MASTER'S THESIS

Affine Term Structure Models with Martingale Estimating Functions

Michal Vala[†]

Under the supervision of:
Dr. Michel van der Wel [‡]

Co-read by:
Dr. Rutger-Jan Lange[§]

April 9, 2017

Abstract

Affine term structure models (ATSM) are used to model bond prices and yields. In this thesis, I propose to estimate the term structure parameters by martingale estimating functions (MEF). I analyze the feasibility of this method in a simulation study to better understand the behavior of the objective function with respect to parameters, as well as the actual convergence to the true population parameters using different search algorithms. MEF works very well in the simulated environment but requires a rescaling of the optimal MEF weights for the most robust results. The method is also applied to an empirical dataset where MEF provide good parameter estimates for the one- and three-factor settings and can directly compete with the minimum chi-squared estimation (MCSE) and the Kalman filter.

Keywords: bond yields, affine term structure models, martingale estimating functions, minimum chi-squared estimation, kalman filter

[†]Graduate student, Econometrics and Management Science: Quantitative finance, Erasmus School of Economics, Erasmus University Rotterdam. Contact: mic.vala@gmail.com. Student number: 429722.

[‡]Associate Professor, Department of Econometrics, Erasmus School of Economics, Erasmus University Rotterdam. Contact: vanderwel@ese.eur.nl.

[§]Assistant Professor, Department of Econometrics, Erasmus School of Economics, Erasmus University Rotterdam. Contact: lange@ese.eur.nl.

Contents

1 Introduction

Yields on risk-free assets form the foundation of finance and macroeconomics. They provide a degree of certainty in future returns and form a fundamental component of the capital asset-pricing model, the Black–Scholes formula, or the modern portfolio theory. These risk-free assets come with different maturities and studying their relationship and dynamics is referred to as the term structure interest rate modeling. It allows us to understand how bonds are priced, how the yield curve is constructed from observable prices, and what variation are we exposed to with time or policy changes.

Market participants face the uncertainty in the form of changing states of economies, implying time-varying values of bond yields. They require a compensation for this uncertainty which is priced by the market by a no-arbitrage assumption. Additionally, the affine term structure describes bond yields in the sense that they are constant and linear functions of some state vector, or factors. Such specification allows for nicer analytical solutions. However, the modeling is numerically challenging and the estimation of structural parameters on multi-dimensional non-linear likelihood surfaces is difficult.

Therefore, researchers develop and improve models that aim to identify and estimate the structural parameters from the observed bond yields, while imposing the cross-sectional restriction of no-arbitrage. The arbitrage-free term structure models date back to Vasicek (1977) and Cox et al. (1985) who focused on estimating one-factor models. Later research shows the need to adjust for multiple — two, three, or more — factors. Dai and Singleton (2000) offer a categorized view on various multi-dimensional affine models. With more complicated models, issues arise with the estimation and identification. There are several approaches to tackle these difficulties.

Joslin et al. (2011) study the concepts of estimation of Gaussian dynamic term structure models with the pricing factors being observable portfolios of yields, offering a lesser degree of computational complexity than the previously used approaches. Hamilton and Wu (2012) introduce the minimum chi-squared estimation (MCSE) method, which has become the baseline estimation for the term structure models. Hamilton and Wu combine the findings of Joslin et al. while pointing out the limitations of the observed portfolio approach, when compared to the reduced-form approach of the MCSE. Complications in the estimation process arise from the number of subjective choices that need to be made. Researchers need to impose parameter restrictions, choose proper starting values for numerical algorithms, or deal with identification issues. New methods are being explored to ease the burden of the estimation.

Recently, there has been a progress in the martingale estimating functions (MEF) that base its theory on the martingale probability theory by Hall and Heyde (2014) with its first drafts dating back to 1973. This area of probability theory kept growing rapidly with the most recent application in the form of MEF, which allow us to fit the continuous-time structures, the mixed frequency data, as well as the discrete-time models. These types of models produce martingale increments and the MEF utilizes these increments by minimizing them under the optimal set of time-varying weights.

Brix and Lunde (2013) apply the concepts of MEF to estimating stochastic volatility (SV) models. They compare the SV estimation to the GMM-type[1] estimation and conclude that — for some cases — there are gains in terms of lower bias and root mean square error. Christensen et al. (2016) apply martingale estimating functions to the dynamic stochastic general equilibrium models, also implying the efficiency of the MEF estimator when compared to GMM, while providing the theoretical framework on how to approach this type of estimation.

MEF is a new attempt to improve the efficiency of estimations, succeeding under some conditions and beating the known methods, such as GMM. MEF focuses on modeling the martingale increments and this paper uncovers the applicability of the method to the affine term structure models. Its approach is based on minimizing the optimally-weighted error terms in each time increment over the whole dataset, with respect to the model-dependent parameters. For ATSM we need to account for the dynamics of the latent explanatory factors and also the no-arbitrage cross-sectional relation of yields. MEF is flexible enough to account for both measures in its estimation process. This allows us to identify the time-series structural parameters that drive the dynamics of the factors as well as the risk-premium that investors require to hold long-maturity bonds over short periods.

For the benchmark estimation method, we consider the commonly used minimum chi-squared estimation (MCSE) method introduced by Hamilton and Wu (2012). The MCSE method allows for a quick parameter estimation of Gaussian class of ATSM. It is — similarly to the MEF — a minimum-distance type of estimation, which minimizes the quadratic form of differences between the restricted and unrestricted statistics. Hamilton and Wu address the issues of identification with previously used model representations and point out the possibility to estimate the parameters with OLS, while providing the asymptotic efficiency with the reduced-form representation. MCSE is also flexible enough to estimate the structural parameters from any Gaussian model representation.

The main objective of this paper is to apply the MEF method to the commonly used mean reverting Vasicek model representation. The main challenges are how to analytically derive the martingale increments for the model, how to optimally weigh them, how to treat the latent factors, and how to calibrate the numerical optimizer. The MEF derivation is feasible for models with closed form analytical solutions to the affine structure of yields, such as the one-factor Vasicek model. We extend the Vasicek model to include more factors, as well as consider the model with separate time-varying and risk-neutral coefficients.

We treat the latent factors as observable through the yields themselves following Hamilton and Wu (2012). They assume for the yields of $N = N_l + N_e$ different maturities with N_l yields observed without and N_e observed with the observational error. They assume for N_l to be equal to the number of pricing factors M, such that the factors become observable given the set of structural parameters. This implies that their reduced form representation is a restricted VAR that can be estimated with OLS. We adopt the same approach to back out the factors for the MEF from the observed yields, but use the martingale increment minimization for the estimation itself. This approach simplifies the estimation procedure, since factors are uniquely determined for each set of parameters.

Since this is a new area of modeling, the numerical optimization, the behavior of the ob-

[1]Generalized method of moment (GMM) estimation, Hansen (1982).

jective function, the dependence on starting parameter values, and the overall goodness of the convergence to the true parameter values are explored in a simulation study. We consider the MEF, the MCSE, and the Kalman filter methods. We include the Kalman filter to the comparison (and assume all yields observed with errors), because it allows us to use the same functional forms for the measurement and the transition equations that we use for the MEF, implying the equal model representation. On the other hand, while the benchmark MCSE incorporates the cross-sectional factor dependence, it is still the fastest and the most robust method to estimate the structural parameters with. Thus we keep the MCSE methodology intact while noting the differences in model representations.

We start the simulation by examining the behaviour of the objective function in a simulated term structure with the Vasicek model specification and estimating the structural parameter values with MEF. We examine the ability of MEF to find the structural parameters separately by varying the values of each respective parameter and fixing others to their population values. This approach uncovers the issue of MEF for systems with too little observational noise, which implies very low variance around the population parameter value and tremendous numerical difficulties. We modify the methodology by rescaling the optimal weights to ensure the convergence even in these systems.

For the numerical optimization we consider four different Matlab algorithms (fmincon, fminunc, fminsearch, patternsearch) to minimize the same MEF functional form. Fminunc and patternseach behave as expected, while fmincon and fminsearch turn out to be unsuitable for this type of estimation. In the last part of the simulation study, we focus on the large sample convergence. With the adjusted MEF specification, the best search algorithm, and other estimation methods for the direct comparison, we show the efficiency of MEF by estimating 1000 simulated term structures. Similarly to the traditional estimation methods, MEF also properly finds the true parameters in the simulation, while providing a faster convergence than the Kalman filter.

Empirical fit is examined on Fama and Bliss (1987) dataset. The one-factor model allows us to match the MEF specification exactly to the MCSE and directly compare both approaches. Empirically, both methods provide us with similar results. The three-factor model contains the information from the first three principal components of yields that are interpreted as the level, the slope and the curvature of the term structure and provides us with additional challenges. The necessary analytical solutions for MEF matrices are found in the diagonal form, whereas MCSE incorporates the cross-dependence of factors. This implies that the three-factor MEF and MCSE specification do not match perfectly, but the empirical fits are still comparable.

An advantage of MEF approach comes from the parameter identification, as all the parameters are just identified. In MCSE, the parameters are just-identified if exactly one time-series of yields is observed with observational error ($N_e = 1$), while several others (equal to the number of factors) are observed without any ($N_l = M$). By including the remaining unobserved maturities, MCSE method faces over-identification (this can be dealt with, for example, by imposing a weighting scheme for the additional unobserved yields). MEF handles the optimal weighting by construction. Nevertheless, even with the just-identified system Hamilton and Wu (2012) are able to estimate the structural parameters with little computational effort and robust results.

MEF offers an alternative to other estimation methods. It is computationally heavy due to the necessity of computing the martingale increments (time-varying matrices) separately and minimizing them over time. With increasing number of factors the size of the optimal matrices increases proportionally and the computational time increases as well. Different search algorithms work for this setup, but some do work better than other which is addressed in the simulation study.

Computational time is not the only difficulty the method faces. MEF relies on analytical solutions for the optimal time-varying matrices that are only available for some models. More complicated models work with recursive solutions to the affine structure of bond prices that make the derivations of the optimal MEF matrices impractical. However, there are models with analytical solutions to the affine structure such as the Vasicek model. This paper provides the evidence that the MEF method works in this type of setting and can directly compete with other methods of estimation.

The remainder of the thesis is organized as follows. Section 2 introduces the term structure modeling and explains the modeling framework used in the subsequent sections. Section 3 summarizes and derives the estimation procedures used in this paper. Section 4 uncovers concepts of the MEF estimation in the simulation study, whereas Section 5 focuses on empirical findings on the Fama and Bliss (1987) dataset. Section 6 concludes the findings. The text is supported by derivations, additional tables and figures in the Appendices.

2 Models

This section starts by introducing the mathematics involved in the affine term structure modeling in Subsection 2.1, with its direct application to a discrete time framework of Hamilton and Wu (2012) in Subsection 2.2, and continues by a more general continuous time framework in Subsection 2.3 which is necessary for martingale estimating functions. We also derive the model specifications for the one- and three-factor Vasicek model specifications directly from the continuous time framework which is used for MEF.

2.1 Introduction to interest rate term structure

Fixed income bases its theory on a fundamental security, termed *the zero-coupon bond* (also the risk-free discount bond or the pure discount bond) that pays one unit of currency at the maturity date with certainty. Let us denote the price of a such bond as $P(t,n) = P_t^{(n)}$, with t representing the current time and n the number of periods[2] till the maturity date. Note that t is used in the context of the price-variation in time for bonds of fixed duration n left until maturity. Additionally to the boundary condition, prices $P_t^{(n)} < 1$ for positive values of interest rates. That is the bond maturing in n is not worth more than its face value of 1.

Continuously compounded yield on a zero-coupon bond maturing at n (*the spot rate of interest*) is inverted directly from its price

$$y_t^{(n)} = -\frac{ln\left(P_t^{(n)}\right)}{n}. \tag{1}$$

Since we assume that prices are smaller than 1, the yields are positive for all maturities. Negative yields are only possible for bond prices that are greater than their face value, which would imply negative values of interest rates.

The short rate of interest is the yield on a very short bond and can be expressed as $r_t = \lim_{n\to 0^+} y_t^{(n)}$. This short rate process forms the foundation of the interest rate modeling and is the main determinant of prices of bonds. If the structure of these prices is

$$P_t^{(n)} = exp(-A - Br_t), \tag{2}$$

with $A = A^{(n)}$ and $B = B^{(n)}$ being deterministic functions, then the model is said to posses an affine term structure (Björk, 2009). In our setting, we assume that the short rate depends on some unobserved factors F_t. We assume an affine specification for these factors, such that

$$\underset{(1\times1)}{r_t} = \underset{(1\times1)}{\delta_0} + \underset{(1\times M)}{\delta_1^\top}\underset{(M\times1)}{F_t}. \tag{3}$$

Thus the bond price also depends on these factors and we can respecify the relation (2) to

[2]The units are set to correspond to the periodicity of data, that is observing the monthly periodic data we conclude that the one unit of n corresponds to the time period of the one month. If we observe the yearly periodic data the one unit of n would corresponds to the one year, e.t.c.

depend on the factors themselves, that is

$$P_t^{(n)} = exp(-\bar{A} - \bar{B}F_t), \tag{4}$$

where $\bar{A} = A^{(n)} + B^{(n)}\delta_0$ and $\bar{B} = B^{(n)}\delta_1^\top$ are deterministic functions. Moreover, it follows from (1) and (4) that the yields can be represented in the affine form as

$$\underset{(1\times1)}{y_t^{(n)}} = \underset{(1\times1)}{\frac{\bar{A}}{n}} + \underset{(1\times M)}{\frac{\bar{B}}{n}} \underset{(M\times1)}{F_t}, \tag{5}$$

where F_t is the vector of factors. This can be viewed as the change of the coordinate system and the given state variables F_t produce the same short rate process r_t.

The price process needs to satisfy the pricing relation such that prices of bonds with different maturities do not allow for a risk-free profit by trading these securities. There are two popular ways to enforce this internal consistency. First, assuming that the price $P_t = \mathbb{E}_t[P_{t+1}M_{t,t+1}]$ with the pricing kernel for affine term structure models $M_{t,t+1}$. This approach follows directly from the asset pricing theory and is used, for example, in the benchmark estimation by Hamilton and Wu (2012). Second, we can assure that the Feynman–Kac representation of the term structure equation holds, following Björk (2009). We adopt the second approach for the derivations of solutions for the continuous time processes.

2.2 Discrete time framework

In this section we introduce the discrete time model formulation following Hamilton and Wu (2012), which we generalize to the continuous time in subsequent Subsection 2.3. We denote F_t to be an $(M \times 1)$ vector of variables, characterized by a vector autoregression

$$F_{t+1} = c + \rho F_t + \Sigma u_{t+1}, \tag{6}$$

where $u_t \sim i.i.d.\mathcal{N}(0, I_m)$. Therefore $F_{t+1}|F_t, \dots F_1 \sim \mathcal{N}(\mu_t, \Sigma\Sigma^\top)$ with $\mu_t = c + \rho F_t$. Let r_t denote the risk-free one period interest rate. Zero coupon bond price can then be expressed as a function of factors F_t. Following the pricing kernel approach, they show that a risk-averse investor holds $P_t = \mathbb{E}_t[P_{t+1}M_{t,t+1}]$ with $M_{t,t+1} = exp(-r_t - (1/2)\lambda_t^\top\lambda_t - \lambda_t^\top u_{t+1})$ with λ_t being an $(M \times 1)$ vector of investor's risk attitude. The risk neutral investor faces $\lambda_t = 0$.

The relation (6) shows how risk-averse investors value assets and that the same holds for a risk-neutral investor under the Q-measure VAR given by

$$F_{t+1} = c^Q + \rho^Q F_t + \Sigma u_{t+1}^Q, \tag{7}$$

with a vector $c^Q = c - \Sigma\lambda_0$, a matrix $\rho^Q = \rho - \Sigma\Lambda$, a vector u_{t+1}^Q of independent standard normal variables, and $\lambda_t = \lambda_0 + \Lambda F_t$ is the market price of risk in its affine form. λ_0 is an $(M \times 1)$ vector and Λ is an $(M \times M)$ matrix.

Let one period risk-free yield also have an affine structure given by (3), where δ_0 is a constant, δ_1 a $(M \times 1)$ vector, and F_t a $(M \times 1)$ vector of factors. Then the yield on a risk-free

n-period zero-coupon bond from (5) can be calculated recursively by

$$y_t^{(n)} = \frac{\bar{A}}{n} + \frac{\bar{B}}{n} F_t = a_n + b_n^\top F_t, \tag{8}$$

where

$$b_n = \frac{1}{n}[I_M + \rho^{Q\top} + \ldots + (\rho^{Q\top})^{n-1}]\delta_1, \tag{9}$$

$$\begin{aligned} a_n = \delta_0 &+ (b_1^\top + 2b_2^\top + \ldots + (n-1)b_{n-1}^\top)c^Q/n \\ &- (b_1^\top \Sigma\Sigma^\top b_1 + 2^2 b_2^\top \Sigma\Sigma^\top b_2 + \ldots + (n-1)^2 b_{n-1}^\top \Sigma\Sigma^\top b_{n-1})/2n, \end{aligned} \tag{10}$$

as was demonstrated by Ang and Piazzesi (2003). Therefore by knowing c^Q, ρ^Q, δ_0, δ_1 and Σ we can compute the yield of zero-coupon bond for any maturity. The normalization of Hamilton and Wu (2012) impose $\Sigma = I_{N_l}$, $\delta_1 \geq 0$, $c = 0$ and ρ^Q upper triangular.

For yields with $N = N_l + N_e$ different maturities, Hamilton and Wu (2012) estimate N_l linear combinations of yields as being observed with no error, and N_e linear combinations to contain a measurement error. Denoting the first vector Y_t^1 and second Y_t^2 we specify the measurement as

$$\begin{bmatrix} \underset{(N_l\times 1)}{Y_t^1} \\ \underset{(N_e\times 1)}{Y_t^2} \end{bmatrix} = \begin{bmatrix} \underset{(N_l\times 1)}{A_1} \\ \underset{(N_e\times 1)}{A_2} \end{bmatrix} + \begin{bmatrix} \underset{(N_l\times M)}{B_1} \\ \underset{(N_e\times M)}{B_2} \end{bmatrix} F_t + \begin{bmatrix} \underset{(N_l\times N_e)}{0} \\ \underset{(N_e\times N_e)}{\Sigma_e} \end{bmatrix} \underset{(N_e\times 1)}{u_t^e}, \tag{11}$$

where $\Sigma_e = \sigma_e I_{N_e}$ is typically taken to be diagonal, determining the deviation of the measurement error $u_t^e \sim \mathcal{N}(0, I_{N_e})$, A_i and B_i are a vector and a matrix respectively, containing elements (10) and (9). Hamilton and Wu numerically compute the structural coefficients from (11) by minimizing the value of the chi-squared statistic for testing whether the restrictions are consistent with the observed reduced-form estimates. Details on the estimation procedure are provided in Subsection 3.1.

2.3 Continuous time framework

A more general approach is to define the processes in the continuous time and utilize the benefits of the stochastic calculus. Note that we observe the bond prices at some frequency, therefore even with the continuous framework we discretize the equations and map them to the discrete time equivalent. Let the unobserved M-dimensional state variables F_t follow a diffusion process

$$dF_t = \underbrace{\mu^P(F_t)dt}_{\text{non-random drift term}} + \underbrace{\sigma(F_t)dW_t^P}_{\text{random diffusion term}}, \tag{12}$$

where μ^P is an $(M \times 1)$ vector, $\sigma(F_t)$ an $(M \times M)$ matrix, and dW_t^P an M dimensional standard Brownian motion under the real world probability measure P.

The instantaneous drift of each state variable is assumed to be an affine function of F_t

$$\mu^P(F_t) = a^P + b^P F_t, \tag{13}$$

with a^P an $(M \times 1)$ vector, and b^P an $(M \times M)$ matrix. Instantaneous covariance between two pairs of factors F_t is an affine function of F_t

$$\left[\sigma(F_t)\sigma^\top(F_t)\right]_{ij} = \alpha_{ij} + \beta_{ij}^T F_t,$$

with ij denoting the row i and the column j, α_{ij} is a scalar, and β_{ij} is an $(M \times 1)$ vector.

There also exists an equivalent risk-neutral measure Q, such that

$$dF_t = \mu^Q(F_t)dt + \sigma(F_t)dW_t^Q,$$

where μ^Q is an $(M \times 1)$ vector, $\sigma(F_t)$ an $(M \times M)$ matrix, and dW_t^Q an M dimensional standard Brownian motion under the risk-neutral probability measure Q. The drift of the factors is then an affine function of F_t

$$\mu^Q(F_t) = a^Q + b^Q F_t,$$

with a^Q an $(M \times 1)$ vector, and b^Q an $(M \times M)$ matrix.

The existence of the solution to the above mentioned setup was explored before by Feller (1951) in the univariate case and by Duffie and Kan (1996) in the multivariate case. Cheridito et al. (2007) generalize the existence of the solution for the cases when the state vector F_t remains in the areas with $\sigma(F_t)\sigma(F_t)^\top$ being positive semidefinite. Cheridito et al. additionally show that there are no separate conditions for uniqueness and if a solution to (12) exists, it is unique. However, restrictions are needed to ensure parameter identification, such as $\sigma(F_t)$ to be diagonal (Dai and Singleton, 2000).

The probability measure Q measure is linked to the P measure by defining the market price of risk being proportional to the volatility term. Since this paper considers Gaussian models (that this $\left[\sigma(F_t)\sigma^\top(F_t)\right]_{ii} = \alpha_{ii}$) the market price of risk is also available in its affine form

$$\lambda(F_t) = [\sigma(F_t)]^{-1}(\mu^P(F_t) - \mu^Q(F_t)) = \lambda_0 + \Lambda F_t$$

with an $(M \times 1)$ vector λ_0 and an $(M \times M)$ matrix Λ. Dai and Singleton (2000) further distinguish between several classes of models. Models with unrestricted $\lambda(F_t)$ are referred to as the essentially affine models. Alternatively, the completely affine class of models restricts the matrix Λ to be equal to zero, making the market price of risk constant.

2.3.1 One-factor model

With the general continuous time framework established in Subsection 2.3, we continue by examining its relation to the model specification introduced by Vasicek (1977). This is a one-factor model specification and therefore assumed matrices become scalars. Let us start from

(3) by assuming an affine form of the short rate which depends on factors

$$\underset{(1\times1)}{r_t} = \underset{(1\times1)}{\delta_0} + \underset{(1\times1)}{\delta_1}\underset{(1\times1)}{F_t}, \tag{14}$$

and continue with (12) to illustrate the necessary restrictions for unique identification.

