ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Comparing Estimation Methods for Non-Gaussian Affine Term Structure Models

Econometrics and Management Science

Quantitative Finance

Master Thesis

Author:
Ruben Beurskens
476605

Supervisor:
Prof.dr. Michel van der Wel
Second assessor:
dr. Maria Grith

Abstract

This thesis investigates two methods used to estimate the yield curve using affine term structure models (ATSMs) with non-Gaussian factors. ATSMs describe the dynamics of the yield curve using affine transformations of latent factors. The factors behave according to Gaussian or non-Gaussian dynamics. The methods of Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015) both promise quick and efficient estimation of ATSMs with non-Gaussian factors, but it is unclear which model performs better. The method of Aït-Sahalia and Kimmel (2010) approximates the likelihood of the latent state variables through Hermite expansions, while the method of Creal and Wu (2015) approximates the entire ATSM by a discrete-time version. Ex ante, it is not clear which of these approximations performs better. This research performs a sensitivity analysis to the amount of starting values and observations. A comparison between the two methods is done using an efficient amount of starting values and observations, as using too much of either increases computation time without significantly increasing performance. The Creal and Wu (2015) method results in a lower root-mean-square error (RMSE). This lower RMSE is mostly noticeable in the yields corresponding to the lowest maturity. The RMSE is comparable between the two methods in the yields corresponding to the higher maturities. The lower RMSE comes at the cost of an increased computation time. An empirical estimation of the parameters using real-world data is performed for both methods, which supports the conclusion that the CW method outperforms the ASK method for real-world data.

March 20, 2020

The content of this thesis is the sole responsibility of the author and does not reflect the view of either Erasmus School of Economics or Erasmus University.

Contents

1 Introduction

The yield curve is one of the most important instruments of an economy. Current yield curve values allow for forecasting future yields, as current long-term yields carry an expectation of future short yields. Pricing of derivatives and hedging of risk also crucially depends on an understanding of what moves the yield curve (Piazzesi, 2010).

The yield curve describes the relation between yield on a treasury investment and time-to-maturity. The yield curve can be described using Gaussian factors of Vasicek (1977), or non-Gaussian factors of Cox, Ingersoll, and Ross (1985). Later, it was shown by Duffie and Kan (1996) that both approaches can be generalized in affine term structure models (ATSMs). ATSMs are able to consistently fit both the cross-section and time-series properties of the yield curve (Piazzesi, 2010).

The Vasicek model has a fixed volatility of the short rate, while the CIR model allows for varying volatility of the short rate, which is also observed empirically. Despite this, models with only Gaussian factors have been very popular with practitioners. In 2011, the US bond yields have occasionally become zero at the lower maturities. The lowest values the yield curve can attain is called the zero lower bound, which is determined by policy. As the name implies, the zero lower bound is mostly held at zero. Negative yields are a last resort of policymakers, and do not happen without special circumstances. The Gaussian models do not have a built-in protection to prevent overshooting the zero lower bound, resulting in negative yield forecasts. Given the nature of the zero lower bound as determined by policy, it is unreasonable to forecast negative yields, leading to inaccurate forecasts. This discrepancy between the reality of the yield curve and the inability of the Gaussian models has resulted in a vast body of literature which aims to deal with this overshooting of the zero lower bound. An overview of this literature can be found in Krippner (2015).

Instead of trying to modify the Gaussian ATSMs, the ATSMs with non-Gaussian factors are unable to reach negative yields. While this is great for modeling the zero lower bound when it is at zero, this also proves to be a limitation when the yield becomes negative. Negative yields are a last resort of policymakers, which rarely occurs. In the US the yields turned slightly negative in 2015, after which yields turned positive again. This is not a big problem, as it simply results in small estimation errors in short periods with negative yields. This should not drastically affect estimation outcomes, when looking at a large enough sample.

The model which includes non-Gaussian factors thus seems an interesting avenue of research, but there have been only a few extensions on this model, in contrast with the large body of research on the Gaussian models. This is likely due to the lack of a general estimation method for non-Gaussian models. Closed-form likelihood expressions are known for only a few special cases, due to the complicated dynamics of the non-Gaussian state variable in continuous time (Piazzesi, 2010). The interaction between Gaussian and non-Gaussian factors which prevents closed-form likelihoods.

Combining Gaussian and non-Gaussian factors results in mixture models in the class of ATSMs, introduced in Duffie and Kan (1996). The definition of these models is refined in Dai and Singleton (2000), where additional restrictions are given to create a 'canonical' definition of the ATSMs. These additional restrictions allow for admissable models, resulting in positive conditional variance, while imposing the minimum amount of identifying restrictions (Dai & Singleton, 2000).

Estimation of ATSMs is done by maximizing the likelihood function, which requires being able to compute the density of the latent variables. Closed-form densities can be computed in the case of Gaussian variables or the one-factor non-Gaussian model. Using multi-factor non-Gaussian models or mixture models does not allow for a closed-form density. Estimation of multi-factor non-Gaussian or mixture models can be done using quasi-maximum likelihood (Lund, 1997; De Jong, 2000), simulated maximum likelihood (Brandt & He, 2006; Piazzesi, 2005), generalized method of moments (Hansen, 1985; Singleton, 2001), and Hermite expansions (Aït-Sahalia, 2008; Aït-Sahalia & Kimmel, 2010).

Aït-Sahalia and Kimmel (2010) uses Hermite expansions aim to approximate the likelihood function. Hermite expansions allow an approximation which is similar to a Taylor expansion. This allows for quick and efficient estimation of the parameters. All estimation methods have to translate the discrete observations of the yield curve to fit into the continuous-time framework of ATSMs. Creal and Wu (2015) instead provides a discrete-time framework which is observationally equivalent to discrete observations in ATSMs. No additional assumptions are needed, when compared to ATSMs. The dynamics between subsequent observations in the discrete-time framework of Creal and Wu (2015) are identical to the dynamics between two observations in the continuous-time framework of ATSMs. This discrete-time framework allows the computation of a closed-form likelihood for mixture models, which is not possible in the continuous-time framework. Still, this closed-form likelihood leads to slow optimization which easily gets stuck in local optima. Combined with a reduction in parameter space by clever use of regressions, Creal and Wu (2015) is able to create a recipe for quick and efficient estimation.

This paper compares the methods of Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015). Both attempt to find a fast and efficient way to estimate the parameters of the yield curve, but it is impossible to know ex ante which method leads to more lower errors, this has not been studied before. This paper uses a simulation study to compare the two methods. Both methods describe a single optimization trial. A single random starting point can either obtain the global optimum or get stuck in a local optimum. In a probabilistic sense, increasing the amount of starting values allows for a larger chance to find the global optimum, but each starting value also increases computation time. Sensitivity analyses are performed to determine the sensitivity to varying the amount of random starting points and the sensitivity to varying the amount of observations to estimate from. The sensitivity analysis with respect to the amount of observations is done to determine whether the two methods benefit from a large amount of observations, or whether they also perform well on smaller datasets. These sensitivity analyses are also used to inform the amount of starting values and observations of the comparison between the two methods. This is done to prevent unnecessary computations which do not significantly increase performance. Lastly, the parameters of the two methods are estimated on empirical data.

The sensitivity analysis regarding the amount of starting points shows that both methods do not significantly benefit from using more than 60 starting points. The average computation time of the Aït-Sahalia and Kimmel (2010) method is around 9 seconds, while that of the Creal and Wu (2015) method is around 43 seconds.

The sensitivity analysis regarding the amount of observations does not provide a clear result for the Aït-Sahalia and Kimmel (2010) method. It is unclear what amount of observations performs best, or whether there is a significant difference between them at all. The sensitivity analysis of the Creal and Wu (2015) method shows that increasing the amount of observations also increases performance. The increase in performance is no longer significant by using more than 500 observations. In the final comparison, the Aït-Sahalia and Kimmel (2010) method uses 300 observations. The method of Aït-Sahalia and Kimmel (2010) needs less observations and evaluates quicker, but leads to larger errors on the lower maturities. The method of Creal and Wu (2015) needs more observations and computation time, but leads to the lower errors on the lower maturities. The performance in the higher maturities is similar. This conclusion is supported by the empirical estimation of the parameters.

The remainder of this paper is structured as follows. The theoretical framework of ATSMs and the methods of Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015) are laid out in section 2. The setup of the Monte Carlo methods is laid out in section 3, after which the results are discussed in section 4. An empirical estimation of the parameters using the Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015) methods is given in section 5. Finally, a discussion and conclusion is given in section 6.

2 Estimation Methods

The main goal of the paper is to compare the existing performance of Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015). First, the theoretical framework of risk-neutral dynamics is described in section 2.1. The physical dynamics and the general construction of the likelihood are given in section 2.2. The specifics of the Aït-Sahalia and Kimmel (2010) methodology is discussed in section 2.3 and the Creal and Wu (2015) methodology is discussed in section 2.4. Finally, the selection of which ATSM to use in estimation is described in section 2.5.

2.1 Risk-neutral dynamics

This section of the paper follows the presentation as given in Dai and Singleton (2000). Without arbitrage opportunities, the price $P_t(\tau)$ at time t of a zero-coupon bond with time to maturity τ, is given by

$$P_t(\tau) = E_t^Q \left[\exp\left(-\int_t^{t+\tau} r_s ds \right) \right], \tag{1}$$

where E^Q denotes the expectation under the risk-neutral measure and r_t is the instantaneous short rate at time t. To obtain an N-factor affine term structure model, we consider the instantaneous short rate as an affine function of the state variables, and the risk-neutral dynamics of the state variables X_t must be affine in X_t. Let us assume the instantaneous short rate r_t is an affine function of the $N \times 1$ vector of unobserved state variables X_t, such that

$$r_t = \delta_0 + \delta_1' X_t, \tag{2}$$

where δ_0 is a scalar and δ_1 is an $N \times 1$ vector. The risk neutral dynamics of the state variables X_t follow an affine diffusion, given by

$$dX_t = \tilde{\mathcal{K}}(\tilde{\Theta} - X_t)dt + \Sigma\sqrt{S_t}dW_t^Q, \tag{3}$$

where dW_t^Q is an N-dimensional independent Brownian motion under Q, $\tilde{\Theta}$ is an $N \times 1$-vector. Both $\tilde{\mathcal{K}}$ and Σ are $N \times N$ matrices, and S_t is an $N \times N$ diagonal matrix with the ith diagonal element given by $\alpha_i + \beta_i' X_t$. Each α_i is a scalar and each β_i is an $N \times 1$ vector, for $1 \leq i \leq N$. We now have a system of dynamics describing bond prices. A solution for this system is then given in Duffie and Kan (1996), who find that

$$P_t(\tau) = \exp\left[A(\tau) + B(\tau)' X_t\right], \tag{4}$$

where $A(\tau)$ and $B(\tau)$ are the scalar and $N \times 1$ solutions, respectively, to the ODEs

$$\begin{aligned} \frac{\partial A(\tau)}{\partial \tau} &= -\delta_0 + B(\tau)'\tilde{\mathcal{K}}\tilde{\Theta} + \frac{1}{2}\sum_{i=1}^{N}\left[\Sigma' B(\tau)\right]_i^2 \alpha_i, \\ \frac{\partial B(\tau)}{\partial \tau} &= -\delta_1 - \tilde{\mathcal{K}}' B(\tau) + \frac{1}{2}\sum_{i=1}^{N}\left[\Sigma' B(\tau)\right]_i^2 \beta_i, \end{aligned} \tag{5}$$

with initial conditions $A(0) = 0$ (scalar) and $B(0) = 0$ $(N \times 1)$. Here $\left[\Sigma' B(\tau)\right]_i$ indicates the i'th element of the $N \times 1$ vector $\Sigma' B(\tau)$. The yields of a zero-coupon bond are then given by

$$Y_t(\tau) = -\frac{\log[P_t(\tau)]}{\tau} = \mathcal{A}(\tau) + \mathcal{B}(\tau)' X_t, \tag{6}$$

where $\mathcal{A}(\tau) = -A(\tau)/\tau$ and $\mathcal{B}(\tau) = -B(\tau)/\tau$.

