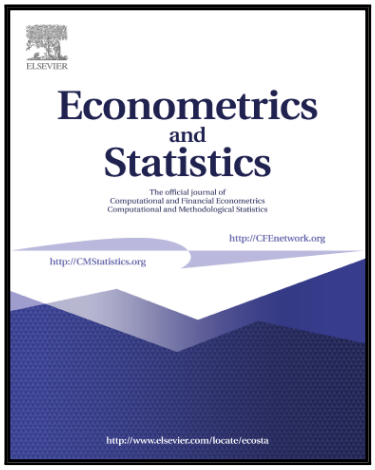
Journal Pre-proof



GMM Estimation of Affine Term Structure Models

Jaroslava Hlouskova, Leopold Sögner

PII: S2452-3062(19)30062-0
DOI: https://doi.org/10.1016/j.ecosta.2019.10.001
Reference: ECOSTA 156

To appear in: *Econometrics and Statistics*

Received date: 16 November 2016
Revised date: 17 October 2019
Accepted date: 17 October 2019

Please cite this article as: Jaroslava Hlouskova, Leopold Sögner, GMM Estimation of Affine Term Structure Models, *Econometrics and Statistics* (2019), doi: https://doi.org/10.1016/j.ecosta.2019.10.001

GMM Estimation of Affine Term Structure Models

Jaroslava Hlouskova

Institute for Advanced Studies, Vienna, Austria

International Institute for Applied Systems Analysis, Laxenburg, Austria

Leopold Sögner*

Institute for Advanced Studies, Vienna, Austria

Vienna Graduate School of Finance (VGSF), Vienna, Austria

Abstract

Parameter estimation of affine term structure models by means of the generalized method of moments is investigated. Exact moments of the affine latent process as well as of the yields are obtained by using results derived for $p-$polynomial processes. Then the generalized method of moments, combined with multi-start random search and Quasi-Bayesian methods, is used to get reliable parameter estimates and to perform inference. After a simulation study, the estimation procedure is applied to empirical interest rate data.

Keywords: Affine term-structure models, GMM, Quasi-Bayesian methods

*Corresponding author. *Email address:* soegner@ihs.ac.at

1 Introduction

This article is concerned with parameter estimation and inference in affine term structure models. We use results of Cuchiero et al. (2012) on $p-$polynomial processes to obtain the exact conditional moments of a latent affine process driving the term structure. By assuming a stationary affine process, we obtain not only the exact moments of yields with various maturities but also the first-order auto-covariance matrices of the yields and the squared yields. Then we estimate the model parameters by means of the *Generalized Method of Moments* (GMM) introduced in Hansen (1982), without the need to estimate the affine latent process driving the yields. Multi-start random search method combined with Quasi-Bayesian approach is used to estimate the model parameters (see, e.g. Törn and Zilinskas, 1989) while Quasi-Bayesian approach is used to estimate the asymptotic covariance matrix of the estimator (see Chernozhukov and Hong, 2003). A further contribution of this paper is a rigorous study on testing market price of risk specifications discussed in quantitative finance literature. By considering a Wald-type test, we observe that test statistics obtained from Quasi-Bayesian methods strongly outperform test statistics which are obtained by standard procedures with respect to power and size.

Affine term structure models have their origin in the univariate models of Vasicek (1977) and Cox et al. (1985). The performance of these models and similar univariate setups were already investigated for example in Aït-Sahalia (1996a) and Aït-Sahalia (1996b). The articles show that these univariate parametric models inadequately describe the interest rate dynamics. Based on this finding Aït-Sahalia (1996a), Aït-Sahalia (1996b) as well as Stanton (1997) propose non-parametric interest rates models. As an alternative, Dai and Singleton (2000) and Dai and Singleton (2003) favor multivariate settings to circumvent the shortcomings of univariate models. This alternative modeling approach has the advantage that a mathematical framework is available, where bonds and derivatives can be priced in a straightforward way.