Let us derive the dynamics of r_t in the univariate case by combining (12) with (13) and (14). It follows that

$$\underset{(1\times1)}{dr_t} = (\underset{(1\times1)}{\delta_1}\underset{(1\times1)}{a^P} - \underset{(1\times1)}{b^P}\underset{(1\times1)}{\delta_0} + \underset{(1\times1)}{b^P}\underset{(1\times1)}{r_t})dt + \underset{(1\times1)}{\delta_1}\underset{(1\times1)}{\sigma(F_t)}\underset{(1\times1)}{dW_t^P}.$$

It is apparent that such a model does not uniquely determine the values of the intercept $\delta_1 a^P - b^P\delta_0$ and the volatility term $\delta_1\sigma(F_t)$. Thus without the loss on generality we can normalize either $\delta_1 = 1$, $a^P = 0$ (denoted $\mathcal{S}_1$) corresponding to for example de Jong (2000); $\delta_0 = 0$, $\delta_1 = 1$ (denoted $\mathcal{S}_2$) corresponding to Bolder (2001); or $\sigma(F_t) = 1$, $a^P = 0$ (denoted $\mathcal{S}_3$) corresponding to Hamilton and Wu (2012). I parametrize consistently with the benchmark parametrization $\mathcal{S}_3$ that is $\sigma(F_t) = 1$, $a^P = 0$ such that

$$\underset{(1\times1)}{dF_t} = \underset{(1\times1)}{b^P}\underset{(1\times1)}{F_t}\,dt + \underset{(1\times1)}{dW_t^P}. \tag{15}$$

The equivalent Q measure then follows

$$\underset{(1\times1)}{dF_t} = (\underset{(1\times1)}{a^Q} + \underset{(1\times1)}{b^Q}\underset{(1\times1)}{F_t})dt + \underset{(1\times1)}{dW_t^Q}, \tag{16}$$

implying for the market price of risk to be in the form of

$$\underset{(1\times1)}{\lambda} = [\sigma(F_t)]^{-1}(\mu^P(F_t) - \mu^Q(F_t)) = -a^Q + (b^P - b^Q)F_t = \underset{(1\times1)}{\lambda_0} + \underset{(1\times1)}{\Lambda}\underset{(1\times1)}{F_t}.$$

This specification belongs to the essentially affine class of models. Vasicek (1977) specifies his model to be completely affine that is the Λ matrix being restricted to zero with only the constant λ_0 determining the market price of risk.

Let us proceed by deriving the dynamics of the short rate from (14), (15), and (16) with the parametrization $\mathcal{S}_3$ to get

$$dr_t = (-b^P\delta_0 + b^P r_t)dt + \delta_1 dW_t^P, \tag{17}$$
$$dr_t = (\delta_1 a^Q - b^Q\delta_0 + b^Q r_t)dt + \delta_1 dW_t^Q. \tag{18}$$

Recall that (17) captures the short rate dynamics under the subjective measure which we directly observe, whereas (18) denotes the risk-neutral measure under which we price the bonds.

Additionally, Vasicek (1977) specifies his model with dynamics of the short rate labeled as $dr_t = \mu r_t dt + \sigma dW_t$, with $\mu = \kappa\big(\gamma^{P(Q)} - r_t\big)$ and $\sigma = \eta$, where κ represents the speed of reversion, γ the target rate, and η the constant volatility term. Specifications (17), (18) coincide with the Vasicek model and can be relabeled as

$$dr_t = \kappa^P(\gamma^P - r_t)dt + \eta dW_t^P, \tag{19}$$
$$dr_t = \kappa^Q(\gamma^Q - r_t)dt + \eta dW_t^Q, \tag{20}$$

where $\gamma^P = \delta_0, \eta = \delta_1, \kappa^P = -b^P$ for the P measure, the market price of risk $\lambda = \lambda_0 = -a^Q$, the risk-neutral mean $\gamma^Q = \delta_0 - \frac{\delta_1 a^Q}{b^Q} = \gamma^P - \frac{\lambda\eta}{\kappa}$, and the same $\kappa^Q = -b^Q = \kappa^P$ and the volatility term η for the Q measure. Note that since the model is completely affine, market price of risk λ does not vary with time and the only parameter affected by the change of measure is the γ^Q.

Under the Vasicek parametrization, for some price process (2) with dynamics (19) and (20), we can analytically solve for $A^{(n)}$ and $B^{(n)}$ to be in the form

$$B^{(n)} = \frac{1}{\kappa}(1 - e^{-\kappa n}), \tag{21}$$
$$A^{(n)} = \left(\frac{\eta^2}{2\kappa^2} - \gamma^P + \frac{\lambda\eta}{\kappa}\right)(B^{(n)} - n) + \frac{\eta^2}{4\kappa}(B^{(n)})^2, \tag{22}$$

derivided previously for example in Vasicek (1977). The full derivation of the solution to the Feynman–Kac theorem for this parametrization can be found in Appendix B2.

Alternatively, for a price process (4) that depends directly on factors F_t with normalization $\mathcal{S}_3$ we can write in the one-factor case

$$\bar{B}^{(n)} = B^{(n)}\eta, \tag{23}$$
$$\bar{A}^{(n)} = A^{(n)} + B^{(n)}\gamma^P. \tag{24}$$

By assuming the dependence of r_t on some factor F_t in an affine form and introducing additional parameters, we can express the same dynamics of r_t with different parametrizations and relabel them to the same functional forms (19) and (20). Therefore by taking a different parametrization route, the solutions (21), (22) would still hold but the mapping to the factors would depend on the chosen parametrization.[3]

Taking the route of Hamilton and Wu (2012), we suppose that there exist a yields vector Y_t^1 observed without the observational error. For the single factor case we assume the number of such yields $N_l = 1$, with the remaining $N_e = N - 1$ yields to be observed with the observational error. We can therefore follow with (11)

$$\begin{bmatrix} \underset{(1\times1)}{Y_t^1} \\ \underset{(N_e\times1)}{Y_t^2} \end{bmatrix} = \begin{bmatrix} \underset{(1\times1)}{A_1} \\ \underset{(N_e\times1)}{A_2} \end{bmatrix} + \begin{bmatrix} \underset{(1\times1)}{B_1} \\ \underset{(N_e\times1)}{B_2} \end{bmatrix} F_t + \begin{bmatrix} \underset{(1\times N_e)}{0} \\ \underset{(N_e\times N_e)}{\Sigma_e} \end{bmatrix} \underset{(N_e\times1)}{u_t^e}, \tag{25}$$

with $\Sigma_e = \sigma_e I_{N_e}$ diagonal and its value determines the deviation of the measurement error $u_t^e \sim N(0, I_{N_e})$. Vectors A_i and B_i contain stacked elements $\bar{A}^{(n)}/n$, $\bar{B}^{(n)}/n$ for different times till maturity n.

[3] For completeness and comparison of different specifications — $\mathcal{S}_1$: $y_t^{(n)} = \frac{\bar{A}^{(n)}}{n} + \frac{\bar{B}^{(n)}}{n}F_t$, with $\bar{B}^{(n)} = B^{(n)}$, $\bar{A}^{(n)} = A^{(n)} + B^{(n)}\gamma^P$; $\mathcal{S}_2$: $y_t^{(n)} = \frac{\bar{A}^{(n)}}{n} + \frac{\bar{B}^{(n)}}{n}F_t$, with $\bar{B}^{(n)} = B^{(n)}$, $\bar{A}^{(n)} = A^{(n)}$.

Since we assume Y_t^1 to be observed without observational error, we can obtain the values of factors F_t directly from yields[4]

$$F_t = [B_1]^{-1}(Y_t^1 - A_1), \tag{26}$$

following the dynamics

$$dF_t = -\kappa F_{t-\Delta}dt + dW_t^P, \tag{27}$$

which can be solved to obtain in the form

$$F_t = e^{-\kappa\Delta}F_{t-\Delta} + e^{-\kappa\Delta}\int_{t-\Delta}^{t} e^{\kappa(s-(t-\Delta))}dW_s^P, \tag{28}$$

with its conditional expectation and variance

$$\begin{aligned} E(F_t|F_{t-\Delta}) &= e^{-\kappa\Delta}F_{t-\Delta}, \\ Var(F_t|F_{t-\Delta}) &= \frac{1-e^{-2\kappa\Delta}}{2\kappa}. \end{aligned}$$

For the estimation purposes we further rewrite the system (25) as

$$\begin{bmatrix} \underset{(1\times1)}{F_t} \\ \underset{(N_e\times1)}{Y_t^2} \end{bmatrix} = \begin{bmatrix} \underset{(1\times1)}{0} \\ \underset{(N_e\times1)}{A_2} \end{bmatrix} + \begin{bmatrix} \underset{(1\times1)}{e^{-\kappa\Delta}F_{t-\Delta}} \\ \underset{(N_e\times1)}{B_2F_t} \end{bmatrix} + \begin{bmatrix} \underset{(1\times1)}{\sigma_\eta} & \underset{(1\times N_e)}{0} \\ \underset{(1\times1)}{0} & \underset{(N_e\times N_e)}{\Sigma_e} \end{bmatrix} \underset{(N_e+1\times1)}{u_t^e}, \tag{29}$$

where we substitute the first observed yield by the solution to the short rate time-series, with its expectation $e^{-\kappa\Delta}F_{t-\Delta}$ and variance $\sigma_\eta^2 = \frac{1-e^{-2\kappa\Delta}}{2\kappa}$. Having the model specification (29) we can capture both the time-series dynamics (P measure) as well as the cross-sectional relation (Q measure) at the same time. With these specifications in place, we can apply the estimation method of choice to numerically solve for the parameter vector $\phi^\top = (\gamma^P, \kappa, \lambda, \eta, \sigma_e)$.[5] More specifically for a given parameter vector ϕ we can observe the previously unobservable time-series of factor F_t from (26) and compute all elements of A_2, B_2 from (29) that describe the whole term structure and compare the goodness of fit with the observed data.

Table 1 maps the parameters from the Vasicek specification from equations (19), (20) to the more general affine continuous time framework from equations (17), (18), as well as the Hamilton and Wu (2012) specification from Subsection 2.2. Note that do not match the Hamilton and Wu (2012) specification exactly since Vasicek (1977) restricts $\kappa^Q = \kappa^P$ with the matrix $\Lambda = 0$. By estimating both the P and Q dynamics together we expect for the κ coefficient to lean either to the Q-side, or the P-side dynamics, depending on how we approach the estimation, which is further analyzed in a simulation and an empirical study.

[4]Factors become observable through the yields themselves (Dai and Singleton, 2000; Duffee, 2002).

[5]Note the equivalence in finding the information about vector $\phi_1 = (\gamma^P, \kappa, \lambda, \eta, \sigma_e)^\top$ and a vector $\phi_2 = (\gamma^P, \kappa, \gamma^Q, \eta, \sigma_e)^\top$, with $\gamma^Q = \gamma^P - \frac{\lambda\eta}{\kappa}$.

Table 1: Parameter mapping for one-factor Vasicek
Table compares three different specifications used in the modeling. Vasicek specification is chosen to be as close to the original paper as possible. This is then mapped to the continuous time framework from Subsection 2.3. Additionally, we can match the coefficients to Hamilton and Wu from Subsection 2.2. Note that then we have essentially affine model, whereas Vasicek uses the completely affine model.

Vasicek	Continuous time	Hamilton and Wu
κ^P	$-b^P$	$(1-\rho)/\Delta$
$\kappa^Q = \kappa^P$	$-b^Q = -b^P$	$(1-\rho^Q)/\Delta$
γ^P	δ_0	δ_0
$\lambda = \lambda_0$	$-a^Q$	$-c^Q$
0	0	Λ
η	δ_1	δ_1
σ_e	Σ_e	Σ_e

2.3.2 Three-factor model

So far we have derived the solutions to the one-factor model. However, multi-factor model is more complicated. In order to use the same analytical solutions for the dynamics of factors that are implied by Itô's lemma, we need to diagonalize the involved matrices. For clarity, we repeat the same derivation steps in the three-factor setting. Three factors are selected to represent the information contained in the first three principal components of yields and represent the level, the slope and the curvature of the term structure. Following Hamilton and Wu (2012) we take the n = 3-, 12-, and 60-month maturities to be the representative yields for factors.

Let us start from the short rate equation (3)

$$\underset{(1\times1)}{r_t} = \underset{(1\times1)}{\delta_0} + \underset{(1\times3)}{\delta_1^\top}\underset{(3\times1)}{F_t},$$

followed up by the P dynamics from (12)

$$\underset{(3\times1)}{dF_t} = \underset{(3\times3)}{b^P}\underset{(3\times1)}{F_t}\,dt + \underset{(3\times3)}{I}\underset{(3\times1)}{dW_t^P}, \tag{30}$$

where we — similarly to the one-factor case — normalize the mean ($a^P = 0$, equivalent to normalizing $c = 0$ in the discrete case) and volatility ($\sigma(F_t) = I$, equivalent to $\Sigma = I$ in the discrete case) in order to assure the parameter identification.

The equivalent Q measure then follows

$$dF_t = (a^Q + b^Q F_t)dt + IdW_t^Q. \tag{31}$$

Implying for the market price of risk to be in the form of

$$\underset{(3\times1)}{\lambda} = [\sigma(F_t)]^{-1}(\mu^P(F_t) - \mu^Q(F_t)) = -a^Q + (b^P - b^Q)F_t = \underset{(3\times1)}{\lambda_0} + \underset{(3\times3)}{\Lambda}\underset{(3\times1)}{F_t}.$$

This specification belongs to the essentially affine class of models. Again, we assume that the market price of risk is constant such that we can use the analytical solutions for $\bar{A}^{(n)}$ and $\bar{B}^{(n)}$ from the one-factor case.

To summarize the model specification — the P dynamics is restricted and described by (30) and its coefficients can be obtained from the time-series of factors, while the Q dynamics used for pricing bonds follow (31). The two measures are linked through the unobserved variable λ, which can be computed from the P and Q estimates, or estimated directly with one of the two measures. Moreover it holds that $a^Q = a^P - \sigma(F_t)\lambda_0$, $b^Q = b^P - \sigma(F_t)\Lambda$, again with Λ restricted to 0.

Similarly to the one-factor case, we re-define the system (11) as

$$\begin{bmatrix} \underset{(3\times1)}{Y_t^1} \\ \underset{(N_e\times1)}{Y_t^2} \end{bmatrix} = \begin{bmatrix} \underset{(3\times1)}{A_1} \\ \underset{(N_e\times1)}{A_2} \end{bmatrix} + \begin{bmatrix} \underset{(3\times3)}{B_1} \\ \underset{(N_e\times3)}{B_2} \end{bmatrix} \underset{(3\times1)}{F_t} + \begin{bmatrix} \underset{(3\times N_e)}{0} \\ \underset{(N_e\times N_e)}{\Sigma_e} \end{bmatrix} \underset{(N_e\times1)}{u_t^e}, \tag{32}$$

with $N_e = N - 3$, with A_i, B_i containing stacked elements $\bar{A}^{(n)}/n$ and row vectors $\bar{B}^{(n)}/n$ defined in (38), (37), with n being the respective time till maturity. We back out the factors of interest from yields as

$$F_t = B_1^{-1}(Y_t^1 - A_1). \tag{33}$$

Since we need to capture the F_t dynamics in P measure, we need analytical solutions for (30), obtainable with the Itô's lemma for a diagonal matrix e^{b^P} with $b^P = diag(-\kappa_1, \ldots, -\kappa_n)$ in the form

$$F_t = e^{b^P\Delta}F_{t-\Delta} + e^{b^P\Delta}\int_{t-\Delta}^{t} e^{-b^P(s-(t-\Delta))}IdW_s^P, \tag{34}$$

with its expectation and variance

$$E(F_t|F_{t-\Delta}) = e^{b^P\Delta}F_{t-\Delta} = e^{-\kappa\Delta}F_{t-\Delta},$$
$$Var(F_t|F_{t-\Delta})_{ii} = \frac{1-e^{-2\kappa_i\Delta}}{2\kappa_i}.$$

From this derivation it is apparent that for this problem formulation we had to further simplify the estimation problem such that the loadings on factors are diagonalized that is independent of each other. With this restriction in place we can apply the same analytical solutions for the factor loadings in the yield equations similarly to the one-factor case.

Thus by recalling the one-factor analytical solutions for $\bar{A}^{(n)}$ and $\bar{B}^{(n)}$ for a price process (4), we can generalize the three-factor case

$$\underset{(1\times1)}{y_t^{(n)}} = \underset{(1\times1)}{\frac{\bar{A}^{(n)}}{n}} + \underset{(1\times3)}{\frac{\bar{B}^{(n)}}{n}} \underset{(3\times1)}{F_t}, \tag{35}$$

$$\underset{(1\times3)}{\bar{B}^{(n)}} = \underset{(1\times3)}{B^{(n)}} \underset{(3\times3)}{diag(\eta)}, \tag{36}$$

$$\bar{A}^{(n)} = \gamma^P n + \sum_{i=1}^{3}\left(\frac{\eta_i^2}{2\kappa_i^2} + \frac{\lambda_i\eta_i}{\kappa_i}\right)(B_i^{(n)} - n) + \frac{\eta_i^2}{4\kappa_i}(B_i^{(n)})^2, \tag{37}$$

with elements of $\bar{B}^{(n)}$

$$B_i^{(n)} = \frac{1}{\kappa_i}(1 - e^{-\kappa_i n}). \tag{38}$$

$$\tag{39}$$

Finally, we arrive at the system

$$\begin{bmatrix} \underset{(1\times1)}{F_t^1} \\ \underset{(1\times1)}{F_t^2} \\ \underset{(1\times1)}{F_t^3} \\ \underset{(N_e\times1)}{Y_t^2} \end{bmatrix} = \begin{bmatrix} \underset{(1\times1)}{0} \\ \underset{(1\times1)}{0} \\ \underset{(1\times1)}{0} \\ \underset{(N_e\times1)}{A_2} \end{bmatrix} + \begin{bmatrix} \underset{(3\times3)}{\begin{pmatrix} e^{-\kappa_1\Delta} & 0 & 0 \\ 0 & e^{-\kappa_2\Delta} & 0 \\ 0 & 0 & e^{-\kappa_3\Delta} \end{pmatrix}} \underset{(3\times1)}{F_{t-\Delta}} \\ \underset{(N_e\times3)}{B_2} \underset{(3\times1)}{F_t} \end{bmatrix} + \begin{bmatrix} \underset{(3\times3)}{\begin{pmatrix} \sigma_{\eta_1} & 0 & 0 \\ 0 & \sigma_{\eta_2} & 0 \\ 0 & 0 & \sigma_{\eta_3} \end{pmatrix}} & \underset{(3\times N_e)}{0} \\ \underset{(N_e\times3)}{0} & \underset{(N_e\times N_e)}{\Sigma_e} \end{bmatrix} \underset{(3+N_e\times1)}{u_t^e}, \tag{40}$$

from which we model the martingale increments (the error terms) and minimize them with MEF.

3 Methods

With the model representations derived in the previous section we continue by explaining how do the considered methods of estimation work. Firstly, the MCSE is shortly summarized in Subsection 3.1, followed by the method of the Kalman filter in Subsection 3.2, and lastly, a greater emphasis is put on the estimation with MEF in Subsection 3.3. That is explaining the proper MEF methodology as well as applying it to the one-factor and three-factor cases.

3.1 Minimum-Chi-Squared Estimation

MCSE is — similarly to GMM — a minimum-distance estimation method in which we minimize the quadratic difference between restricted and unresticted statistics. When we consider the unrestricted maximum likelihood estimation (MLE) for the unrestricted statistics and weights from their asymptotic variance, MCSE is asymptotically efficient (Hamilton and Wu, 2012). This is an advantage over the GMM estimation, which is not. Additionally, MCSE allows us to estimate the structural coefficients without the computational complexity of other methods, while solving their shortcomings as well (the presence of the unit root with highly persistent data, or its applicability to any model representation).

For a latent 3 factor model, there are still 37 parameters to be estimated. Following restrictions are enforced to assure parameter identification — $\Sigma = I_M$, ρ is lower triangular, $c = 0$, $\delta_1 \geq 0$ — making it 14 restrictions, with remaining 23 identifiable parameters. That is 3 in $c^{\mathbb{Q}}$, 6 in $\rho^{\mathbb{Q}}$, 9 in ρ, 1 in δ_0, 3 in δ_1, 1 in σ_e (the case of a just identified structure and $N_e = 1$). Following directly from (11), F_t become observable

$$F_t = B_1^{-1}(Y_t^1 - A_1).$$

Moreover, an invariant transformation of (6) yields a reduced form VAR

$$\begin{aligned}
Y_t^1 &= A_1^* + \phi_{11}^* Y_{t-1}^1 + u_{1t}^*,\\
A_1^* &= A_1 - B_1 \rho B_1^{-1} A_1,\\
\phi_{11}^* &= B_1 \rho B_1^{-1},\\
Y_t^2 &= A_2^* + \phi_{21}^* Y_t^1 + u_{2t}^*,\\
A_2^* &= A_2 - B_2 B_1^{-1} A_1,\\
\phi_{21}^* &= B_2 B_1^{-1},\\
\begin{bmatrix} u_{1t}^* \\ u_{2t}^* \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega_1^* & 0 \\ 0 & \Omega_2^* \end{bmatrix} \right),\\
\Omega_1^* &= B_1 B_1^\top = \frac{1}{T}\sum_{t=1}^{T} (Y_t^1 - A_1^* - \phi_{11}^* Y_{t-1}^1)(Y_t^1 - A_1^* - \phi_{11}^* Y_{t-1}^1)^\top,\\
\Omega_2^* &= \Sigma_e \Sigma_e^\top = \frac{1}{T}\sum_{t=1}^{T} (Y_t^2 - A_2^* - \phi_{21}^* Y_t^1)(Y_t^2 - A_2^* - \phi_{21}^* Y_t^1)^\top.
\end{aligned}$$

The estimation procedure involves guessing $(\rho^{\mathbb{Q}}, \delta_1)$, calculating respective $B(\rho^{\mathbb{Q}}, \delta_1)$ and defining $\hat{\pi} = (vec(\hat{\phi}^*_{21}\hat{\Omega}^*_1)^\top, vech(\hat{\Omega}^*_1)^\top)^\top$, $g(\rho^Q, \delta_1) = (vec(B_2 B_1^\top)^\top, vech(B_1 B_1^\top)^\top)^\top$, while numerically finding $(\hat{\rho}^Q, \hat{\delta}_1) = \underset{\rho^Q,\delta_1}{argmin}(\hat{\pi} - g_2(\rho^Q, \delta - 1))^\top(\hat{\pi} - g_2(\rho^Q, \delta - 1))$, which is just identified for $N_e = 1$, and overidentified for $N_e > 1$. With an estimate of (ρ^Q, δ_1), we find $\hat{\rho} = \hat{B}_1^{-1}\hat{\rho}^*_{11}\hat{B}_1$. Furthermore, we find (δ_0, c^Q) from $\hat{A}^*_1 = (I - \hat{B}_1\hat{\rho}\hat{B}_1^{-1})\hat{A}_1$, $\hat{A}^*_2 = \hat{A}_2 - \hat{B}_2\hat{B}_1^{-1}\hat{A}_1$. That is we guess (δ_0, c^Q) and calculate $A(c^Q, \delta_0, \rho^Q, \delta_1)$ and minimize the squared differences in the equations. That is the minimum chi-square estimation.

Asymptotic standard errors can be obtained by assuming the hypothesis $\pi = g(\phi)$ and defining the Fisher information matrix

$$R = -\frac{1}{T}\mathbb{E}\left[\frac{\partial^2 logL(\pi, Y)}{\partial\pi\partial\pi^\top}\right],$$

with π being the vector of reduced-form parameters and ϕ the structural parameter vector. A linear approximation $g(\phi) \simeq \gamma + \Gamma\phi$ for $\Gamma = \partial g(\phi)/\partial\phi^\top|_{\phi=\phi_0}$ and $\gamma = g(\phi_0) - \Gamma\phi_0$, where we assume that there exists a true value of ϕ_0 for which $g(\phi_0) = \pi_0$. Hamilton and Wu (2012) show that since $\sqrt{T}(\hat{\pi} - \pi_0) \overset{L}{\sim} \mathcal{N}(0, R^{-1})$, it follows

$$\hat{\phi}_{MCSE} \overset{L}{\sim} \mathcal{N}\left(\phi_0, \frac{(\Gamma^\top R\Gamma)^{-1}}{T}\right),$$

with the var-covariance matrix approximated with $(\hat{\Gamma}^\top\hat{R}\hat{\Gamma})^{-1}/T$ with $\hat{\Gamma} = \partial g(\phi)/\partial\phi^\top|_{\phi=\hat{\phi}_{MCSE}}$.