2.2 Physical dynamics and likelihood construction

So far we have only considered risk-neutral dynamics. These do not depend on the physical dynamics or the market price of risk. When estimating the models, we also need to understand the dynamics under the physical measure P. Using the simple market price of risk found in Dai and Singleton (2000), the market price of risk Λ_t is specified as

$$\Lambda_t = \sqrt{S_t}\lambda, \tag{7}$$

where λ is a $N \times 1$ vector of constants. Let us define the P-measure dynamics analogous to the Q-dynamics,

$$dX_t = \mathcal{K}(\Theta - X_t)dt + \Sigma\sqrt{S_t}dW_t^P, \tag{8}$$

where dW_t^P now represents an N-dimensional independent Brownian motion under P. The drift parameters are different under the P-measure, now defined as

$$\begin{aligned} \mathcal{K} &= \tilde{\mathcal{K}} - \Sigma\phi, \\ \Theta &= \mathcal{K}^{-1}\left(\tilde{\mathcal{K}}\tilde{\Theta} + \Sigma\psi\right), \end{aligned} \tag{9}$$

where ϕ is a $N \times N$ matrix with the i'th row given by $\lambda_i \beta_i'$ and ψ is a $N \times 1$ vector with the i'th element given by $\lambda_i \alpha_i$. This completes the dynamics of the model. The canonical $A_m(N)$ representation of Dai and Singleton (2000) contains N factors, m of which are volatility factors. To follow the specification, the normalized form is given by

$$\begin{aligned}
\Theta &= \begin{bmatrix} \Theta_{m\times 1} \\ 0_{(N-m)\times 1} \end{bmatrix}, \quad \mathcal{K} = \begin{bmatrix} \mathcal{K}_{m\times m} & 0_{m\times(N-m)} \\ \mathcal{K}_{(N-m)\times m} & \mathcal{K}_{(N-m)\times(N-m)} \end{bmatrix}, \\
\alpha &= \begin{bmatrix} 0_{m\times 1} \\ \alpha_{(N-m)\times 1} \end{bmatrix}, \quad \beta = \begin{bmatrix} I_{m\times m} & \beta_{m\times(N-m)} \\ 0_{(N-m)\times m} & 0_{(N-m)\times(N-m)} \end{bmatrix},
\end{aligned}$$

as well as Σ to be a $N \times N$ identity matrix. There are several additional parameter restrictions, given by

$$\begin{aligned}
\delta_{1,i} &\geq 0, && m+1 \leq i \leq N, \\
\sum_{j=1}^{m} \mathcal{K}_{ij}\Theta_j &> 0, && 1 \leq i \leq m, \\
\mathcal{K}_{ij} &\leq 0, && 1 \leq j \leq m, \qquad j \neq i, \\
\Theta_i &\geq 0, && 1 \leq i \leq m, \\
\beta_{ij} &\geq 0, && 1 \leq i \leq m, \qquad m+1 \leq j \leq N.
\end{aligned}$$

These restrictions are needed to ensure the process attains positive conditional variance $\sqrt{S_t}$ over all possible state variables.

The model to be estimated can now be summarized as a state space system with an observation equation, given in Equation 6, which relates the observed yields to the state vector, and a state equation, given in Equation 8, which describes the physical dynamics of the state. The observation equation allows the determination of N state vectors from N observed yields by inverting Equation 6.

When there are more yields than state variables, the observations equation can not be inverted. Inverting the observation equation can then be facilitated by allowing errors. Two alternatives are possible. Either all observations are observed with error, or a subset of the yields is observed with error. When all observations are assumed with error, Equation 6 can not be inverted to yield the state variables from the observed yields, as none of the observations is certain. In this case, a Kalman filter can be used to filter out this uncertainty and obtain the state variables (Piazzesi, 2010). Instead, this paper assumes only N yields to be observed without errors, and the remaining yields are observed with Gaussian errors. This allows the inversion of Equation 6. The time to maturity $\tau^{(1)}$ corresponds to the yields observed without error, and the time to maturity $\tau^{(2)}$ corresponds to the yields observed with error. Following the notation of Creal and Wu (2015), the observation equation can be split as

$$Y_t(\tau^{(1)}) = \mathcal{A}(\tau^{(1)}) + \mathcal{B}(\tau^{(1)})' X_t, \tag{10}$$

$$Y_t(\tau^{(2)}) = \mathcal{A}(\tau^{(2)}) + \mathcal{B}(\tau^{(2)})' X_t + \eta_t, \qquad \eta_t \sim N(0, \Omega). \tag{11}$$

The dynamics in Equation 5 are used to obtain $\mathcal{A}(\tau)$ and $\mathcal{B}(\tau)$. The N state variables are obtained from inverting Equation 10, resulting in

$$X_t = \mathcal{B}(\tau^{(1)})^{-1}\left[Y_t(\tau^{(1)}) - \mathcal{A}(\tau^{(1)})\right]. \tag{12}$$

Once the state variables are obtained, we can find the observation errors η_t of Equation 11 by

$$\eta_t = Y_t(\tau^{(2)}) - \mathcal{A}(\tau^{(2)}) - \mathcal{B}(\tau^{(2)})' X_t. \tag{13}$$

Estimating the parameters is done through maximizing the likelihood of the observed yields, by varying the parameter vector θ. The likelihood of the observed yields is given by $p(Y_{1:T}|\theta)$. This can now be expanded in the elements containing $Y_t(\tau^{(1)})$ and $Y_t(\tau^{(2)})$. As the elements of $Y_t(\tau^{(1)})$ are an affine transformation of the state variables, we can instead consider the conditional likelihood of the state variables, multiplied by the Jacobian determinant. By considering the likelihood via the prediction error decomposition, we obtain

$$\begin{aligned} p\left(Y_{1:T}|\theta\right) &= p\left(Y_{1:T}(\tau^{(2)})|Y_{1:T}(\tau^{(1)});\theta\right) p\left(Y_{1:T}(\tau^{(1)})|\theta\right) \\ &= \prod_{t=1}^{T} p\left(Y_t(\tau^{(2)})|Y_t(\tau^{(1)});\theta\right) \prod_{t=1}^{T} p\left(X_t|\mathcal{I}_t;\theta\right)|J(\theta)|^{-T}, \end{aligned} \tag{14}$$

where $\mathcal{I}_t$ denotes set of available information at time t. The estimation of both the Aït-Sahalia and Kimmel (2010) method and the Creal and Wu (2015) method roughly follows the same steps. For a given parameter vector θ, the complete likelihood is then constructed as follows:

1. For a given θ, calculate the bond loadings $\mathcal{A}$ and $\mathcal{B}$. Then, use Equation 12 to find the state variables X_t.
2. Given X_t, obtain the likelihood of the state variables $\prod_{t=1}^{T} p\left(X_t|\mathcal{I}_t;\theta\right)$. This step differs significantly between the two methods.
3. Using Equation 13, calculate $\hat{\Omega} = \frac{1}{T-1}\sum_{t=2}^{T}\eta_t\eta_t'$.
4. By multiplying the result of step 2 with the Jacobian determinant, the likelihood of the yields observed without error is obtained. Using the variance found in step 3, we can compute the likelihood of the yields observed with error. Thus the complete likelihood can now be calculated, for a given parameter vector θ.

To allow for easier computation, the log likelihood is considered. The maximization of the log likelihood is done using standard optimization techniques. The specifics of the two methods are laid out in the following sections. Aït-Sahalia and Kimmel (2010) approximate the density of the state variables using a likelihood expansion. Creal and Wu (2015) instead approximate the entire process with a discrete version, which then allows for complete analytical solutions to the transition densities.

2.3 Likelihood expansion of the state variables

The method of Aït-Sahalia and Kimmel (2010) uses Hermite expansions to approximate the density of the state variables, $\prod_{t=1}^{T} p\left(X_t|\mathcal{I}_t;\theta\right)$. This method will be referred to as the ASK method. This method is based on the expansions described in Aït-Sahalia (2008). Realizing the likelihood of the state variables is Markovian, the information set $\mathcal{I}_t$ consists of the information available at time t and contains the previous observation X_{t-1}. Suppose the time between yield observations is given by Δ, denoted in years. The log conditional likelihood of the state variables can be approximated using a likelihood expansion, which has the form of a Taylor expansion in the dimension Δ at order K,

$$l_X^{(K)}(\Delta, X_t|X_{t-1};\theta) = -\frac{N}{2}\ln(2\pi\Delta) - D_v + \frac{C_{X_t}^{(-1)}(X_t|X_{t-1};\theta)}{\Delta} + \sum_{k=0}^{K} C_{X_t}^{(k)}(X_t|X_{t-1};\theta)\frac{\Delta^k}{k!}, \tag{15}$$

with $D_v = \frac{1}{2}\ln\left(\text{Det}\left[\sigma(X_t;\theta)\sigma(X_t;\theta)'\right]\right)$. As the yield data is observed monthly, this means $\Delta = \frac{1}{12}$. The coefficients $C_{X_t}^{(k)}$ for $k = -1, 0, \ldots, K$ can often not be computed in closed form for mixture models. By performing a Taylor expansion in $(X_t - X_{t-1})$ of each coefficient $C_{X_t}^{(k)}$ at order j_k, the coefficients can be calculated in closed-form. Let such a Taylor expansion be denoted by $C_{X_t}^{(j_k,k)}$. The order $j_k = 2(k-K)$ The approximation then becomes

$$\tilde{l}_X^{(K)}(\Delta, X_t|X_{t-1};\theta) = -\frac{N}{2}\ln(2\pi\Delta) - D_v + \frac{C_{X_t}^{(j_{-1},-1)}(X_t|X_{t-1};\theta)}{\Delta} + \sum_{k=0}^{K} C_{X_t}^{(j_k,k)}(X_t|X_{t-1};\theta)\frac{\Delta^k}{k!}, \tag{16}$$

The closed-form likelihood expansions $C_{X_t}^{(j_k,k)}$ within the context of the physical dynamics of Equation 8 are provided by Aït-Sahalia and Kimmel (2010). The parameter vector θ consists of all free parameters $\{\mathcal{K}, \Theta, \alpha, \beta, \delta_0, \delta_1, \lambda, \sigma\}$. For a given parameter vector θ, the computation of the log likelihood is as follows:

1. For a given θ, the ODEs in Equation 5 are solved numerically to obtain $\mathcal{A}(\tau)$ and $\mathcal{B}(\tau)$. Then, use Equation 12 to find the state variables X_t.

2. Evaluate the joint likelihood of the state variables X_t, using the expansions of the likelihood in Equation 16.

3. Using Equation 13, calculate $\hat{\Omega} = \frac{1}{T-1}\sum_{t=2}^{T}\eta_t\eta_t'$.

4. Add together the terms from step 2 and 3, and the Jacobian determinant to find the complete data log likelihood, for a given parameter vector θ.

2.4 Discretization of the entire process

Creal and Wu (2015) consider a discrete-time process with the same dynamics as the continuous-time model we have already defined. This method will be referred to as the CW method. The state variable $X_t = (g_t', h_t')'$ consists of a $G \times 1$ vector of conditionally Gaussian state variables g_t, whose volatilities are captured by a $H \times 1$ vector of positive state variables h_t. In the $A_m(N)$ notation of Dai and Singleton (2000), this corresponds to $G = (N-m)$ and $H = m$. By splitting the state variables, the instantaneous short rate of Equation 2 now becomes

$$r_t = \delta_0 + \delta_{1,h}' h_t + \delta_{1,g}' g_t. \tag{17}$$

Under the risk-neutral measure Q, the Gaussian state variables g_t follow a vector autoregression with conditional heteroskedasticity

$$\begin{aligned} g_{t+1} &= \mu_g^Q + \Phi_g^Q g_t + \Phi_{gh}^Q h_t + \Sigma_{gh}\varepsilon_{h,t+1}^Q + \varepsilon_{g,t+1}^Q, \qquad \varepsilon_{g,t+1}^Q \overset{Q}{\sim} N(0, \Sigma_{g,t}\Sigma_{g,t}'), \qquad (18) \\ \Sigma_{g,t}\Sigma_{g,t}' &= \Sigma_{0,g}\Sigma_{0,g}' + \sum_{i=1}^{H} \Sigma_{i,g}\Sigma_{i,g}' h_{i,t}, \qquad \varepsilon_{h,t+1}^Q = h_{t+1} - \mathbb{E}^Q(h_{t+1}|\mathcal{I}_t), \end{aligned}$$

where $\mathcal{I}_t$ denotes the information set at time t. The volatility factors follow an affine transformation of the discrete-time equivalent of a multivariate Cox et al. (1985) process

$$h_{t+1} = \mu_h + \Sigma_h w_{t+1}, \tag{19}$$

$$w_{i,t+1} \sim \text{Gamma}(\nu_{h,i}^Q + z_{i,t+1}^Q, 1), \qquad i = 1, \ldots, H \tag{20}$$

$$z_{i,t+1}^Q \sim \text{Poisson}(e_i' \Sigma_h^{-1} \Phi_h^Q \Sigma_h w_t), \qquad i = 1, \ldots, H \tag{21}$$

where e_i denotes the ith column of the $H \times H$ identity matrix I_H. The Gaussian state variables g_{t+1} are a function of the non-Gaussian state variables h_t through both the autoregressive term $\Phi_{gh}^Q h_t$ and the covariance term $\Sigma_{gh}\varepsilon_{h,t+1}^Q$. A process using only non-Gaussian state variables would only require Equations 19 - 21, while a process using only Gaussian state variables would vastly simplify Equation 18, and not require Equations 19 - 21. Mixture models using both Gaussian and non-Gaussian factors require Equations 18 - 21.