Let us briefly discuss some literature on the performance of different estimation approaches: Regarding parameter estimation, Zhou (2001) study the efficient method of moments (EMM), the GMM, the quasi-maximum likelihood estimation (QMLE) and the maximum likelihood estimation (MLE) for the Cox et al. (1985) model. In his study the author assumes that the instantaneous interest rate, driven by a square

root process, can be observed. The most efficient results are observed for the MLE, which is followed by the QMLE and the EMM. (For stochastic volatility models Andersen et al. (1999) show that the EMM estimator has almost the same efficiency as the maximum likelihood estimator.) Regarding the GMM, this method performs well if the sample size is sufficiently large. In addition, Zhou (2003) construct a GMM estimator by deriving moments for univariate latent processes by applying Ito's formula. This estimator has been compared to the ML estimator. In contrast to Zhou (2001), in this setup the GMM estimator performs quite well in the finite sample compared to the maximum likelihood estimator.

More recent literature proposes different frequentist and Bayesian approaches to estimate the parameters of multivariate affine term structure models. Bayesian methods have been applied in Chib and Ergashev (2009) while an earlier application is e.g. Frühwirth-Schnatter and Geyer (1996). Regarding Bayesian estimation methods, Jones (2003) points out that strong priors are necessary to estimate the parameters in the case of a low degree of mean reversion (i.e. high persistence) of the stochastic process. MLE has been performed in a three factor Gaussian model (an $\mathbb{A}_0(3)$ model in the terminology of Dai and Singleton, 2000) by Hamilton and Wu (2012).

Additional articles on parameter estimation for affine models are e.g. Diebold et al. (2006), Duffee (2011), Aït-Sahalia and Kimmel (2010), Egorov et al. (2011), Joslin et al. (2011) and Creal and Wu (2015). An overview is provided in Piazzesi (2010). A further approach is to approximate the transition density of the affine process via approximations of the Chapman/Kolmogorov forward equation. This approach is explored in series of papers by Aït-Sahalia (see, e.g. Aït-Sahalia, 2002; Aït-Sahalia and Kimmel, 2010).

Almost recently Creal and Wu (2015) introduce a new procedure to estimate the model parameters by means of maximum likelihood. In particular, the authors decompose the estimation problem into maximizing a concentrated likelihood function and running a generalized least squares regression. The main difference in terms of the model is that in this article *all* yields are observed with noise, while Creal and Wu (2015) assume that the number of yields observed without noise is equal to the dimension of the affine latent process driving the term structure. While the latter approach directly allows to extract the latent process from the yields observed without noise (see, in particular Creal and Wu, 2015, Procedure 1.(i.)), the latent process driving the term structure cannot be obtained reliably by an affine

linear transformation if all yields are observed with noise. Even if the noise is small for all maturities, we observe in our study that the differences between a process obtained from an affine transformation and the (true) latent process can be substantial.

Unlike our study, papers by de Los Rios (2015), Hamilton and Wu (2012, 2014), Joslin et al. (2011) as well as Joslin et al. (2013) deal with Gaussian affine term structure models where (some of the) factors are assumed to be observable. Except for Joslin et al. (2013) these studies assume that certain yields are measured without errors, while we do not impose this assumption. de Los Rios (2015) proposes the asymptotic least-square estimator that can be obtained without applying any numerical optimization techniques and thus is relatively easy-to-compute (while our approach suffers a bit from numerical optimization as we deal with non-Gaussian term structure models). This asymptotic least-square estimator of de Los Rios (2015) is asymptotically equivalent to the maximum likelihood estimator of Joslin et al. (2011). On the other hand, the minimum-chi-square estimator proposed in Hamilton and Wu (2012) that bypasses some numerical challenges, is also asymptotically equivalent to the maximum likelihood estimator and can be viewed as a special case of minimum distance estimator. Their assumption that certain specified yields are priced without errors is testable as shown in Hamilton and Wu (2014). Finally, Joslin et al. (2013) explore the maximum likelihood estimates for Gaussian macro-finance term structure models when yields are priced imperfectly. By contrast, we assume an affine term structure model where in addition to two Gaussian factors one square-root component shows up and all yields are subject to noise. Our parameter estimation approach uses the generalized method of moments and does not require to estimate the latent factors.