3.2 Kalman filtering

A more traditional method of the model likelihood maximization is provided by the Kalman filter (applied to finance, for example, by Hamilton (1994)), which allows us to model the term structure with the state-space model representation with unobserved values of factors. In the state-space system we consider a measurement equation for the yields of different maturities and a transition equation for the latent factors. We follow the methodology of de Jong (2000) and allow for all measurement equations to have the measurement error and integrate out the latent factors with the Kalman filter. With the use of the quasi-likelihood method we can construct the conditional mean and variance of the latent factors and compute the value of the likelihood.

We follow with the de Jong (2000) specification

$$\begin{aligned} y_t &= A + BF_t + e_t, & \text{Var}(e_t) &= H, \\ F_t &= C + DF_{t-\Delta} + \eta_t, & \text{Var}(\eta_t) &= Q, \end{aligned}$$

along with some initial conditions

$$\begin{aligned} F_0 &= \mathbb{E}(F_t) \\ P_0 &= \text{Var}(F_t). \end{aligned}$$

The prediction step is defined as

$$F_{t|t-\Delta} = C + DF_{t-\Delta},$$
$$P_{t|t-\Delta} = DP_{t-\Delta}D^{\top} + Q,$$

with the likelihood contributions

$$u_t = y_t - A - BF_{t|t-\Delta},$$
$$V_t = BF_{t|t-\Delta}B^{\top} + H,$$
$$-2\text{ln}L_t = \text{ln}|V_t| + u_t^{\top}V_t^{-1}u_t,$$

and the updating step, including computing the Kalman gain, adjust conditional variance and factor value

$$K_t = P_{t|t-\Delta}B^{\top}V_t^{-1},$$
$$F_t = F_{t|t-\Delta} + K_t u_t,$$
$$P_t = (I - K_t B)P_{t|t-\Delta}.$$

We therefore minimize the sum of the $-2\text{ln}L_t$ to obtain a set of parameters that maximizes the model likelihood.

Asymptotic standard errors can be obtained from the Fisher information matrix evaluated at the maximum likelihood estimates of the parameter vector ϕ. Since we minimize the negative log-likelihood $-lnL$, the hessian matrix

$$H(\phi) = \frac{\partial^2}{\partial\phi_i\partial\phi_j}(-lnL),$$

is exactly equal to the Fisher information matrix. We therefore estimate the covariance matrix of the optimal $\hat{\phi}_{ML}$ with

$$Var(\hat{\phi}_{ML}) = [H(\hat{\phi}_{ML})]^{-1},$$

and the estimates are then asymptotically normally distributed

$$\hat{\phi}_{ML} \overset{a}{\sim} \mathcal{N}(\phi_0, Var(\hat{\phi}_{ML})),$$

with standard errors obtained as square-roots of the diagonal elements of $Var(\hat{\theta}_{ML})$.

3.3 Martingale estimating functions

In both the discrete and the continuous time models generate martingale increments. The main explanatory variable in ATSM are the values of unobserved factors. These factors follow some stochastic processes that produce martingale increments. This can be further generalized for the prices of the yields themselves as they are affine functions of the factors and also produce martingale increments. Therefore, ATSM form a feasible setup for MEF and allowing the identification and estimation of the structural parameter vector ϕ.

We need to capture two separate measures — the subjective P measure and the risk-neutral Q measure. The P measure can be inferred from the dynamics of the unobserved factors (more specifically from the discrete solution to the stochastic process for the factor dynamics). Recall that the assumed approach allows for factors to become observable through yields themselves. The Q measure then needs to hold for the pricing relation and is obtained by fitting the observed bond prices with the P parameters while letting the model to determine the market price of risk. The measures combined allow to specify a vector $m_t(\phi)$ of martingale increments and minimize it over time with MEF.

Martingale estimating functions are defined as weighted sums of martingale increments. MEF assume the same number of estimating equations as the number of parameters, allowing for a just identified structure. Additionally, Christensen et al. (2016) point out that *"optimal weights are time-varying matrices in the information set one period earlier and depend on the conditional variance of the martingale increment and the conditional mean of the parameter derivatives"*. Intuitively, we minimize the time-varying error terms given the available information set at respective time-increment, subject to optimal weights. The optimal weights can be viewed as instruments that satisfy the first order optimality conditions (first order partial derivatives) and are further weighted by "the precision matrix" (the inverse of the covariance matrix), which approximates the amount of certainty that the optimal parameters are indeed optimal.

Formally, we define martingale estimating function following Christensen et al. (2016) as

$$M_T = \sum_{t=1}^{T} w_t m_t, \tag{41}$$

which is a zero-mean martingale for any w_t, that depends on the data up to $t-1$. MEF is given by specifying w_t as a series of $\dim(\phi) \times \dim(m)$ matrices. Parameter vector ϕ is estimated by solving $M_T(\phi) = 0$ and $\mathbb{E}(M_T) = 0$ at the true value of ϕ. The conditional moment restriction is then defined as $\mathbb{E}_{t-1}(m_t) = 0$.

We obtain w_t by computing

$$w_t = \psi^\top (\Psi_t)^{-1}, \tag{42}$$

where Ψ_t is the conditional variance matrix of the vector martingale increment

$$\Psi_t = \mathrm{Var}_{t-1}(m_t) = \mathbb{E}_{t-1}(m_t m_t^\top)\ , \tag{43}$$

and ψ_t the matrix of conditional means of its parameter derivatives

$$\psi_t = \mathbb{E}_{t-1}\Big(\frac{\partial m_t}{\partial \phi^\top}\Big). \tag{44}$$

Then the optimal MEF estimator is the MEF estimator using instruments $w_t = \psi_t^\top \Psi_t^{-1}$.

The optimal MEF estimator is consistent and asymptotically normally distributed

$$\hat{\phi} \overset{a}{\sim} \mathcal{N}\left(\phi_0, \frac{1}{T} V_{MEF}\right),$$

with asymptotic covariance matrix

$$V_{MEF} = \left(\mathbb{E}[\psi_t^T (\Psi_t)^{-1} \psi_t]\right)^{-1},$$

and consistently estimated by the inverted sample average

$$\hat{V}_{MEF} = \left(\frac{1}{T} \sum_{t=1}^{T} \psi_t^T (\Psi_t)^{-1} \psi_t\right)^{-1}.$$

With the derived model representation — which for ATSM is obtainable in a purely analytical form — we apply the MEF methodology and derive the vector of martingale increments $m_t(\phi)$ (depending on the parameter vector ϕ) and the optimal weighting matrices $\psi_t^\top$ (consisting of partial derivatives of $m_t(\phi)$ with respect to each element of ϕ) and Ψ_t^{-1} that scales the $\psi_t^\top$ weights by the inverse of the amount of the implied noise ("the precision matrix") in each of the martingale increments. This system is time-varying in the sense that in each time point we need to compute these matrices separately, implying a degree of numerical complexity.

3.3.1 MEF with one-factor

The one-factor setting allows us to track the estimation process closely. We therefore explain the procedure thoroughly with all necessary steps, together with stating all relevant matrices that enter the MEF. A similar setup is used for the three-factor model with highlighted differences in the subsequent section. We proceed with following steps.

Step 1. We solve for F_t from (26), with A_1, B_1 being solved analytically and depending on parameter vector $\phi = (\gamma^P, \kappa, \lambda, \eta)$. That is for a given parameter vector we can directly compute the values or F_t form the chosen yield Y_t^1. This also allows us to compute the implied volatility σ_e^2 of the observational noise.

Step 2. Solve the SDE from (28) to arrive at the discrete time formulation of F_t. This time-series is used to estimate the P dynamics of F_t.

Step 3. Apply Martingale Estimating Functions for (29) with estimated F_t from Step 1, along with the result from Step 2. We derive the (5×1)[6] vector m_t (identifying the P dynamics of F_t time-series as well as the Q measure thanks to the cross-sectional no-arbitrage pricing dependence) but also the $\psi_t^\top$ and Ψ_t^{-1} for the weighting vector w_t.

[6] Elements of ϕ extended by the deviation of the observational noise σ_e.

Martingale increments under Vasicek parametrization are defined from (29) as

$$m_t(\phi) = \left(\begin{bmatrix} \underset{(1\times 1)}{F_t} \\ \underset{(N_e\times 1)}{Y_t^2} \end{bmatrix} - \begin{bmatrix} \underset{(1\times 1)}{0} \\ \underset{(N_e\times 1)}{A_2} \end{bmatrix} - \begin{bmatrix} \underset{(1\times 1)}{e^{-\kappa\Delta}F_{t-\Delta}} \\ \underset{(N_e\times 1)}{B_2 F_t} \end{bmatrix} \right)$$

$$= \begin{pmatrix} F_t - e^{-\kappa\Delta}F_{t-\Delta} \\ y_t^{(n_2)} - \bar{A}^{(n_2)}/n_2 - \bar{B}^{(n_2)}/n_2 \cdot F_t \\ \vdots \\ y_t^{(n_N)} - \bar{A}^{(n_N)}/n_N - \bar{B}^{(n_N)}/n_N \cdot F_t \end{pmatrix}.$$

We derive the components of the weighting matrix $w_t = {\psi_t}^\top {\Psi_t}^{-1}$ as

$${\psi_t}^\top = \begin{pmatrix} 0 & -1 & \cdots & -1 \\ \Delta F_{t-\Delta} e^{-\kappa\Delta} & \begin{array}{c} -\Big[\Big(\frac{(e^{-\kappa n_2}-1)}{\kappa^2} + \frac{n_2 e^{-\kappa n_2}}{\kappa}\Big)\Big(\frac{\eta^2}{2\kappa^2} - \gamma^P \\ + \frac{\eta\lambda}{\kappa}\Big) + \Big(n_2 + \frac{(e^{-\kappa n_2}-1)}{\kappa}\Big)\Big(\frac{\eta^2}{\kappa^3} + \frac{\eta\lambda}{\kappa^2}\Big) \\ - \frac{3\eta^2(e^{-\kappa n_2}-1)^2}{4\kappa^4} + \frac{\gamma^P(e^{-\kappa n_2}-1)}{\kappa^2} + \\ \frac{\gamma^P n_2 e^{-\kappa n_2}}{\kappa} - \frac{\eta^2 n_2 e^{-\kappa n_2}(e^{-\kappa n_2}-1)}{2\kappa^3}\Big]\frac{1}{n_2} \\ - \frac{F_t \eta e^{-\kappa n_2}}{\kappa} - \frac{F_t \eta (e^{-\kappa n_2}-1)}{\kappa^2 n_2} \end{array} & \cdots & \begin{array}{c} -\Big[\Big(\frac{(e^{-\kappa n_N}-1)}{\kappa^2} + \frac{n_N e^{-\kappa n_N}}{\kappa}\Big)\Big(\frac{\eta^2}{2\kappa^2} - \gamma^P \\ + \frac{\eta\lambda}{\kappa}\Big) + \Big(n_N + \frac{(e^{-\kappa n_N}-1)}{\kappa}\Big)\Big(\frac{\eta^2}{\kappa^3} + \frac{\eta\lambda}{\kappa^2}\Big) \\ - \frac{3\eta^2(e^{-\kappa n_N}-1)^2}{4\kappa^4} + \frac{\gamma^P(e^{-\kappa n_N}-1)}{\kappa^2} + \\ \frac{\gamma^P n_N e^{-\kappa n_N}}{\kappa} - \frac{\eta^2 n_N e^{-\kappa n_N}(e^{-\kappa n_N}-1)}{2\kappa^3}\Big]\frac{1}{n_N} \\ - \frac{F_t \eta e^{-\kappa n_N}}{\kappa} - \frac{F_t \eta (e^{-\kappa n_N}-1)}{\kappa^2 n_N} \end{array} \\ 0 & \frac{\eta}{\kappa n_2}\Big(n_2 + \frac{e^{-\kappa n_2}-1}{\kappa}\Big) & \cdots & \frac{\eta}{\kappa n_2}\Big(n_2 + \frac{e^{-\kappa n_2}-1}{\kappa}\Big) \\ 0 & \begin{array}{c} \Big[\Big(n_2 + \frac{(e^{-\kappa n_2}-1)}{\kappa}\Big)\Big(\eta\kappa^2 + \lambda\kappa\Big) \\ - \frac{\eta(e^{-\kappa n_2}-1)^2}{2\kappa^3}\Big]\frac{1}{n_2} + \frac{F_t(e^{-\kappa n_2}-1)}{\kappa n_2} \end{array} & \cdots & \begin{array}{c} \Big[\Big(n_N + \frac{(e^{-\kappa n_N}-1)}{\kappa}\Big)\Big(\eta\kappa^2 + \lambda\kappa\Big) \\ - \frac{\eta(e^{-\kappa n_N}-1)^2}{2\kappa^3}\Big]\frac{1}{n_N} + \frac{F_t(e^{-\kappa n_N}-1)}{\kappa n_N} \end{array} \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$\Psi_t^{-1} = \begin{pmatrix} \frac{1}{\sigma_\eta^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_e^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_e^2} \end{pmatrix}$$

with $\sigma_\eta^2 = \frac{1-e^{-2\kappa\Delta}}{2\kappa}$ representing the volatility of the factor process, σ_e^2 represents the constant volatility term of all yields observed with observational error that is $\Sigma_e = \sigma_e I_{N_e}$.[7]

Since we know the analytical representation of $M_T = \sum_{t=1}^{T} {\psi_t}^\top {\Psi_t}^{-1} m_t$, we can solve for the optimal parameter vector ϕ, such that $M_T(\phi) = 0$. This is done by a numerical minimizer in Matlab. Note that we have the set of five equations with the equal number of parameters. We therefore minimize the objective function defined as a scalar

$$f(\phi) = M_T^\top M_t.$$

Moreover, even though σ_e enters the estimation it is not identified by the system by construc-

[7] The derivations from this section can be found in the Appendix B3.

tion. We use the implied volatility σ_e for each set of parameters $(\gamma^P, \kappa, \lambda, \eta)$ instead and continue with the estimation as described.

Based on the simulated and empirical findings we additionally extend the methodology as follows. First, next to estimating the completely affine Vasicek model we also consider the unrestricted essentially affine model that lets $\kappa^Q \neq \kappa^P$. This way we match perfectly the one-factor estimation of Hamilton and Wu (2012) and can directly compare results. Note the necessity to adjust the $m_t, \psi_t^\top, \Psi_t^{-1}$ matrices to account for the two separate κ coefficients.

Secondly, we propose an extension to the standard MEF methodology (labeled estimation $\mathcal{E}_1$), because of the uncertainty in estimated values coming from "the precision matrix" $\Psi_t{}^{-1}$, which is prone to exploding. We consider a second estimation setting (labeled estimation $\mathcal{E}_2$) with rescaled Ψ_t^{-1} matrix in such a way that the diagonal elements sum up to one. That way we keep the property of shifting the weights ratios for different processes, but provide them with a certain boundary. As an example consider the estimation of just two yields — we label $\Psi_{t,11}^{-1} = 1/\sigma_\eta^2 = w_1$ and $\Psi_{t,22}^{-1} = 1/\sigma_e^2 = w_2$. The proposed method rescales these weights such that $\Psi_{t,11}^{-1} = w_1/(w_1 + w_2)$ and $\Psi_{t,22}^{-1} = w_2/(w_1 + w_2)$.

3.3.2 MEF with three-factors

The three-factor model in more challenging than its one-factor equivalent. By solving the term structure equation for a price process $P = exp(-\bar{A} - \bar{B}F_t)$ we arrive at a set of Ricatti equations. These have — under the affine model setting discussed in this paper — known solutions. An n-period zero coupon bond yield from (5)

$$y_t^{(n)} = a_n + b_n^\top F_t,$$

has the coefficients a_n, b_n given recursively by (9) and (10).

We quickly realize that deriving the optimal weights for MEF under this setting is not be an easy task to do, especially assuming the longest maturity of 120 months which translates to a b_n consisting of a polynomial of degree 119 of which we need to take derivatives with respect to every coefficient (that is every member of each matrix). Such derivation is not feasible. Nevertheless, when we diagonalize the loadings on factor dynamics, we can proceed with the same one-factor analytical solutions. This allows us to analytically specify all MEF matrices, define proper martingale increments and minimize them.[8]

The estimation steps are identical to the one-factor steps taken in Subsubsection 3.3.1. The parameter vector is extended to account for multiple factors, that is

$$\phi^\top = (\gamma^P, \kappa_1, \kappa_2, \kappa_3, \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2, \eta_2),$$

along with the observational noise σ_e. Similarly to the one-factor case we follow with Matlab minimizer of the objective function $f(\phi) = M_T^\top M_t$. We also consider the same extensions — estimating the κ^Q and κ^P coefficients separately and extending the MEF methodology by estimation $\mathcal{E}_2$ with rescaled matrix Ψ_t^{-1} such that diagonal elements sum up to one.

[8]The derived three-factor matrices can be found in the Appendix C.

4 Simulation Study

In order to better understand the MEF method and its applicability to the ATSM we perform an extensive simulation study for the one-factor model. In Subsection 4.1, we explain the simulation setup. In Subsection 4.2, we focus on numerical convergence of different Matlab search algorithms. In Subsection 4.3, we discuss values of the objective function with respect to changes in each coefficient in a great detail. Lastly in Subsection 4.4, we report findings about the objective function and resulting convergence from different Matlab algorithms, while also comparing the results to other estimation methods.

4.1 Setup

We repeatedly simulate the term structure of 480 monthly observations (equal to the size of the empirical dataset) of maturities equal to 3- and 36-months. These maturities are chosen because 3-month is the shortest yield available in our empirical dataset and 36-month is the only unobserved yield used by Hamilton and Wu (2012). Nevertheless, MEF already handles optimal weighting of multiple yields so an extension of the framework is possible. We report all results under the Vasicek parametrization for the direct comparability.

The first 3-month maturity is assumed to be observed with no observational error, whereas 36-month is simulated with the observational error σ_e. We simulate the factor time-series from $dF_t = -\kappa^P F_t dt + dW_t^P$ which implies the conditional normal distribution

$$F_{t|t-\Delta} \sim \mathcal{N}\left(e^{-\kappa^P \Delta} F_{t-\Delta}, \frac{1 - e^{-2\kappa^P \Delta}}{2\kappa^P}\right),$$

and fit the term structure in the form $y_t^{(n)} = \bar{A}^{(n)}/n + \bar{B}^{(n)}/n \cdot F_t$. Note that the short rate is defined as an affine function of the factor in form $r_t = \gamma^P + \eta F_t$. Additionally, we consider the Vasicek model which assumes $\kappa^P = \kappa^Q$ and we therefore drop the upper index for now.

The parameter estimates for the simulation are matched to the one-factor empirical values of monthly periodic data[9]

$$\phi = \begin{pmatrix} \gamma^P = 0.0038 \\ \kappa = 0.0141 \\ \lambda = -0.1179 \\ \eta = 0.0005 \\ \sigma_e \end{pmatrix},$$

with σ_e denoting the observational noise and we consider two values of this noise — a very small value of 1E-6 and 0.0006 which is close to the empirical value. These true population values are also referred to as the data generating process (DGP).

For the MEF starting values we pick "smart" values with an OLS regression on the shortest yield time-series (that is by a slight abuse of notation we set $y_t^{(3)} = r_t$ and estimate the P

[9] If we were to assume annualized data all coefficients would be estimated at higher levels. We use monthly periodic data similarly to Hamilton and Wu (2012).

dynamics of r_t by OLS from $y_t^{(3)}$). This provides a good proxy for γ^P, κ, η to start our search algorithms from. For the market price of risk λ which in the one-factor setting is set to a really high value we take the negative average of the longest yield available and multiply it by 20. This way we consistently start in a reasonable vicinity to the population values. This is necessary especially for the gradient-based solvers.

4.2 Numerical minimization in Matlab

Since MEF is a new methodology it is unclear which numerical optimizers behaves well with the objective function and does not have a tendency to get stucked in local optima. We therefore consider four popular Matlab search algorithms — fmincon, fminunc, fminsearch and patternsearch. Fmincon and fminunc are gradient based algorithms that work with numerically computed Hessian matrices. Fminsearch and patternsearch do not make any assumptions on the gradients of the objective surface and cleverly search in the parameter space for values that may fit the objective function. Fminsearch draws $n+1$ sized simplexes around the n-dimensional starting vector, whereas patternsearch draws four directional steps away from each of the starting parameter values and adjusts the step size when no better value exists until the convergence is achieved.

Fmincon and patternsearch allow us to directly specify parameter bounds that are enforced by a direct comparison of parameter values and not by the transformation of the parameter space.[10] This is an important quality as the part of our optimal weighting matrix ${\psi_t}^\top$ needs to be adjusted by the Jacobian of the given transformation. This is well illustrated with fminsearch and patternsearch for which we assume an exponential transformation of parameters γ^P, κ, and η to ensure their positive values. This transformation scales each column of our optimal weighting matrix ${\psi_t}^\top$ by the J matrix

$$J(\phi) = \begin{pmatrix} \frac{\partial e^{\gamma^P}}{\partial \gamma^P} = e^{\gamma^P} \\ \frac{\partial e^{\kappa}}{\partial \kappa} = e^{\kappa} \\ \frac{\partial \lambda}{\partial \lambda} = 1 \\ \frac{\partial e^{\eta}}{\partial \eta} = e^{\eta} \\ \frac{\partial \sigma_e}{\partial \sigma_e} = 1 \end{pmatrix}.$$

Note that for the σ_e the transformation is unnecessary as this coefficient is not identified by the system by construction and is backed out as implied volatility of the yields observed with error when given a parameter vector $(\gamma^P, \kappa, \lambda, \eta)$.

4.3 Objective function and the sensitivity to parameter changes

Matlab optimizers operate by minimizing an objective function that is a scalar value, thus we optimize $f(\phi) = M_T^\top M_T$. Let us start by examining this objective function in a single simulated term structure. Note that for replicability we set rng(1) in Matlab to fix the pseudo

[10] We set reasonable bounds for coefficients — $\gamma^P \in [0, 0.5]$, $\kappa \in [0, 0.5]$, $\lambda \in [-0.5, 0]$ and $\eta \in [0, 0.5]$.

number generator which assures the same term structure each time we re-run the code with modified settings.

We analyze two settings for the term structure of two maturities 3- and 36- months. Estimation $\mathcal{E}_1$ assumes the proper MEF methodology — the part of the optimal weights in Ψ^{-1} is as stated in the methodology, that is $\Psi_{11}^{-1} = 1/\sigma_\eta^2$ and $\Psi_{22}^{-1} = 1/\sigma_e^2$ with σ_e^2 being computed as implied volatility for the given parameter vector ϕ. Due to troubles in finding the true population values under some conditions we alter the setting by fixing σ_e to the true population value. It turns out that this part of the weighting scheme is causing our objective function to misbehave and such a restriction fixes the issue. Nevertheless, we cannot know the true population value of the observational noise in the empirical setting and that is why we suggest a solution in the second estimation method.

In estimation $\mathcal{E}_2$ we realize that $\Psi_{11}^{-1} = 1/\sigma_\eta^2 = w_1$ approaches one, whereas $\Psi_{22}^{-1} = 1/\sigma_e^2 = w_2$ is of a much bigger magnitude close to 1E+12, 2.78E+6 for σ_e = 1E-6 and $\sigma_e = 0.0006$ respectively. This causes our objective function to blow up. Therefore even in the undesirable case of increasing the observational noise σ_e, while not producing a better fit, we can in fact decrease the value of the objective function, because not enough weight is being put on the factor dynamics equation. We limit this property by rescaling the optimal weights.