The discrete-time equivalent to Equation 4, linking the price and the state variables, is given by

$$P_t(\tau) = \exp\left[A(\tau) + B_h(\tau)' h_t + B_g(\tau)' g_t\right]. \tag{22}$$

While the factor loadings in the continuous-time framework are found by solving the ODEs of Equation 5, the loadings can in the discrete-time framework are computed in closed form through the recursions

$$\begin{aligned} A(\tau) &= -\delta_0 + A(\tau-1) + \mu_g^{Q\prime} B_g(\tau-1) + \left[\mu_h + \Phi_h^Q \mu_h + \Sigma_h \nu_h^Q\right]' B_h(\tau-1) \\ &\quad + \frac{1}{2} B_g(\tau-1)\Sigma_{0,g}\Sigma_{0,g}' B_g(\tau-1) - \nu_h^{Q\prime}\left[\log(\iota_H - \Sigma_h' B_{gh}(\tau-1)) + \Sigma_h' B_{gh}(\tau-1)\right] \\ &\quad + \mu_h' \Phi_h^{Q\prime} \Sigma_h^{-1\prime} \left(I_H - \left[\text{diag}(\iota_H - \Sigma_h' B_{gh})\right]^{-1}\right) \Sigma_h' B_{gh}(\tau-1), \qquad (23) \\ B_h(\tau) &= -\delta_{1,h} + \Phi_{gh}^{Q\prime} B_g(\tau-1) + \Phi_h^{Q\prime} B_h(\tau-1) \\ &\quad + \frac{1}{2}(I_H \otimes B_g(\tau-1)')\Sigma_g\Sigma_g'(I_H \otimes B_g(\tau-1)) \\ &\quad - \Phi_h^{Q\prime}\Sigma_h^{-1\prime}\left(I_H - \left[\text{diag}(\iota_h - \Sigma_h' B_{gh}(\tau-1))\right]^{-1}\right)\Sigma_h' B_{gh}(\tau-1), \qquad (24) \\ B_g(\tau) &= \delta_{1,g} + \Phi_g^{Q\prime} B_g(\tau-1), \qquad (25) \end{aligned}$$

with initial values $A(0) = 0$, $B_h(0) = 0$, and $B_g(0) = 0$. The matrix $\Sigma_g \Sigma_g'$ is a $(G \times H) \times (G \times H)$ block diagonal matrix with elements $\Sigma_{i,g} \Sigma_{i,g}'$ for $i = 1, \ldots, H$ and $B_{gh}(\tau - 1) = \Sigma_{gh}' B_g(\tau - 1) + B_h(\tau - 1)$. The bond yields can then be expressed as

$$Y_t(\tau) = \mathcal{A}(\tau) + \mathcal{B}_h(\tau)' h_t + \mathcal{B}_g(\tau)' g_t = \mathcal{A}(\tau) + \mathcal{B}(\tau)' X_t, \tag{26}$$

where $\mathcal{A}(\tau) = A(\tau)/\tau$, $\mathcal{B}_h(\tau) = B_h(\tau)/\tau$, and $\mathcal{B}_g(\tau) = B_g(\tau)/\tau$. The final equality holds if $\mathcal{B}(\tau)' = (\mathcal{B}_h(\tau)', \mathcal{B}_g(\tau)')$, this is needed to invert the relation of Equation 12.

This concludes the discrete-time analogue of the continuous-time dynamics under Q laid out in section 2.1. Similar to the continuous-time model, the discrete-time analogue needs to take into account the P-dynamics to arrive at a model which can be estimated from data.

In the continuous-time model the P-dynamics have the same functional form as under Q, though with distinct parameter values. In the discrete-time model, this also holds. Under the P-measure the dynamics have the same functional form as under Q,

$$g_{t+1} = \mu_g + \Phi_g g_t + \Phi_{gh} h_t + \Sigma_{gh} \varepsilon_{h,t+1} + \varepsilon_{g,t+1}, \qquad \varepsilon_{g,t+1} \sim N(0, \Sigma_{g,t}, \Sigma_{g,t}'), \tag{27}$$

$$\Sigma_{g,t} \Sigma_{g,t}' = \Sigma_{0,g} \Sigma_{0,g}' + \sum_{i=1}^{H} \Sigma_{i,g} \Sigma_{i,g}' h_{i,t}, \qquad \varepsilon_{h,t+1} = h_{t+1} - \mathbb{E}(h_{t+1} | \mathcal{I}_t),$$

The parameters controlling the conditional mean are different between P and Q, while the scale parameters Σ_{gh} and $\Sigma_{i,g}$ for $i = 1, \ldots, H$ are the same. The dynamics of the volatility factors are given by

$$h_{t+1} = \mu_h + \Sigma_h w_{t+1}, \tag{28}$$

$$w_{i,t+1} \sim \text{Gamma}(\nu_{h,i} + z_{i,t+1}, 1), \qquad i = 1, \ldots, H \tag{29}$$

$$z_{i,t+1} \sim \text{Poisson}(e_i' \Sigma_h^{-1} \Phi_h \Sigma_h w_t), \qquad i = 1, \ldots, H \tag{30}$$

The splitting of the observations in those with and without errors remains the same as with.

Creal and Wu (2015) notice the parameter vector θ can be simplified in those parameters that enter the bond loadings and those that do not. This fact can be used to reduce the amount of free parameters to be estimated. Thus, given the parameters that enter the bond loadings of Equations 23 - 25, the factors can be extracted from Equation 12. The parameters of the P-dynamics can now be extracted by running generalized least squares (GLS) in the form of

$$g_{t+1} - \Sigma_{gh} \varepsilon_{h,t+1} = \mu_g + \Phi_g g_t + \Phi_{gh} h_t + \Sigma_{g,t} \varepsilon_{g,t+1}. \tag{31}$$

The parameters μ_g, Φ_g, Φ_{gh} are now obtained. By using Equation 11 and 13, Ω is calculated as $\hat{\Omega} = \frac{1}{T-1} \sum_{t=2}^{T} \eta_t \eta_t'$. Estimating these these parameters through GLS allows for a reduced dimensionality of the parameter space, making optimization easier. The parameter vector θ only contains the free parameters $\{\delta_0, \Phi_g^Q, \Phi_h, \Phi_h^Q, \Sigma_{0,g}, \Sigma_{i,g}, \Sigma_{gh}, \Sigma_h, \nu_h, \nu_h^Q\}$. The method of evaluating the log likelihood for a given parameter vector θ is then as follows:

1. For a given θ, calculate the values for the bond loadings through the recursions in Equations 23 - 25. Then, use Equation 12 to find the state variables g_t and h_t.

2. Given g_t and h_t, run the GLS regression of Equation 31 to obtain $\hat{\mu}_g, \hat{\Phi}_g$ and $\hat{\Phi}_{gh}$.

3. Using Equation 11 and 13, calculate $\hat{\Omega} = \frac{1}{T-1}\sum_{t=2}^{T} \eta_t \eta_t'$.

4. Use the estimates of step 2 and 3 in the complete log likelihood function.

The strength of the method lies in the ability to create a concentrated likelihood, which contains less parameters. By creating a reduced parameter space, the optimization technique converges faster and has less numerical instability. Several restrictions are needed for identification. $\delta_{1,h}$ is a $H \times 1$ vector of ones, $\delta_{1,g}$ is a $G \times 1$ vector of ones. μ_g^Q is a $G \times 1$ vector of zeros, μ_h is a $H \times 1$ vector of zeros. $\nu_{h,i} > 1$ and $\nu_{h,i}^Q > 1$ for $i = 1, \ldots, H$. $\Phi_{gh}^Q = 0$. The matrices $\Sigma_h, \Sigma_h^{-1}\Phi_h\Sigma_h$, and $\Sigma_h^{-1}\Phi_h^Q\Sigma_h$ must be positive and Σ_h is diagonal.

2.5 Model selection

Following the Dai and Singleton (2000) notation, the $A_m(N)$ model contains N factors, m of which are volatility factors. Dai and Singleton (2000) note that the $A_1(3)$ model seems to fit the real-world data better than the $A_2(3)$ model. This is one of the main reasons Creal and Wu (2015) conclude the $A_1(3)$ model is the benchmark non-Gaussian ATSM. This paper follows this recommendation, and all data is simulated as such. Thus this paper uses the $A_1(3)$ model. This model contains 3 factors, with 1 of those factors being a volatility factor.

3 Monte Carlo Set-up

This section outlines the methods used to evaluate the performance of the ASK and CW estimation methods. The general setup of the simulated data is given in section 3.1, after which the parameters used for the simulated data using the ASK and CW methods are discussed in sections 3.1.1 and 3.1.2, respectively. The evaluation method of the estimation methods is discussed in section 3.2.

This paper first performs the sensitivity analyses, before arriving at the final comparison. This is mostly due to the computation time needed to optimize the models. Lowering the amount of trials or the amount of observations can significantly reduce the computational burden of any following steps. The sensitivity analyses to determine the right amount of starting values reduced computation time linearly, and is therefore performed first. The sensitivity analysis regarding the amount of observations, does not reduce the computation time linearly, and is thus performed second. The sensitivity analyses are discussed in section 3.3. Finally, the method of comparing the two methods, using the resulting amount of trials and observations, is outlined in section 3.4.

3.1 Simulated data and parameters

The simulated data data is generated through the continuous-time framework of Aït-Sahalia and Kimmel (2010), discussed in section 2.2, using an Euler discretization, and the discrete-time framework of Creal and Wu (2015), discussed in section 2.4. The data simulated through the continuous-time framework of Aït-Sahalia and Kimmel (2010) will be referred to as the ASK data, while the data simulated through the discrete-time framework of Creal and Wu (2015) will be referred to as the CW data. The simulated datasets contain $T = 800$ monthly in-sample observations. For both methods, first the underlying state process is simulated, after which the state process is transformed into the yield process. To simulate the state variables in the continuous-time framework for a given parameter vector θ, an Euler discretization of Equation 3 is used. The Euler discretization uses 30 intervals per month. Of these 30 steps, 29 are discarded, leaving only the observations at the monthly frequency. After simulating the state variables, they are converted into a yield curve.

Each dataset contains the yields for the 1, 3, 12, 24, 36, 48, and 60 month maturities, similar to the datasets used by Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015). As there are three state variables, three of the yields are assumed to be observed without errors, following Equation 10, corresponding to the maturities in months $\tau^{(1)} = \{1, 12, 60\}$. The selection of these maturities is in line with those in both Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015). The maturities observed without error correspond to the shortest maturity, longest maturity and a maturity in the middle which is skewed towards the shorter end, as the shorter maturities experience more volatility, and thus could contain more information. The remaining maturities $\tau = \{3, 24, 36, 48\}$ are observed with error, following Equation 11.

A single dataset thus consists of 800 observations of 7 maturities. To obtain a distribution of results, 100 datasets are simulated using the same parameter vector θ. The values of the parameters used to simulate the data are displayed in section 3.1.1. The resulting data follows the stylized facts of the yield curve.

The simulated data following the discrete-time framework is similar in construction, but does not need an Euler discretization, as the dynamics of the state variables are already given in discrete steps in Equations 27 and 28. These are used to simulate $T = 800$ observations of 3 state variables. Construction of the yields from the state variables is the same as before. Similarly, 100 datasets are simulated to allow for a distribution of results using the same parameter vector θ. The values of the parameters used to simulate the data are given in section 3.1.2. These parameter values are obtained from the code of Creal and Wu (2015). The resulting simulated data follows the stylized facts of the yield curve.