In this article, we use the *exact* moments of the yields observed, arising from a multivariate affine term structure model. Neither an approximation of the moments (such as an approximation via the solution of the stochastic differential equation) nor an approximation of the likelihood is required. Since we have to minimize a GMM distance function in more than twenty parameters, the parameter estimation is nontrivial. To account for this problem, we combine multi-start random search method with Quasi-Bayesian methods developed in Chernozhukov and Hong (2003). We observe that, in contrast to standard optimization routines, multi-start random search method combined with Quasi-Bayesian approach improve

both parameter estimation and inference (see, e.g., Chernozhukov and Hong, 2003). By contrast, when using standard routines to estimate the asymptotic covariance matrix of the unknown parameter vector, the performance of the Wald-type test, measured in terms of power and size, is very poor. Thus, we use our methodology to test for *the extended affine market price of risk specification* as proposed and analytically investigated in Cheridito et al. (2007) (for an extension in discrete time term structure models see, e.g., Le et al. (2010)). In an empirical interest rates data set significant market prices of risk are observed for the parameters driving the level of the interest rates as well as for the parameters driving the speed of mean reversion.

This paper is organized as follows: Section 2 describes the model assumptions and obtains the moments of the latent process as well as the yields observed. Section 3 describes the small sample properties of the GMM estimator, while Section 4 applies the estimator to empirical data. Finally, Section 5 offers conclusions.

2 Model

We follow Filipović (2009) and consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ as well as a continuous time stochastic process $(\mathbf{X}(t))_{t\geq 0}$, $\mathbf{X}(t) \in \mathbb{R}^d$, defined on the state space $\mathscr{S} = \mathbb{R}_+^m \times \mathbb{R}^n \subset \mathbb{R}^d$, where $m, n \geq 0$, $m + n = d$. The stochastic process $(\mathbf{X}(t))_{t\geq 0}$ is generated by the affine stochastic differential equation

$$d\mathbf{X}(t) = \left(\mathbf{b}^P + \boldsymbol{\beta}^P \mathbf{X}(t)\right) dt + \boldsymbol{\rho}(\mathbf{X}(t)) d\mathbf{W}^P(t) \ , \tag{1}$$

where $\mathbf{b}^P$ is a $d-$dimensional vector and $\boldsymbol{\beta}^P$ and $\boldsymbol{\rho}(\mathbf{x})$ are $d \times d$ matrices. The $d \times d$ diffusion term $\mathbf{a}(\mathbf{x})$ is defined such that $\mathbf{a}(\mathbf{x}) = \boldsymbol{\rho}(\mathbf{x})\boldsymbol{\rho}(\mathbf{x})' = \mathbf{a} + \sum_{i=1}^d x_i \ \boldsymbol{\alpha}_i$, where $\mathbf{a}$, $\boldsymbol{\alpha}_i$, $i = 1, \ldots, d$, are $d \times d$ matrices, $\mathbf{W}^P(t)$ is a $d-$dimensional standard Brownian motion and $\mathbb{P}$ is the empirical measure.

In this article we apply the following notation: For vectors and matrices we use boldface. If not otherwise stated, the vectors considered are column vectors. Given a $r_M \times c_M$ matrix $\mathbf{M}$, the term $\mathbf{M}_{r_a:r_b, c_a:c_b}$ stands for "from row r_a to row r_b and from column c_a to column c_b of matrix $\mathbf{M}$". The abbreviation $\mathbf{M}_{r_a:r_b,:}$ stands for "for all columns from row r_a to row r_b of matrix $\mathbf{M}$", and $\mathbf{M}_{r_a:r_b, c_a}$

extracts the elements r_a to r_b of the column c_a. In addition, β_{ij} stands for $[\boldsymbol{\beta}]_{ij}$; $\mathbf{0}_{a\times b}$ and $\mathbf{e}_{a\times b}$ stand for $a \times b$ matrices of zeros and ones; $\mathbf{0}_a$ and $\mathbf{e}_a$ is used to abbreviate $\mathbf{0}_{a\times 1}$ and $\mathbf{e}_{a\times 1}$; $\mathbf{I}_a$ is the $a \times a$ identity matrix, while $\mathbb{I}_{(.)}$ stands for an indicator function. Given a vector $\mathbf{x} \in \mathbb{R}^n$, $diag(\mathbf{x})$ transforms $\mathbf{x}$ into a $n \times n$ diagonal matrix.

