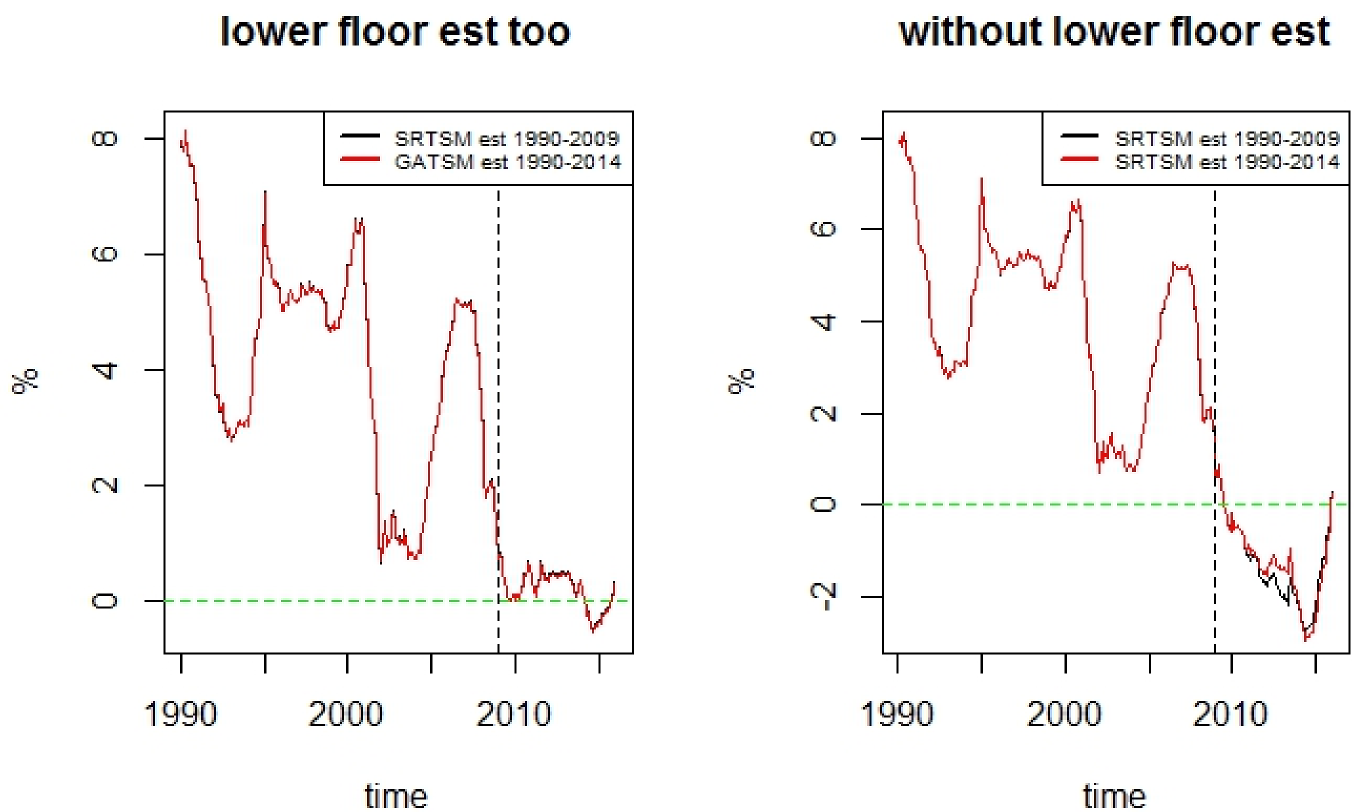

Figure 17: Plot of short rates obtained by the SRTSM and GATSM when using different estimation periods.

anced panel of 97 macroeconomic variables. The data set is from the Global Insight Basic Economics data base and is the same data set used by Wu and Xia [5]. After estimating the first three principal components $pc_t \in \mathbb{R}^3$ of Y, we remove the influence of the rate s_t (take the part that is orthogonal to the rate see Bernanke [16]).
This is done by estimating the following linear regression model:

$$pc_t = b_{pc}pc_t^* + b_s s_t + \nu_t^{pc} \tag{48}$$

that is estimated by OLS. Where pc_t^* are the first three principal components of as subset of the data set Y. This subset is denoted by $Y^* \in \mathbb{R}^{l \times T}$ with $l < 97$ and represents slow moving variables. That means they do not react fast to the policy rate (i.e. Consumer price index CPI, Unemployment rate, Consumption expenditures, Industrial production, ...) in contrast to fast moving ones like various interest and exchange rates. The list of the variables used and which data transformations are applied to ensure stationarity can be found in the Online Appendix to Wu and Xia "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound" [46]. There you can also see which variables are considered slow moving. The macro factors for the VAR are then given by: $x_t = pc_t - \hat{b}_s s_t$. We remove the influence of

[46] Can be found on the internet: https://sites.google.com/site/jingcynthiawu/ or http://faculty.chicagobooth.edu/jing.wu/

component MSE	1	2	3	4	5	6	7	mean MSE
GATSM 1990-2016	0.043	0.069	0.097	0.110	0.100	0.092	0.119	0.090
SRTSM 1990-2016	0.043	0.067	0.093	0.111	0.096	0.091	0.110	0.087
GATSM 2009-2016	0.010	0.021	0.039	0.058	0.133	0.122	0.216	0.086
SRTSM 2009-2016	0.013	0.016	0.0254	0.065	0.120	0.121	0.186	0.078
GATSM 2014-2016	0.009	0.045	0.062	0.039	0.111	0.053	0.358	0.097
SRTSM 2014-2016	0.021	0.027	0.024	0.041	0.073	0.053	0.261	0.071
GATSM 1990-2009	0.056	0.086	0.119	0.129	0.090	0.084	0.089	0.093
SRTSM 1990-2009	0.056	0.086	0.119	0.129	0.091	0.084	0.089	0.093

Table 4: Comparison SRTSM and GATSM with parameters estimated from 1990 to 2014 period. MSE for different periods.

the policy rate from the factors because we want to study the effect of the policy rate in the VAR model where it is explicitly incorporated. Therefore to identify the influence correctly the macro factors should represent all forces, except for the policy rate, active in the economy.

After the macro factors x_t are created we estimate the following linear regression model:

$$Y_t^m = a_m + b_x x_t + b_s s_t + \nu_t^m \tag{49}$$

this is done to get the loadings $\hat{b}_x \in \mathbb{R}^{97\times 3}$ and $\hat{b}_s \in \mathbb{R}^{97\times 1}$ of the macro factors and policy rate on the panel of macroeconomic variables Y_t^m. With these loadings we can construct an impulse response function for all 97 variables and not only the factors used in the VAR. The estimation of (49) is done with 97 OLS estimations for each $Y_t^{m,i}$ $i \in \{1, ..., 97\}$[47]. If $Y_t^{m,i}$ is part of the slow moving variables we set $\hat{b}_s^i = 0$.