3.1.1 ASK data

The parameter values used to simulate the data in the continuous-time framework are inspired by those provided in Aït-Sahalia and Kimmel (2010), which in turn have been inspired by Cheridito, Filipović, and Kimmel (2010). The simulated data of Aït-Sahalia and Kimmel (2010) however fixes several off-diagonal elements of $\mathcal{K}$ at 0, as well as the free elements of β fixed at 0 to allow the computation of the exact density of the dataset. Knowing the exact density is not necessary for the methods of this paper. Instead of fixing these values at 0, these parameters have been assigned reasonable values, allowing for more complex behaviour of the dataset. The selected values have been found by trial and error, but the resulting dataset has been tested to follow the stylized facts of the yield curve.

It should be noted the dataset is very sensitive to changing these parameters, as only small changes can lead to drastically different shapes of the yield curve, which do not follow the stylized facts of the yield curve. Trying to estimate the parameter values from the real-world dataset and using these parameters to simulate data, does not yield data which behaves according to the stylized facts of the yield curve.

The parameters of Aït-Sahalia and Kimmel (2010), with minor adjustments used in this paper, are given by

$$
\begin{aligned}
&\alpha = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} 1 & 0.4 & 0.1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{K} = \begin{bmatrix} 0.5 & 0 & 0 \\ -0.8 & 0.04 & 0.007 \\ -1.3 & -0.9 & 1.8 \end{bmatrix}, \\
&\Theta = \begin{bmatrix} 0.01 \\ 0 \\ 0 \end{bmatrix}, \quad \lambda = \begin{bmatrix} -0.1 \\ -0.25 \\ -0.35 \end{bmatrix}, \quad \delta_1 = \begin{bmatrix} 0.002 \\ 0.005 \\ 0.001 \end{bmatrix}, \\
&\Omega = \begin{bmatrix} 3\text{e}{-08} & -1\text{e}{-09} & -3\text{e}{-09} & -1.5\text{e}{-09} \\ -1\text{e}{-09} & 1\text{e}{-08} & 6\text{e}{-09} & 4\text{e}{-09} \\ -3\text{e}{-09} & 6\text{e}{-09} & 1\text{e}{-08} & 6\text{e}{-09} \\ -1.5\text{e}{-09} & 4\text{e}{-09} & 6\text{e}{-09} & 7.5\text{e}{-09} \end{bmatrix}, \\
&\delta_0 = 0.001.
\end{aligned}
$$

3.1.2 CW data

The parameter values used to simulate the data in the discrete-time framework are provided in the code of Creal and Wu (2015). These parameters have been estimated on a dataset of zero coupon bond yields between June 1952 and June 2012. The parameters can be split in three categories; free parameters, GLS parameters, and fixed parameters. The free parameters are those that are directly manipulated by the optimization procedure. The GLS parameters depend on the fixed and free parameters, but are not allowed to vary freely. The value of the GLS parameters is dictated by the values of the other paramaters during the optimization procedure. The fixed parameters are fixed with given values throughout the entire optimization procedure to allow for econometric identification.

The free parameters are given by

$$
\begin{aligned}
&\nu_h = 1.934, && \nu_h^Q = 2.637, && \delta_0 = -0.001, \\
&\Phi_g^Q = \begin{bmatrix} 0.951 & 0 \\ 0 & 0.536 \end{bmatrix}, && \Phi_h = 0.994, && \Phi_h^Q = 0.996, \\
&\Sigma_{0,g} = \begin{bmatrix} -3.07\text{e}{-10} & 0 \\ 3.36\text{e}{-10} & 1.65\text{e}{-12} \end{bmatrix}, && \Sigma_{i,g} = \begin{bmatrix} 0.006 & 0 \\ -0.004 & 0.005 \end{bmatrix}, && \Sigma_{gh} = \begin{bmatrix} -0.893 \\ 0.054 \end{bmatrix}, \\
&\Sigma_h = 1.55\text{e}{-05}.
\end{aligned}
$$

The GLS parameters are given by

$$\Phi_g = \begin{bmatrix} 0.985 & 0.0657 \\ -0.073 & 0.643 \end{bmatrix}, \quad \Phi_{gh} = \begin{bmatrix} 0.008 \\ -0.041 \end{bmatrix},$$

$$\mu_g = \begin{bmatrix} -1.34\text{e}{-05} \\ 3.32\text{e}{-05} \end{bmatrix}, \quad \Omega = \begin{bmatrix} 3\text{e}{-08} & -1\text{e}{-09} & -3\text{e}{-09} & -1.5\text{e}{-09} \\ -1\text{e}{-09} & 1\text{e}{-08} & 6\text{e}{-09} & 4\text{e}{-09} \\ -3\text{e}{-09} & 6\text{e}{-09} & 1\text{e}{-08} & 6\text{e}{-09} \\ -1.5\text{e}{-09} & 4\text{e}{-09} & 6\text{e}{-09} & 7.5\text{e}{-09} \end{bmatrix}.$$

The fixed parameters are given by

$$\Phi_g^Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Phi_{gh}^Q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Phi_h = 1, \quad \Phi_h^Q = 1,$$

$$\mu_g^Q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mu_h = 0, \quad \delta_{1,g} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \delta_{1,h} = 1.$$

3.2 Evaluating performance through RMSE

Evaluation of the estimation methods has some roadblocks. As there is no closed-form method to calculate the likelihood of the data, it is impossible to compare the likelihood resulting from the estimation methods to any true likelihood of the data. Because of the complex interactions between the parameters, it is possible that multiple parameter values result in similar performance of the evaluation methods. Evaluating model performance by the ability to find the parameters used in the data-generating process are thus not precise.

Both methods contain an error term, given in Equation 11. To evaluate performance, this error term can be evaluated. Using, for example, $T = 800$ observations, this results in 3200 error observations, 800 for each maturity observed without error. As discussed in section 3.1, the maturities observed with error are given by $\tau = \{3, 24, 36, 48\}$. To summarize these error observations into a single number for easy comparison, the root-mean-squared error (RMSE) is computed. For most of the results, a total RMSE is used, resulting in a single result computed over all 3200 error observations, equally weighing each maturity. In the sensitivity analyses, the total RMSE is used. In the final comparison between the two methods, the RMSE is split for each maturity, resulting in four numbers corresponding to a single result.

Using the error term in-sample would not be desirable to compare performance. For example, during the sensitivity analysis with respect to the amount of observations needed, the model trained on 12 observations is evaluated only on those 12 observations, while a model trained on 800 observations is evaluated over those 800 observations. The performance evaluation of both methods should be done over the same data, to allow for a fair comparison. Using the same in-sample amount of data becomes a problem, as the smallest amount of observations in the sensitivity only uses 6 observations, which is not enough to perform a thorough observation.

Thus, the estimation methods should be evaluated on the same out-of-sample dataset. Out-of-sample evaluation can be done for multiple time horizons, using 24, 120, or 600 observations. These represent short-, medium- and long-term goals of fitting the parameters to the model.

In graphs all out-of-sample horizons can be easily shown together. In tables, combining all out-of-sample horizons would clutter the results. Only the tables using 120 out-of-sample observations are placed in the main text, while the results of using 24 or 600 out-of-sample observations are placed in the Appendix B.

The simulated data then takes the largest amount of in-sample data, 800 observations, and adds the largest amount of out-of-sample data, 600 observations, resulting in a single dataset containing $T = 1400$ observations. The in-sample and out-of-sample data is always connected, to ensure the data do not represent vastly differing periods.

3.3 Sensitivity analysis and Wilcoxon rank sum test

A single random starting point of θ does not guarantee the global optimum is reached. Multiple random starting values are used to allow a larger chance of getting close to the global optimum, this amount of random starting values is referred to as the amount of trials. The optimum amount of trials needed is determined through a sensitivity analysis. In this case, optimum refers to a point where adding more trials does not significantly increase estimation performance, while it does increase computation time. In this sensitivity analysis the same datasets are estimated using $\{10, 20, 40, 60, 80, 100\}$ trials. In the sensitivity analysis regarding the amount of trails, each dataset uses 800 observations. Intuitively, this is because the largest amount of observations allows the most accurate parameter estimates. To allow for a distribution of results, the sensitivity analysis uses 100 datasets. A single trial then simply estimates the given dataset with a new starting value of the parameter vector θ.

Of the total amount of trials, the one with the highest likelihood is the best observation. The remaining trials with sub-optimal likelihoods are discarded. With 100 simulated datasets, this results in 100 parameter vector observations. The RMSE is computed over the 24, 120, and 600 out-of-sample observations. For the sensitivity analyses, only the total RMSE is used for evaluation. Ex ante, it is impossible to know which random starting value performs well. In a probabilistic sense, each trial has a chance to reach the global optimum, and thus increasing the amount of trials increases the chance of finding the optimum, which would result in the lowest RMSE. To find the optimal amount of trials, a balance needs to be found between the lowest RMSE and the lowest computation time. Increasing the amount of trials linearly increases the computation time.

To compare whether two distributions of RMSE's (for two different amounts of trials) significantly differ, a two-sided wilcoxon rank sum test is performed between all combinations of amounts of trials. The Wilcoxon rank sum test is a nonparametric test for unequal locations of the distributions between two samples. Combined with visually inspecting boxplots of the distribution of RMSE, this allows identifying the optimal amount of trials while distinguishing between non-significant differences.

The amount of observations used to estimate the parameters is fixed in most practical cases, as the largest amount of available data is used. In the case of Aït-Sahalia and Kimmel (2010) the Monte Carlo data used to evaluate performance consists of $T = 500$ observations, while the empirical estimation is performed on $T = 372$ observations. Creal and Wu (2015) use $T = 721$ observations for the empirical estimation. In both estimation methods, it is unclear how the estimation methods perform if they are employed with more or less observations. To determine the behaviour in case of varying amounts of data, a sensitivity analysis is performed. Does increasing the amount of trials increase performance, or will performance stay the same? Do the methods perform well with less data, or does performance drop when there are less observations? In the context of efficient estimation, it is interesting to see whether an optimum can be found, similar to the amount of trials, where performance can no longer significantly increase by adding more observations.

The sensitivity analysis regarding the amount of observations is performed with the amount of trials which have been determined as optimal. The amount of observations within each dataset is given by $T = \{6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800\}$. The computation time does not increase linearly with the amount of observations, such that deciding an optimal amount of observations is less straight-forward. The evaluation again uses the total RMSE over all maturities, computed by fitting the estimated parameter vector $\hat{\theta}$ over 24, 120, and 600 out-of-sample observations. The two-sided Wilcoxon rank sum test is used to evaluate differences in the location of the distribution of RMSE's.

3.4 Performance Comparison

After the right amount of trials and observations has been determined, the two methods are compared. Up until this point, the ASK method has been applied to the simulated data of the continuous-time framework, while the CW method has been applied to the discrete-time framework simulated data. The parameter values used within each framework have no clear counterparts in the other framework, resulting in data which behaves vastly different. Both datasets follow the stylized facts of the yield curve, but the steepness, volatility, and absolute size of the datasets differs. Thus it would not provide much insight to compare the results of the sensitivity analyses between the methods.

To allow a fair comparison, the continuous-time data is estimated using both methods, and these results can be compared. Similarly, the discrete-time data is estimated using both methods, allowing a comparison between the methods. In this final comparison both the total RMSE, computed over all maturities of errors, and the RMSE's computed over each maturity are displayed. Decomposing the RMSE into the maturities allows for a more granular look at the performance of the models.

4 Monte Carlo Results

In this section, the results of the Monte Carlo methods are discussed. The results of the sensitivity analysis with regards to the amount of trials and the amount of observations are discussed in sections 4.1 and 4.2, respectively. The result of the comparison between the two estimation methods are discussed in section 4.3.

4.1 Amount of trials

To determine the amount of trials needed for each estimation method, each simulated dataset is estimated with either $\{10, 20, 40, 60, 80, 100\}$ trials. For now, 800 observations are used in all datasets. Of the total amount of trials, the one with the highest likelihood will be the best observation. With 100 simulated datasets, this results in 100 best observations. The RMSE of these observations is then used to evaluate performance. In Figure 1.A the distributions of RMSE's are displayed of estimating 100 datasets using the ASK method with 10, 20, 40, 60, 80, or 100 trials. The RMSE is evaluated over 24, 120, or 600 out-of-sample observations. The distribution of RMSE's using the CW method are displayed in Figure 1.B. It should be stressed that the numerical values of the RMSE can not be compared between methods. The RMSE should only be compared between the varying number of trials within a given method. This is due to the difference in simulated data used between the two methods.