The *instantaneous interest rate (short rate, $r(t) \in \mathbb{R}$)* follows from

$$r(t) = \gamma_0 + \boldsymbol{\gamma}_x' \mathbf{X}(t) \ , \tag{2}$$

where γ_0 is a scalar and $\boldsymbol{\gamma}_x$ is a $d-$dimensional vector. Consider an arbitrage free market, where $\mathbb{Q}$ is an equivalent martingale measure to the empirical $\mathbb{P}$ measure. We assume that the process $(\mathbf{X}(t))_{t\geq 0}$ is affine also in the measure $\mathbb{Q}$, such that

$$d\mathbf{X}(t) = \left(\mathbf{b}^Q + \boldsymbol{\beta}^Q \mathbf{X}(t)\right) dt + \boldsymbol{\rho}(\mathbf{X}(t)) d\mathbf{W}^Q(t) \ , \tag{3}$$

where $\mathbf{W}^Q(t)$ is a $d-$dimensional standard Brownian motion under $\mathbb{Q}$ measure. By equations (1) and (3), the stochastic process $(\mathbf{X}(t))_{t\geq 0}$ is affine in both measures. While the diffusion parameters ($\mathbf{a}$, $\boldsymbol{\alpha}_i$, $i = 1, \ldots, d$) remain the same under both measures, we have to consider parameters $\mathbf{b}^P$, $\boldsymbol{\beta}^P$, $\mathbf{b}^Q$ and $\boldsymbol{\beta}^Q$, in both measures $\mathbb{P}$ and $\mathbb{Q}$. This specification, namely equations (1) and (3), is called *the extended affine market price of risk specification*, and its mathematical foundation is provided in Cheridito et al. (2007). These authors also show by means of the Girsanov theorem that $\mathbf{W}^Q(t) = \mathbf{W}^P(t) + \int_0^t \boldsymbol{\phi}(\mathbf{X}(s))ds$, where $\boldsymbol{\phi}(\mathbf{X}(t)) \in \mathbb{R}^d$ is given by $\boldsymbol{\phi}(\mathbf{X}(t)) = \left(\boldsymbol{\rho}\left(\mathbf{X}(t)\right)\right)^{-1} \left(\mathbf{b}^P - \mathbf{b}^Q + \left(\boldsymbol{\beta}^P - \boldsymbol{\beta}^Q\right) \mathbf{X}(t)\right)$. To connect the market price of risk to risk premia see Cochrane (2005)[p. 339].

In the remaining part of this article we apply the following assumption.

Assumption 1. *The background driving process $(\mathbf{X}(t))$ is stationary and admissible (under both measures). In addition, $\mathbb{E}\left(\exp(-\int_0^{\bar{\tau}} r(z)dz)\right) < +\infty$, for some $\bar{\tau} \in \mathbb{R}_+$.*

Sufficient conditions for a stationary process $(\mathbf{X}(t))$ are provided in Glasserman and Kim (2010) and in the context of $\mathbb{A}_m(d)$ models that we use here they are also reported in Aït-Sahalia and Kimmel (2010) and in Online-Appendix A-6. If admissibility holds, the process $(\mathbf{X}(t))$ does not leave the state space $\mathscr{S}$.

Sufficient conditions for an admissibility follow from Theorem 10.2 in Filipović (2009).