The impulse response functions for all macroeconomic variables are given by:

$$\Psi_j^{MP,i} = \frac{\partial Y_{t+j}^{m,i}}{\partial \epsilon_t^{MP}} = \hat{b}_x^i \frac{\partial x_{t+j}}{\partial \epsilon_t^{MP}} + \hat{b}_s^i \frac{\partial s_{t+j}}{\partial \epsilon_t^{MP}}$$

where $\frac{\partial x_{t+j}}{\partial \epsilon_t^{MP}}$ and $\frac{\partial s_{t+j}}{\partial \epsilon_t^{MP}}$ are the standard impulse response functions generated by the VAR model. Since we are only interested in the influence of monetary policy shocks and do not know how to interpret the macro factors x_t we state the impulse response function only for the influence of ϵ_t^{MP}.
The impact of monetary policy after the financial crisis on the variable $Y_t^{m,i}$ can be charac-

[47]For this I standardize the variables $Y_t^{m,i}$ $\forall i \in \{1, ..., 97\}$ by demeaning and normalizing to unit variance. Therefore the loads reported in Figure 8 are between -1 and 1. They match the ones reported by Wu and Xia up to a scaling factor.

terised through[48]

$$\sum_{\tau=t_0}^{t} \Psi_{t-\tau}^{MP,i} \epsilon_{\tau}^{MP} \tag{50}$$

8.2 Results FAVAR

In this section the results for the FAVAR are reported. I have implemented the factor augmented vector auto regression like described above. In Figure 18 I plot the loadings b_x and b_s obtained by the regression of equation (49). They match the ones reported by Wu and Xia [5] almost perfectly up to a rescaling factor[49]

In Figure 18 it can bee seen that that the first macro factor ($x_{t,1}$) loads most heavily on the real activity measures (index 1 to 60, various industrial production indices, labour force, consumption and housing starts measures), the second factor loads most heavily on the price level indices (index 80 to 95), and the third loads on unemployment measures (index 19 to 30) and price measures (index 80 to 95). The short rate loads mostly on the various interest rate related macro variables (index 60 to 70), this is per design since we set $\hat{b}_s^i = 0$ for the slow moving macro variables (real activity measures 1 to 46 and price measures 80 to 96). The ordering of the variables is according to the online appendix published by Wu and Xia [5] (Can be found on the internet: https://sites.google.com/site/jingcynthiawu/ or http://faculty.chicagobooth.edu/jing.wu/). There you also see which indices correspond to slow moving variables. To keep the analysis tractable we focus on six different macro variables in the data set, namely the shadow rate, industrial production, consumer price index, capacity utilization, unemployment rate and housing starts.

Furthermore I report a summary of the 97 R^2 (coefficient of determination) values of the 97

[48] Since in a VAR(p) model we have $a(B)Y_t = \Sigma\epsilon_t$ with $\epsilon \sim N(0, I)$ and $a(B) := (I - A_1B - ... - A_pB^p)$ where B stands for the backshift operator ($B^n x_t = x_{t-n}$ $\forall n \in \mathbb{N}$) and the matrices A_i are the coefficients for lag i, if $a(B)$ is invertible (corresponding to Y being stationary) we have ($t > t_0$, $t - j = \tau \in \{t, ..., -\infty\}$):

$$Y_t = a^{-1}(B)\Sigma\epsilon_t = \sum_{j=0}^{\infty} \Phi_j \Sigma\epsilon_{t-j} = \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j} = \sum_{\tau=-\infty}^{t} \Psi_{t-\tau}\epsilon_\tau = \sum_{\tau=-\infty}^{t_0-1} \Psi_{t-\tau}\epsilon_\tau + \sum_{\tau=t_0}^{t} \Psi_{t-\tau}\epsilon_\tau$$

with $\Psi_j = \Phi_j\Sigma$ where Σ is the cholesky decomposition of the covariance matrix $\Omega = \Sigma\Sigma'$ of $\Sigma\epsilon$. This means we use the orthoganalized impulse response function Ψ which corresponds to the recursiveness assumption to identify monetary policy shocks. The second sum on the right side stands for the influence of shocks after t_0. If we pick out the monetary policy shocks we get (49).

The impulse responses Φ are calculated by a comparison of coefficients of the right and left side of $a^{-1}(B)a(B) = I$ with $(a^{-1}(B) = (\Phi_0 + \Phi_1B + \Phi_2B^2 + ...)$ and $a(B) = (I - A_1B - ... - A_pB^p)$. We get $\Phi_0 = I$, $\Phi_s = \sum_{j=1}^{p} A_j\Phi_{s-j}$(putting $\Phi_s = 0$ for $s < 0$).

After we obtained the estimates $\hat{A}_i$ for the coefficients by OLS, we can calulate the estimated impulse response functions $\hat{\Phi}_s$ by the formula given above. Then they are orthogonalized using the cholesky decomposition $\hat{\Sigma}$ of the estimated covariance matrix $\hat{\Omega} = \frac{1}{T}\sum_{t=0}^{T} \hat{u}_t\hat{u}_t' = \hat{\Sigma}\hat{\Sigma}'$ where $\hat{u}_t$ are the estimated residuals obtained by OLS.

The confidence bands are bootstrapped using a recursive scheme: To produce a new draw of $(Y_t)_{t=1}^{T}$ we take a random permutation π of the set $\{1, ...T\}$ and set $Y_t^* = \sum_{j=1}^{p} \hat{A}_j Y_{t-j}^* + \hat{u}_{\pi(t)}$ $\forall t$ where the first starting values are given by the original time series ($Y_t^* = Y_t$ for $t \in \{1, ..., p\}$). With this procedure a large number of new draws is generated from which the confidence intervals are generated.

[49] This could be due to numeric problems or that the principal components are differently normalized or calculated.

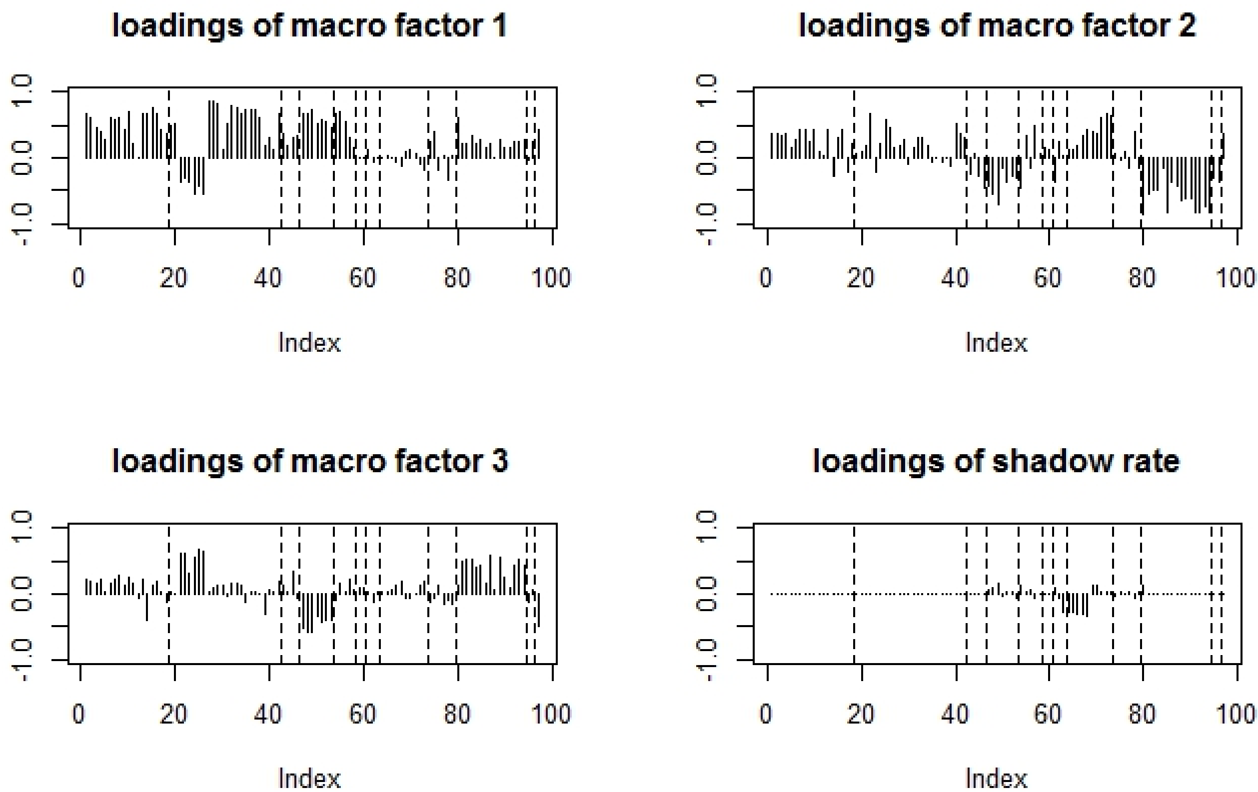

Figure 18: loadings b_x and b_s on the macro variable Y_t, obtained by regression of equation (49), dashed lines indicate the different categories of the macro variables (i.e. real output, employment, price measures) using the same ordering as Wu and Xia [5].