We note that this part of optimal weights are inverted volatilities which are always positive. We rescale the Ψ^{-1} part of weights in such a way that the elements sum up to one that is $\Psi_{11}^{-1} = w_1/(w_1 + w_2)$ and $\Psi_{22}^{-1} = w_2/(w_1 + w_2)$ for the setting of just two yields. This has the same effect as the original weights — the more observational noise there is the smaller weight is assigned to that part of MEF equations, giving it a smaller importance weight in the estimation and *vice versa*. It also smoothens out the previously troublesome shape of our objective function, as is illustrated in figures in Subsection 4.4. Additionally, it turns out that estimation $\mathcal{E}_2$ results in the same shape of the objective function in the simulation as the estimation method $\mathcal{E}_1$ with σ_e fixed to its population value, which is exactly what we tried to achieve.

4.4 Results

In Subsubsection 4.4.1, we focus on a one simulated term structure and analyze the behaviour of the MEF objective function under different settings. In Subsubsection 4.4.2, we report the MEF, MCSE and Kalman filter convergence in larger samples. We first repeat the simulation 100 times and estimate the structural parameters with four search algorithms to pick the best one. With the best search algorithm we simulate 1000 times and report the results.

4.4.1 Objective function and the sensitivity to parameter changes

For the analysis of the single simulated term structure we fix all coefficients to their true population values, except for the one that we vary and evaluate our objective function repeatedly. This way we get the information on how do the marginal changes in one of the parameters impact the objective function for MEF. We illustrate the importance of signal-to-noise ratio for MEF methodology by considering two noise settings — extremely low noise σ_e = 1E-6, and moderate (close to empirical) noise level $\sigma_e = 0.0006$.

Figure 1 introduces the behavior of the objective function with respect to changes in all coefficients that enter the estimation $\mathcal{E}_1$, that is the proper MEF weighting. Part **(a)** of the figure shows the behavior of the objective function in a low noise setting. We can observe that we correctly find the values of γ^P, λ, η, whereas the coefficient κ is not possible to locate, because the objective function is lower for κ-values extremely positive, or close to zero. This causes our search algorithms to slide to the boundaries and fail to find all population parameters. The variance around the optimal coefficient κ gets so small that the optimal coefficient is effectively a single point, yet the objective function is defined at broader vicinity, which causes tremendous numerical challenges for the proper MEF methodology.

Part **(b)** of Figure 1 considers the same setting as **(a)** but introduces a noisier environment with $\sigma_e = 0.0006$. We can already see a dramatic improvement over the initial setting in the shape of the objective function. Here we are more likely to find the optima, provided a good starting point. The issue remains that we can never know how much noise-to-signal we are dealing with. The methodology proposed by Christensen et al. (2016) is highly dependent on this property in the ATSM setting. That is why we propose an extension to the original methodology by rescaling the optimal weighting matrix, that is estimation $\mathcal{E}_2$.

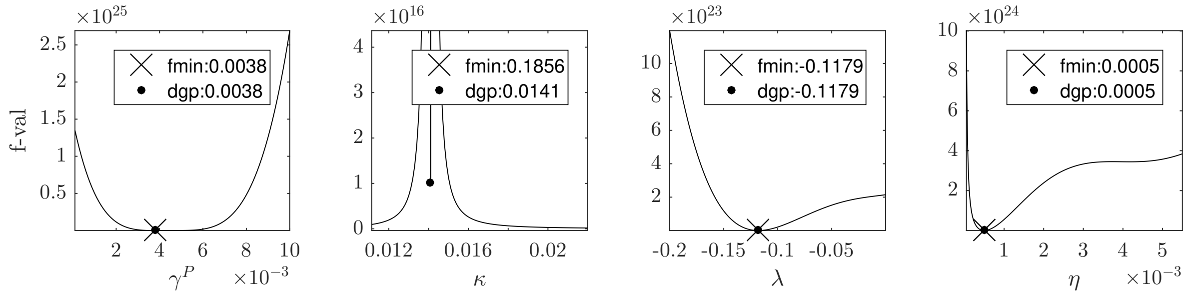
Part **(c)** of Figure 1 captures the low-noise environment with rescaled part of optimal weights according to estimation $\mathcal{E}_2$. We can observe that the convergence in parameters remains the same in all parameters except for κ, where we observe the flattening of the objective function. This way we are much more likely to find the true optimal values of all parameters without the dependence on the amount of noise in the data.

Part **(d)** of Figure 1 adds further evidence in favor of $\mathcal{E}_2$ in a moderately noisy setting. We can again see that the objective function keeps the same shape as in the original $\mathcal{E}_1$ setting except for κ coefficient that is much easier to find. We also note that the functional minima in the noisy setting $\mathcal{E}_1$ **(b)** and $\mathcal{E}_2$ **(d)** are exactly matched. The estimation $\mathcal{E}_2$ seems to be much more robust, while keeping all the properties of MEF methodology that cleverly adjusts the weights based on the amount of the implied signal-to-noise ratio.

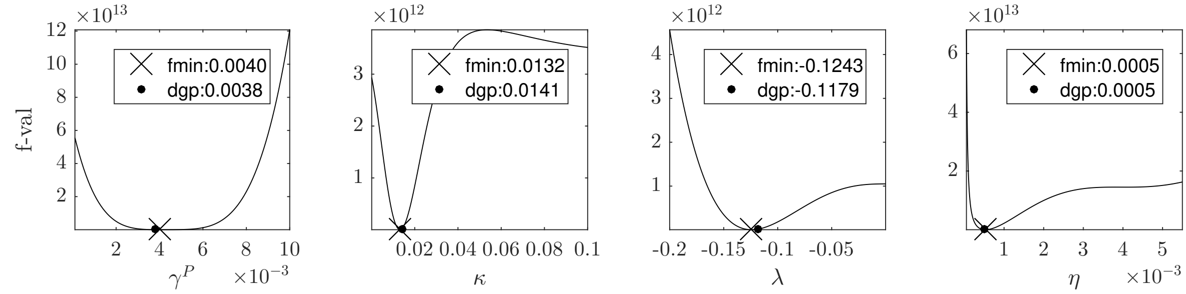
4.4.2 Large sample convergence

After examining the behavior of the objective function we look at MEF convergence in large samples. We start by simulating a representative sample of 100 term structures and estimate $\mathcal{E}_1$, $\mathcal{E}_2$ with four suitable Matlab algorithms to see their behavior. Additionally, we estimate the same term structures with Kalman filter and MCSE for a direct comparison. Note that by construction, MCSE separates estimated measures and assumes a time-varying market price of risk which implies that $\kappa^P \neq \kappa^Q$. This also has an effect on the estimated long run mean of short rate γ^P and the market price of risk λ. Nevertheless, the pricing κ^Q coefficient corresponds to κ of our setting closely and we therefore report this value for a quick comparison of methods. We again consider a low noise environment (σ_e = 1E-6) and a moderately noise setting (σ_e = 0.0006).

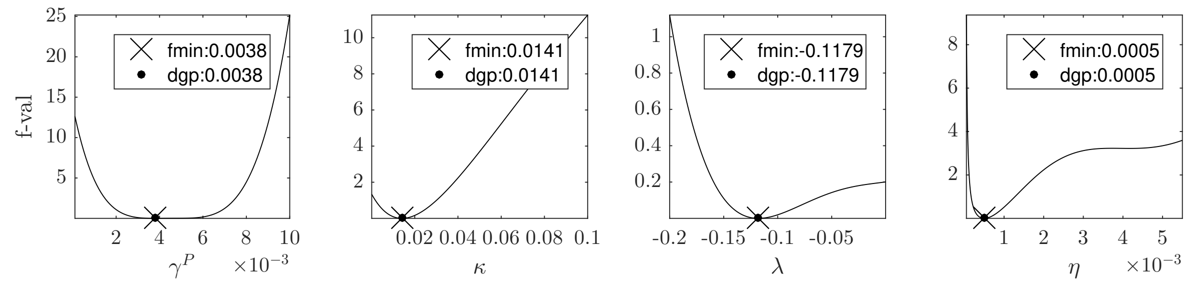
Table 2 and Table 3 summarize the results of the algorithm comparison for all considered settings. Out of the four different search algorithms `fminunc` performs the best (in terms of the closes mean and the lowest average variance of estimates) and in the quickest time for MEF



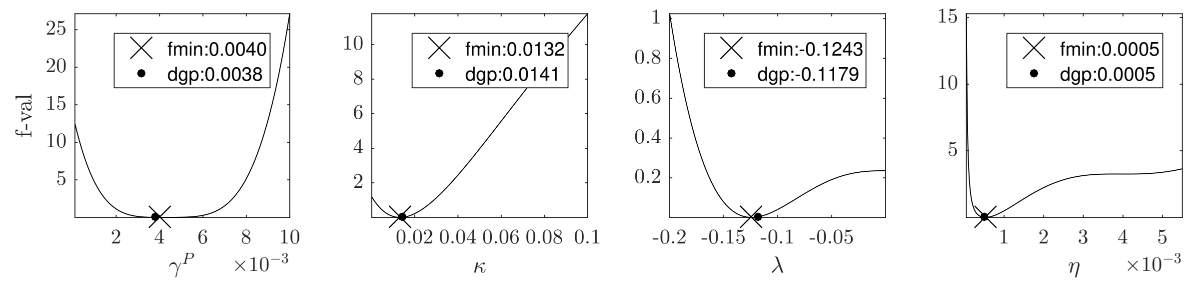
(a) $\mathcal{E}_1$, population with $\sigma_e = $ 1E-6



(b) $\mathcal{E}_1$, population with $\sigma_e = 0.0006$



(c) $\mathcal{E}_2$, population with $\sigma_e = $ 1E-6



(d) $\mathcal{E}_2$, population with $\sigma_e = 0.0006$

Figure 1: MEF estimation $\mathcal{E}_1$ and $\mathcal{E}_2$, f-val denotes the value of objective function $f(\phi) = M_T^\top M_T$
MEF $\mathcal{E}_1$ uses the proper MEF weights, whereas MEF $\mathcal{E}_2$ uses the rescaling of Ψ^{-1} such that diagonal elements sum up to one.

estimations $\mathcal{E}_1$, $\mathcal{E}_2$ in both the noisy and the low-noise environments. Second best algorithm for this type of problem is patternsearch which provides similarly good estimates and is able to jump over local optima and can in some situations be more robust. The computational time is much longer and we therefore prefer fminunc. Fminsearch provides good estimates only in a moderately noisy environments with estimation $\mathcal{E}_2$ and still struggles to properly identify the volatility η of factors, while taking much larger computational time than other algorithms. We conclude fminsearch to be not sufficiently robust for this type of estimation. Fmincon has the largest tendency to slide to boundary values and provides the worst estimates of all considered

Table 2: DGP_1, Kalman filter, MCSE and MEF estimates, 100 simulations

This table reports the coefficients used for creating the population (DGP), "smart" OLS start values for MEF, and the resulting mean, median and variance of 100 term structures estimated by Kalman filter, MCSE, and MEF. MEF estimation $\mathcal{E}_1$ assumes the proper weighting, whereas $\mathcal{E}_2$ assumes the sum of weighting elements Ψ^{-1} to be rescaled to one. MEF estimations with `fminunc`, `fminsearch` Matlab algorithms use the exponential transformation of γ^P, κ, and η. MEF estimations with `fmincon`, `patternsearch` Matlab algorithms use the parameter bounds $\gamma^P \in [0, 0.5]$, $\kappa \in [0, 0.5]$, $\lambda \in [-0.5, 0]$ and $\eta \in [0, 0.5]$. For MCSE we report κ^Q as it corresponds to the κ used for pricing.

		γ^P	κ	λ	η	σ_e	Iterations	Time [s]
DGP_1	true value	0.0038	0.0141	−0.1179	0.0005	0.000 001	-	-
Kalman filter	mean	0.0035	0.0141	−0.1195	0.0005	0.000 001	40	13
	median	0.0038	0.0141	−0.1163	0.0005	0.000 001	39	12
	var	1.70E-6	2.67E-12	1.07E-3	5.60E-10	6.04E-16	-	-
MCSE	mean	0.0035	0.0140	−0.1282	0.0005	0.000 001	-	<1
	median	0.0036	0.0140	−0.1244	0.0005	0.000 001	-	<1
	var	1.95E-6	2.62E-12	1.58E-3	2.71E-10	9.65E-16		-
OLS start	mean	0.0035	0.0242	−0.0892	0.0005	-	-	-
	median	0.0037	0.0210	−0.0928	0.0005	-	-	-
				Panel A: Estimation $\mathcal{E}_1$				
MEF patternsearch	mean	0.0031	0.1138	−0.5000	0.0005	0.001 125	1891	124
	median	0.0000	0.0322	0.0429	0.0001	0.000 000	81 032	403
	var	1.02E-2	1.92E-1	−1.15E-1	9.05E-3	9.69E-4	-	-
MEF fmincon	mean	0.0027	0.0999	−0.0548	0.0054	0.000 911	100	11
	median	0.0005	0.0454	0.0234	0.0001	0.000 000	921 641	3878
	var	3.19E-3	6.86E-2	−9.40E-1	6.23E-4	7.90E-4	-	-
MEF fminunc	mean	0.0002	0.0611	−1.2561	0.0003	0.000 746	63	6
	median	0.0001	0.0066	1.0267	0.0000	0.000 000	937	7
	var	5.01E-1	5.49E-1	3.76E-1	5.98E-2	1.00E-3	-	-
MEF fminsearch	mean	0.1100	0.0639	0.0585	0.0029	0.000 858	2096	123
	median	1.8503	0.5882	1.6164	0.0108	0.000 000	5 021 747	2287
	var	3.47E-3	1.40E-2	−1.28E-1	4.93E-4	1.00E-6	-	-
				Panel B: Estimation $\mathcal{E}_2$				
MEF patternsearch	mean	0.0036	0.0142	−0.1261	0.0010	0.000 006	473	36
	median	0.0037	0.0141	−0.1205	0.0005	0.000 001	321	23
	var	2.17E-6	2.14E-6	1.76E-3	4.04E-6	1.18E-9	-	-
MEF fmincon	mean	0.0386	0.1557	−0.1714	0.0414	0.000 901	129	11
	median	0.0194	0.1001	−0.1051	0.0166	0.000 908	115	11
	var	5.00E-3	2.00E-2	2.07E-2	2.74E-3	3.58E-7	-	-
MEF fminunc	mean	0.0035	0.0141	−0.1304	0.0005	0.000 001	18	2
	median	0.0037	0.0141	−0.1260	0.0005	0.000 001	18	2
	var	1.88E-6	2.69E-12	1.69E-3	3.77E-10	9.81E-16	-	-
MEF fminsearch	mean	0.0052	0.0138	0.0400	0.0004	0.000 012	140	4
	median	0.0052	0.0141	−0.0911	0.0004	0.000 001	117	3
	var	2.52E-6	3.90E-6	1.60+00	7.78E-8	5.13E-9	-	-

algorithms.

With the best algorithm in hand — that is `fminunc` — we continue by simulating 1000 term structures in the low noise environment and 1000 term structures in the noisy environment. Table 4 summarizes the results. We report the mean, median and variance of estimates for the Kalman filter, MCSE and MEF methods. We can see in Panel A of Table 4 that the proper MEF methodology $\mathcal{E}_1$ in the low noise environment fails to identify the population optima. Note that both Kalman filter and MCSE had no such struggle with the estimation. However, when estimating with the adjusted weights $\mathcal{E}_2$ we observe a large improvement in the convergence and obtain very good results.

Table 3: DGP_2, Kalman filter, MCSE and MEF estimates, 100 simulations

This table reports the coefficients used for creating the population (DGP), "smart" OLS start values for MEF, and the resulting mean, median and variance of 100 term structures estimated by Kalman filter, MCSE, and MEF. MEF estimation $\mathcal{E}_1$ assumes the proper weighting, whereas $\mathcal{E}_2$ assumes the sum of weighting elements Ψ^{-1} to be rescaled to one. MEF estimations with fminunc, fminsearch Matlab algorithms use the exponential transformation of γ^P, κ, and η. MEF estimations with fmincon, patternsearch Matlab algorithms use the parameter bounds $\gamma^P \in [0, 0.5]$, $\kappa \in [0, 0.5]$, $\lambda \in [-0.5, 0]$ and $\eta \in [0, 0.5]$. For MCSE we report κ^Q as it corresponds to the κ used for pricing.

		γ^P	κ	λ	η	σ_e	Iterations	Time [s]
DGP_2	true value	0.0038	0.0141	−0.1179	0.0005	0.0006	-	-
Kalman filter	mean	0.0037	0.0124	−0.1289	0.0005	0.000 45	35	11
	median	0.0037	0.0124	−0.1293	0.0005	0.000 45	35	10
	var	8.27E-8	1.28E-6	2.07E-4	6.84E-10	2.00E-10		
MCSE	mean	0.0035	0.0140	−0.1282	0.0005	0.000 60	-	<1
	median	0.0036	0.0139	−0.1251	0.0005	0.000 60	-	<1
	var	1.95E-6	9.48E-7	1.61E-3	2.72E-10	3.47E-10		
OLS start	mean	0.0035	0.0242	−0.0892	0.0005	-	-	-
	median	0.0037	0.0210	−0.0928	0.0005	-	-	-
				Panel A: Estimation $\mathcal{E}_1$				
MEF patternsearch	mean	0.0038	0.1109	−0.0962	0.0052	0.000 85	741	60
	median	0.0038	0.0145	−0.0973	0.0005	0.000 61	460	38
	var	1.89E-6	3.18E-2	1.65E-3	6.99E-5	2.15E-7		
MEF fmincon	mean	0.0091	0.0315	−0.0761	0.0024	0.000 68	341	27
	median	0.0040	0.0141	−0.0710	0.0010	0.000 61	31	5
	var	4.60E-4	4.19E-3	1.73E-3	1.63E-5	4.13E-8	-	-
MEF fminunc	mean	0.0038	0.0141	−0.1205	0.0005	0.000 60	28	4
	median	0.0038	0.0140	−0.1153	0.0005	0.000 60	18	3
	var	1.91E-6	1.16E-6	1.71E-3	1.57E-9	4.03E-10	-	-
MEF fminsearch	mean	0.0626	0.0136	0.0107	0.0008	0.000 61	2947	72
	median	0.0038	0.0140	−0.0823	0.0007	0.000 60	2325	56
	var	2.33E-1	7.42E-6	1.61E+0	2.95E-7	4.05E-9	-	-
				Panel B: Estimation $\mathcal{E}_2$				
MEF patternsearch	mean	0.0039	0.0139	−0.1214	0.0009	0.000 60	404	38
	median	0.0038	0.0140	−0.1202	0.0005	0.000 60	264	25
	var	2.40E-6	5.19E-6	1.73E-3	4.30E-6	4.74E-10	-	-
MEF fmincon	mean	0.0647	0.1798	−0.0817	0.0709	0.001 29	148	16
	median	0.0243	0.2165	−0.0836	0.0647	0.001 28	122	13
	var	9.18E-3	1.63E-2	3.24E-4	3.88E-3	2.48E-7	-	-
MEF fminunc	mean	0.0035	0.0141	−0.1286	0.0005	0.000 60	23	3
	median	0.0037	0.0140	−0.1253	0.0005	0.000 60	19	2
	var	1.87E-6	9.78E-7	1.62E-3	1.05E-9	3.53E-10	-	-
MEF fminsearch	mean	0.0037	0.0141	−0.0833	0.0009	0.000 61	2787	77
	median	0.0037	0.0140	−0.0801	0.0007	0.000 60	2046	51
	var	2.38E-6	1.19E-5	2.24E-3	3.49E-7	1.42E-9	-	-

Panel B of Table 4 reports the results from the noisy environment. We can see that MEF methodology $\mathcal{E}_1$ catches up and provides a better convergence, but still has a few very extreme outliers. On the other hand, the estimation with the adjusted weights $\mathcal{E}_2$ results in similar parameter distributions (can be examined in reported histograms) close to the population values while removing extreme outliers. Note how do Kalman filter and MCSE deal with the noisy environment. MCSE locates the true population values very well. On the other hand, the Kalman filter tries to de-noise the measurement and the transition equations and it indeed succeeds, estimating the measurement noise with $\hat{\sigma}_e = 0.00045$ which is lower than for other estimating methods and the true population value $\sigma_e = 0.0006$. Similar findings are observed

Table 4: $\text{DGP}_{1,2}$, Kalman filter, MCSE and MEF estimates, 1000 simulations

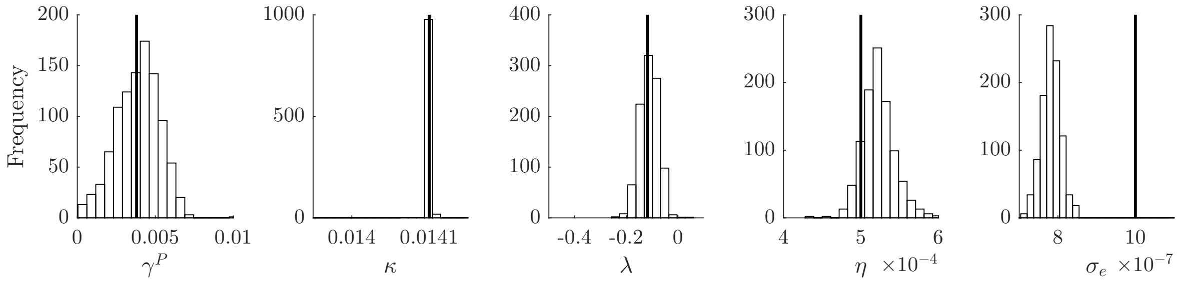
This table reports the coefficients used for creating the population (DGP), "smart" OLS start values for MEF, and the resulting mean, median and variance of 1000 term structures estimated by Kalman filter, MCSE, and MEF. MEF estimation $\mathcal{E}_1$ assumes the proper weighting, whereas $\mathcal{E}_2$ assumes the sum of weighting elements Ψ^{-1} to be rescaled to one. MEF estimations use fminunc Matlab algorithms with the exponential transformation of γ^P, κ, and η. For MCSE we report κ^Q as it corresponds to the κ used for pricing.