Next, we define the index sets $I = \{1, \ldots, m\}$ and $J = \{m+1, \ldots, m+n\}$, where $m + n = d$. Let $\mathbf{b}_I = (b_1, \ldots, b_m)'$ and $\boldsymbol{\beta}_{II} = \boldsymbol{\beta}_{1:m,1:m}$. The admissibility restrictions, the short-rate model (2) and the condition $\mathbb{E}\left(\exp(-\int_0^{\bar{\tau}} r(z)dz)\right) < +\infty$, for some $\bar{\tau} \in \mathbb{R}_+$, imply that there exists a unique solution $(\Phi(t,\mathbf{u}), \boldsymbol{\Psi}(t,\mathbf{u})')' \in \mathbb{C} \times \mathbb{C}^d$ of the following system of Riccati differential equations

$$\begin{aligned}
\partial_t \Phi(t,\mathbf{u}) &= \tfrac{1}{2}\left(\boldsymbol{\Psi}_J(t,\mathbf{u})\right)' \mathbf{a}_{JJ} \boldsymbol{\Psi}_J(t,\mathbf{u}) + \left(\mathbf{b}^Q\right)' \boldsymbol{\Psi}(t,\mathbf{u}) - \gamma_0; \qquad \Phi(0,\mathbf{u}) = 0\ , \\
\partial_t \Psi_i(t,\mathbf{u}) &= \tfrac{1}{2}\left(\boldsymbol{\Psi}(t,\mathbf{u})\right)' \boldsymbol{\alpha}_i \boldsymbol{\Psi}(t,\mathbf{u}) + \left(\boldsymbol{\beta}_i^Q\right)' \boldsymbol{\Psi}(t,\mathbf{u}) - \gamma_{xi}; \qquad \text{for } i \in I\ , \\
\partial_t \boldsymbol{\Psi}_J(t,\mathbf{u}) &= \left(\boldsymbol{\beta}_{JJ}^Q\right)' \boldsymbol{\Psi}_J(t,\mathbf{u}) - \boldsymbol{\gamma}_{xJ}; \qquad \boldsymbol{\Psi}(0,\mathbf{u}) = \mathbf{u}\ ,
\end{aligned} \tag{4}$$

where $t \in [0, \bar{\tau}]$, $\mathbf{u} \in \imath\mathbb{R}^d$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$, with $\boldsymbol{\beta}_i$ being a d–dimensional vector, $i = 1, \ldots, d$ (see Filipović, 2009, Theorem 10.4). This system of ordinary differential equations is used to calculate the time t *price of a zero coupon bond*, $\pi^0(t,\tau)$, with time to maturity τ. The arbitrage free zero coupon model prices $\pi^0(t,\tau)$ and the *model yields* $y^0(t,\tau)$ follow from Filipović (2009)[Corollary 10.2]. That is

$$\begin{aligned}
\pi^0(t,\tau) &= \exp\left(\Phi(\tau,\mathbf{0}) + \boldsymbol{\Psi}(\tau,\mathbf{0})'\mathbf{X}(t)\right) \text{ and} \\
y^0(t,\tau) &= -\frac{1}{\tau}\log\left(\pi^0(t,\tau)\right) = -\frac{1}{\tau}\left(\Phi(\tau,\mathbf{0}) + \boldsymbol{\Psi}(\tau,\mathbf{0})'\mathbf{X}(t)\right).
\end{aligned} \tag{5}$$

The time to maturity, τ, and $\mathbf{u}$ are the arguments of functions $\Phi(t,\mathbf{u})$ and $\boldsymbol{\Psi}(t,\mathbf{u})$ described in (4). Note that parameters under $\mathbb{Q}$ are necessary for derivations of functions $\Phi(\tau,\mathbf{0})$ and $\boldsymbol{\Psi}(\tau,\mathbf{0})$ by means of which the model yields are calculated, see (5).

In the following we consider an $\mathbb{A}_1(3)$ model of the Dai and Singleton (2000) class, where $m = 1$ and $d = 3$. Let $\boldsymbol{\theta}^Q = (\theta_1^Q, \theta_2^Q, \theta_3^Q)'$ and $\boldsymbol{\theta}^P = (\theta_1^P, \theta_2^P, \theta_3^P)'$, such that $\mathbf{b}^Q = -\boldsymbol{\beta}^Q\boldsymbol{\theta}^Q$ as well as $\mathbf{b}^P = -\boldsymbol{\beta}^P\boldsymbol{\theta}^P$.