OLS estimations in Table 5 to get a feeling how well the three macro factors $x_t \in \mathbb{R}^3$ and the shadow rate $s_t \in \mathbb{R}$ describe the full macro data set Y_t^m. 35% of the R^2 are greater than 0.6, 50% are greater than 0.5. Only 24% of them are lower than 0.2 and the average R^2 over the 97 regressions is 0.45700.

min	1 quantile	median	mean	3 quantile	max
0.00493	0.22910	0.49440	0.45700	0.66770	0.99400

Table 5: Summary statistics of the 97 R^2 values of (49).

In Figure 19 I plot the R^2 values:

To summarize, the four variables $(x_t', s_t)' \in \mathbb{R}^4$ explain on average about halve the variation of the large macroeconomic panel Y_t. Furthermore the variables $Y_t^{m,i}$ with $i \in \{11, 12\} \cup \{55, ...62\} \cup \{74, ..., 79\}$ and some other are not captured well by the regression (index 11,12 correspond to industrial production of Oil&Gas and residential utilities, 55 to 62 are real inventories, stock price indices and foreign exchange rates, 74 to 79 are money and credit quantity aggregates). Therefore we can conclude that the four regressors used summarize the economy quite well apart for the above mentioned sectors. This is quite good

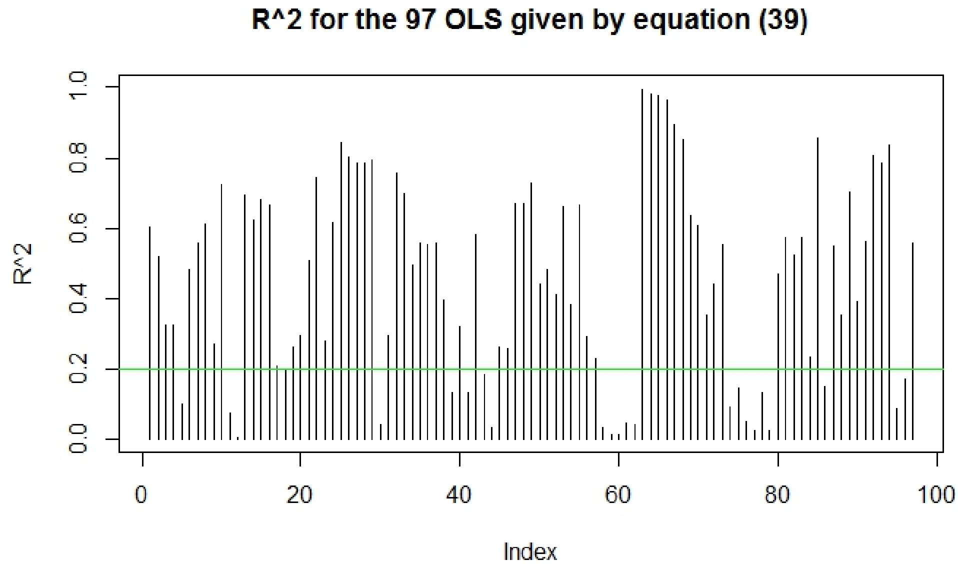

Figure 19: R^2 for the 97 macro variables of equation (49).

when you think that we reduced the dimension from 97 to 4.

Next I follow Wu and Xia [5] and look at how the macro economic variables would have evolved when the monetary policy shocks since the financial crisis are shut off. Practically this is implemented by subtracting the influence of monetary policy on the variable $Y_t^{m,i}$ given by (50). This is done to get the picture what would have happened if the central bank did nothing in the aftermath of the great recession. By comparing how the variables evolved with how they would have behaved when the MP shocks are turned off, we get an indication of impact of the measures taken by central banks after the crisis.
Furthermore we make the thought experiment how the macro variables would have behaved if the central banks were forced to hold the shadow rate at the zero lower bound. This is done by calculating the shocks needed to hold the policy rate steadily at the lower bound $\bar{r} = 0.25$ and then using these shocks and equation (50)[50] to create an alternative trajectory of the macro variables Y_t^m. This experiment is plotted in Figure 20.

In Figure 20 we see that the shadow rate with turned off MP shocks (red dashed line in top-left plot, counter-factual 1 shadow rate) first goes below the shadow rate (black line) before crossing it around the beginning of 2011. After that the counter-factual 1 rate stays above the actual shadow rate. On average the counter-factual 1 rate is 0.15% higher than the

[50] This is done by calculating new policy shocks $\tilde{\epsilon}_t^{MP}$ such that the new policy rate $\tilde{s}_t = s_t - \sum_{\tau=t_0}^{t} \Psi_{t-\tau}^{MP,i} \tilde{\epsilon}_\tau^{MP} = s_{t_0+j} - \sum_{k=1}^{j+1} \Psi_{j-k+1}^{MP,i} \tilde{\epsilon}_{k+t_0-1}^{MP} = 0.25 = \bar{r}$ $(k = \tau - t_0 + 1 \in \{1, ..., t - t_0 + 1 = j + 1\})$ $\forall t = t_0 + j \geq t_0$ for the $i \in \{1, ..., 97\}$ such that $Y_t^{m,i} = s_t$. This I have done recursivly. For $t = t_0$ we get $\tilde{\epsilon}_{t_0}^{MP} = \frac{s_{t_0} - \bar{r}}{\Psi_0^{MP,i}}$. For the rest we calculate $(t = t_0 + j)$:

$$\tilde{\epsilon}_t^{MP} = \frac{s_{t_0+j} - \bar{r} - \sum_{k=1}^{j} \Psi_{j-k+1}^{MP,i} \tilde{\epsilon}_{k+t_0-1}^{MP}}{\Psi_0^{MP,i}}$$