		γ^P	κ	λ	η	σ_e	Iterations	Time [s]
				PANEL A: LOW NOISE ENVIRONMENT $\sigma_e = 1\text{E-}6$				
DGP_1	true value	0.0038	0.0141	−0.1179	0.000 50	0.000 001 0	-	-
OLS start	mean	0.0039	0.0238	−0.0939	0.000 49	-	-	-
	median	0.0039	0.0212	−0.0941	0.000 49	-	-	-
Kalman filter	mean	0.0039	0.0141	−0.1108	0.000 52	0.000 000 8	41	12
	median	0.0040	0.0141	−0.1089	0.000 52	0.000 000 8	39	11
	var	1.89E-6	3.75E-9	1.33E-3	1.53E-9	3.24E-15	-	-
MCSE	mean	0.0039	0.0140	−0.1171	0.000 49	0.000 001 0	-	<1
	median	0.0039	0.0140	−0.1165	0.000 49	0.000 001 0	-	<1
	var	2.46E-6	2.95E-12	2.01E-3	2.50E-10	9.95E-16	-	-
MEF $\mathcal{E}_1$	mean	0.0074	0.0855	−0.7330	0.001 92	0.000 860 4	67	7
	median	0.0003	0.0630	−1.1739	0.000 33	0.000 774 4	65	7
	var	6.95E-4	9.50E-3	1.11E+0	3.50E-5	1.14E-7	-	-
MEF $\mathcal{E}_2$	mean	0.0039	0.0141	−0.1191	0.000 48	0.000 001 0	18	1
	median	0.0039	0.0141	−0.1178	0.000 48	0.000 001 0	18	1
	var	2.40E-6	3.04E-12	2.28E-3	4.37E-10	9.94E-16	-	-
				PANEL B: MODERATE NOISE ENVIRONMENT $\sigma_e = 0.0006$				
DGP_2	true value	0.0038	0.0141	−0.1179	0.000 50	0.000 600	-	-
OLS start	mean	0.0039	0.0238	−0.0939	0.000 49	-	-	-
	median	0.0039	0.0212	−0.0941	0.000 49	-	-	-
Kalman filter	mean	0.0038	0.0125	−0.1264	0.000 46	0.000 454	35	9
	median	0.0038	0.0126	−0.1260	0.000 46	0.000 454	35	9
	var	8.70E-8	1.43E-6	1.88E-4	5.83E-10	2.00E-10	-	-
MCSE	mean	0.0039	0.0140	−0.1172	0.000 49	0.000 600	-	<1
	median	0.0039	0.0140	−0.1167	0.000 49	0.000 600	-	<1
	var	2.46E-6	1.07E-6	2.04E-3	2.50E-10	3.58E-10	-	-
MEF $\mathcal{E}_1$	mean	0.0039	0.0141	−0.1610	0.000 49	0.000 602	28	4
	median	0.0039	0.0141	−0.1154	0.000 48	0.000 600	18	3
	var	3.14E-6	2.02E-6	2.34E+0	3.03E-9	1.52E-9	-	-
MEF $\mathcal{E}_2$	mean	0.0039	0.0141	−0.1169	0.000 49	0.000 600	25	2
	median	0.0040	0.0141	−0.1150	0.000 49	0.000 600	18	2
	var	2.26E-6	1.10E-6	1.92E-3	1.81E-9	3.58E-10	-	-

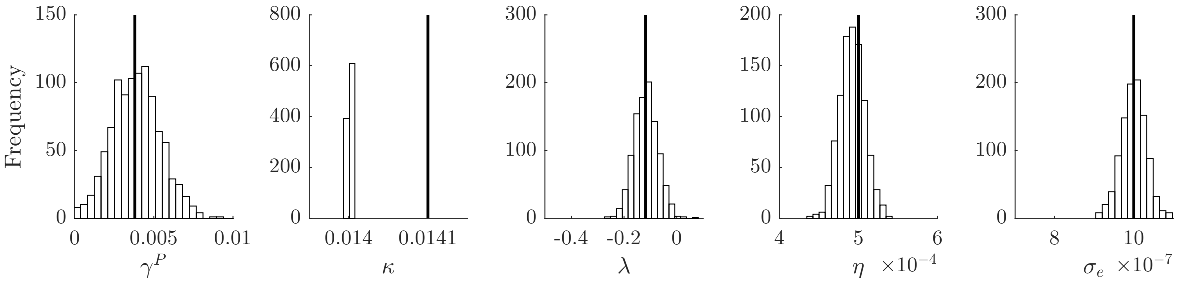
for the transition noise η.

Further evidence is provided by added histograms in Figure 2 and Figure 3 that show how well does the MEF estimation $\mathcal{E}_2$ compare to the best estimation methods. The proper MEF estimation $\mathcal{E}_1$ fails to estimate the structural parameters in the environment with low observational noise (Figure 2). For the noisy environment (Figure 3), we can observe that the estimations $\mathcal{E}_1$ (part **(c)**) and the estimation $\mathcal{E}_2$ (part **(d)**) both hold very similar distributions for all structural parameters, while $\mathcal{E}_1$ having extreme outliers that affect the mean, median, and variance measures reported in Table 4.

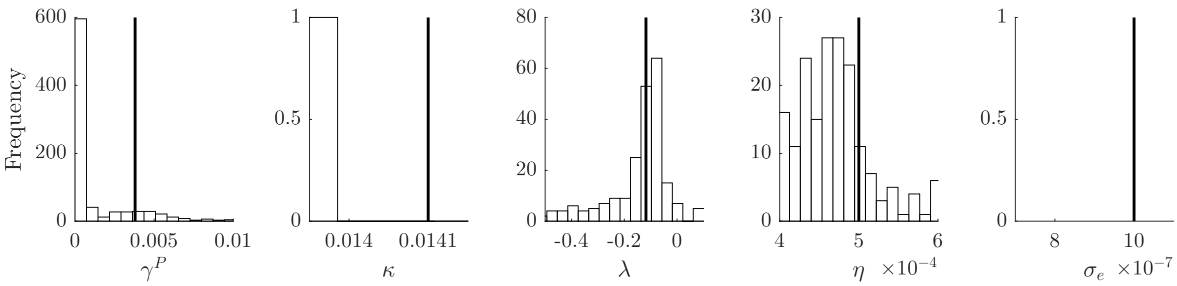
To summarize, we have estimated two MEF estimation settings $\mathcal{E}_1$, $\mathcal{E}_2$ in a low noise and noisy environment next to the traditional estimation methods of Kalman filtering and MCSE. Out of four different search algorithms for MEF the fminunc with the exponential transforma-



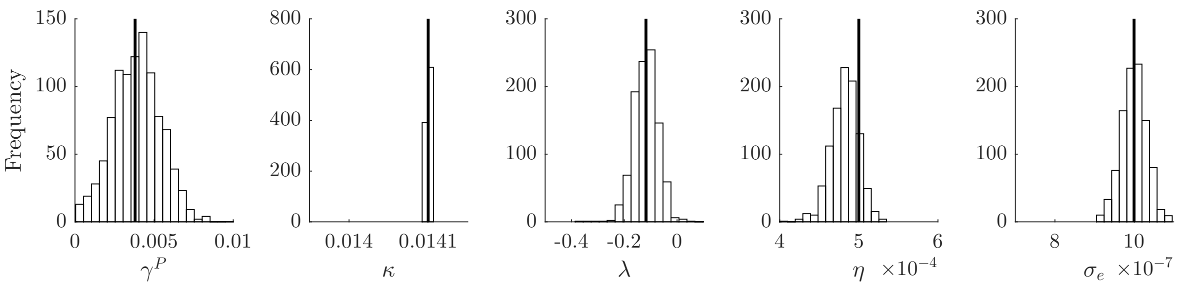
(a) Kalman filter



(b) MCSE



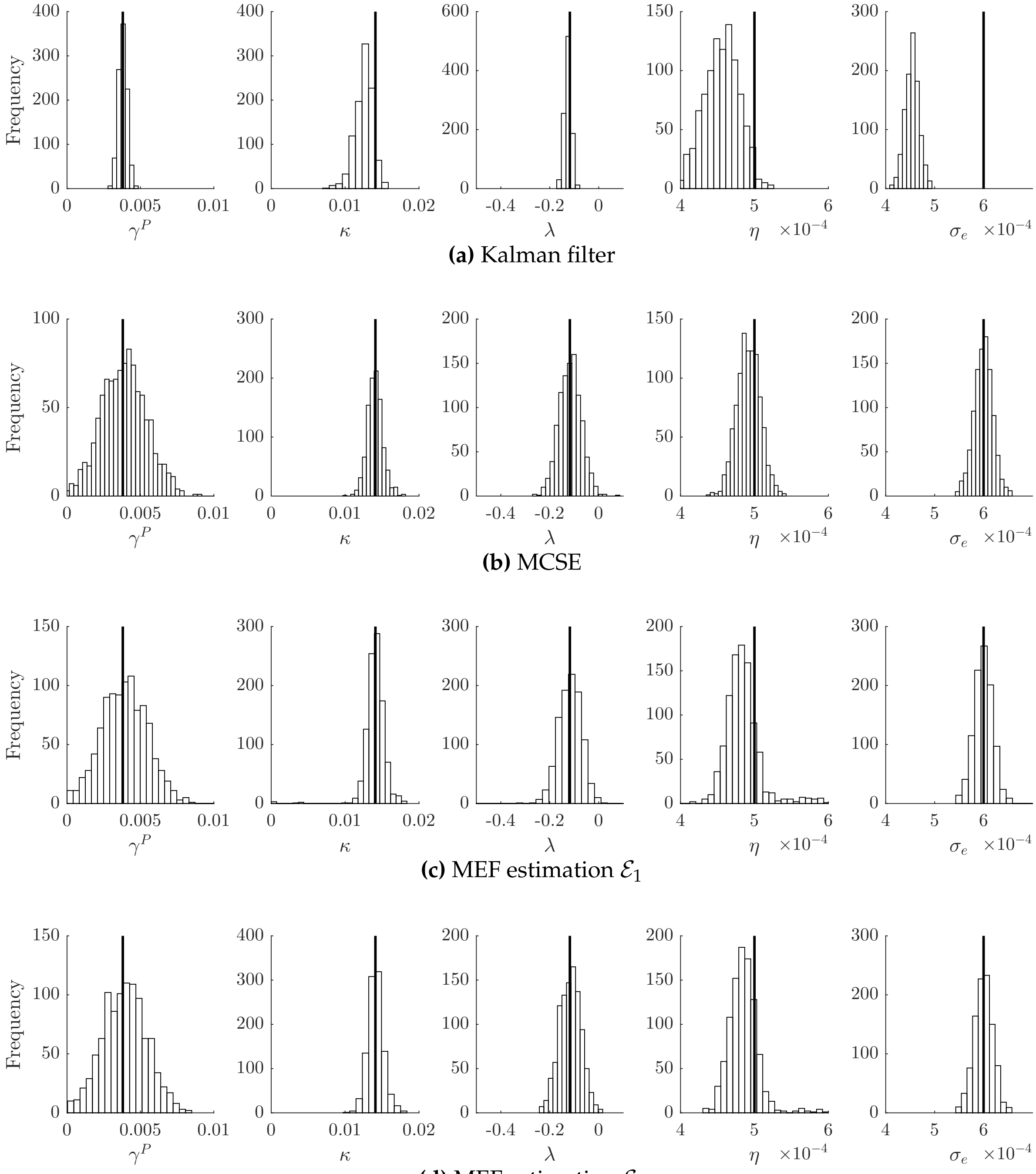
(c) MEF estimation $\mathcal{E}_1$



(d) MEF estimation $\mathcal{E}_2$

Figure 2: Histograms for 1000 simulations with low noise $\sigma_e = $ 1E-6

We report results for Kalman filter, MCSE, MEF estimation $\mathcal{E}_1$ with proper weights, MEF estimation $\mathcal{E}_2$ with adjusted weights. The thick vertical line corresponds to population values $\gamma^P = 0.0038$, $\kappa = 0.0141$, $\lambda = -0.1179$, $\eta = 0.0005$, $\sigma_e = $ 1E-6. Note that x-axis is normalized for each estimation method to the same lower and upper limit and thus may not include extreme outliers.



(a) Kalman filter

(b) MCSE

(c) MEF estimation $\mathcal{E}_1$

(d) MEF estimation $\mathcal{E}_2$

Figure 3: Histograms for 1000 simulations with moderate noise $\sigma_e = 0.0006$

We report results for Kalman filter, MCSE, MEF estimation $\mathcal{E}_1$ with proper weights, MEF estimation $\mathcal{E}_2$ with adjusted weights. The thick vertical line corresponds to population values $\gamma^P = 0.0038$, $\kappa = 0.0141$, $\lambda = -0.1179$, $\eta = 0.0005$, $\sigma_e = 0.0006$. Note that x-axis is normalized for each estimation method to the same lower and upper limit and thus may not include extreme outliers.

tion of γ^P, κ, and η performs the best (in terms of the closest mean and the lowest average variance of estimated coefficients) and in the quickest time, beating the average computational time of Kalamn filtering by 2-4 times, yet still staying behind the quickest MCSE approach. MEF estimation $\mathcal{E}_2$ is strictly faster in convergence than the proper MEF estimation $\mathcal{E}_1$ which can be addressed to the tendency of $\mathcal{E}_1$ to occasionally slide to extreme values.

We identify the issue of MEF methodology $\mathcal{E}_1$ in the low noise environment and suggest the solution to rescale the optimal weights with estimation $\mathcal{E}_2$ which solves the convergence issue in the low noise environment. Both Kalman filter and MCSE in the low noise environment behave very well. Comparing the results of all methods, we conclude that MCSE is the quickest and the most robust method of estimation, closely followed by MEF $\mathcal{E}_2$ with adjusted weights. Kalman filter also produces good estimates but does undervalue certain coefficient values to obtain a better theoretical fit, which however does not match the population values as closely as other methods. The proper MEF $\mathcal{E}_1$ is not sufficiently robust with respect the amount of observed noise and is be placed last in this comparison.[11]

[11] Similar simulation results hold for the three-factor model and can be examined in the Appendix C2.

5 Empirical Study

We have established the validity of MEF in Section 4. We continue by examining the empirical applicability. In Subsection 5.1, we summarize the setup used for the empirical estimation. In Subsection 5.2, we describe the chosen dataset. Lastly in Subsection 5.3, we report one- and three-factor structural parameters and the corresponding empirical and theoretical fit.

5.1 Setup

Following the findings from the simulation study we fit the one-factor model in the empirical setting in four ways — using the Kalman filter, MCSE, MEF method $\mathcal{E}_1$ with proper MEF weights and MEF method $\mathcal{E}_2$ which rescales weights of Ψ^{-1} to sum up to one and is more robust. Moreover, following the Hamilton and Wu (2012) we only consider one additional yield next to the yields that correspond to factors, even though the methodology already handles the proper optimal weighting for multiple additional yield equations. We use the 3-month yield observed with no observational error and 36-month yield observed with the observational error.

Next to the one-factor model we also estimate the three-factor model with the Kalman filter, the MCSE and the MEF estimations $\mathcal{E}_1$ and $\mathcal{E}_2$. We diagonalize κ matrices such that our analytical results apply. Of the observed yields we again consider just one additional yield next to the three yields representing factors. Following Hamilton and Wu (2012) we use the set of 3-, 12-, 60- months observed yields with no observational error and 36-month maturity for the yield with the observational noise.

For obtaining the empirical estimates we use the same setting as in the simulation. That is for Kalman filter we minimize $-2\ln L$ with minimizer fminunc with exponential transformation of γ^P, κ, and η parameters. MCSE method follows the same estimation procedure as in the original paper with fsolve algorithms and manually enforced restrictions. For MEF methodology we use the fminunc[12] with exponential transformation of γ^P, κ, and η parameters. Note that since our optimal weighting matrix $\psi^\top$ contains partial derivatives, we need to properly scale this matrix by the part of the Jacobian of the transformation.[13] We start the search algorithms from the "smart" OLS coefficients.

For the model comparison we use the mean absolute error (MAE) metric of the difference between the observed data and the theoretical curve fitted by the estimated parameters. The lower the value of MAE the closer to the empirical surface we are and the better fit we have obtained. Asymptotic standard errors are also reported for each method of estimation. However, we note that these errors are valid if and only if the model specifications are correct, which remains uncertain. Noting that all used methods are slightly different in specifications, we avoid drawing conclusions based solely on these standard error terms. For estimation $\mathcal{E}_2$ we use the proper asymptotic theory derived for $\mathcal{E}_1$, since we view the change of involved matrices as a type of rescaling that helps with the numerical convergence.

[12]The following set of options was used for the algorithm options = optimoptions('fminunc','Algorithm','quasi-newton','HessUpdate','bfgs','MaxFunctionEvaluations',10000,'MaxIterations',5000,'Disp','off','TolFun', 1e-20, 'TolX', 1e-20);. Note that these are mostly default fminunc options.

[13]Details are summarized in Subsection 4.2.

5.2 Data

To verify the methodology introduced in Section 3 we examine the empirical fit on the chosen dataset. The data contains monthly time-series of zero yields Fama and Bliss (1987) forward rates, constructed from the CRSP Monthly Treasury Cross-Sectional File by Dijk et al. (2014).[14] The data contains 480 time-series observations for 17 different maturities, ranging from 3 to 120 months. The oldest yields are from 30 January 1970, while the latest from 31 December 2009. We normalize the yields to contain periodic monthly observations.[15]

5.3 Results

5.3.1 One-factor results

Table 5 captures the results from the one-factor estimation. We estimate the one-factor parameters with the Kalman filter, the MCSE, the MEF method $\mathcal{E}_1$, the MEF method $\mathcal{E}_2$ and also the extended MEF methods that unrestrict the condition $\kappa^Q = \kappa^P$. We therefore denote κ^Q for the pricing coefficient and κ^P for the risk-neutral factor dynamics coefficient and have the model directly comparable to the MCSE. We note that both the $\mathcal{E}_1$ and the $\mathcal{E}_2$ produce identical estimates, because we are estimating in a moderately noisy environment as was explored in the simulation study. Still, $\mathcal{E}_2$ is computationally faster than $\mathcal{E}_1$.

With the parametrization $\mathcal{S}_3$ the weighting scheme puts much bigger weight on the Q dynamics for the MEF and that is why we report the κ coefficient for $\mathcal{E}_1$ and $\mathcal{E}_2$ in the κ^Q column. When we relax this condition and let the factor dynamics κ^P to vary independently on the pricing κ^Q and estimate with our $\mathcal{E}_{1,2}$ methods we reproduce the same coefficients as Hamilton and Wu (2012) with the MCSE. Note that we can directly compete in terms of computational times with MCSE as well. This sort of estimation is numerically very challenging and MEF holds its ground next to the MCSE method very well. The one-factor fit can be also examined in the Figure 4.

5.3.2 Three-factor results

The empirical estimation is much more dependent on the chosen starting point than in the simulated setup. We therefore adopt the "smart" OLS start where we estimate the $\gamma_i, \kappa_i, \eta_i$ from the yields themselves and use them for starting points. For the market price of risk we assume the negative of the mean of 60-month yields.

Table 6 captures the results from three-factor estimation similarly to the one-factor results. We note that both $\mathcal{E}_1$ and $\mathcal{E}_2$ yield slightly different estimates and the proposed $\mathcal{E}_2$ outperforms the $\mathcal{E}_1$ in terms of the MAE. The computational times in the three-factor setting raise up to 53s, 170s, 160s for the Kalman filter, the MEF $\mathcal{E}_1$ and the MEF $\mathcal{E}_2$ respectively, while the MCSE is still able to solve the three-factor problem instantly.

We note that our weighting scheme with parametrization $\mathcal{S}_3$ puts much bigger weight on the Q dynamics in our MEF setup and that is why we report the κ coefficient for $\mathcal{E}_1$ and $\mathcal{E}_2$ in

[14] Available: `http://qed.econ.queensu.ca/jae/datasets/van_dijk002/`

[15] That is having annualized yields we divide each observation by 12 to get monthly periodic yields.

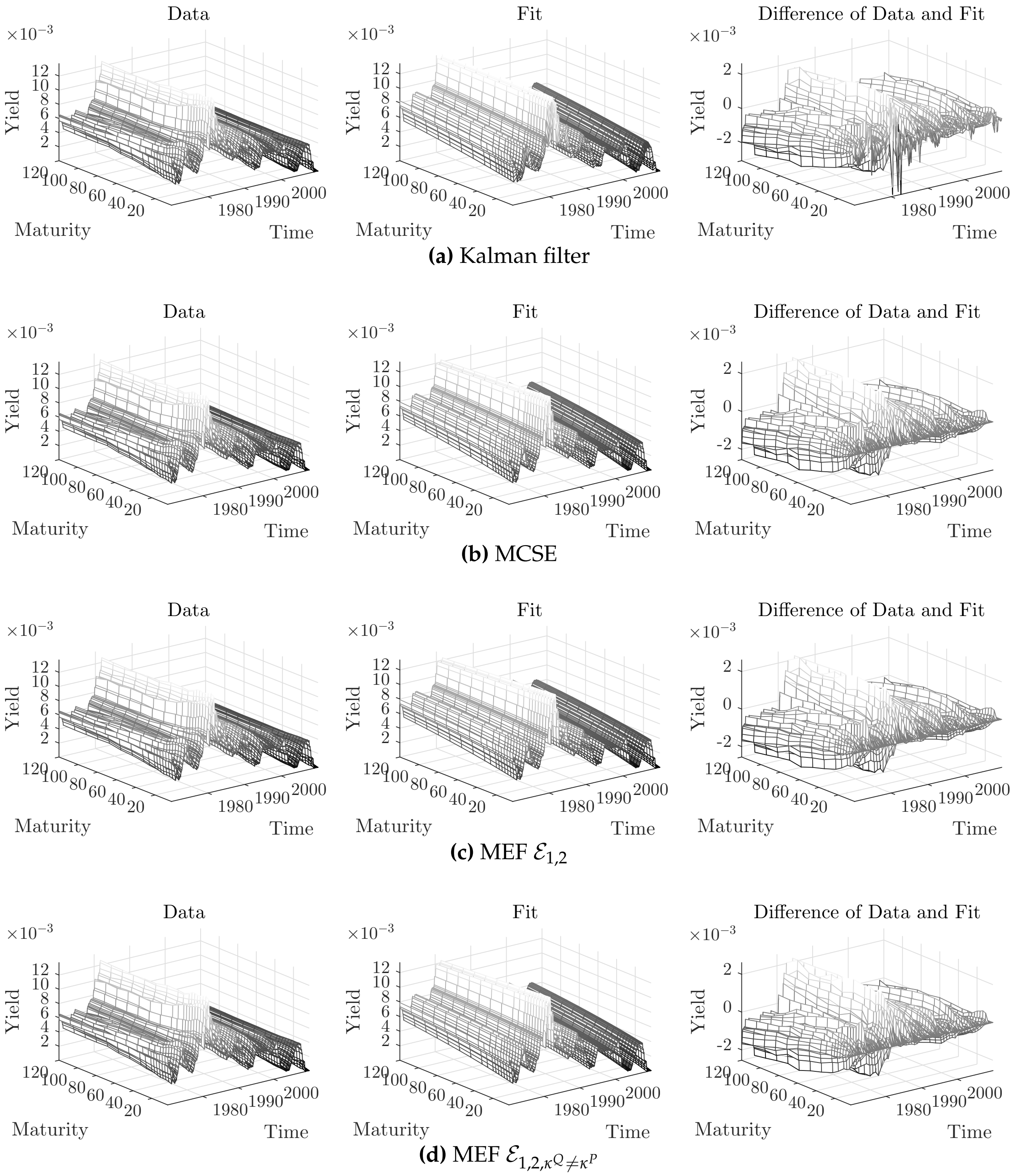


Figure 4: One-factor fit on empirical dataset
Coefficients are estimated with Kalman filter, MCSE, MEF $\mathcal{E}_{1,2}$ (plotted together due to the same coefficient values), and MEF $\mathcal{E}_{1,2,\kappa^Q\neq\kappa^P}$ (plotted together due to the same coefficient values).

Table 5: One-factor model estimates

The table that reports "smart" OLS start values for MEF, and the resulting γ^P, κ^P, κ^Q, λ, η and σ_e estimated by the Kalman filter, the MCSE, and the MEF. MEF $\mathcal{E}_1$ assumes the proper weighting, whereas $\mathcal{E}_2$ assumes the sum of weighting elements Ψ^{-1} to be rescaled to one. MEF use fminunc Matlab algorithm with the exponential transformation of γ^P, κ, and η. The overall fit is measured as the average MAE across all observed yields. Asymptotic standard errors are provided in the parentheses.