Under the measure $\mathbb{Q}$,

$$d\mathbf{X}(t) = \left(\mathbf{b}^Q + \boldsymbol{\beta}^Q\mathbf{X}(t)\right)dt + \boldsymbol{\rho}(\mathbf{X}(t))d\mathbf{W}^Q(t) = \boldsymbol{\beta}^Q\left(-\boldsymbol{\theta}^Q + \mathbf{X}(t)\right)dt + \boldsymbol{\rho}(\mathbf{X}(t))d\mathbf{W}^Q(t)\ ,\ \text{where}$$

$$\boldsymbol{\beta}^Q = \begin{pmatrix} \beta_{11}^Q < 0 & 0 & 0 \\ \beta_{21}^Q \geq 0 & \beta_{22}^Q & \beta_{23}^Q \\ \beta_{31}^Q \geq 0 & \beta_{32}^Q & \beta_{33}^Q \end{pmatrix},\quad \boldsymbol{\theta}^Q = \begin{pmatrix} \theta_1^Q > 0 \\ 0 \\ 0 \end{pmatrix},\quad \mathbf{b}^Q = -\boldsymbol{\beta}^Q\boldsymbol{\theta}^Q = \begin{pmatrix} b_1^Q = -\beta_{11}^Q\theta_1^Q > 0 \\ b_2^Q = -\beta_{21}^Q\theta_1^Q \leq 0 \\ b_3^Q = -\beta_{31}^Q\theta_1^Q \leq 0 \end{pmatrix},$$

$$\text{and}\quad \boldsymbol{\rho}(\mathbf{X}(t)) = \begin{pmatrix} \Sigma_1\sqrt{X_1(t)} & & \\ & \Sigma_2\sqrt{1+\mathcal{B}_{12}^x X_1(t)} & \\ & & \Sigma_3\sqrt{1+\mathcal{B}_{13}^x X_1(t)} \end{pmatrix}, \tag{6}$$

where under admissibility conditions, as discussed in Dai and Singleton (2000), the following restrictions apply: $\theta_1^Q > 0$, $\beta_{11}^Q < 0$, $\mathcal{B}_{12}^x, \mathcal{B}_{13}^x \geq 0$, and $\Sigma_1, \Sigma_2, \Sigma_3 > 0$. Note that (6) has 13 parameters while under $\mathbb{Q}$ we can identify 14 parameters. These parameters are the thirteen parameters in (6) and γ_0 arising in (2). In more detail: $\boldsymbol{\beta}^Q$ (7 parameters), θ_1^Q (1 parameter), $\boldsymbol{\Sigma}$ (3 parameters, only the elements in the main diagonal are positive, the other parameters are zero), $\mathcal{B}_{12}^x \geq 0$ and $\mathcal{B}_{13}^x \geq 0$. The same structure is assumed under $\mathbb{P}$. That is, the elements of matrix $\boldsymbol{\beta}^P$ are $\beta_{11}^P \leq 0$, $\beta_{12}^P = 0$, $\beta_{13}^P = 0$, $\beta_{21}^P \geq 0$, $\beta_{31}^P \geq 0$, β_{22}^P, β_{32}^P, β_{23}^P, β_{33}^P, while the coordinates of $\boldsymbol{\theta}^P$ are $\theta_1^P \geq 0$, $\theta_2^P = 0$ and $\theta_3^P = 0$. Since $\boldsymbol{\theta}_{2:3}^P = \boldsymbol{\theta}_{2:3}^Q = \mathbf{0}_2$ for the $\mathbb{A}_1(3)$ model considered, we write θ^Q and θ^P instead for θ_1^Q and θ_1^P in the following. Based on Cheridito et al. (2007) the *extended affine market price of risk* specification is mathematically well defined given that $b_I^P = b_1^P \geq 0$, $b_J^P = (b_2^P, b_3^P)' \leq 0$, which is satisfied in the model we consider. The non-zero restricted elements of these matrices are collected in $\bar{\boldsymbol{\beta}}^Q = \left(\beta_{11}^Q, \beta_{21}^Q, \beta_{31}^Q, \beta_{22}^Q, \beta_{32}^Q, \beta_{23}^Q, \beta_{33}^Q\right)'$ and $\bar{\boldsymbol{\beta}}^P = \left(\beta_{11}^P, \beta_{21}^P, \beta_{31}^P, \beta_{22}^P, \beta_{32}^P, \beta_{23}^P, \beta_{33}^P\right)'$. By collecting these parameters (not subject to an equality restriction), we obtain the vector of model parameters $\boldsymbol{\vartheta}_{\mathbb{A}_{1(3)}} \in \mathbb{R}^{22}$.