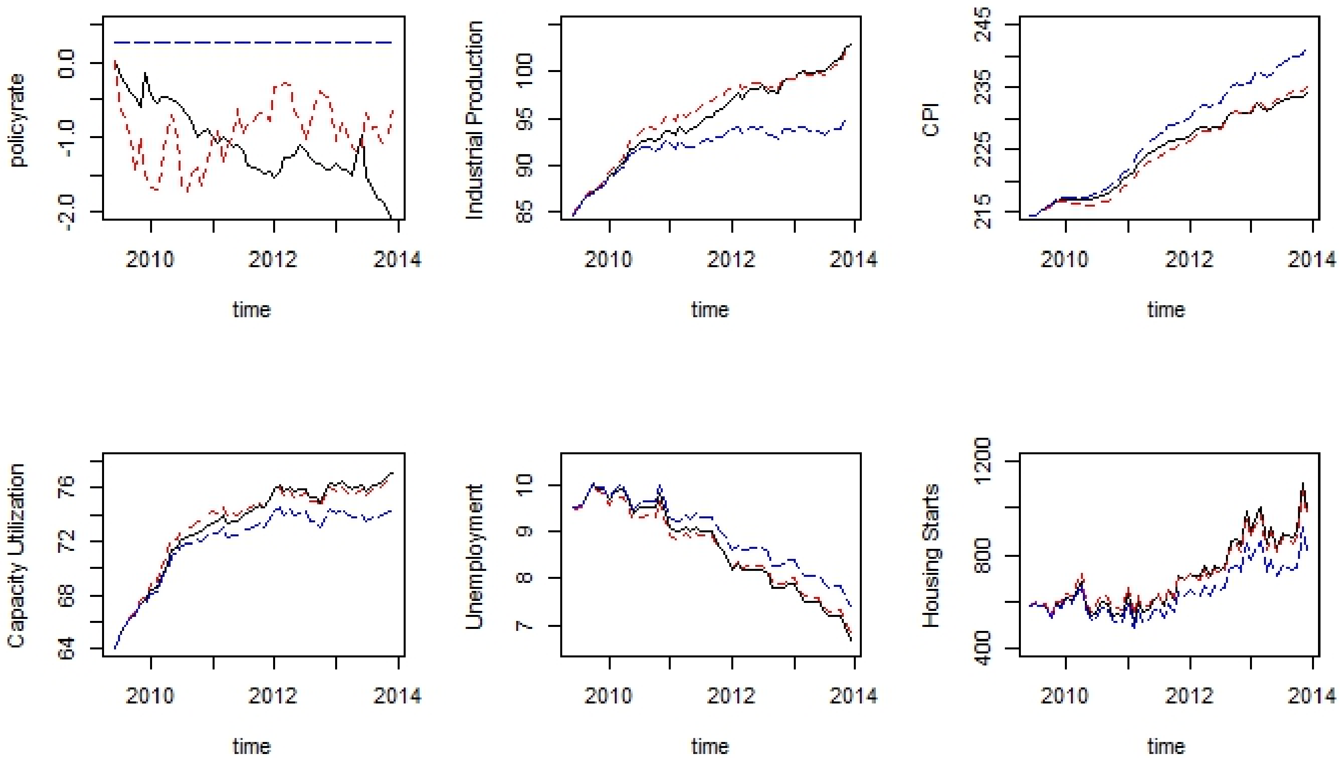

Figure 20: black normal trajectories. red dashed trajectories with turned of monetary policy influence (counter-factual 1), blue dashed line trajectories with shocks that keep the policy rate at the lower bound (counter-factual 2).

actual shadow rate. This indicates that the monetary policy actions after the financial crisis lowered the shadow rate. This is what one would expect given that central banks used a lot of unprecedented tools, like quantitative easing and forward guidance, to loosen monetary policy and give the economy a boost.
Somewhat counter-intuitively the counter-factual 1 rate undershoots the actual shadow rate from the start of 2009 to 2011. This can be attributed to the temporary structural break caused by the financial crisis. Maybe the shock of the great recession to the system we try to model was so great that the system needs some time to stabilize and therefore the results immediately after the crisis are not really meaningful. Another interpretation of this undershooting is that the central banks were first caught by surprise by the severity of the financial crisis. And even though they loosened monetary policy a lot in the aftermath of the crisis it was not enough and they actually tightened monetary policy compared to what the market expected or needed. (The shadow rate and the MP shocks are calculated using the forward rates that are observed in the real world markets (through the zero coupon yields). After the crisis an extremely strong risk off move occurred in the markets pushing yields of secure bonds (US government bonds) lower. So they are relative to what the market expected or needed, therefore we could say that monetary policy was too tight in the aftermath of the financial crisis. Some economists also share this view. The most prominent one is the noble price winner Paul Krugman). Compared to the findings of Wu and Xia [5] in their Figure

6 my Figure 20 is extremely similar, confirming their findings. At the beginning of 2014 the actual shadow rate was -2.133% and the counter-factual 1 rate was at -0.649%. This confirms that the central banks succeeded in lowering monetary policy conditions through unconventional policy tools since without the MP shocks the rate would have been 1.484% $= -0.649 - (-2.133)$% higher.

Furthermore we see that per construction the counter-factual 2 rate (blue dashed line) stays at 0.25. Next we analyse the remaining 5 pictures of Figure 18. We see that the counter-factual 1 trajectories match the actual trajectories relatively closely like in Wu and Xia. Looking at industrial production where first the counter-factual 1 is above the actual and after 2013 it is a little bit below the actual path. Compared to the counter-factual 2 trajectory the actual industrial production index is much higher. When looking at the CPI plot we see the price puzzle that the counter-factual 2 CPI path is higher than the actual one. This is strange since the shocks needed to push the shadow rate to the lower bound can be interpreted as restrictive monetary policy that should according to theory reduce inflation and therefore the growth of CPI. But this counter-intuitive result is also present in Wu and Xia [5] and other related literature[51]. The pattern that emerges when comparing counter-factual 1 and actual is that at the start of our observation period (2009 to 2011) the counter-factual 1 displays looser monetary conditions therefore the paths for it are in the beginning a bit better and after 2013 a bit worse than the actual (i.e. counter-factual 1 unemployment rate being lower in the beginning and then higher, or utilization first higher than lower). This pattern is closely related to the fact that the counter-factual 1 rate first undershoots the actual before getting higher.

For the other three variables the picture matches the one found for industrial production. Actual and counter-factual 1 closely match each other and the counter-factual 2 paths being worse from an economic viewpoint (i.e. lower capacity utilization, higher unemployment and lower housing starts).

To summarize Figure 20 we can say that the counter factual 1 is very similar to the actual path but in the end (2013 onwards) behaves like expected (i.e. lower industrial production, utilization, housing starts and higher unemployment rate) but the differences are relatively minor. When comparing the actual to the counter factual 2 paths the picture is clearer, the counter-factual 2 trajectories are worse from an economic viewpoint (lower industrial production, capacity utilization, unemployment rate, housing starts and higher unemployment). So all in all using both thought experiments we can conclude that monetary policy after the crisis achieved its goal of supporting the economy. Nevertheless for both the CPI price puzzle emerges, namely that restrictive MP actions increased the CPI instead of reducing it.

Next we look at the Impulse response functions (the entries of the matrices Ψ_s). This is a standard time series analysis tool that can help answer the question what happens if the central bank loosens monetary policy through a -0.25 MP shock. It is plotted in Figure 21.

In Figure 21 we see that all the variables except CPI react as expected. A monetary policy shock of -0.25 decreases the policy rate, CPI and Unemployment rate and increases the industrial production capacity utilization and housing starts. Apart from CPI this is exactly what is expected. The effects are strongest in the first year and then slowly die off. The

[51]Examples include Sims (1992)[18] and Eichenbaum (1992)[19].