It is immediately clear that increasing the amount of trials lowers the RMSE of the best observations. Ex ante, it is unknown whether a trial with a random starting value will reach a low RMSE, or whether the optimization gets stuck in a local minimum.[1] Thus, increasing the amount of trials allows for more opportunities to reach the optimum. By increasing the amount of trials, the distribution of RMSE's becomes more compact and shrinks towards zero. However, increasing the amount of trials linearly increases the computation time.

The computation times of a single trial for both methods is shown in Figure 2. The average ASK trial, displayed in Figure 2.A, takes around 9 seconds to compute. The average CW trial, displayed in Figure 2.B takes around 43 seconds. Both methods show trials which take only fractions of a second, these trials cancel as the parameters do not follow the parameter restrictions. There are only a few of these invalid attempts using the ASK method, indicating a relatively simple parameter space. A larger fraction of trials cancels using the CW method, indicating a more complex parameter space.

To ensure no time is spent on slightly lowering the distribution of RMSE's, an optimum needs to be found between lowering the distribution of RMSE's and the computation time. The results of the Wilcoxon rank sum test on the ASK data, evaluated with 120 out-of-sample observations, are displayed in Table 1.A. Additional results of performing the Wilcoxon rank sum test with 24 and 600 out-of-sample observations can be found in Appendix B. The results for all amounts of out-of-sample observations are similar. To illustrate, consider the first row. Starting at 10 trials, the distribution of RMSE's can be significantly improved at the 5% confidence level by increasing the amount of trials to 20, 40, 60, 80, or 100 trials. The same holds for the second and third row, using 20 or 40 trials, respectively. Any higher amount of trials is a significant improvement. In the fourth row, using 60 trials, there is no significant improvement in the distribution of results by moving towards 80 or 100 trials. Using 60 trials appears to be the optimum between computation time and performance. The following results concerning the ASK method will all use 60 trials.

The results of performing the Wilcoxon rank sum test on the distributions of RMSE using 120

[1] If starting the optimization from the data-generating process (DGP) values, instead of random starting values, the CW method remains close to the DGP values, with some small variance because of the random nature of the data. Roughly 2 in 3 trials of the ASK method remains close to the DGP values, while 1 in 3 finds another optimum with vastly different parameters, though with similar likelihood and RMSE.

out-of-sample observations are displayed in Table 1.B. Additional results from using 24 and 600 out-of-sample observations can be found in Appendix B. The results show only slight differences, but the conclusion is the same for all amounts of out-of-sample observations. Improvement can again be found by increasing the amount of trials. Increasing the amount of trials from 40 to 60 is does not lead to a significant improvement with 120 out-of-sample observations, but it does for 24 and 600 observations. Increasing the amount of trials beyond 60 is non-significant for any amount of observations. The optimum between performance and computation time again appears to be found by using 60 trials. In the following results, the CW method will be applied with 60 trials.

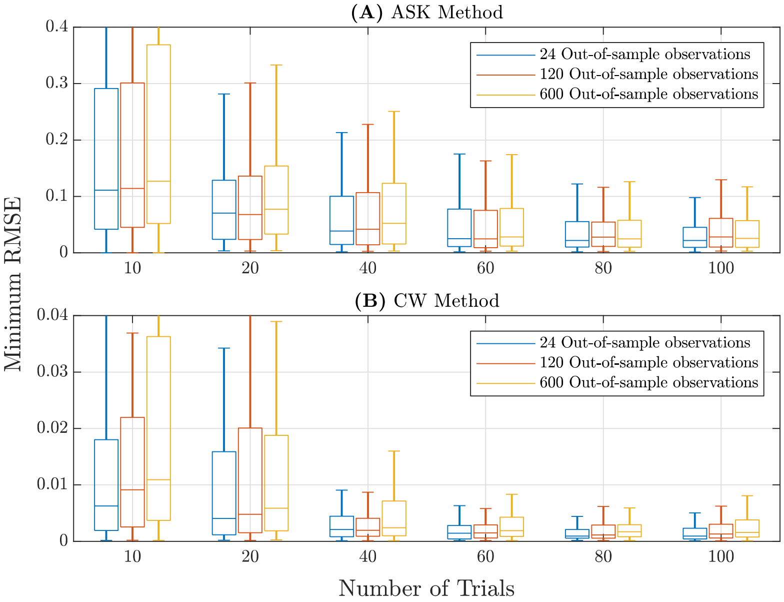

Figure 1: Boxplots of RMSE obtained by optimizing with varying amounts of trials. These boxplots show the resulting minimum RMSE of optimizing the simulated data with the ASK method and CW method. The boxplots show the results of optimizing using 10, 20, 40, 60, 80, and 100 trials, with the RMSE evaluated using 24, 120, and 600 out-of-sample observations. Outliers are omitted for visual clarity. See Section 4.1 for a discussion of the results.

Table 1: Wilcoxon results for different amounts of trials, 120 observations

(A) ASK method

# Trials	20	40	60	80	100
10	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
20		**0.042**	**0.000**	**0.000**	**0.000**
40			**0.038**	**0.011**	**0.010**
60				0.856	0.836
80					0.911

(B) CW method

# Trials	20	40	60	80	100
10	0.073	**0.000**	**0.000**	**0.000**	**0.000**
20		**0.000**	**0.000**	**0.000**	**0.000**
40			0.072	**0.007**	**0.023**
60				0.490	0.767
80					0.675

The tables display the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using either the ASK method or the CW method, evaluated over 120 out-of-sample observations. The RMSE's of optimizing using either 10, 20, 40, 60, 80, or 100 trials are compared. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

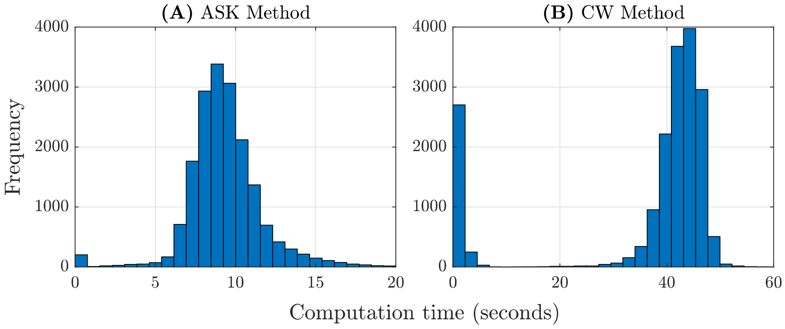

Figure 2: Histograms of computation time per trial for both estimation methods.
The computation time for both methods is displayed, registered over the complete sensitivity analysis with respect to the amount of trials. The computation times are registered with both methods using 800 observations per dataset.

4.2 Amount of observations

To determine the amount of observations to use in the final estimation, each of the trials is run with datasets consisting of 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, or 800 observations. The expected result is to find a reduced RMSE for a larger amount of observations. This would be due to the increased amount of observations allowing the algorithm to determine more precise parameter estimates.

The distributions of RMSE using varying amounts of observations are displayed in Figure 3. Again, it should be stressed the numerical values can not be compared between the two methods, due to the two methods using different datasets. Only a comparison within a method, using different amounts of observations is possible. The results of performing the sensitivity analysis using the ASK method are displayed in Figure 3.A. Comparing the distributions of RMSE, an increase in observations does not seem to lower the RMSE, there is no clear result from increasing the amount of observations. The results of performing the sensitivity analysis for the amount of observations on the CW method are displayed in in Figure 3.B. It is clear that increasing the amount of observations, generally leads to a lower RMSE.

The results of the Wilcoxon rank sum test on the RMSE of 120 out-of-sample observations of the ASK method are displayed in Table 2. The results of using 24 and 600 observations can be found in Appendix B. There is no clear story to be found. No single amount of observations results in a distribution of RMSE which significantly differs from all others on all evaluation amounts. The results of performing a Wilcoxon rank sum test using 120 out-of-sample observations of the CW method are displayed in Table 3. The results using 24 and 600 out-of-sample observations can be found in Appendix B. Using 24 out-of-sample observations, there seems to be little significant improvement beyond using 200 observations. Using 120 out-of-sample observations, there is no clear point after which there is no more significant improvement. Using 600 out-of-sample observations, the longest evaluation period, shows that there is no significant improvement beyond using 500 observations. To determine the right amount of observations to use in the final comparison, the computation time needs to be considered.

The computation times using each amount of observations are displayed in Figure 4, with the ASK method and CW method displayed in Figures 4.A and 4.B, respectively. For the ASK method, it is clear that increasing the amount of observations does not strongly influence the computation time. Using 6 observations, the average trial lasts 6 seconds, while using 800 observations still uses less than 10 seconds. Though the difference is significant, both are short enough that they should not influence the decision on which amount of observations to proceed with. For the CW method, the computation is more dependent on the amount of observations. Using 6 observations, the average trial takes 5 seconds, while using 800 observations it takes 43 seconds.

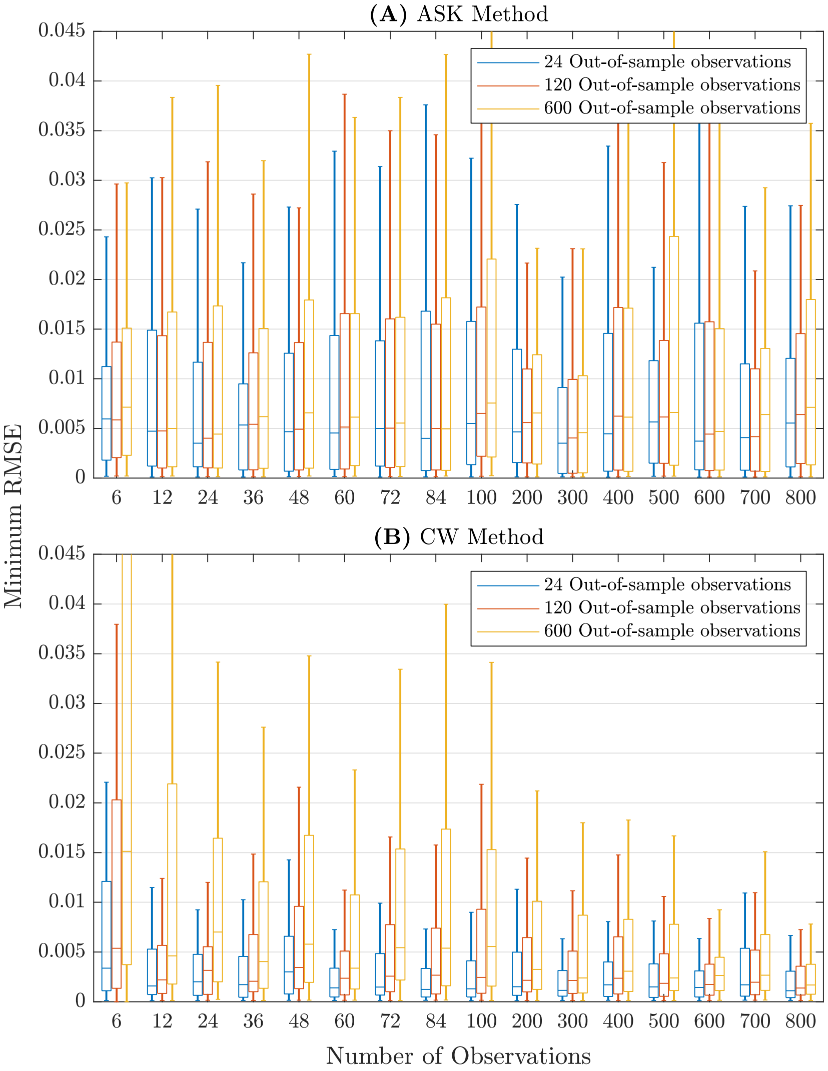

Figure 3: Boxplots of optimizing using with varying amounts of observations.
These boxplots show the resulting minimum RMSE of optimizing the simulated data with the CW method. The boxplots show the results of optimizing using datasets with lengths of 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, and 800 observations, with the RMSE evaluated using 24, 120, and 600 out-of-sample observations. Outliers are omitted for visual clarity. See Section 4.2 for a discussion of the results.