	γ^P	κ^Q	κ^P	λ	η	σ_e	Iter	Time [s]	MAE
OLS start	0.0038	0.0070	0.0140	-0.1100	0.000 47	-	-	-	-
Kalman filter	0.0060	0.0050	-	-0.1172	0.000 34	0.000 45	117	30	5.57E-4
	(4.3E-4)	(7.6E-4)	-	(1.0E-2)	(1.8E-5)	(1.2E-5)			
MCSE	0.0038	0.0077	0.0140	-0.1192	0.000 47	0.000 68	-	<1	5.21E-4
	(1.7E-3)	(8.8E-4)	(8.4E-3)	(2.8E-2)	(1.5E-5)	(2.2E-5)			
MEF $\mathcal{E}_1$	0.0038	0.0077	-	-0.1220	0.000 46	0.000 68	19	3	5.37E-4
	(5.6E-5)	(7.7E-3)	-	(4.5E-7)	(6.3E-5)	-			
MEF $\mathcal{E}_2$	0.0038	0.0077	-	-0.1213	0.000 47	0.000 68	18	1	5.37E-4
	(5.6E-5)	(7.7E-3)	-	(4.5E-7)	(6.4E-5)	-			
MEF $\mathcal{E}_{1,\kappa^Q \neq \kappa^P}$	0.0038	0.0070	0.0142	-0.1173	0.000 47	0.000 68	17	4	5.36E-4
	(3.8E-5)	(2.3E-7)	(7.9E-3)	(3.1E-7)	(6.1E-6)	-			
MEF $\mathcal{E}_{2,\kappa^Q \neq \kappa^P}$	0.0038	0.0077	0.0140	-0.1195	0.000 47	0.000 68	16	2	5.37E-4
	(3.8E-5)	(2.4E-7)	(7.9E-3)	(3.1E-7)	(6.1E-6)	-			

the κ^Q column. Relaxing the restriction that $\kappa^Q = \kappa^P$ does not provide us with the better fit. In fact the MEF $\mathcal{E}_2$ is the best of MEF methods in terms of the MAE from the four considered MEF approaches. All estimation methods provide us with the overall good fit to the empirical term structure which is well observable from Figure 5 of the fitted and the observed term structures.

Table 6: Three-factor model estimates

The table that reports "smart" OLS start values for MEF, and the resulting γ^P, κ^P, κ^Q, λ, η and σ_e estimated by the Kalman filter, the MCSE, and the MEF. MEF $\mathcal{E}_1$ assumes the proper weighting, whereas $\mathcal{E}_2$ assumes the sum of weighting elements Ψ^{-1} to be rescaled to one. MEF use fminunc Matlab algorithm with the exponential transformation of γ^P, κ, and η. The overall fit is measured as the average MAE across all observed yields. Asymptotic standard errors are provided in the parentheses.

γ^P	κ^Q			κ^P			λ	η	σ_e
				OLS start					
0.0038	0.0070	0	0	0.0140	0	0	−0.0057	0.0005	-
	0	0.0061	0	0	0.0122	0	−0.0057	0.0004	
	0	0	0.0047	0	0	0.0094	−0.0057	0.0003	
				Kalman filter, MAE: 2.92E-4					
0.0052	0.3430	0	0	-	-	-	−0.2391	6.76E-4	6.60E-5
(4.5E-4)	(2.4E-2)	-	-	-	-	-	(4.9E-2)	(3.4E-5)	(2.0E-6)
	0	0.0495	0	-	-	-	−0.0090	4.91E-4	
	-	(2.1E-3)	-	-	-	-	(3.7E-2)	(2.2E-5)	
	0	0	0.0050	-	-	-	−0.0422	4.03E-4	
	-	-	(4.9E-4)	-	-	-	(5.2E-3)	(1.7E-5)	
				MCSE, MAE: 7.32E-5					
0.0034	0.0043	1	1	0.0226	1.0097	0.9470	−0.0656	1.98E-4	8.23E-5
(2.1E-3)	(5.9E-3)	-	-	(3.0E-2)	(1.1E-2)	(2.5E-2)	(5.3E-2)	(3.0E-5)	(2.7E-6)
	0.9885	0.0728	1	1.0047	0.1489	0.9011	−0.0339	1.25E-4	
	(3.7E-3)	(4.6E-4)	-	(2.5E-2)	(1.0E-2)	(2.5E-2)	(2.8E-2)	(2.2E-5)	
	0.9619	0.7741	0.2053	0.9768	0.8675	0.1697	−0.2113	4.44E-4	
	(1.3E-2)	(1.5E-2)	(2.6E-2)	(2.5E-2)	(1.3E-2)	(2.9E-2)	(1.2E-1)	(1.4E-5)	
				MEF $\mathcal{E}_1$, MAE: 8.62E-5					
0.0013	0.2070	0	0	-	-	-	−0.6276	3.48E-4	1.00E-4
(6.1E-4)	(1.1E-2)	-	-	-	-	-	(8.7E-7)	(2.2E-5)	-
	0	0.0377	0	-	-	-	−3.5621	1.15E-3	
	-	(4.6E-4)	-	-	-	-	(8.4E-6)	(9.4E-6)	
	0	0	0.0007	-	-	-	−0.6504	9.96E-5	
	-	-	(3.0E-4)	-	-	-	(1.1E-6)	(5.8E-7)	
				MEF $\mathcal{E}_2$, MAE: 7.67E-5					
0.0023	0.2395	0	0	-	-	-	−0.5552	3.46E-4	8.23E-5
(1.3E-4)	(3.3E-3)	-	-	-	-	-	(1.6E-7)	(1.3E-5)	-
	0	0.0735	0	-	-	-	0.3802	5.68E-4	
	-	(1.1E-3)	-	-	-	-	(6.4E-7)	(8.3E-6)	
	0	0	0.0043	-	-	-	−0.2211	1.73E-4	
	-	-	(1.9E-3)	-	-	-	(3.8E-7)	(5.6E-6)	
				MEF $\mathcal{E}_{1,\kappa^Q \neq \kappa^P}$, MAE: 8.24E-5					
0.0038	0.6612	0	0	0.0152	0	0	−0.0634	6.68E-4	1.00E-4
(7.7E-6)	(6.7E-8)	-	-	(2.3E-2)	-	-	(7.5E-9)	(6.7E-5)	-
	0	0.0350	0	0	0.0062	0	−0.0171	1.10E-3	
	-	(8.2E-8)	-	-	(2.7E-2)	-	(1.0E-7)	(5.2E-6)	
	0	0	0.0003	0	0	0.0278	−0.0460	6.66E-4	
	-	-	(3.3E-8)	-	-	(6.2E-3)	(9.2E-8)	(1.5E-6)	
				MEF $\mathcal{E}_{2,\kappa^Q \neq \kappa^P}$, MAE: 7.68E-5					
0.0038	0.0737	0	0	0.0139	0	0	−0.1091	4.73E-4	8.99E-5
(5.6E-6)	(2.6E-8)	-	-	(6.6E-3)	-	-	(2.3E-8)	(4.1E-6)	-
	0	0.1552	0	0	0.0122	0	0.0137	5.68E-4	
	-	(4.5E-8)	-	-	(9.8E-3)	-	(1.7E-8)	(1.3E-5)	
	0	0	0.0002	0	0	0.0069	−0.0527	4.22E-4	
	-	-	(1.5E-8)	-	-	(1.2E-2)	(4.2E-8)	(8.8E-7)	

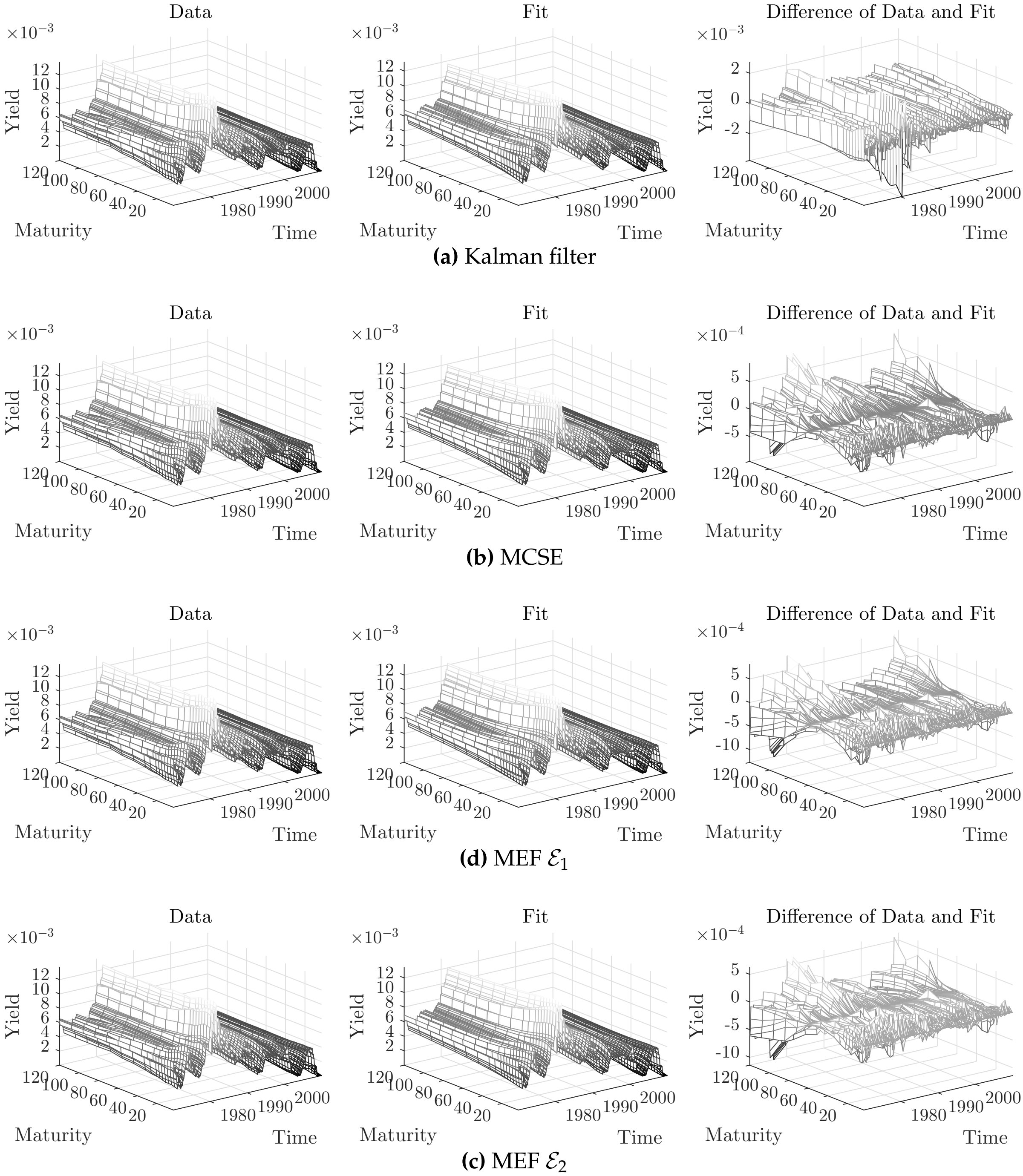


Figure 5: Three-factor fit on empirical dataset
Coefficients are estimated with Kalman filter, MCSE, MEF $\mathcal{E}_1$, and MEF $\mathcal{E}_2$. MEF $\mathcal{E}_{1,2,\kappa^Q \neq \kappa^P}$ are not plotted, because the fit very similar to $\mathcal{E}_{1,2}$.

6 Conclusion

Affine term structure models (ATSM) provide us with attractive model representations that can be used for the identification and the estimation of structural parameters. Popular estimation methods that work efficiently are the Kalman filtering and the Minimum Chi-Squared Minimization (MCSE). This paper provides the evidence that the method of martingale increment minimization using the martingale estimating functions (MEF) works as well. MEF is a new method of estimation of stochastic and non-stochastic processes proposed by Christensen et al. (2016). In the combination with Vasicek (1977) model we can estimate the structural parameters in the one- and multi-factor settings with the fit comparable to the MCSE.

The optimal MEF weights can be derived from the Vasicek model specification. With analytical affine coefficients for bond prices the MEF estimation is very intuitive and works in both the simulated and the empirical settings. The simulation implies that not all search algorithms are equally efficient when searching on badly-behaved multi-dimensional surfaces. Nevertheless, with the right search algorithm the convergence is reasonably quick and robust in the large sample simulation. The simulation study also uncovers the shortcomings of MEF, which are the behavior in low-noise environment the dependence on analytical solutions.

MEF struggles when identifying coefficients in the setting with very low observational noise. This can be resolved by providing a boundary for the Ψ^{-1} weighting matrix of MEF such that its diagonal elements sum up to one. Both the one- and three-factor simulation studies provide us with the evidence in favor of this rescaling. The multi-factor setting estimates the coefficients in much longer time that approaches the three minute mark. A lot of computational power is required with the extension of the optimal MEF matrices. Nevertheless, the estimated coefficients fit the term structure very well even in the empirical setting.

Further researchers may put focus on deriving the model with cross-correlated factors to properly match the model specification of MCSE in the multi-factor setting. Additionally, the emphasis should be put on dealing with the recursive solutions to the affine coefficients of yields. Up until now these recursions were the best method for calculating the discretized solutions to the Ricatti equations from the pricing constrains in ATSM. If we can find an alternative solution or a good method of computing the partial derivatives of these recursions the MEF method would be applicable to a much wider range of models.

MEF is a new estimation method that can directly compete with the MCSE and the Kalman filter. MEF can offer more flexibility by extending its time-varying matrices (for example adding conditional moment restrictions) while providing the asymptotic efficiency, but the numerical burden increases proportionally. With the increase in computational power the MEF method can become one of the best methods for many types of estimations that work by cleverly minimizing the residuals of properly defined models.

References

Ang, A. and M. Piazzesi (2003). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary economics 50*(4), 745–787.

Björk, T. (2009). *Arbitrage theory in continuous time*. Oxford university press.

Bolder, D. J. (2001). Affine term-structure models: Theory and implementation. *Available at SSRN 1082826*.

Brix, A. F. and A. Lunde (2013). Estimating stochastic volatility models using prediction-based estimating functions. *Research paper 23*, 2013–23.

Cheridito, P., D. Filipović, and R. L. Kimmel (2007). Market price of risk specifications for affine models: Theory and evidence. *Journal of Financial Economics 83*(1), 123–170.

Christensen, B. J., O. Posch, and M. van der Wel (2016). Estimating dynamic equilibrium models using mixed frequency macro and financial data. *Journal of Econometrics 194*(1), 116–137.

Cox, J. C., J. E. Ingersoll Jr, and S. A. Ross (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, 385–407.

Dai, Q. and K. J. Singleton (2000). Specification analysis of affine term structure models. *The Journal of Finance 55*(5), 1943–1978.

de Jong, F. (2000). Time series and cross-section information in affine term-structure models. *Journal of Business & Economic Statistics 18*(3), 300–314.

Dijk, D., S. J. Koopman, M. Wel, and J. H. Wright (2014). Forecasting interest rates with shifting endpoints. *Journal of Applied Econometrics 29*(5), 693–712.

Duffee, G. R. (2002). Term premia and interest rate forecasts in affine models. *The Journal of Finance 57*(1), 405–443.

Duffie, D. and R. Kan (1996). A yield-factor model of interest rates. *Mathematical finance 6*(4), 379–406.

Fama, E. F. and R. R. Bliss (1987). The information in long-maturity forward rates. *The American Economic Review*, 680–692.

Feller, W. (1951). Two singular diffusion problems. *Annals of mathematics*, 173–182.

Hall, P. and C. C. Heyde (2014). *Martingale limit theory and its application*. Academic press.

Hamilton, J. D. (1994). *Time series analysis*, Volume 2. Princeton university press Princeton.

Hamilton, J. D. and J. C. Wu (2012). Identification and estimation of gaussian affine term structure models. *Journal of Econometrics 168*(2), 315–331.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Joslin, S., K. J. Singleton, and H. Zhu (2011). A new perspective on gaussian dynamic term structure models. *Review of Financial Studies 24*(3), 926–970.

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

Appendices

A Conditional moments of factors

Starting from the factor dynamics in the P measure in the most general unrestricted form

$$dF_t = a^P + b^P F_t dt + \sigma(F_t) dW_t^P,$$

with $[\sigma(F_t)\sigma(F_t)^\top]_{ij} = \alpha_{ij} + \beta_{ij}^\top F_t$ diagonal.We follow by premultiplying the dynamics equation by $e^{-b^P t} = diag(e^{-b_1^P t}, \ldots, e^{-b_M^P t})$ with $-b^P = diag(-b_1^P, \ldots, -b_M^P)$ that is

$$e^{-b^P t} dF_t = e^{-b^P t}(a^P + b^P F_t)dt + e^{-b^P t}\sigma(F_t)dW_t^P,$$

noting the product rule $d(u \cdot v) = du \cdot v + u \cdot dv$

$$\begin{aligned} e^{-b^P t} dF_t - e^{-b^P t} b^P F_t dt &= e^{-b^P t} a^P dt + e^{-b^P t}\sigma(F_t)dW_t^P, \\ d(e^{-b^P t} F_t) &= e^{-b^P t} a^P dt + e^{-b^P t}\sigma(F_t)dW_t^P, \end{aligned}$$

taking the Itô integral from $t - \Delta$ to t we solve for F_t as

$$\begin{aligned} \int_{t-\Delta}^{t} d(e^{-b^P s} F_s) ds &= \int_{t-\Delta}^{t} e^{-b^P s} a^P ds + \int_{t-\Delta}^{t} e^{-b^P s}\sigma(F_s) dW_s^P, \\ \left[e^{-b^P s} F_s\right]_{t-\Delta}^{t} &= a^P \left[\frac{e^{-b^P s}}{-b^P}\right]_{t-\Delta}^{t} + \int_{t-\Delta}^{t} e^{-b^P s}\sigma(F_s) dW_s^P, \\ F_t &= e^{b^P \Delta} - \frac{a^P}{b^P}(1 - e^{b^P \Delta}) + e_t, \end{aligned}$$

where $e_t = \int_{t-\Delta}^{t} e^{-b^P(s-t)} \sqrt{\alpha + \beta^\top F_t} dW_s^P$ is the the martingale increment. It is easy to derive the conditional expectation in the form

$$\mathbb{E}(F_t | F_{t-\Delta}) = e^{b^P \Delta} F_{t-\Delta} - \frac{a^P}{b^P}(1 - e^{b^P \Delta}), \tag{A1}$$

and variance

$$\begin{aligned} Var(F_t|F_{t-\Delta})_{ij} &= \int_{t-\Delta}^{t} e^{-(b_i^P + b_j^P)(s-t)} (\alpha_{ij} + \beta_{ij}^\top \mathbb{E}_{t-s}[F_t]) ds, \\ &= \int_{t-\Delta}^{t} e^{-(b_i^P + b_j^P)(s-t)} \left(\alpha_{ij} - \beta_{ij}^\top \frac{a^P}{b^P}\right) ds + \int_{t-\Delta}^{t} e^{-(b_i^P + b_j^P)(s-t)} \beta_{ij}^\top e^{b^P s} \left(F_{t-\Delta} + \frac{a^P}{b^P}\right) ds, \\ &= \frac{1 - e^{(b_i + b_j)\Delta}}{-b_i^P - b_j^P} \left(\alpha_{ij} - \beta_{ij}^\top \frac{a^P}{b^P}\right) + \sum_k \frac{e^{b_k^P \Delta} - e^{(b_i^P + b_j^P)\Delta}}{-b_i^P - b_j^P + b_k^P} \beta_{ij,k} (F_{t-\Delta,k} + \frac{a^P}{b^P}). \end{aligned} \tag{A2}$$

Unconditional variance is then in the simple form

$$Var(F_t) = \frac{\alpha_{ij} - \beta_{ij}^\top \left(\frac{a^P}{b^P}\right)}{-b_i^P - b_j^P}. \tag{A3}$$

B MEF with one-factor Vasicek

Considering the Vasicek parametrization and normalizations $\mathcal{S}_3$ and following the solutions from (A1), (A2) and (A3) we write

$$\begin{aligned} \mathbb{E}(F_t|F_{t-\Delta}) &= e^{-\kappa\Delta}F_{t-\Delta}, \\ Var(F_t||F_{t-\Delta}) &= \frac{1-e^{-2\kappa\Delta}}{2\kappa}, \\ Var(F_t) &= \frac{1}{2\kappa}. \end{aligned}$$

B1 Feynman–Kac solution

Let us define the price of the zero coupon bond $P = e^{-A-Br}$. We denote $A(t,T)$ and $B(t,T)$ for the bond $P(t,T)$ at time t maturing at T. It holds that the price at maturity is exactly equal to the payoff of the bond that is $P(T,T) = 1$, from which it directly follows that the affine structure needs to satisfy the conditions $A(T,T) = B(T,T) = 0$. We assume that the short rate is an affine function of factors $r_t = \delta_0 + \delta_1^\top F_t$ and that the factors follows the P and Q dynamics

$$\begin{aligned} dF_t &= b^P F_t dt + dW_t^P, \\ dF_t &= (a^Q + b^Q F_t)dt + dW_t^Q. \end{aligned}$$

The short rate therefore follows

$$\begin{aligned} dr_t &= -b^P(\delta_0 - r_t)dt + \delta_1 dW_t^P, \\ dr_t &= (a^Q\delta_1 - b^Q\delta_0 + b^Q r_t)dt + \delta_1 dW_t^Q. \end{aligned}$$

Define the term structure equation under the arbitrage-free bond market as

$$\begin{aligned} \frac{\partial P(t,T)}{\partial t} + \mu^Q \frac{\partial P(t,T)}{\partial r} + \frac{\sigma^2}{2}\frac{\partial^2 P(t,T)}{\partial r^2} - rP(t,T) &= 0, \\ P(T,T) &= 1. \end{aligned} \tag{B1}$$

The first and second order partial derivatives of P yield

$$\begin{aligned} \frac{\partial P}{\partial t} &= P \cdot \Big(-\frac{\partial A}{\partial t} - \frac{\partial B}{\partial t} r \Big), \\ \frac{\partial P}{\partial r} &= P \cdot \big(-B \big), \\ \frac{\partial^2 P}{\partial t^2} &= P \cdot \big(B^2 \big). \end{aligned}$$

Continuing again from (B1) with solutions for partial derivatives

$$
\begin{aligned}
\Big(-\frac{\partial A}{\partial t}-\frac{\partial B}{\partial t}r\Big)-\mu^{Q}B+\frac{\sigma^{2}}{2}B^{2}-r=0,\\
\Big(-\frac{\partial A}{\partial t}-\mu^{Q}B+\frac{\sigma^{2}}{2}B^{2}\Big)-\Big(\frac{\partial B}{\partial t}+1\Big)r=0.
\end{aligned}
$$

Additionally, to match our specification, we plug for the mean and volatility terms

$$
\begin{aligned}
\mu^{Q}&=a^{Q}\delta_1-b^{Q}\delta_0+b^{Q}r,\\
\sigma&=\delta_1,
\end{aligned}
$$

and continue

$$
\Big(-\frac{\partial A}{\partial t}-a^{Q}\delta_1B+b^{Q}\delta_0B+\frac{\delta_1^2}{2}B^{2}\Big)-\Big(\frac{\partial B}{\partial t}+1+b^{Q}B\Big)r=0,
$$

where each of the respective brackets need to equate to zero such that the condition is true for all values of r. We can therefore solve separately

$$
\begin{aligned}
\Big(\frac{\partial B(t,T)}{\partial t}+1+b^{Q}B(t,T)\Big)&=0,\\
\int_t^T\frac{dB(s,T)}{-1-b^{Q}B(s,T)}&=\int_t^T ds,\\
\frac{1}{b^{Q}}\big[-ln|1+b^{Q}B(s,T)|\big]_t^T&=T-t,\\
ln(1+b^{Q}B(t,T))&=b^{Q}(T-t),\\
B(t,T)&=\frac{-1}{b^{Q}}\big(1-e^{b^{Q}(T-t)}\big).
\end{aligned}
$$

Similarly for the first term

$$
\begin{aligned}
\Big(-\frac{\partial A(t,T)}{\partial t}-a^{Q}\delta_1B(t,T)+b^{Q}\delta_0B(t,T)+\frac{\delta_1^2}{2}B^{2}(t,T)\Big)&=0,\\
dA(t,T)=(-a^{Q}\delta_1+b^{Q}\delta_0)B(t,T)dt+\frac{\delta_1^2}{2}B^{2}(t,T)dt,\\
-A(t,T)=(-a^{Q}\delta_1+b^{Q}\delta_0)\int_t^T B(s,T)ds+\frac{\delta_1^2}{2}\int_t^T B^{2}(s,T)ds,
\end{aligned}
$$

where

$$
\begin{aligned}
\int_t^T B(s,T)ds&=\frac{-1}{b^{Q}}\big[s+e^{b^{Q}(T-s)}\big]_t^T=\frac{B(t,T)-(T-t)}{b^{Q}},\\
\int_t^T B^2(s,T)ds&=\frac{1}{b^{Q2}}\Big[\frac{e^{2b^{Q}(T-s)}}{-2b^{Q}}+s+\frac{2e^{b^{Q}(T-s)}}{b^{Q}}\Big]_t^T=\frac{B^2(t,T)}{2b^{Q}}-\frac{B(t,T)-(T-t)}{b^{Q2}},
\end{aligned}
$$

and therefore

$$A(t,T) = \left(\frac{a^Q\delta_1}{b^Q} - \delta_0 + \frac{\delta_1^2}{2b^{Q2}}\right)(B(t,T) - (T-t)) - \frac{\delta_1^2}{4b^Q}B^2(t,T),$$

thus the price process P with analytical A and B is indeed a solution to the term structure equation.