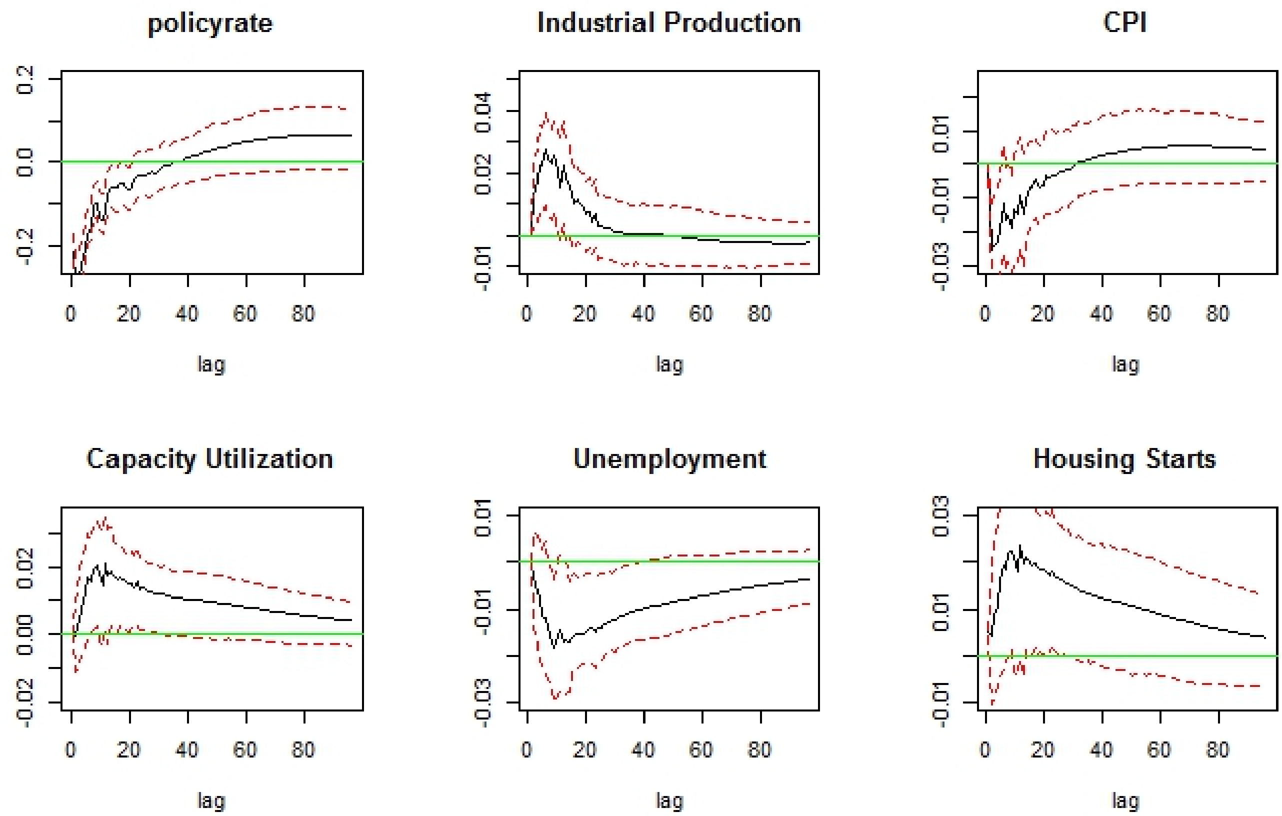

Figure 21: Impulse response functions for the VAR given by (47) with 90% confidence bands.

graphic matches the one reported by Wu and Xia quite well except for some rescaling.
Last but not least we compare the impulse response function to the impulse response function generated when using only a sub sample. The sub sample is the time period when the economy is at the zero lower bound from 01/01/2009 forward. The sub sample has only 59 observations, so we can not fit a VAR(13) model since the number of parameters to be estimated would be to big for the small sample to allow a good estimation. Therefore I compare the full sample impulse response to the ones generated when fitting a VAR(1) model to the sub sample. We expect them to be similar since the hypothesis is that the shadow rate can replace the effr when it is stuck at the zero lower bound. We have seen that the full sample model impulse response functions behaves like expected. If the sub sample one look similar that is a further validation of the shadow rate.

In Figure 22 we see that both are similar except for CPI, there the sub sample one behaves a little bit strange shooting up first before coming down again. But overall they are qualitatively the same confirming the hypothesis that the economy interacts with the shadow rate like the effr interacts with it in normal times.
As a last test I compare the impulse response functions of a VAR(1) model using 1990 to 2009 with a VAR(1) model using 2009 to 2013. In the first model the economy is away from the zero lower bound, therefore this model represents the normal dynamics between effr and the economy. In the second model the economy is at the zero lower bound and we want to see that the interactions between the shadow rate and the economy represented by the

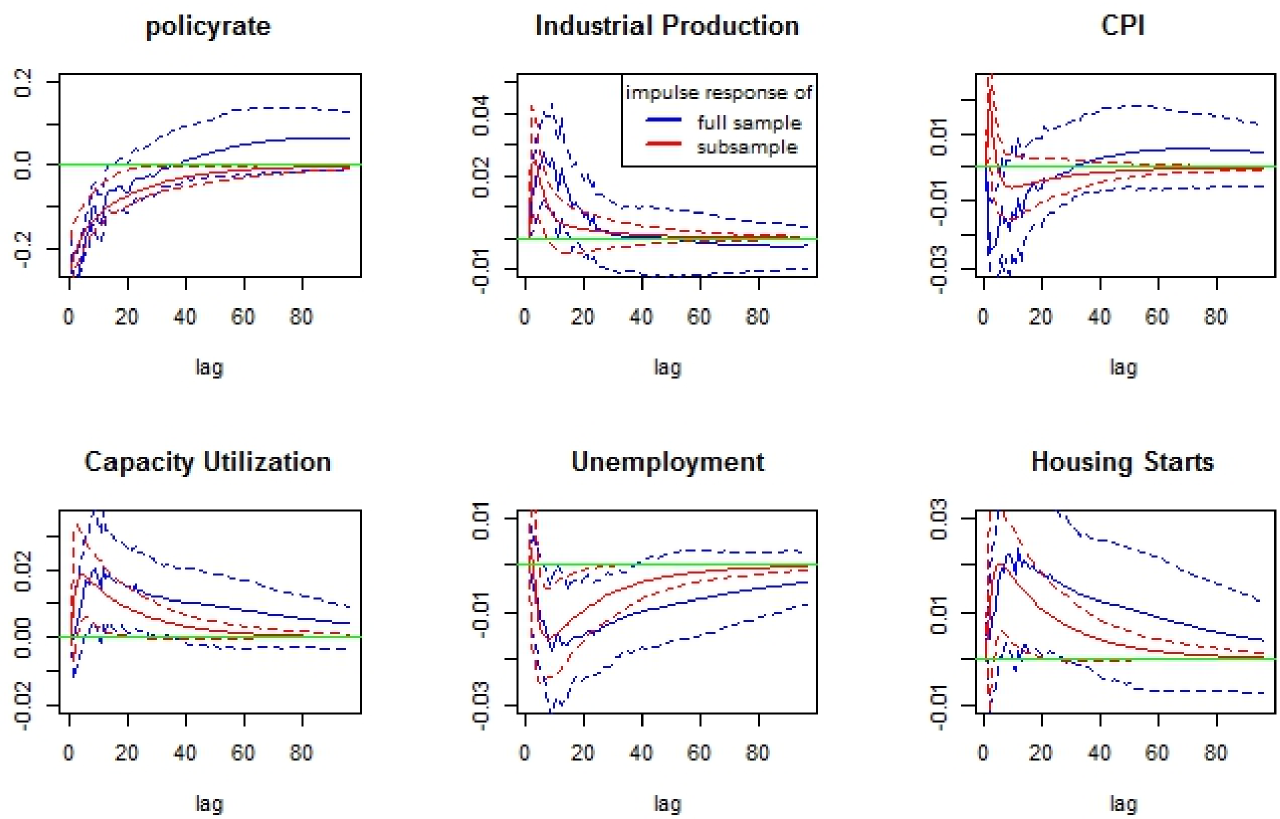

Figure 22: Impulse response functions for the VAR given by (47) with 90% confidence bands: blue VAR(13) using the full sample 1960-2014, red VAR(1) using sub sample 2009-2014.

impulse response function are similar to model 1.

In Figure 23 we see that the two impulse response functions behave very similar confirming that the shadow rate can replace the effr in economic models.