To determine an amount of observations to use for the final comparison, both performance and computation time have to be considered. For the ASK method, looking only at performance, using 300 observations seems to contain not only the lowest medians, but also the lowest 75th percentile, implying a denser and lower distribution of results. This can be due to three possibilities. The first possibility for this is that using 300 observations actually provides a better result than any other amount, in which case the right choice has been made. The second possibility for the apparently denser distribution of 300 observations, is due to pure luck, while the actual performance is no better or worse than the other amounts of observation. In that case, it does not hurt to proceed with 300 observations. In the third case, the performance is actually worse than the other amounts of observations, but due to bad luck it appears to be better than the other choices. This result is unlikely, and shall not be given much weight. Thus it is fine to proceed with 300 observations for the rest of the comparison, with an average trial taking 8 seconds to compute.

For the CW method, it appears that using more than 500 observations does not increase performance, while it does increase computation time. Using 500 observations takes on average 28 seconds per trial. Using less than 500 observations lowers both computation time and performance, though the computation time does not scale linearly. It makes sense to prefer performance over speed, and thus the CW method will use 500 observations in the final comparison.

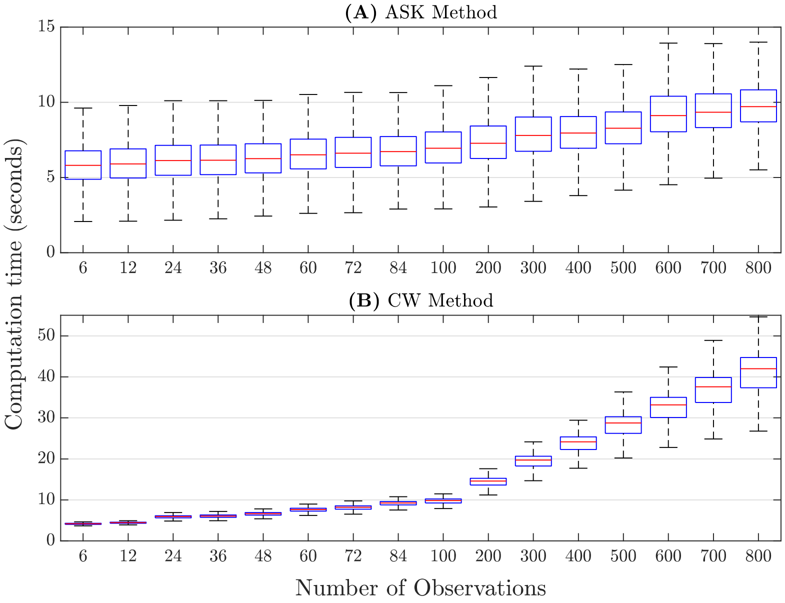

Figure 4: Boxplots of computation time per trial with varying amounts of observations.
The computation times for the ASK method and CW method are displayed, registered over the sensitivity analysis with respect to the amount of observations. The computation time is displayed in seconds per trial. Note the vastly different scales on the y-axis. Outliers are omitted for visual clarity. See Section 4.2 for a discussion of the results

Table 2: Wilcoxon results for ASK method with different amounts of training observations, 120 out-of-sample observations

# Trials	12	24	36	48	60	72	84	100	200	300	400	500	600	700	800
6	**0.023**	**0.035**	0.070	0.127	0.539	0.359	**0.012**	0.252	0.376	0.170	0.622	0.956	0.296	0.215	0.737
12		0.865	0.768	0.458	0.291	0.282	0.755	**0.005**	0.341	0.750	0.153	0.077	0.279	0.562	**0.030**
24			0.735	0.528	0.275	0.304	0.695	**0.006**	0.399	0.782	0.194	0.087	0.336	0.558	**0.030**
36				0.717	0.331	0.407	0.478	**0.007**	0.396	0.778	0.232	0.103	0.490	0.659	**0.034**
48					0.688	0.717	0.318	**0.037**	0.836	0.763	0.539	0.279	0.783	0.929	0.135
60						0.852	0.123	0.177	0.741	0.415	0.825	0.539	0.737	0.534	0.316
72							0.152	0.071	0.948	0.577	0.666	0.442	0.900	0.681	0.199
84								**0.001**	0.146	0.364	0.073	**0.031**	0.184	0.309	**0.010**
100									0.075	**0.021**	0.248	0.435	0.065	**0.039**	0.733
200										0.567	0.550	0.378	0.962	0.670	0.190
300											0.235	0.160	0.631	0.894	0.062
400												0.780	0.647	0.422	0.421
500													0.346	0.270	0.552
600														0.768	0.181
700															0.094

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the ASK method, with either 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800 observations, evaluated over 120 out-of-sample observations. The values reported are the *p*-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low *p*-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table 3: Wilcoxon results for CW method with different amounts of training observations, 120 out-of-sample observations

# Trials	12	24	36	48	60	72	84	100	200	300	400	500	600	700	800
6	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
12		0.297	0.083	0.339	**0.013**	0.083	**0.004**	**0.025**	**0.010**	**0.014**	0.328	0.119	0.204	0.563	0.179
24			0.392	0.910	0.123	0.385	**0.050**	0.182	0.091	0.138	0.908	0.631	0.761	0.675	0.848
36				0.380	0.481	0.886	0.270	0.787	0.492	0.650	0.399	0.856	0.575	0.250	0.555
48					0.131	0.354	**0.044**	0.190	0.074	0.134	0.991	0.470	0.659	0.782	0.739
60						0.412	0.649	0.755	0.983	0.902	0.102	0.349	0.210	0.052	0.191
72							0.203	0.672	0.346	0.537	0.399	0.913	0.628	0.285	0.605
84								0.389	0.661	0.487	**0.040**	0.145	0.087	**0.022**	**0.090**
100									0.642	0.838	0.193	0.534	0.321	0.104	0.292
200										0.804	0.072	0.296	0.158	**0.041**	0.171
300											0.129	0.396	0.235	0.060	0.212
400												0.461	0.730	0.719	0.717
500													0.774	0.372	0.802
600														0.512	0.991
700															0.445

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800 observations, evaluated over 120 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

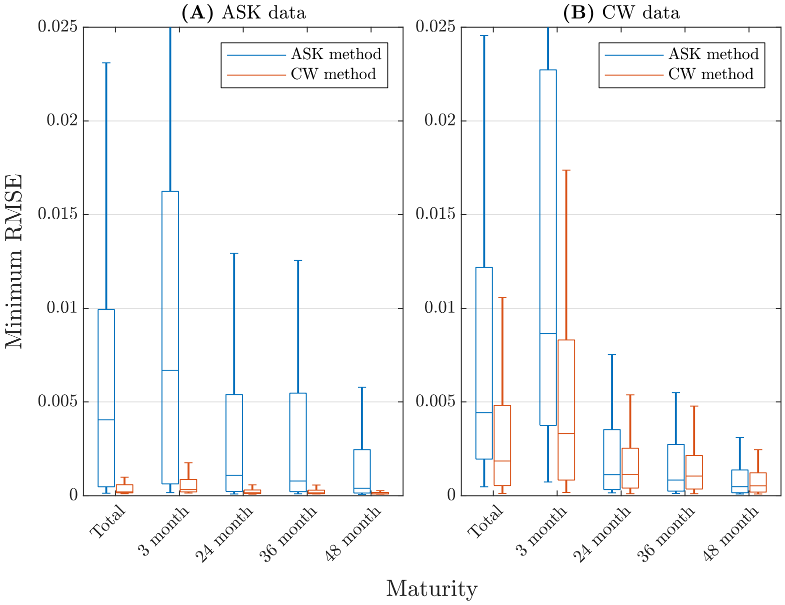

Figure 5: Boxplots of optimizing using both methods over both datasets, 120 observations. The RMSE resulting from optimizing both datasets with both methods are displayed. Both methods use 60 trials per dataset. The ASK method uses 300 observations per dataset, while the CW method uses 500 observations. The total RMSE, computed over all maturities, and the RMSE decomposed in individual maturities are displayed. The RMSE is evaluated over 120 out-of-sample observations.

4.3 Performance comparison

An efficient amount of amount of trials and observations has been determined for both methods, which allows for a representative performance, while keeping computation time as low as possible. The ASK method uses 300 observations and 60 trials, while the CW method uses 500 observations and 60 trials. The performance of the two methods will now be compared. Until now, the ASK method has only been used to estimate ASK data, and the CW method has only been used to estimate CW data. Because the parameters between these simulated datasets have no clear counterparts, the datasets behave differently. The estimates are found on differing data, and putting these side-by-side would not allow for a valid comparison.

Now both of the methods are used to estimate both datasets. Comparing the results of estimating the two methods on the same data leads to a valid comparison. The results of estimating the datasets with both methods can be found in Figure 5. The RMSE is evaluated over 120 out-of-sample observations. Additional results concerning 24 and 600 observations can be found in Appendix A. The results using 24 and 600 observations is in line with those using 120 observations. Both the total RMSE and the RMSE for each maturity is shown. As expected, the RMSE over a low maturity is higher than over the high maturity. The yield curve displays more variance in the low end than in the high end, as expected.

It is clear the CW method performs much better than the ASK method, for both simulated datasets. By applying the ASK estimation method to the CW data, a slight deterioration in results is found. This can be explained by the different parameters corresponding to different behaviour of the data. The CW method shows vastly different results between the datasets. The distribution of RMSE using the ASK data is roughly 10 times as small as the RMSE using the CW data. When comparing the two methods using CW data, the difference in performance is mostly found in the 3 month maturity errors, as the higher maturity yields show little difference between the two methods. As a matter of fact, the 36 and 48 month maturities show a lower median RMSE using the ASK method. When comparing the two methods using the ASK data, the CW method clearly outperforms the ASK method.

This increased performance is not without a cost. The CW method takes over three times as long to compute. A single optimization run, consisting of 60 trials for both methods, takes 28 minutes using the CW method. The ASK method only takes 8 minutes. If the real world data behaves more like the CW data, the added performance is mostly found in the 3 month maturity. The CW data is simulated using parameter values estimated from the real-world data of Creal and Wu (2015), while the parameter values of the ASK data are simply those that seem to work in behaving like the real-world data. Thus it seems reasonable to assume the real-world data behaves like the CW simulated data, and the difference in performance is mostly found in the 3-month maturity. The ASK method is able to create a quick estimate which is capable of fitting the higher maturities, though lacks precision in the 3 month maturity. If the real world data, against common sense, behaves more like the ASK data, the RMSE resulting from the ASK method is roughly 15 times that of the CW method. This clearly outweighs the 3 times increased computation cost. In this case, the CW method is the clear winner.

5 Empirical Parameter Estimation

This section estimates the parameters of both the Aït-Sahalia and Kimmel (2010) and Creal and Wu (2015) models on real-world data. The construction of the data is first discussed in section 5.1. The estimated parameters of the models using the ASK and CW methods are provided in sections 5.3 and 5.3, respectively.

5.1 Real-world yield data

The dataset of the real-world yields consists of two parts. First, Fama and Bliss zero coupon bond data is extracted from CRSP. The data consists of monthly US Fama and Bliss (1987) zero coupon bond yields. The data spans from June 1952 through December 2018, for a total of $T = 799$ observations. For each month, the zero coupon bond yields for maturities of 12, 24, 36, 48, and 60 months are available.

The second part consists of the monthly riskfree treasury series, also extracted from CRSP. This series contains the 1 and 3 month maturity zero coupon bond yields for the same date range of June 1952 through December 2018. Combined, these datasets replicate and expand on the dataset used in Creal and Wu (2015). The real-world dataset is used in the empirical estimation of the parameters.

As discussed in the introduction, non-Gaussian ATSMs are a good fit for estimating data which incorporates a zero lower bound, but only if the zero lower bound is set such that the yields never reach zero. When the yields become exactly zero or negative, a non-Gaussian ATSM is unable to estimate the parameters correctly. Keeping the observations which do not fit the model would invalidate entire regions of the parameter vector, possibly leading to entirely different parameter estimates. The data of the US yields contains four observations of exactly zero at the end of 2011. Three of these observations are on the 1-month yield and one on the 3-month yield. In 2015 there are six observations of negative yields, five of which on the 1-month yields and one on the 3-month yield. These ten observations prevent the estimation of the real-world data using non-Gaussian ATSMs.