2.1 Moments of the Latent Process $(\mathbf{X}(t))$

To obtain the first moments of order p of the latent process $(\mathbf{X}(t))$, let $\mathbf{x}^k = \left(x_1^k, x_1^{k-1}x_2, \ldots, x_3^k\right)'$, where $k = 0, 1, \ldots, p$. In addition, $\tilde{\mathbf{x}} = (1, (\mathbf{x}^1)', (\mathbf{x}^2)', \ldots, (\mathbf{x}^p)')' \in \mathbb{R}^N$, while $\tilde{\mathbf{x}}_{2:N} = ((\mathbf{x}^1)', (\mathbf{x}^2)', \ldots, (\mathbf{x}^p)')' \in \mathbb{R}^{N-1}$. The processes $\tilde{\mathbf{X}}(t)$ and $\tilde{\mathbf{X}}(t)_{2:N}$ are defined in the same way. The number of all moments of the

latent process $\mathbf{X}(t)$, denoted by N, follows from the corresponding multinomial coefficients. To obtain conditional moments $\mathbb{E}(\tilde{\mathbf{X}}(t)|\mathbf{X}(s) = \mathbf{x})$, $t > s$, we apply results derived in Cuchiero et al. (2012) on $p-$polynomial Markov processes, resulting in $\mathbb{E}(\tilde{\mathbf{X}}(t)|\mathbf{X}(s) = \mathbf{x}) = \exp((t-s)\mathbf{A})\,\tilde{\mathbf{x}}$, where $\mathbf{A}$ is an $N \times N$ matrix. Appendices A-2.2 and A-3 present matrix $\mathbf{A}$ for an affine model with $d \leq 3$ components and moments of order $p = 4$.

As will be discussed in Section 3, the first and the second order moments of the yields will be used to perform GMM-parameter estimation. To obtain these moments of the yields (as described in Section 2.2), we derive the first and the second conditional moments of $\mathbf{X}(t)$. In particular, to obtain the first and the second conditional moments of $\mathbf{X}(t)$ for $\mathbb{A}_1(3)$ model (i.e., $p = 2$) we derive the following matrix, where $N = 10$ and $\tilde{\mathbf{x}} = (1, (\mathbf{x}^1)', (\mathbf{x}^2)')' \in \mathbb{R}^{10}$ (note that $\tilde{\mathbf{x}}$ contains the $d = 3$ dimensional vector $\mathbf{x}^1$ and the $\frac{d(d+1)}{2} = 6$ dimensional vector $\mathbf{x}^2$),

$$\mathbf{A} = \left(\begin{array}{c|ccc|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline b_1^P & \beta_{11}^P & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_{21}^P & \beta_{22}^P & \beta_{23}^P & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_{31}^P & \beta_{32}^P & \beta_{33}^P & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 2b_1^P + \Sigma_1^2 & 0 & 0 & 2\beta_{11}^P & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & b_1^P & 0 & \beta_{21}^P & \beta_{11}^P + \beta_{22}^P & \beta_{23}^P & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1^P & \beta_{31}^P & \beta_{32}^P & \beta_{11}^P + \beta_{33}^P & 0 & 0 & 0 \\ \Sigma_2^2 & \Sigma_2^2\mathcal{B}_{12}^x & 0 & 0 & 0 & 2\beta_{21}^P & 0 & 2\beta_{22}^P & 2\beta_{23}^P & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{31}^P & \beta_{21}^P & \beta_{32}^P & \beta_{22}^P + \beta_{33}^P & \beta_{23}^P \\ \Sigma_3^2 & \Sigma_3^2\mathcal{B}_{13}^x & 0 & 0 & 0 & 0 & 2\beta_{31}^P & 0 & 2\beta_{32}^P & 2\beta_{33}^P \end{array}\right). \quad (7)$$