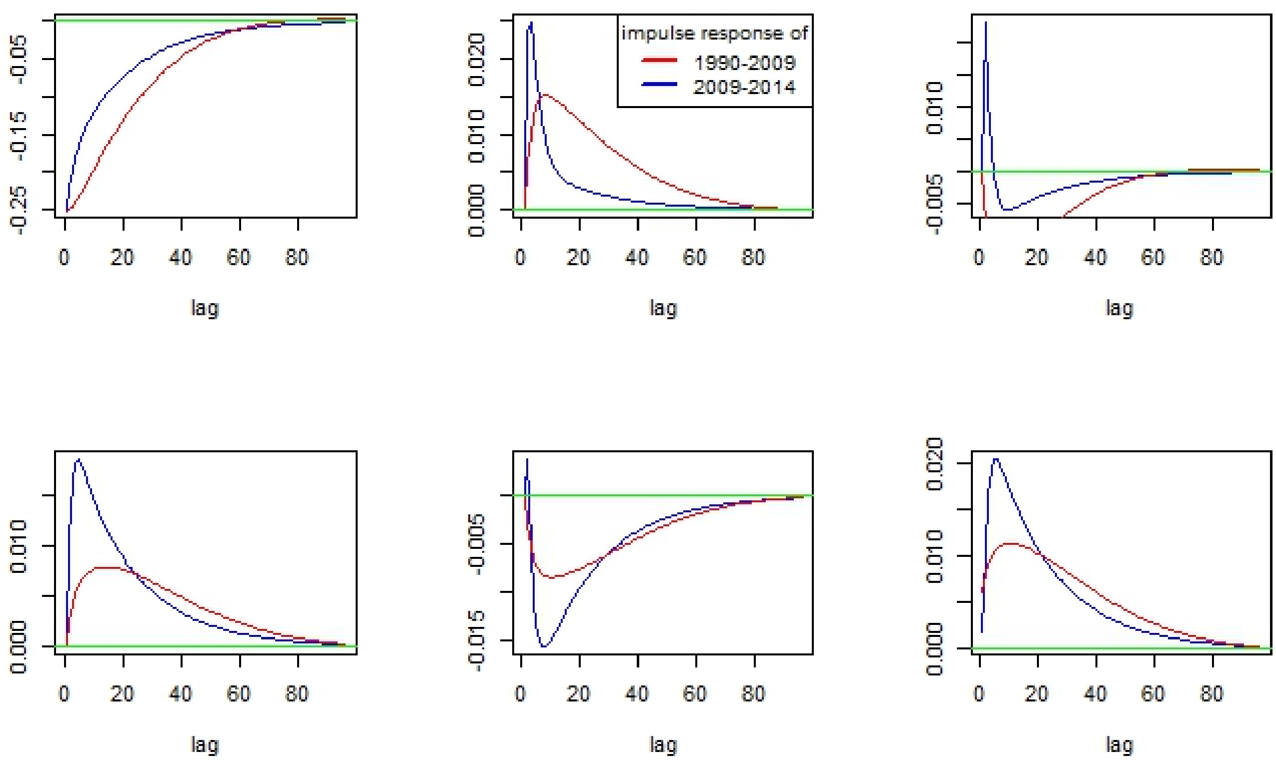

Figure 23: blue VAR(1) using the sample 2009-2014, red VAR(1) using sample 1960-2009.

8.3 Structural break test

In this section we are interested whether the great financial crisis in 2008 was a permanent or temporary structural break. That means we use a statistical test to look if the influence of the policy rate on the macro factors before and after the financial crisis is significantly different. This is done by analysing a new slightly extended VAR model and comparing it with the restricted VAR model. A likelihood ratio test is used, the $H_0 : \theta \in \Theta_0 \subseteq \Theta$ and corresponding $H_1 : \theta \in \Theta_0^c$:

$$\lambda := \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_R)}{L(\hat{\theta}_U)}$$

where $L(\theta)$ stands for the likelihood function and $\hat{\theta}_R$ for the estimated (using the maximum likelihood technique) parameters under the restriction H_0. $\hat{\theta}_U$ is the unrestricted estimator. The test statistic is then given by:

$$D = -2\ln(\lambda) = 2(\ln(L(\theta_U)) - \ln(L(\theta_R))) \approx \chi^2_{d_U - d_R}$$

The statistic is asymptotic chi-squared distributed with $d_U - d_R$ degrees of freedom where d_U and d_R stand for the degrees of freedom in the unrestricted and restricted model. For a VAR model we can simplify further since we assume that the residuals are i.i.d. normal distributed, therefore the likelihood, with the same reasoning of the conditional distribution as in the section Parameter estimation of Kalman filter[52], is given by:

$$\ln(L(\hat{\theta})) = -\frac{dT}{2}\ln(2\pi) - \frac{T}{2}(\ln(\det(\hat{\Omega})) - \frac{1}{2}\sum_{t=1}^{T} \hat{u}_t'\hat{\Omega}^{-1}\hat{u}_t \quad (51)$$

where $\hat{u}_t \in \mathbb{R}^d$ (d stands for the dimension of the VAR) are the residuals of the VAR and $\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t\hat{u}_t'$. The last term can be rewritten since for a scalar $a = tr(a)$ holds:

$$\sum_{t=1}^{T} \hat{u}_t'\hat{\Omega}^{-1}\hat{u}_t = tr(\sum_{t=1}^{T} \hat{u}_t'\hat{\Omega}^{-1}\hat{u}_t) = tr(\sum_{t=1}^{T} \hat{\Omega}^{-1}\hat{u}_t\hat{u}_t') =$$

using the linearity and $A, B' \in \mathbb{R}^{n\times m}$ $tr(AB) = tr(BA)$ of the trace operator.

$$= tr(\hat{\Omega}^{-1}\sum_{t=1}^{T} \hat{u}_t\hat{u}_t') = tr(\hat{\Omega}^{-1}\hat{\Omega}T) = tr(TI_d) = dT$$

Therefore we can write the log likelihood for a VAR model as:

$$\ln(L(\hat{\theta})) = -\frac{dT}{2}\ln(2\pi) - \frac{T}{2}(\ln(\det(\hat{\Omega})) + \frac{Td}{2} \quad (52)$$

Inserting (52) into the formula for the test statistic D gives (using that for a VAR(p) model the actual number of residuals is T-p, the model invariant term $-\frac{dT}{2}\ln(2\pi) + \frac{Td}{2}$ drops out because of the difference):

$$D = (T-p)(\ln(\det(\hat{\Sigma}_R\hat{\Sigma}_R')) - \ln(\det(\hat{\Sigma}_U\hat{\Sigma}_U'))) \approx \chi^2_{d_U - d_R}$$

[52] In a VAR(p) model $x_t = \rho_1^m x_{t-1} + ... + \rho_p^m x_{t-p} + \epsilon_t$ with $\epsilon_t \sim N(0, \Sigma^m(\Sigma^m)')$ an i.i.d sequence. It holds $x_t|(x_{t-1}, ..., x_{t-p}) \sim N(\sum_{j=1}^{p} \rho_j^m x_{t-j}, \Sigma^m(\Sigma^m)')$ due to the gaussian i.i.d assumption. Using the same reasoning as in the section Parameter estimation of Kalman filter we obtain (50) as log likelihood function

where $\hat{\Sigma}_U\hat{\Sigma}_U'$ is the estimated covariance matrix of the unrestricted model and $\hat{\Sigma}_R\hat{\Sigma}_R'$ the one from the restricted.

We use this test for the model given by (47), where we alter the first block corresponding to the macro factors x_t to:

$$x_t = \mu_x + \rho^{xx}X_{t-1} + 1_{\{t<t_0\}}\rho_1^{xs}S_{t-1} + 1_{\{t_0<t<t_1\}}\rho_2^{xs}S_{t-1} + 1_{\{t>t_1\}}\rho_3^{xs}S_{t-1} + \Sigma^{xx}\epsilon_t^m \qquad (53)$$

where $t_0 \hat{=} 12/2007$ and $t_1 \hat{=} 06/2009$. That means we split the time horizon into three parts, before, during and after the financial crisis. We want to test whether the influence of the shadow rate before and after the crisis on the macro factors is the same. We split into three periods because we acknowledge that we can not explain the financial crisis with this model and therefore the crisis period from 12/2007 to 06/2009 is a structural break. The question is whether it is a temporary or permanent break.