In the code of Creal and Wu (2015), observations of exactly zero are dealt with by changing them to a very small, positive value. This will introduce a new kind of error. This error mostly manifests in the 3-month maturity observations, inflating the parameter estimates of the error term variance Ω slightly. The parameter estimates of all model parameters are also slightly changed. Still, this changing of the data to fit the model allows the model to estimate the rest of the data, which would not be possible otherwise. As such, the model can still be used to create a good estimate of the parameters. To accomodate the zero and negative data, the practice of Creal and Wu (2015) is maintained, by changing these values to small, positive values. Again, this increases the error of the model, but is necessary to allow the real-world data to be estimated.

5.2 ASK parameters

Estimating the real-world data using the ASK method, results in the following parameter estimates. To be clear, all parameters that show exactly 0 or 1 are fixed parameters during the estimation method, but shown here to create a clear picture of the full parameters, instead of only displaying the parameter values which are allowed to vary.

$$\alpha = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \qquad \beta = \begin{bmatrix} 1 & 0.477 & 0.924 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathcal{K} = \begin{bmatrix} 0.890 & 0 & 0 \\ 0.330 & -0.152 & -0.662 \\ 0.937 & 1.117 & 1.383 \end{bmatrix},$$

$$\Theta = \begin{bmatrix} 0.574 \\ 0 \\ 0 \end{bmatrix}, \qquad \lambda = \begin{bmatrix} -0.878 \\ -0.681 \\ 0.819 \end{bmatrix}, \qquad \delta_1 = \begin{bmatrix} 0.720 \\ 0.008 \\ 0.931 \end{bmatrix},$$

$$\operatorname{diag}(\Omega) = \begin{bmatrix} 0.584 \\ 0.111 \\ 0.079 \\ 0.026 \end{bmatrix}, \qquad \delta_0 = 0.638.$$

It is hard to provide meaningful interpretations to the parameters, due to the way the parameter values interact with each other to create the dynamics of the state variables. The variance of the errors diag(Ω) show that the lower maturity observations exhibit larger error variance than the higher maturities, with the lowest maturity observation being 20 times as large as the highest maturity.

5.3 CW parameters

Estimating the real-world data using the CW method, results in the following parameter estimates. The parameters which show exactly 0 are fixed during the optimization procedure. As explained in section 3.1.2, the parameters can be split in free parameters, GLS parameters and fixed parameters. In estimation, the fixed parameters do not change, and remain the same as provided in section 3.1.2. The free parameters are directly manipulated by the optimization procedure, these are given by

$$\begin{aligned}
\nu_h &= 1.765, & \nu_h^Q &= 2.766, & \delta_0 &= -0.0001,\\
\Phi_g^Q &= \begin{bmatrix} 0.109 & 0 \\ 0 & 0.886 \end{bmatrix}, & \Phi_h &= 0.946, & \Phi_h^Q &= 1.329,\\
\Sigma_{0,g} &= \begin{bmatrix} -1.08\text{e}{-08} & 0 \\ 6.40\text{e}{-08} & 5.82\text{e}{-08} \end{bmatrix}, & \Sigma_{1,g} &= \begin{bmatrix} 0.301 & 0 \\ -0.313 & 0.375 \end{bmatrix}, & \Sigma_{gh} &= \begin{bmatrix} 0.695 \\ 0.123 \end{bmatrix},\\
\Sigma_h &= 0.001.
\end{aligned}$$

The GLS parameters are not allowed to vary freely, but the values are dictated by the free parameters. The GLS parameters are given by

$$\begin{aligned}
\Phi_g &= \begin{bmatrix} 0.653 & 0.020 \\ -0.144 & 0.981 \end{bmatrix}, & \Phi_{gh} &= \begin{bmatrix} -0.149 \\ 0.148 \end{bmatrix},\\
\mu_g &= \begin{bmatrix} -0.019 \\ 0.030 \end{bmatrix}, & \text{diag}(\Omega) &= \begin{bmatrix} 0.071 \\ 0.037 \\ 0.036 \\ 0.022 \end{bmatrix}.
\end{aligned}$$

Again, it is hard to give reasonable interpretations to the specific values of the parameters, due to the complexities of the interaction between the variables. The variance of the errors diag(Ω) displays the same decreasing pattern when moving from the lower maturities to the higher maturities, but the difference is less extreme than exhibited on the ASK method. Here, the 3-month maturity displays 3 times as much variance as the 48-month maturity.

In the comparison of the Monte Carlo results, there were two scenarios. If the real-world data behaved more like the ASK data, the performance increase is found on all maturities. If the real-world data behaved more like the CW data, the performance increase of the CW data is mostly found in the 3-month maturity. Comparing the variance of the errors between the two empirical estimations, it seems the conclusion is somewhere in the middle. There is a large difference between the two methods on the 3-month maturity. But the CW method also outperforms on the 24- and 36-month maturities. It is only on the 48-month maturity that the two methods seem to perform roughly the same, with the CW method slightly outperforming the ASK method.

6 Conclusion

This research looks at two methods to estimate non-Gaussian affine term structure models, both allow parameter estimation by approximating the dynamics of the yield curve. The Aït-Sahalia and Kimmel (2010) method approximates the dynamics of the yield curve by applying a Taylor expansion to the dynamics of the state variables. The Creal and Wu (2015) method instead approximates the dynamics of the entire yield curve by considering a discrete-time variation of an ATSM with identical properties. Creal and Wu (2015) uses the insight that not all parameters are independent of others, a couple of parameter values can be trimmed from the parameter space but still be recovered using a GLS regression. Ex ante, it is unclear which of these two approximates allows for a better approximation.

Moreover, both methodology papers do not talk about the whether the methods are sensitive to the amount of observations to put into the model, or the amount of trials after which improvement is unlikely. While the amount of observations in most practical applications is fixed, it is useful to know whether the methods are able to efficiently estimate parameters with a certain amount of observations. The optimal amount of trials is needed to find a balance between computation time and performance. After the sensitivity analyses, the methods are compared using the selected amounts of trials and observations to find out which of the two models is superior.

Increasing the amount of trials linearly increases the computation time. A significant increase in the computation time should come with a significant improvement in the resulting RMSE. A significant difference in performance is tested by means of the Wilcoxon rank sum test. Both estimation methods show that the performance keeps increasing as the amount of trials increases. This is as expected, as each trial has a chance to find the global maximum or get stuck in local maxima. Increasing the amount of trials thus leads to a higher chance to find the global optimum. The Wilcoxon rank sum test shows both methods have an optimum at 60 trials, any increase in trials beyond 60 does not yield a significant increase in performance.

Increasing the amount of observations should increase performance, as the model has more observations to train, which should allow for a better fit in the out-of-sample evaluation. The increased amount of observations also comes with an increase in computation time, though the CW method has a much steeper increase in computation time by adding observations than the ASK method. The two method also differ in how they respond to an increase in observations. While there is no clear result from increasing the observations for the ASK method, the CW method clearly benefits with as much observations as possible. The ASK method is chosen to be tested with 300 observations as this results in the lowest RMSE when inspecting the results visually. These results are not significant, as there are no significant differences between the different amount of observations, across any of the evaluation horizons. The lowest RMSE is mostly due to the upper tail being more compact than the other distributions of RMSE. Using the CW method, the method clearly benefits from as much observations as possible, but inspecting the Wilcoxon rank sum test results shows there is no significant increase possible by using more than 500 observations. Thus the CW method uses 500 observations in the final comparison.

In the final comparison, the two methods are applied to both simulated datasets. In the sensitivity analyses, the ASK method has only been applied to the ASK data and the CW method has only been applied to the CW data. By applying the two methods to the same dataset, the RMSE between the methods can be compared. In both datasets, the CW method performs better. In the CW dataset, the increased performance of the CW method is mostly due to the lowest maturity of 3 months being much better, while the other maturities show relatively similar performance. In the ASK dataset, the RMSE resulting from the ASK method is roughly 15 times as large as that of the CW method.

The parameter values used to simulate the data of the CW method are estimated from the real-world data, the parameter values of the ASK data are not estimated from real data. It is a reasonable assumption that the real-world data behaves similarly to the CW data, and not like the ASK data. In this case, the performance is similar between the two estimation methods, apart from the lowest maturity, where the RMSE of the CW method is roughly half that of the ASK method. The difference between the two methods can then be reduced to an increase in performance in the low maturity, at the cost of a 3 times increase in computation time.

The empirical estimation of the parameters using both methods also supports the conclusion of smaller errors in the lower maturities using the CW method. As there is no out-of-sample data to evaluate the RMSE, the variance of the in-sample errors can be compared. The lower and middle maturities show significantly lower variance in the CW method, while only the highest maturity shows similar performance between the two methods.

References

Aït-Sahalia, Y. (2008). Closed-Form Likelihood Expansions for Multivariate Diffusions. *The Annals of Statistics*, *36*(2), 906–937. doi: 10.1214/009053607000000622

Aït-Sahalia, Y., & Kimmel, R. L. (2010). Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions. *Journal of Financial Economics*, *98*, 113–144. doi: 10.2139/ssrn.1283741

Brandt, M. W., & He, P. (2006). *Simulated Likelihood Estimation of Affine Term Structure Models from Panel Data* (Tech. Rep.). doi: http://dx.doi.org/10.2139/ssrn.885682

Cheridito, P., Filipović, D., & Kimmel, R. L. (2010, 7). A note on the dai-singleton canonical representation of affine term structure models. *Mathematical Finance*, *20*(3), 509–519. doi: 10.1111/j.1467-9965.2010.00408.x

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A Theory of the Term Structure of Interest Rates. *Econometrica*, *53*(2), 385–407.

Creal, D., & Wu, J. C. (2015). Estimation of Affine Term Structure Models with Spanned or Unspanned Stochastic Volatility. *Journal of Econometrics*, *185*, 60–81. doi: 10.3386/w20115

Dai, Q., & Singleton, K. J. (2000). Specification Analysis of Affine Term Structure Models. *The Journal of Finance*, *55*(5), 1943–1978. doi: 10.1111/0022-1082.00278

De Jong, F. (2000). Time Series and Cross-section Information in Affine Term-Structure Models. *Journal of Business & Economic Statistics*, *18*(3), 300–314. doi: 10.1080/07350015.2000.10524872

Duffie, D., & Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance*, *6*(4), 379–406. doi: 10.1111/j.1467-9965.1996.tb00123.x

Fama, E. F., & Bliss, R. R. (1987). The Information in Long-Maturity Forward Rates. *The American Economic Review*, *77*(4), 680–692.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics*, *30*(1-2), 203–238. doi: 10.1016/0304-4076(85)90138-1

Krippner, L. (2015). *Zero Lower Bound Term Structure Modeling: A practitioner's guide.* New York: Palgrave Macmillan US. doi: 10.1057/9781137401823

Lund, J. (1997). *Non-Linear Kalman Filtering Techniques for Term-Structure Models* (Tech. Rep.).

Piazzesi, M. (2005). *Bond Yields and the Federal Reserve* (Vol. 113; Tech. Rep. No. 2).

Piazzesi, M. (2010). Affine Term Structure Models. In *Handbook of financial econometrics, vol 1* (pp. 691–766). Elsevier Inc. doi: 10.1016/B978-0-444-50897-3.50015-8

Singleton, K. J. (2001). *Estimation of affine asset pricing models using the empirical characteristic function* (Vol. 102; Tech. Rep.).