To summarize and re-label consistently with the Vasicek parametrization we can write for the price process $P_t^{(n)} = e^{-A^{(n)}-B^{(n)}r_t}$ with $dr_t = \kappa(\gamma^P - \frac{\lambda\eta}{\kappa} - r_t)dt + \eta dW_t^Q$ that prices zero coupon bonds under the Q measure to be solved analytically by

$$\begin{aligned} B^{(n)} &= \frac{1}{\kappa}(1-e^{-\kappa n}), \\ A^{(n)} &= \left(\frac{\eta^2}{2\kappa^2} - \gamma^P + \frac{\lambda\eta}{\kappa}\right)(B^{(n)} - n) + \frac{\eta^2}{4\kappa}(B^{(n)})^2, \end{aligned}$$

with n denoting the period till maturity. Note that the t indicates the price-variation of bonds in time which have the full duration $n = T$.

B2 MEF derivation—Vasicek 1 factor

For a price $P_t^{(n)} = e^{-\bar{A}^{(n)}-\bar{B}^{(n)}F_t}$, the n-period yield $y_t^{(n)}$ is defined as $y_t^{(n)} = -\frac{lnP_t^{(n)}}{n} = \frac{\bar{A}^{(n)}}{n} + \frac{\bar{B}^{(n)}}{n}F_t$, that is an affine function of factors linked to the short rate by the relation $r_t = \gamma^P + \eta F_t$ (Vasicek parametrization). We can write down the martingale increment matrix for our specification

$$\begin{aligned} m_t(\phi) &= \left(\begin{bmatrix} \underset{(1\times1)}{F_t} \\ \underset{(N_e\times1)}{Y_t^2} \end{bmatrix} - \begin{bmatrix} \underset{(1\times1)}{0} \\ \underset{(N_e\times1)}{A_2} \end{bmatrix} - \begin{bmatrix} \underset{(1\times1)}{e^{-\kappa\Delta}F_{t-\Delta}} \\ \underset{(N_e\times1)}{B_2F_t} \end{bmatrix} \right) \\ &= \begin{pmatrix} F_t - e^{-\kappa\Delta}F_{t-\Delta} \\ y_t^{(n_2)} - \bar{A}^{(n_2)}/n_2 - \bar{B}^{(n_2)}/n_2 \cdot F_t \\ \vdots \\ y_t^{(n_N)} - \bar{A}^{(n_N)}/n_N - \bar{B}^{(n_N)}/n_N \cdot F_t \end{pmatrix}, \end{aligned}$$

where $\bar{B}(n_i)$, $\bar{A}(n_i)$ come from (23) and (24) respectively for different times till maturity n_i. We take partial derivatives of m_t with respect to each parameter of $\phi = \begin{pmatrix}\gamma^P & \kappa & \lambda & \eta & \sigma_e\end{pmatrix}^\top$ to get the parameter derivatives of the vectors martingale increments with σ_e being the constant deviation term for all yields observed with observational error, that is $\Sigma_e = \sigma_e I$. We therefore write[16]

[16]Differentiation was checked with Matlab symbolics toolbox to ensure the correct derivation from $m_t(\phi)$.

$$
\left(\frac{\partial m_t}{\partial \phi_t^\top}\right)^\top = \begin{pmatrix}
0 & -1 & \cdots & -1 \\
\Delta F_{t-\Delta} e^{-\kappa\Delta} & \begin{matrix} -\Big[\Big(\frac{(e^{-\kappa n_2}-1)}{\kappa^2}+\frac{n_2 e^{-\kappa n_2}}{\kappa}\Big)\Big(\frac{\eta^2}{2\kappa^2}-\gamma^P \\ +\frac{\eta\lambda}{\kappa}\Big)+\Big(n_2+\frac{(e^{-\kappa n_2}-1)}{\kappa}\Big)\Big(\frac{\eta^2}{\kappa^3}+\frac{\eta\lambda}{\kappa^2}\Big) \\ -\frac{3\eta^2(e^{-\kappa n_2}-1)^2}{4\kappa^4}+\frac{\gamma^P(e^{-\kappa n_2}-1)}{\kappa^2}+ \\ \frac{\gamma^P n_2 e^{-\kappa n_2}}{\kappa}-\frac{\eta^2 n_2 e^{-\kappa n_2}(e^{-\kappa n_2}-1)}{2\kappa^3}\Big]\frac{1}{n_2} \\ -\frac{F_t\eta e^{-\kappa n_2}}{\kappa}-\frac{F_t\eta(e^{-\kappa n_2}-1)}{\kappa^2 n_2} \end{matrix} & \cdots & \begin{matrix} -\Big[\Big(\frac{(e^{-\kappa n_N}-1)}{\kappa^2}+\frac{n_N e^{-\kappa n_N}}{\kappa}\Big)\Big(\frac{\eta^2}{2\kappa^2}-\gamma^P \\ +\frac{\eta\lambda}{\kappa}\Big)+\Big(n_N+\frac{(e^{-\kappa n_N}-1)}{\kappa}\Big)\Big(\frac{\eta^2}{\kappa^3}+\frac{\eta\lambda}{\kappa^2}\Big) \\ -\frac{3\eta^2(e^{-\kappa n_N}-1)^2}{4\kappa^4}+\frac{\gamma^P(e^{-\kappa n_N}-1)}{\kappa^2}+ \\ \frac{\gamma^P n_N e^{-\kappa n_N}}{\kappa}-\frac{\eta^2 n_N e^{-\kappa n_N}(e^{-\kappa n_N}-1)}{2\kappa^3}\Big]\frac{1}{n_N} \\ -\frac{F_t\eta e^{-\kappa n_N}}{\kappa}-\frac{F_t\eta(e^{-\kappa n_N}-1)}{\kappa^2 n_N} \end{matrix} \\
0 & \frac{\eta}{\kappa n_2}\left(n_2+\frac{e^{-\kappa n_2}-1}{\kappa}\right) & \cdots & \frac{\eta}{\kappa n_N}\left(n_N+\frac{e^{-\kappa n_N}-1}{\kappa}\right) \\
0 & \begin{matrix} \Big[\Big(n_2+\frac{(e^{-\kappa n_2}-1)}{\kappa}\Big)\Big(\eta\kappa^2+\lambda\kappa\Big) \\ -\frac{\eta(e^{-\kappa n_2}-1)^2}{2\kappa^3}\Big]\frac{1}{n_2}+\frac{F_t(e^{-\kappa n_2}-1)}{\kappa n_2} \end{matrix} & \cdots & \begin{matrix} \Big[\Big(n_N+\frac{(e^{-\kappa n_N}-1)}{\kappa}\Big)\Big(\eta\kappa^2+\lambda\kappa\Big) \\ -\frac{\eta(e^{-\kappa n_N}-1)^2}{2\kappa^3}\Big]\frac{1}{n_N}+\frac{F_t(e^{-\kappa n_N}-1)}{\kappa n_N} \end{matrix} \\
0 & 0 & \cdots & 0
\end{pmatrix}
$$

We apply (44), that is we take the expectation and condition on the time $t-1$ to get

$$
\psi_t = \mathbb{E}_{t-1}\left[\left(\frac{\partial m_t}{\partial \phi_t^\top}\right)^\top\right]
$$

With m_t we can derive $(N \times N)$ matrix Ψ_t from (43) as

$$
\Psi_t = \mathrm{Var}_{t-1}(m_t) = \begin{pmatrix} \sigma_\eta^2 & 0 & \dots & 0 \\ 0 & \sigma_e^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_e^2 \end{pmatrix}.
$$

We note that in such a setting, the Ψ_t does not depend on t and thus can be refered to as Ψ, such that

$$
\Psi_t^{-1} = \Psi^{-1} = \begin{pmatrix} \frac{1}{\sigma_\eta^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_e^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_e^2} \end{pmatrix}
$$

By plugging to (41) we get

$$
\begin{aligned}
M_T &= \sum_{t=1}^{T} w_t m_t = \sum_{t=1}^{T} \psi_t^{\top} \Psi^{-1} m_t \ , \\
&= \sum_{t=1}^{T}
\begin{pmatrix}
0 & -1 & \cdots & -1 \\
\Delta F_{t-\Delta} e^{-\kappa\Delta} &
\begin{matrix}
-\Big[\Big(\frac{(e^{-\kappa n_2}-1)}{\kappa^2}+\frac{n_2 e^{-\kappa n_2}}{\kappa}\Big)\Big(\frac{\eta^2}{2\kappa^2}-\gamma^P \\
+\frac{\eta\lambda}{\kappa}\Big)+\Big(n_2+\frac{(e^{-\kappa n_2}-1)}{\kappa}\Big)\Big(\frac{\eta^2}{\kappa^3}+\frac{\eta\lambda}{\kappa^2}\Big) \\
-\frac{3\eta^2(e^{-\kappa n_2}-1)^2}{4\kappa^4}+\frac{\gamma^P(e^{-\kappa n_2}-1)}{\kappa^2}+ \\
\frac{\gamma^P n_2 e^{-\kappa n_2}}{\kappa}-\frac{\eta^2 n_2 e^{-\kappa n_2}(e^{-\kappa n_2}-1)}{2\kappa^3}\Big]\frac{1}{n_2} \\
-\frac{F_t \eta e^{-\kappa n_2}}{\kappa}-\frac{F_t\eta(e^{-\kappa n_2}-1)}{\kappa^2 n_2}
\end{matrix}
& \cdots &
\begin{matrix}
-\Big[\Big(\frac{(e^{-\kappa n_N}-1)}{\kappa^2}+\frac{n_N e^{-\kappa n_N}}{\kappa}\Big)\Big(\frac{\eta^2}{2\kappa^2}-\gamma^P \\
+\frac{\eta\lambda}{\kappa}\Big)+\Big(n_N+\frac{(e^{-\kappa n_N}-1)}{\kappa}\Big)\Big(\frac{\eta^2}{\kappa^3}+\frac{\eta\lambda}{\kappa^2}\Big) \\
-\frac{3\eta^2(e^{-\kappa n_N}-1)^2}{4\kappa^4}+\frac{\gamma^P(e^{-\kappa n_N}-1)}{\kappa^2}+ \\
\frac{\gamma^P n_N e^{-\kappa n_N}}{\kappa}-\frac{\eta^2 n_N e^{-\kappa n_N}(e^{-\kappa n_N}-1)}{2\kappa^3}\Big]\frac{1}{n_N} \\
-\frac{F_t \eta e^{-\kappa n_N}}{\kappa}-\frac{F_t\eta(e^{-\kappa n_N}-1)}{\kappa^2 n_N}
\end{matrix} \\
0 & \frac{\eta}{\kappa n_2}\Big(n_2+\frac{e^{-\kappa n_2}-1}{\kappa}\Big) & \cdots & \frac{\eta}{\kappa n_N}\Big(n_N+\frac{e^{-\kappa n_N}-1}{\kappa}\Big) \\
0 &
\begin{matrix}
\Big[\Big(n_2+\frac{(e^{-\kappa n_2}-1)}{\kappa}\Big)\Big(\eta\kappa^2+\lambda\kappa\Big) \\
-\frac{\eta(e^{-\kappa n_2}-1)^2}{2\kappa^3}\Big]\frac{1}{n_2}+\frac{F_t(e^{-\kappa n_2}-1)}{\kappa n_2}
\end{matrix}
& \cdots &
\begin{matrix}
\Big[\Big(n_N+\frac{(e^{-\kappa n_N}-1)}{\kappa}\Big)\Big(\eta\kappa^2+\lambda\kappa\Big) \\
-\frac{\eta(e^{-\kappa n_N}-1)^2}{2\kappa^3}\Big]\frac{1}{n_N}+\frac{F_t(e^{-\kappa n_N}-1)}{\kappa n_N}
\end{matrix} \\
0 & 0 & \cdots & 0
\end{pmatrix} . \\
&\quad \cdot
\begin{pmatrix}
\frac{1}{\sigma_\eta^2} & 0 & \cdots & 0 \\
0 & \frac{1}{\sigma_e^2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{1}{\sigma_e^2}
\end{pmatrix}
\cdot
\begin{pmatrix}
F_t - e^{-\kappa\Delta} F_{t-\Delta} \\
y_t^{(n_2)} - \bar{A}^{(n_2)}/n_2 - \bar{B}^{(n_2)}/n_2 \cdot F_t \\
\vdots \\
y_t^{(n_N)} - \bar{A}^{(n_N)}/n_N - \bar{B}^{(n_N)}/n_N \cdot F_t
\end{pmatrix} ,
\end{aligned}
\tag{B2}
$$

which is out objective vector that we need to minimize. We proceed by numerical minimization in Matlab where we minimize the objective scalar $f(\phi) = M_T^{\top} M_T$ by a chosen minimizer.

C MEF with three-factor Vasicek

In our multifactor parametrization we negligate the cross-dependencies of factors and because of this generalization the one-factor results still apply. We continue directly with deriving the MEF system for a parameter vector

$$\phi^\top = (\gamma^P, \kappa_1, \kappa_2, \kappa_3, \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2, \eta_2, \sigma_e).$$

C1 MEF derivation—Vasicek 3 factors

Following from the one-factor case we derive the martingale increment matrix for our specification

$$\begin{aligned}
m_t(\phi) &= \left(\begin{bmatrix} \underset{(1\times1)}{F_t^1} \\ \underset{(1\times1)}{F_t^2} \\ \underset{(1\times1)}{F_t^3} \\ \underset{(N_e\times1)}{Y_t^2} \end{bmatrix} - \begin{bmatrix} \underset{(1\times1)}{0} \\ \underset{(1\times1)}{0} \\ \underset{(1\times1)}{0} \\ \underset{(N_e\times1)}{A_2} \end{bmatrix} - \begin{bmatrix} \underset{(3\times3)}{\begin{pmatrix} e^{-\kappa_1\Delta} & 0 & 0 \\ 0 & e^{-\kappa_2\Delta} & 0 \\ 0 & 0 & e^{-\kappa_3\Delta} \end{pmatrix}} \underset{(3\times1)}{F_{t-\Delta}} \\ \underset{(N_e\times3)}{B_2} \underset{(3\times1)}{F_t} \end{bmatrix} \right), \\
&= \begin{pmatrix} F_t^1 - e^{-\kappa_1\Delta} F_{t-\Delta}^1 \\ F_t^2 - e^{-\kappa_2\Delta} F_{t-\Delta}^2 \\ F_t^3 - e^{-\kappa_3\Delta} F_{t-\Delta}^3 \\ y_t^{(n_2)} - \bar{A}^{(n_2)}/n_2 - B_1^{(n_2)}\eta_1 F_t^1/n_2 - B_2^{(n_2)}\eta_2 F_t^2/n_2 - B_3^{(n_2)}\eta_3 F_t^3/n_2 \\ \vdots \\ y_t^{(n_N)} - \bar{A}^{(n_N)}/n_N - B_1^{(n_N)}\eta_1 F_t^1/n_N - B_2^{(n_N)}\eta_2 F_t^2/n_N - B_3^{(n_N)}\eta_3 F_t^3/n_N \end{pmatrix},
\end{aligned}$$

with

$$\begin{aligned}
\underset{(1\times1)}{y_t^{(n)}} &= \underset{(1\times1)}{\frac{\bar{A}^{(n)}}{n}} + \underset{(1\times3)}{\frac{\bar{B}^{(n)}}{n}} \underset{(3\times1)}{F_t}, \\
\underset{(1\times3)}{\bar{B}^{(n)}} &= \underset{(1\times3)}{B^{(n)}} \underset{(3\times3)}{diag(\eta)}, \\
B_i^{(n)} &= \frac{1}{\kappa_i}(1 - e^{-\kappa_i n}), \\
\bar{A}^{(n)} &= \gamma^P n + \sum_{i=1}^{3} \left(\frac{\eta_i^2}{2\kappa_i^2} + \frac{\lambda_i \eta_i}{\kappa_i} \right)(B_i^{(n)} - n) + \frac{\eta_i^2}{4\kappa_i}(B_i^{(n)})^2.
\end{aligned}$$

We proceed by deriving the optimal matrices $\psi_t^\top$ and Ψ_t^{-1}

$$\psi_t^\top = \begin{pmatrix} 0 & 0 & 0 & -1 & \dots & -1 \\ \Delta F^1_{t-\Delta} e^{-\Delta\kappa_1} & 0 & 0 & \psi_{2,4} & \cdots & \psi_{2,N} \\ 0 & \Delta F^2_{t-\Delta} e^{-\Delta\kappa_2} & 0 & \psi_{3,4} & \cdots & \psi_{3,N} \\ 0 & 0 & \Delta F^3_{t-\Delta} e^{-\Delta\kappa_3} & \psi_{4,4} & \cdots & \psi_{4,N} \\ 0 & 0 & 0 & \psi_{5,4} & \cdots & \psi_{5,N} \\ 0 & 0 & 0 & \psi_{6,4} & \cdots & \psi_{6,N} \\ 0 & 0 & 0 & \psi_{7,4} & \cdots & \psi_{7,N} \\ 0 & 0 & 0 & \psi_{8,4} & \cdots & \psi_{8,N} \\ 0 & 0 & 0 & \psi_{9,4} & \cdots & \psi_{9,N} \\ 0 & 0 & 0 & \psi_{10,4} & \cdots & \psi_{10,N} \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

with

$$\begin{aligned}
\psi_{2,4} =& -\left(\frac{\eta_1^2}{2\kappa_1^2} + \frac{\eta_1\lambda_1}{\kappa_1}\left(\frac{e^{-\kappa_1 n_2}-1}{\kappa_1^2} + n_2 e^{-\kappa_1 n_2}\kappa_1\right) + \left(n_2 + \frac{e^{-\kappa_1 n_2}-1}{\kappa_1}\right)\left(\frac{\eta_1^2}{\kappa_1^3} + \frac{\eta_1\lambda_1}{\kappa_1^2}\right)\right. \\
&\left. - \frac{3\eta_1^2(e^{-\kappa_1 n_2}-1)^2}{4\kappa_1^4} - \frac{\eta_1^2 n_2 e^{-\kappa_1 n_2}(e^{-\kappa_1 n_2}-1)}{2\kappa_1^3}\right)/n_2 - \frac{F_t^1 \eta_1 e^{-\kappa_1 n_2}}{\kappa_1} \\
&- F_t^1 \eta_1 (e^{-\kappa_1 n_2}-1)\kappa_1^2 n_2 \\
\psi_{2,N} =& -\left(\frac{\eta_1^2}{2\kappa_1^2} + \frac{\eta_1\lambda_1}{\kappa_1}\left(\frac{e^{-\kappa_1 n_N}-1}{\kappa_1^2} + n_N e^{-\kappa_1 n_N}\kappa_1\right) + \left(n_N + \frac{e^{-\kappa_1 n_N}-1}{\kappa_1}\right)\left(\frac{\eta_1^2}{\kappa_1^3} + \frac{\eta_1\lambda_1}{\kappa_1^2}\right)\right. \\
&\left. - \frac{3\eta_1^2(e^{-\kappa_1 n_N}-1)^2}{4\kappa_1^4} - \frac{\eta_1^2 n_N e^{-\kappa_1 n_N}(e^{-\kappa_1 n_N}-1)}{2\kappa_1^3}\right)/n_N - \frac{F_t^1 \eta_1 e^{-\kappa_1 n_N}}{\kappa_1} \\
&- F_t^1 \eta_1 (e^{-\kappa_1 n_N}-1)\kappa_1^2 n_N
\end{aligned}$$

$$\begin{aligned}
\psi_{3,4} =& -\left(\frac{\eta_2^2}{2\kappa_2^2} + \frac{\eta_2\lambda_2}{\kappa_2}\left(\frac{e^{-\kappa_2 n_2}-1}{\kappa_2^2} + n_2 e^{-\kappa_2 n_2}\kappa_2\right) + \left(n_2 + \frac{e^{-\kappa_2 n_2}-1}{\kappa_2}\right)\left(\frac{\eta_2^2}{\kappa_2^3} + \frac{\eta_2\lambda_2}{\kappa_2^2}\right)\right. \\
&\left. - \frac{3\eta_2^2(e^{-\kappa_2 n_2}-1)^2}{4\kappa_2^4} - \frac{\eta_2^2 n_2 e^{-\kappa_2 n_2}(e^{-\kappa_2 n_2}-1)}{2\kappa_2^3}\right)/n_2 - \frac{F_t^2 \eta_2 e^{-\kappa_2 n_2}}{\kappa_2} \\
&- F_t^2 \eta_2 (e^{-\kappa_2 n_2}-1)\kappa_2^2 n_2 \\
\psi_{3,N} =& -\left(\frac{\eta_2^2}{2\kappa_2^2} + \frac{\eta_2\lambda_2}{\kappa_2}\left(\frac{e^{-\kappa_2 n_N}-1}{\kappa_2^2} + n_N e^{-\kappa_2 n_N}\kappa_2\right) + \left(n_N + \frac{e^{-\kappa_2 n_N}-1}{\kappa_2}\right)\left(\frac{\eta_2^2}{\kappa_2^3} + \frac{\eta_2\lambda_2}{\kappa_2^2}\right)\right. \\
&\left. - \frac{3\eta_2^2(e^{-\kappa_2 n_N}-1)^2}{4\kappa_2^4} - \frac{\eta_2^2 n_N e^{-\kappa_2 n_N}(e^{-\kappa_2 n_N}-1)}{2\kappa_2^3}\right)/n_N - \frac{F_t^2 \eta_2 e^{-\kappa_2 n_N}}{\kappa_2} \\
&- F_t^2 \eta_2 (e^{-\kappa_2 n_N}-1)\kappa_2^2 n_N
\end{aligned}$$