Therefore $H_0 : \rho_1^{xs} = \rho_3^{xs} \in \mathbb{R}^{3\times 13}$. The modified VAR is estimated per OLS for each equation and the residuals are computed to estimate the covariance matrix. This is done for the restricted and unrestricted model. Using the above described test statistic which is asymptotic χ^2_{39} distributed since the difference between the numbers of parameters in the model is $3 \cdot 13 = 39$ (the unrestricted model has one $\rho_3^{xs} \in \mathbb{R}^{3\times 13}$ matrix more, or general $3 \cdot p$ when p denotes the number of lags).

If the test does not reject the H_0 it is a further justification for the shadow rate model. Since before the financial crisis the shadow rate equals the effective federal funds rate and after the crisis the two diverge. But if the influence on the macro factors is not significantly different in both periods, we can conclude that the shadow rate captures the monetary policy stance well. Even in times when the effr is at the zero lower bound and most of the policy is carried out through unconventional monetary policy the shadow rate captures this. In Table 6 I report the p values of this test for different lags[53]:

lag	p value	lag	p value
3	0.04154977	10	0.92539736
4	0.29012416	11	0.70899096
5	0.46468712	12	0.47018998
6	0.48923964	13	0.62886946
7	0.74532115	14	0.51387952
8	0.76192240	15	0.56650534
9	0.77858223		

Table 6: p values for structural break test for different lags for $H_0 : \rho_1^{xs} = \rho_3^{xs}$ using the shadow rate s_t

We see that the H_0 does not get rejected for lags ≥ 4 since the p values are greater than any conventional significance level. Therefore we can conclude that no permanent structural break occurred when using the shadow rate. At least for lags greater 4 but since the observation frequency is monthly it is reasonable to only look at models with more lags since

[53]Wu and Xia [5] use a slightly different test statistic because they use (T-k) where k is the number of regressors on the right hand side of (43). Values in Table 2 are calculated based on this method.
But simply using the number of lags p does not change the picture they match the p values quite well and the two methods never contradict (meaning one rejects while the other does not) each other.

it takes more than four months until the interest rate translates into the broader economy represented by the macro factors x_t. Most economist think that monetary policy needs at least 6 months to gain its full effect on the real economy.

Next we repeat this exercise but instead of the shadow rate s_t we use the effective federal funds rate (before 2009 it holds $s_t = \text{effr}_t$). We expect to detect a structural break using this set up, since after the financial crisis the effr is stuck at zero lower bound and does not display much variation. The p values for all lags up to 15 are zero[54] and I report them in Table 7. Therefore the H_0 is rejected for all lags considered. We conclude that using the effr the financial crisis can be considered as permanent structural break. This indicates that the shadow rate has meaningful information in it and that it captures also the unconventional monetary policy actions. Also the hypothesis that the shadow rate can replace the effr as input for economic models is validated.

lag	3	4	5	6	7	8	9	10	11	12	13	14	15
p value	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7: p values for structural break test for different lags for $H_0 : \rho_1^{xs} = \rho_3^{xs}$ using the effective federal funds rate r_t

As a last check I also test the $H_0 : \rho_1^{xs} = \rho_2^{xs} = \rho_3^{xs}$ using the likelihood ratio test. With this hypothesis the test statistic is asymptotic chisquared distributed with $3 \cdot 13 \cdot 2 = 78$ (or more general $3 \cdot p \cdot 2$ where p stands for the lag) degrees of freedom since in the unrestricted model we have two more matrices with dimension $3 \cdot p$. The p values are reported in Table 8.

lag	p value	lag	p value
3	6.508374e-04	10	2.407974e-09
4	3.826035e-03	11	0
5	1.478694e-03	12	0
6	1.730426e-04	13	0
7	2.411632e-05	14	0
8	1.671907e-05	15	0
9	8.783583e-08		

Table 8: p values for structural break test for different lags for $H_0 : \rho_1^{xs} = \rho_2^{xs} = \rho_3^{xs}$ using the shadow rate s_t

In Table 8 we see that the H_0 is below all conventional significance levels for all lags. Therefore the null hypothesis gets rejected, meaning that a structural break occurred.
To summarize, the test showed the financial crisis caused a structural break. When using the shadow rate it is just a temporary structural break and the influence of the shadow rate on the economy before and after the crisis is not significantly different. When using the effective federal funds rate the crisis is a permanent structural break meaning that the interactions

[54]The p values for both procedures are basically the same, see previous footnote

between the effective federal funds rate and the economy before and after is significantly different.

9 Summary

In this master's thesis I have analysed the shadow rate term structure model (SRTSM) that was recently proposed in the literature (Wu and Xia, 2016 [5]). I can confirm their findings and even extend them a little bit. Furthermore I analysed and summarized all the mathematical concepts used by the SRTSM in this thesis. The SRTSM allows us to estimate the so-called shadow rate from observed bond price data. The shadow rate matches the effective federal funds rate (effr) set by the central bank (i.e. fed) in normal times when the effr is away from the zero lower bound. But when the economy is at the zero lower bound the shadow rate can go negative and displays a lot of variation. Therefore the hope is that the shadow rate can capture the monetary policy conditions better than the effr. Since the effr is basically constant at the zero lower bound and has no meaningful variation in it that can explain other variables.

One interpretation would be that most of the monetary policy was conducted with unconventional tools like forward guidance and quantitative easing and the shadow rate captures these effects well. Whereas the effr is constant and does not contain information about unconventional policy. This means that there is much more information in the variation of the shadow rate than in that of the effr. This hypothesis is confirmed by means of a factor augmented vector auto regression analysis. Consistent with Wu and Xia my thesis clearly shows that the economy interacts with the shadow rate like the effr, even in times when the economy is stuck at the zero lower bound. That means researchers can simply use the shadow rate instead of the effr as input for economic models.

I also analysed the robustness of the model by varying the estimation period. My findings show that the shadow rate can be considered to be relatively robust with respect to different lower floors. Furthermore I tested for the model assumptions of the Kalman filter, especially if the noise processes of the state space model are i.i.d processes. My result demonstrate that this assumption is not met. It is open for further investigation how to extend the model and estimation with a non zero autocorrelation function. Furthermore I have to mention that different estimation periods for the parameters used in calculating the shadow rate, result in quite different shadow rates. Nevertheless all the rates behave qualitatively similar only some display more extreme moves (when the estimation period contains data of 2015).

Furthermore I looked at the behaviour of the shadow rate when the economy exits the zero lower bound. On 16/12/2015 the fed hiked rates by 0.25% for the first time since the financial crisis. The shadow rate reacts to this like expected. In the year up to the first rate hike it rushes back to the zero lower bound and matches the effr since.

All in all the analysis showed that the shadow rate can replace the effective federal funds rate as input for economic models.

10 References

[1] https://www.federalreserve.gov/monetarypolicy/openmarket.htm (accessed on 15/04/2017)

[2] http://www.ecb.europa.eu/stats/policy_and_exchange_rates/key_ecb_interest_rates/html/index.en.html (accessed on 17/04/2017)

[3] Paul Kruehner, Interest Rate Theory, 2016 (lecture script of the course interest rate structure models and derivatives)

[4] Refet S. Gürkaynak, Brian Sack, and Jonathan H. Wright, The U.S. Treasury Yield Curve: 1961 to the Present, 2006, Finance and Economics Discussion Series Divisions of Research & Statistics and Monetary Affairs Federal Reserve Board, Washington, D.C.