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*(2), 177–188. doi: 10.1016/0304-405X(77)90016-2

Appendix A Figures

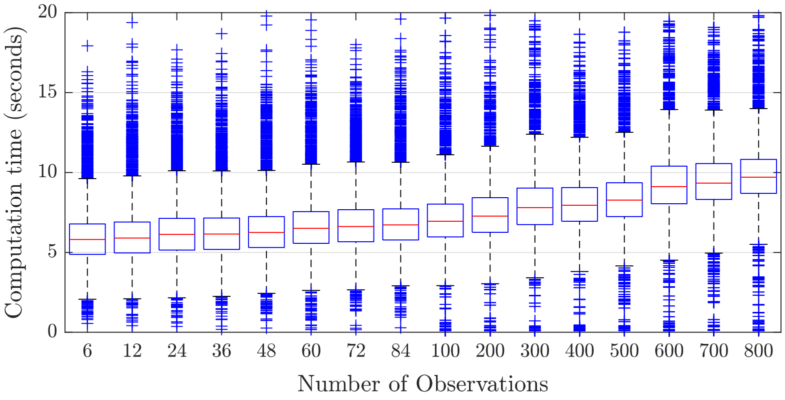

Figure A1: Boxplots of computation time per trial using the ASK method with varying amounts of observations
The computation time for the ASK methods is displayed, registered over the sensitivity analysis with respect to the amount of observations. The computation time is displayed in seconds per trial. Outliers are included. See Section 4.2 for a discussion of the results

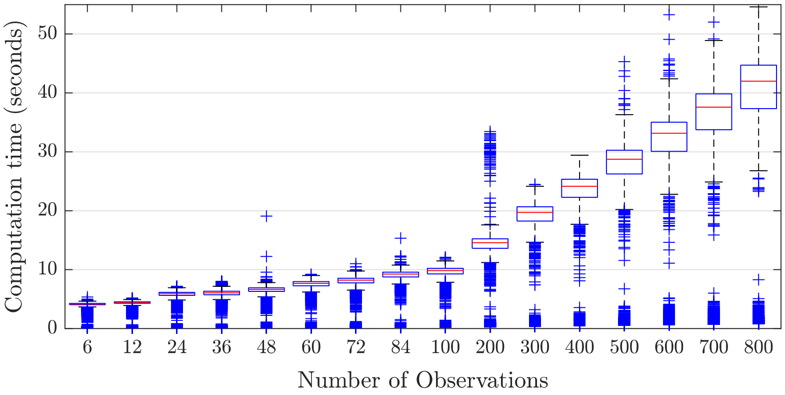

Figure A2: Boxplots of computation time per trial using the CW method with varying amounts of observations
The computation time for the CW methods is displayed, registered over the sensitivity analysis with respect to the amount of observations. The computation time is displayed in seconds per trial. Outliers are included. See Section 4.2 for a discussion of the results

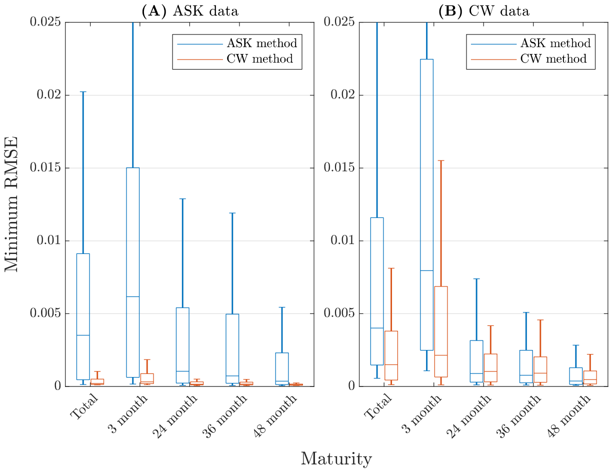

Figure A3: Boxplots of optimizing using both methods over both datasets, 24 observations
The RMSE resulting from optimizing both datasets with both methods are displayed. The total RMSE, computed over all maturities, and the RMSE decomposed in individual maturities are displayed. The RMSE is evaluated over 24 out-of-sample observations.

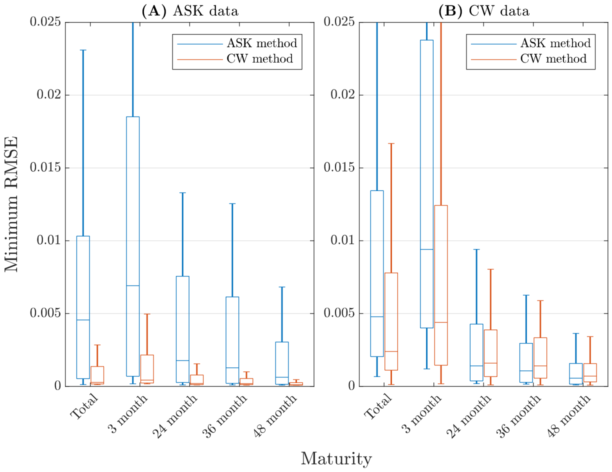

Figure A4: Boxplots of optimizing using both methods over both datasets, 600 observations
The RMSE resulting from optimizing both datasets with both methods are displayed. The total RMSE, computed over all maturities, and the RMSE decomposed in individual maturities are displayed. The RMSE is evaluated over 600 out-of-sample observations.

Appendix B Tables

Table A1: Wilcoxon results for ASK method with different amounts of trials, 24 observations

# Trials	20	40	60	80	100
10	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
20		**0.028**	**0.000**	**0.000**	**0.000**
40			0.093	**0.004**	**0.002**
60				0.296	0.209
80					0.708

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the ASK method, with either 10, 20, 40, 60, 80, or 100 trials, evaluated over 24 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A2: Wilcoxon results for ASK method with different amounts of trials, 600 observations

# Trials	20	40	60	80	100
10	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**
20		**0.025**	**0.000**	**0.000**	**0.000**
40			**0.015**	**0.000**	**0.000**
60				0.142	0.203
80					0.993

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the ASK method, with either 10, 20, 40, 60, 80, or 100 trials, evaluated over 600 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A3: Wilcoxon results for CW method with different amounts of trials, 24 observations

# Trials	20	40	60	80	100
10	0.153	**0.000**	**0.000**	**0.000**	**0.000**
20		**0.001**	**0.000**	**0.000**	**0.000**
40			**0.014**	**0.000**	**0.001**
60				0.336	0.361
80					0.791

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 10, 20, 40, 60, 80, or 100 trials, evaluated over 24 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A4: Wilcoxon results for CW method with different amounts of trials, 600 observations

# Trials	20	40	60	80	100
10	**0.025**	**0.000**	**0.000**	**0.000**	**0.000**
20		**0.000**	**0.000**	**0.000**	**0.000**
40			**0.028**	**0.001**	**0.008**
60				0.360	0.691
80					0.619

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 10, 20, 40, 60, 80, or 100 trials, evaluated over 600 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A5: Wilcoxon results for ASK method with different amounts of training observations, 24 out-of-sample observations

# Trials	12	24	36	48	60	72	84	100	200	300	400	500	600	700	800
6	**0.009**	0.135	0.091	0.297	0.488	0.879	0.199	0.130	0.819	0.768	0.602	0.757	0.810	0.232	0.363
12		0.354	0.392	0.140	0.106	**0.029**	0.275	**0.000**	**0.029**	0.069	**0.005**	**0.009**	**0.007**	0.211	**0.001**
24			0.815	0.582	0.524	0.323	0.960	**0.017**	0.281	0.464	0.062	0.138	0.104	0.915	**0.030**
36				0.517	0.374	0.138	0.697	**0.002**	0.166	0.205	**0.033**	0.058	**0.048**	0.659	**0.010**
48					0.863	0.531	0.681	**0.038**	0.565	0.765	0.197	0.319	0.301	0.791	0.090
60						0.552	0.649	**0.038**	0.588	0.681	0.220	0.248	0.340	0.633	0.093
72							0.295	0.146	0.987	0.873	0.340	0.595	0.614	0.312	0.216
84								**0.012**	0.309	0.431	0.067	0.128	0.102	0.943	**0.027**
100									0.142	0.128	0.626	0.344	0.417	**0.012**	0.871
200										0.819	0.359	0.628	0.565	0.322	0.211
300											0.213	0.419	0.404	0.412	0.141
400												0.704	0.750	0.078	0.726
500													0.944	0.172	0.351
600														0.127	0.524
700															**0.030**

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800 observations, evaluated over 24 out-of-sample observations. The values reported are the *p*-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low *p*-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A6: Wilcoxon results for ASK method with different amounts of training observations, 600 out-of-sample observations

# Trials	12	24	36	48	60	72	84	100	200	300	400	500	600	700	800
6	0.063	0.111	**0.027**	0.285	0.518	0.239	0.066	0.755	0.161	**0.030**	0.179	0.386	0.111	0.138	0.958
12		0.768	0.750	0.458	0.399	0.445	0.884	0.052	0.911	0.472	0.778	0.492	0.902	0.968	0.146
24			0.679	0.730	0.628	0.752	0.619	0.115	0.935	0.321	0.974	0.757	0.873	0.825	0.243
36				0.312	0.240	0.351	0.993	**0.022**	0.536	0.763	0.602	0.327	0.752	0.679	**0.050**
48					0.958	0.993	0.389	0.187	0.549	0.143	0.715	0.956	0.557	0.549	0.405
60						0.879	0.381	0.369	0.448	0.129	0.642	0.948	0.401	0.445	0.573
72							0.400	0.175	0.691	0.175	0.800	0.904	0.572	0.600	0.331
84								**0.040**	0.854	0.531	0.640	0.434	0.848	0.850	0.111
100									0.081	**0.014**	0.130	0.276	0.063	0.093	0.789
200										0.464	0.859	0.582	0.869	0.950	0.162
300											0.318	0.163	0.470	0.487	**0.038**
400												0.768	0.728	0.787	0.229
500													0.469	0.585	0.411
600														0.919	0.148
700															0.146

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800 observations, evaluated over 600 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A7: Wilcoxon results for CW method with different amounts of training observations, 24 out-of-sample observations

# Trials	12	24	36	48	60	72	84	100	200	300	400	500	600	700	800
6	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.001**	**0.000**	**0.018**	0.135	0.292	0.700	0.642	0.448	0.722	0.659
12		0.200	0.052	0.730	0.997	0.457	0.972	0.068	**0.008**	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
24			0.421	0.138	0.256	**0.047**	0.252	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
36				**0.030**	0.060	**0.010**	0.063	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
48					0.744	0.697	0.780	0.171	**0.025**	**0.008**	**0.001**	**0.002**	**0.002**	**0.000**	**0.001**
60						0.501	0.985	0.093	**0.014**	**0.003**	**0.000**	**0.001**	**0.001**	**0.000**	**0.000**
72							0.495	0.269	0.052	**0.016**	**0.001**	**0.003**	**0.005**	**0.000**	**0.002**
84								0.094	**0.012**	**0.003**	**0.000**	**0.001**	**0.001**	**0.000**	**0.000**
100									0.329	0.161	**0.024**	**0.047**	0.055	**0.002**	**0.028**
200										0.638	0.231	0.291	0.363	**0.029**	0.220
300											0.520	0.460	0.649	0.083	0.353
400												0.879	0.806	0.321	0.852
500													0.810	0.274	0.898
600														0.187	0.670
700															0.333

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800 observations, evaluated over 24 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.

Table A8: Wilcoxon results for CW method with different amounts of training observations, 600 out-of-sample observations

# Trials	12	24	36	48	60	72	84	100	200	300	400	500	600	700	800
6	**0.003**	**0.001**	**0.000**	**0.002**	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
12		0.733	0.308	0.944	**0.017**	0.881	0.266	0.621	**0.030**	**0.002**	**0.014**	**0.000**	**0.000**	**0.000**	**0.000**
24			0.495	0.793	**0.030**	0.848	0.438	0.833	**0.026**	**0.002**	**0.016**	**0.000**	**0.000**	**0.000**	**0.000**
36				0.374	0.203	0.382	0.879	0.610	0.209	**0.035**	0.131	**0.000**	**0.000**	**0.000**	**0.000**
48					**0.019**	0.937	0.271	0.587	**0.022**	**0.001**	**0.011**	**0.000**	**0.000**	**0.000**	**0.000**
60						**0.023**	0.241	**0.037**	0.958	0.414	0.884	**0.008**	**0.000**	**0.002**	**0.000**
72							0.311	0.711	**0.022**	**0.001**	**0.012**	**0.000**	**0.000**	**0.000**	**0.000**
84								0.507	0.286	0.063	0.197	**0.000**	**0.000**	**0.000**	**0.000**
100									0.058	**0.003**	**0.021**	**0.000**	**0.000**	**0.000**	**0.000**
200										0.317	0.763	**0.003**	**0.000**	**0.001**	**0.000**
300											0.472	0.065	**0.005**	**0.029**	**0.005**
400												**0.010**	**0.000**	**0.003**	**0.000**
500													0.410	0.929	0.403
600														0.498	0.960
700															0.355

The table displays the results of performing a two-sided Wilcoxon rank sum test on the RMSE's obtained by optimizing using the CW method, with either 6, 12, 24, 36, 48, 60, 72, 84, 100, 200, 300, 400, 500, 600, 700, 800 observations, evaluated over 600 out-of-sample observations. The values reported are the p-values associated with the null hypothesis corresponding to identical locations of the distributions of RMSE's. A low p-value indicates the compared distributions show different locations. Values below 0.05 are displayed in bold, these distributions differ significantly at the 5% level.