$$
\begin{aligned}
\psi_{4,4} =& -\left(\frac{\eta_3^2}{2\kappa_3^2} + \frac{\eta_3\lambda_3}{\kappa_3}\left(\frac{e^{-\kappa_3 n_2}-1}{\kappa_3^2} + n_2 e^{-\kappa_3 n_2}\kappa_3\right) + \left(n_2 + \frac{e^{-\kappa_3 n_2}-1}{\kappa_3}\right)\left(\frac{\eta_3^2}{\kappa_3^3} + \frac{\eta_3\lambda_3}{\kappa_3^2}\right)\right. \\
& \left. - \frac{3\eta_3^2(e^{-\kappa_3 n_2}-1)^2}{4\kappa_3^4} - \frac{\eta_3^2 n_2 e^{-\kappa_3 n_2}(e^{-\kappa_3 n_2}-1)}{2\kappa_3^3}\right)/n_2 - \frac{F_t^3\eta_3 e^{-\kappa_3 n_2}}{\kappa_3} \\
& - F_t^3\eta_3(e^{-\kappa_3 n_2}-1)\kappa_3^2 n_2 \\
\psi_{4,N} =& -\left(\frac{\eta_3^2}{2\kappa_3^2} + \frac{\eta_3\lambda_3}{\kappa_3}\left(\frac{e^{-\kappa_3 n_N}-1}{\kappa_3^2} + n_N e^{-\kappa_3 n_N}\kappa_3\right) + \left(n_N + \frac{e^{-\kappa_3 n_N}-1}{\kappa_3}\right)\left(\frac{\eta_3^2}{\kappa_3^3} + \frac{\eta_3\lambda_3}{\kappa_3^2}\right)\right. \\
& \left. - \frac{3\eta_3^2(e^{-\kappa_3 n_N}-1)^2}{4\kappa_3^4} - \frac{\eta_3^2 n_N e^{-\kappa_3 n_N}(e^{-\kappa_3 n_N}-1)}{2\kappa_3^3}\right)/n_N - \frac{F_t^3\eta_3 e^{-\kappa_3 n_N}}{\kappa_3} \\
& - F_t^3\eta_3(e^{-\kappa_3 n_N}-1)\kappa_3^2 n_N
\end{aligned}
$$

$$
\begin{aligned}
\psi_{5,4} =& \frac{\eta_1(n_2 + (e^{-\kappa_1 n_2}-1)/\kappa_1)}{\kappa_1 n_2} \\
\psi_{5,N} =& \frac{\eta_1(n_N + (e^{-\kappa_1 n_N}-1)/\kappa_1)}{\kappa_1 n_N}
\end{aligned}
$$

$$
\begin{aligned}
\psi_{6,4} =& \frac{\eta_2(n_2 + (e^{-\kappa_2 n_2}-1)/\kappa_2)}{\kappa_2 n_2} \\
\psi_{6,N} =& \frac{\eta_2(n_N + (e^{-\kappa_2 n_N}-1)/\kappa_2)}{\kappa_2 n_N}
\end{aligned}
$$

$$
\begin{aligned}
\psi_{7,4} =& \frac{\eta_3(n_2 + (e^{-\kappa_3 n_2}-1)/\kappa_3)}{\kappa_3 n_2} \\
\psi_{7,N} =& \frac{\eta_3(n_N + (e^{-\kappa_3 n_N}-1)/\kappa_3)}{\kappa_3 n_N}
\end{aligned}
$$

$$
\begin{aligned}
\psi_{8,4} =& \left(\left(n_2 + \frac{e^{-\kappa_1 n_2}-1}{\kappa_1}\right)\left(\frac{\eta_1}{\kappa_1^2} + \frac{\lambda_1}{\kappa_1}\right) - \frac{\eta_1(e^{-\kappa_1 n_2}-1)^2}{2\kappa_1^3}\right)/n_2 + F_t^1(e^{-\kappa_1 n_2}-1)\kappa_1 n_2 \\
\psi_{8,N} =& \left(\left(n_N + \frac{e^{-\kappa_1 n_N}-1}{\kappa_1}\right)\left(\frac{\eta_1}{\kappa_1^2} + \frac{\lambda_1}{\kappa_1}\right) - \frac{\eta_1(e^{-\kappa_1 n_N}-1)^2}{2\kappa_1^3}\right)/n_N + F_t^1(e^{-\kappa_1 n_N}-1)\kappa_1 n_N
\end{aligned}
$$

$$
\begin{aligned}
\psi_{9,4} =& \left(\left(n_2 + \frac{e^{-\kappa_2 n_2}-1}{\kappa_2}\right)\left(\frac{\eta_2}{\kappa_2^2} + \frac{\lambda_2}{\kappa_2}\right) - \frac{\eta_2(e^{-\kappa_2 n_2}-1)^2}{2\kappa_2^3}\right)/n_2 + F_t^2(e^{-\kappa_2 n_2}-1)\kappa_2 n_2 \\
\psi_{9,N} =& \left(\left(n_N + \frac{e^{-\kappa_2 n_N}-1}{\kappa_2}\right)\left(\frac{\eta_2}{\kappa_2^2} + \frac{\lambda_2}{\kappa_2}\right) - \frac{\eta_2(e^{-\kappa_2 n_N}-1)^2}{2\kappa_2^3}\right)/n_N + F_t^2(e^{-\kappa_2 n_N}-1)\kappa_2 n_N
\end{aligned}
$$

$$
\begin{aligned}
\psi_{10,4} =& \left(\left(n_2 + \frac{e^{-\kappa_3 n_2}-1}{\kappa_3}\right)\left(\frac{\eta_3}{\kappa_3^2} + \frac{\lambda_3}{\kappa_3}\right) - \frac{\eta_3(e^{-\kappa_3 n_2}-1)^2}{2\kappa_3^3}\right)/n_2 + F_t^3(e^{-\kappa_3 n_2}-1)\kappa_3 n_2 \\
\psi_{10,N} =& \left(\left(n_N + \frac{e^{-\kappa_3 n_N}-1}{\kappa_3}\right)\left(\frac{\eta_3}{\kappa_3^2} + \frac{\lambda_3}{\kappa_3}\right) - \frac{\eta_3(e^{-\kappa_3 n_N}-1)^2}{2\kappa_3^3}\right)/n_N + F_t^3(e^{-\kappa_3 n_N}-1)\kappa_3 n_N
\end{aligned}
$$

We continue with

$$\Psi_t^{-1} = \Psi^{-1} = \begin{pmatrix} \frac{1}{\sigma_{\eta_1}^2} & 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{\eta_2}^2} & 0 & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sigma_{\eta_3}^2} & 0 & \dots & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma_e^2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \frac{1}{\sigma_e^2} \end{pmatrix},$$

with

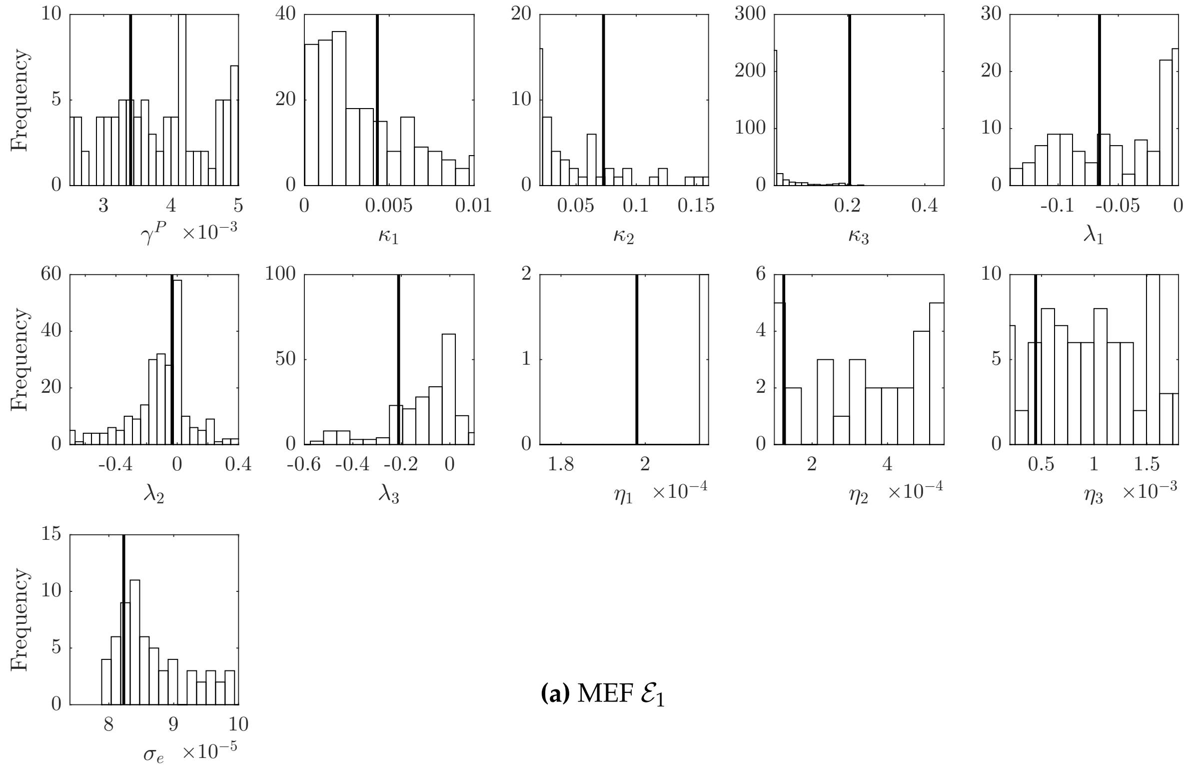
$$\sigma_{\eta_1}^2 = \frac{1 - e^{-2\kappa_1 \Delta}}{2\kappa_1},$$
$$\sigma_{\eta_2}^2 = \frac{1 - e^{-2\kappa_2 \Delta}}{2\kappa_2},$$
$$\sigma_{\eta_3}^2 = \frac{1 - e^{-2\kappa_3 \Delta}}{2\kappa_3}.$$

C2 Additional results

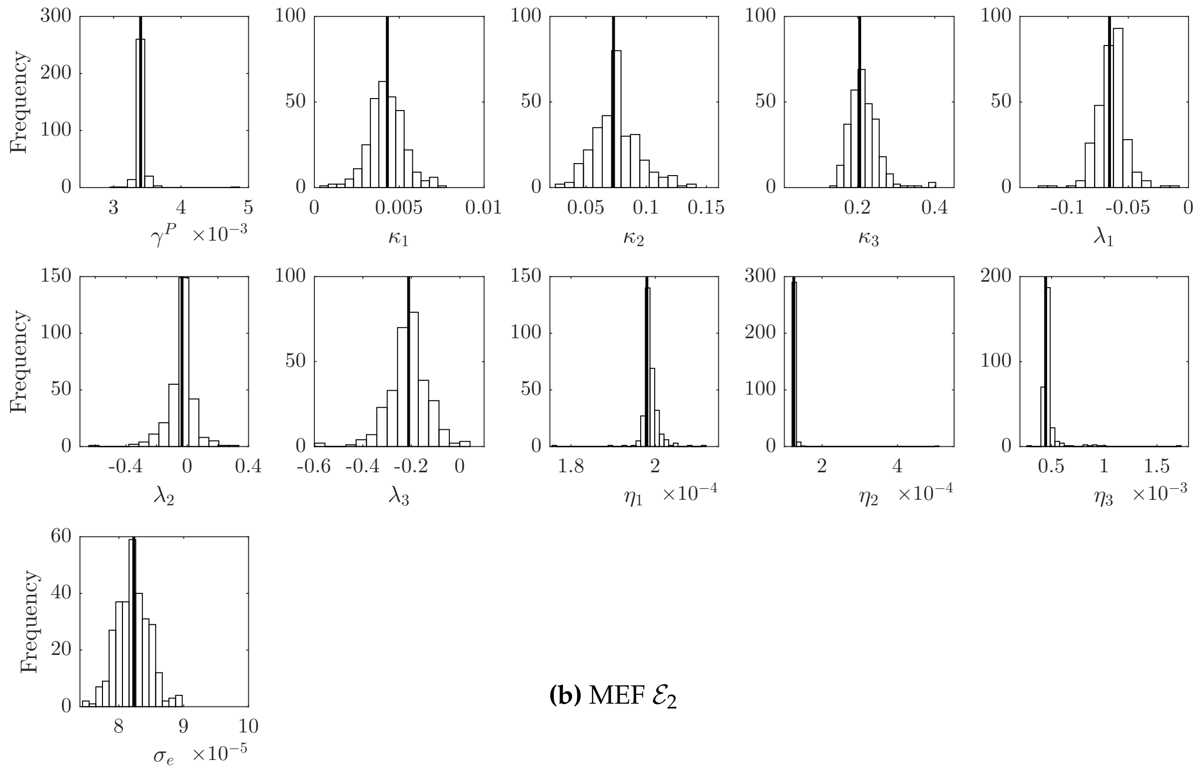
Table C1: DGP, MCSE and MEF estimates, 300 simulations

This table reports the coefficients used for creating the population (DGP), "smart" OLS start values for MEF, and the resulting mean, median and variance of 300 term structures estimated by Kalman filter, MCSE, and MEF. MEF estimation $\mathcal{E}_1$ assumes the proper weighting, whereas $\mathcal{E}_2$ assumes the sum of weighting elements Ψ^{-1} to be rescaled to one. MEF estimations use fminunc Matlab algorithms with the exponential transformation of γ^P, κ, and η. For MCSE we report κ^Q as it corresponds to the κ used for pricing.

γ^P	κ_1	κ_2	κ_3	λ_1	λ_2	λ_3	η_1	η_2	η_3	σ_e
DGP										
0.0034	0.0043	0.0728	0.2053	−0.0656	−0.0339	−0.2113	0.0002	0.0001	0.0004	0.00008
Random Start [mean; median; var]										
0.0048	0.0049	0.0049	0.0048	−0.0053	−0.0052	−0.0050	0.0049	0.0050	0.0049	-
0.0049	0.0051	0.0050	0.0048	−0.0055	−0.0054	−0.0048	0.0049	0.0051	0.0049	-
8.33E-6	7.67E-6	8.10E-6	8.70E-6	8.00E-6	8.31E-6	8.32E-6	7.91E-6	8.49E-6	8.04E-6	-
Kalman filter [mean; median; var], average computational time 64s										
0.0034	0.0033	0.0159	0.1861	−0.0904	0.0724	−0.2134	0.0002	0.0001	0.0004	0.00005
0.0034	0.0040	0.0133	0.1867	−0.0775	0.0067	−0.2152	0.0002	0.0001	0.0004	0.00005
2.55E-8	3.29E-6	8.10E-5	2.57E-5	1.30E-2	6.10E-1	1.89E-3	1.88E-9	2.40E-9	2.75E-10	1.98E-12
MCSE [mean; median; var], average computational time <1s										
0.0033	0.0356	0.0556	0.1821	−0.0742	−0.0646	−0.2045	0.0002	0.0002	0.0003	0.05811
0.0036	0.0055	0.0463	0.1814	−0.0652	−0.0679	−0.2007	0.0002	0.0002	0.0003	0.02978
3.05E-5	1.73E-3	3.07E-3	3.72E-3	2.13E-2	2.32E-2	9.89E-3	4.49E-9	8.54E-9	5.67E-9	2.94E-3
MEF $\mathcal{E}_1$ [mean; median; var], average computational time 42s										
0.0048	0.0201	0.0193	0.0190	−0.0126	−0.1308	−0.0135	0.0038	0.0039	0.0039	0.00014
0.0048	0.0045	0.0043	0.0052	−0.0355	−0.0801	−0.0529	0.0037	0.0033	0.0036	0.00016
8.19E-6	2.40E-3	1.84E-3	1.52E-3	1.19E+0	2.47E-1	1.20E+0	7.36E-6	8.27E-6	7.43E-6	1.32E-9
MEF $\mathcal{E}_2$ [mean; median; var], average computational time 21s										
0.0034	0.0043	0.0752	0.2162	−0.0644	−0.0437	−0.2113	0.0002	0.0001	0.0005	0.00008
0.0034	0.0042	0.0735	0.2108	−0.0638	−0.0336	−0.2115	0.0002	0.0001	0.0004	0.00008
9.86E-9	1.14E-6	3.49E-4	1.41E-3	1.42E-4	8.05E-3	6.59E-3	4.99E-12	4.98E-10	1.10E-8	6.47E-12

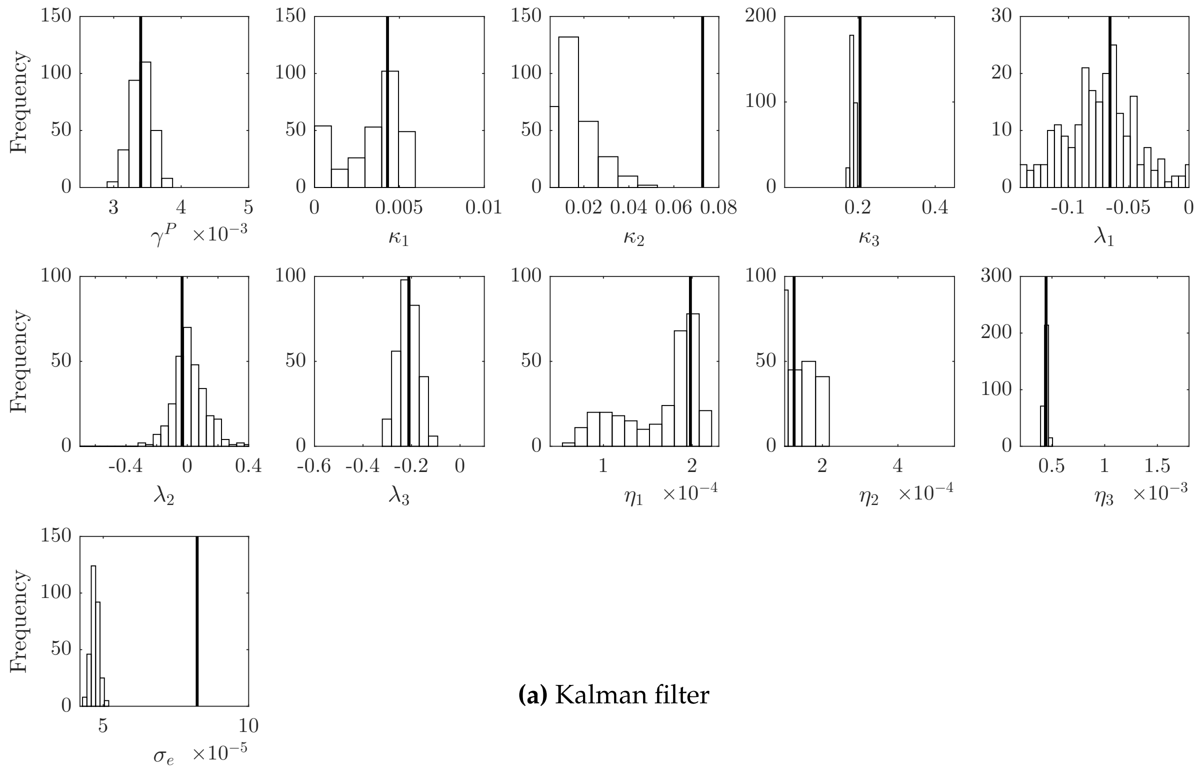


(a) MEF $\mathcal{E}_1$

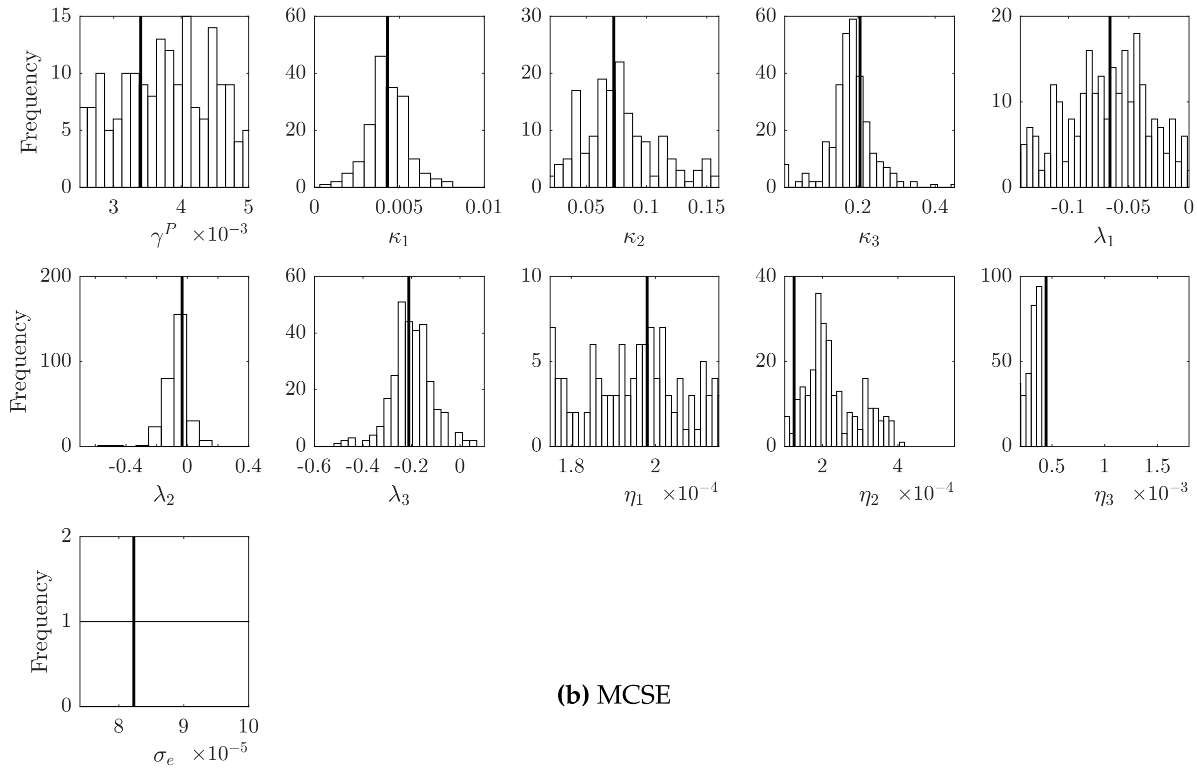


(b) MEF $\mathcal{E}_2$

Figure C1: Histograms for 300 simulations in three-factor setting
The population is simulated from the empirical values. MEF $\mathcal{E}_1$ uses the proper MEF weights, whereas MEF $\mathcal{E}_2$ uses the normalization of Ψ^{-1} such that diagonal elements sum up to one. Both algorithms start from random starting points from uniform distribution $U(0.5/100, 1/100^2)$ with market price of risk starting from negative values. Note that x-axis is normalized for each estimation method to the same lower and upper limit and thus may not include extreme outliers.



(a) Kalman filter



(b) MCSE

Figure C2: Histograms for 300 simulations in three-factor setting
The population is simulated from the empirical values. Note that we diagonalize the factor dynamics equations and that we also restrict $\kappa_i^Q = \kappa_i^P$, whereas MCSE separated the measures. We report κ^Q used for pricing. Bear in mind that some coefficients are off because of the setting.