[5] Jing Cynthia Wu and Fan Dora Xia, Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound (2016), Journal of Money, Credit and Banking

[6] Darrell Duffie, Rui Kan, A yield-factor model of interest rates, 1996, Mathematical Finance

[7] Black and Fischer, Interest Rates as Options, 1995, Journal of Finance, 50, 1371–6

[8] Jens H. E. Christensen, Francis X. Diebold, Glenn D. Rudebusch, The affine arbitrage-free class of Nelson-Siegel term structure models, 2007, National Bureau of economic research

[9] Scott Joslin, Kenneth J. Singleton, Haoxiang Zhu , "A New Perspective on Gaussian Dynamic Term Structure Models" (2011), Oxford university press

[10] James D. Hamilton and Jing Cynthia Wu, "Risk premia in crude oil futures prices" 2014, Journal of international money and finance

[11] Jens H. E. Christensen, Affine Term Structure Models: An Introduction, 2015 Federal Reserve Bank of San Francisco

[12] Francis Diebold and Canlin Li, Forecasting the term structure of government bond yields" (2006), Journal of Econometrics

[13] Drew D. Creal and Jing Cynthia Wu, "Estimation of affine term structure models with spanned or unspanned stochastic volatility", (2015), Journal of Econometrics

[14] Gürkaynak, Sack, and Wright (2007) data set https://www.federalreserve.gov/pubs/feds/2006/200628/200628abs.html (accessed on 10/03/2017)

[15] EFFR data till may 2017 https://fred.stlouisfed.org/series/FEDFUNDS (accessed on 12/03/2017)

[16] Ben S. Bernanke, Jean Boivin, Piotr Eliasz, Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach, 2005, Journal of Economics

[17] Gregory R. Duffee, Term Premia and Interest Rate Forecasts in Affine Models (2002), Haas School of Business

[18] Sims, Christopher A. (1992) “Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy.” European Economic Review, 36, 975–1000

[19] Eichenbaum,Martin.(1992) Comment on Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy. European Economic Review, 36, 1001–11.

[20] Dan Simon, Optimal State Estimation, 2006 Wiley-Interscience

[21] Robert Shumway and David Stoffer, Time series analysis and its applications, 2006 Springer

[22] Rheinländer Thorsten, Sexton Jenny, Hedging Derivatvies 2011 World Scientific

[23] Havlicek Hans TU vienna, Lineare Algebra für Technische Mathematiker Band 16, Heldermann Verlag (lecture script at the TU Vienna)

[24] R. Meise, D. Vogt: Einführung in die Funktionalanalysis 1992, Vieweg

[25] Dean W. Wichern and Richard A. Johnson, Applied Multivariate Statistical Analysis (2013)

11 Appendix

In the Appendix I present some theorems used for the derivation of the Kalman filter, they are based on the book of Shumway and Stoffer [21] and Meise and Vogt [24] (projection theorem).

Projection theorem:
Let H be an Hilbert space and $M \subseteq H$ a closed subspace. Then for each $x \in H$ there exists a unique decomposition:

$$x = \hat{x} + \hat{u}$$

where $\hat{x} \in M$ and $\hat{u} \perp M$. $\hat{x}$ is called the orthogonal projection of x on M. Furthermore $\hat{x}$ is the best approximation of x by an element out of M:

$$\|x - \hat{x}\| < \|x - \tilde{x}\| \qquad \forall \tilde{x} \in M, \tilde{x} \neq \hat{x}$$

In the following we use the projection theorem on L^2. $L^2(\Omega, \mathscr{A}, P)$ is the space of square integrable random variables[55] with the scalar product $< X, Y >:= \mathbb{E}(XY)$.

Furthermore we denote the projection of $x \in L^2$ on a closed subspace M of L^2 by $\text{Pr}_M x$. For the following theorems we define (for given random vectors $y_1, ..., y_n$) $Y_n := span\{1, y_1, ..., y_n\}$

[55] with the equivalence relation $x \sim y \Leftrightarrow P(x = y) = 1$

(all linear combinations) and $M(Y_n) := \{z \in L^2 : z = f(y_1, ..., y_n)$ with f measurable$\}$

Theorem 1: If $(x, y_1, ..., y_n)$ is multivariate Gaussian, then:

$$\mathbb{E}(x|y_1, ..., y_n) = \text{Pr}_{span\{1,y_1,...,y_n\}}x$$

The proof of theorem 1 is based on the projection theorem and the principle that the error made with the best prediction is orthogonal (with respect to $< X, Y >:= \mathbb{E}(XY)$) on the space we are projecting on.

Proof:

The conditional expectation $\mathbb{E}(x|y_1, ..., y_n) := \text{Pr}_{M(Y_n)}x$ is defined as projection on the space $M(Y_n)$ that is the space of all random variables in L^2 that can be written as measurable functions of $y_1, ..., y_n$. Therefore we have due to the projection theorem that the projection is unique and the following holds:

$$\mathbb{E}((x - \mathbb{E}(x|y_1, ..., y_n))w) = 0 \qquad \forall w \in M(Y)$$

Next we show that $\hat{x} = \text{Pr}_{Y_n}x$ is that element. We have from the projection theorem:

$$\mathbb{E}((x - \hat{x})z) = 0 \qquad \forall z \in Y_n$$

Due to $y_i \in Y_n$ and the normality assumption we get that $x - \hat{x}$ is uncorrelated with y_i and therefore independent. Also $1 \in Y_n$ gives $\mathbb{E}(x - \hat{x}) = 0$. Therefore:

$$\mathbb{E}((x - \hat{x})w) = \mathbb{E}(x - \hat{x})\mathbb{E}(w) = 0 \qquad \forall w \in M(Y)$$

since $x - \hat{x}$ is independent to $f(y_1, ..., y_n)$ for all measurable functions. Due to the fact that independence is not compromised by measurable functions. □

Theorem 2: If $(y, x) \in \mathbb{R}^{n+m}$ is multivariate normal distributed. The following holds:

$$\mathbb{E}(x|y) = \mathbb{E}(x) + \mathbb{C}ov(x, y)\mathbb{V}ar(y)^{-1}(y - \mathbb{E}(y))$$
$$\mathbb{V}ar(x|y) = \mathbb{V}ar(x) - \mathbb{C}ov(x, y)\mathbb{V}ar(y)^{-1}\mathbb{C}ov(x, y)$$

Proof:

First we demean all variables involved. We use the theorem 1 and have $\mathbb{E}(x|y) = \text{Pr}_{Y_n}x = \beta' y$ and we know from the projection theorem:

$$\mathbb{E}((x - \beta' y)y') = 0 \Leftrightarrow \mathbb{E}(xy') = \beta'\mathbb{E}(yy')$$

Therefore it follows $\beta' = \mathbb{E}(xy')\mathbb{E}(yy')^{-1} = \mathbb{C}ov(x, y)\mathbb{V}ar(y)^{-1}$ since we demeaned the variables. Therefore it follows the first statement $\mathbb{E}(x|y) = \mathbb{C}ov(x, y)\mathbb{V}ar(y)^{-1}y$ for the demeaned variables. If we reverse the demeaning we get the first statement (just insert $y = y - \mathbb{E}(y)$ and $x = x - \mathbb{E}(x)$ in the last equation by slightly abusing notation).

The second statement follows due to:

$$\mathbb{V}ar(x|y) = \mathbb{V}ar(x - \mathbb{E}(x|y)|y) = \mathbb{V}ar(x - \mathbb{E}(x|y))$$

Note that the error term $x-\mathbb{E}(x|y)$ is orthogonal to the y. The normality gives the independence, so we can drop the condition. Last we insert the first expression for the conditional expectation.

$$\mathbb{V}ar(x|y) = \mathbb{V}ar(x - \mathbb{C}ov(x,y)\mathbb{V}ar(y)^{-1}y)$$

Using $\mathbb{V}ar(y-x) = \mathbb{V}ar(y) - 2\mathbb{C}ov(y,x) + \mathbb{V}ar(x)$ gives the result. □