Since $(\mathbf{X}(t))$ is assumed to be stationary and $\mathbb{E}\left(\tilde{\mathbf{X}}(t)\right) = \mathbb{E}\left(\mathbb{E}\left(\tilde{\mathbf{X}}(t)|\mathbf{X}(s)\right)\right)$, for $0 \leq s < t$, the tower

rule yields

$$\begin{aligned}\mathbb{E}\left(\tilde{\mathbf{X}}(t)\right) &= \begin{pmatrix} 1 \\ \mathbb{E}\left(\tilde{\mathbf{X}}(t)_{2:N}\right) \end{pmatrix} = \mathbb{E}\left(\left[\exp((t-s)\mathbf{A})\right]\tilde{\mathbf{X}}(t)\right) = \left[\exp((t-s)\mathbf{A})\right]\mathbb{E}\left(\tilde{\mathbf{X}}(t)\right) \\ &= \begin{pmatrix} 1 & \mathbf{0}_{1\times N-1} \\ \left[\exp((t-s)\mathbf{A})\right]_{2:N,1} & \left[\exp((t-s)\mathbf{A})\right]_{2:N,2:N} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbb{E}\left(\tilde{\mathbf{X}}(t)_{2:N}\right) \end{pmatrix} . \end{aligned} \tag{8}$$

The $N \times N$ matrix $\exp((t-s)\mathbf{A})$ can be partitioned into four blocks: (i) north-western $\left[\exp((t-s)\mathbf{A})\right]_{11} = 1$, (ii) north-eastern $\left[\exp((t-s)\mathbf{A})\right]_{1,2:N} = \mathbf{0}_{1\times N-1}$, (iii) south-western $\left[\exp((t-s)\mathbf{A})\right]_{2:N,1}$, and (iv) south-eastern $\left[\exp((t-s)\mathbf{A})\right]_{2:N,2:N}$. Hence, the (unconditional) moments of order 1 to p follow from

$$\mathbb{E}\left(\tilde{\mathbf{X}}(t)_{2:N}\right) = \left(\mathbf{I}_{N-1} - \left[\exp((t-s)\mathbf{A})\right]_{2:N,2:N}\right)^{-1} \left[\exp((t-s)\mathbf{A})\right]_{2:N,1} .$$

Note that $\exp((t-s)\mathbf{A})$ and $\mathbf{A}$ are of the same structure. This follows from the power series representation of the matrix exponential $\exp((t-s)\mathrm{A}) = \sum_{v=0}^{\infty} \frac{1}{v} \left((t-s)\mathbf{A}\right)^v$. In addition, the existence of $\left(\mathbf{I}_{N-1} - \left[\exp((t-s)\mathbf{A})\right]_{2:N,2:N}\right)^{-1}$ follows from the properties of the matrix exponential.

2.2 Moments of the Observed Yields

This section deals with the case of empirical data, when the number of yields observed is larger than the dimension of $(\mathbf{X}(t))_{t\geq 0}$ and thus the yields observed cannot be matched exactly with the model yields derived in (5). For an affine term structure model the *model yields* with time to maturity τ are

$$y^0(t,\tau) = -\frac{1}{\tau}\left(\Phi(\tau,\mathbf{0}) + \mathbf{\Psi}(\tau,\mathbf{0})'\mathbf{X}(t)\right).$$

The calculation of the moments also requires to solve the Riccati equations (4). For the Vasicek and the Cox-Ingersol-Ross model closed form solutions are available, as e.g. presented in Filipović (2009)[Chapter 10.3.2]. For $\mathbb{A}_m(d)$ models, however, Φ and $\mathbf{\Psi}$ have to be derived by means of numerical tools in general (see also Duffie and Kan, 1996; Dai and Singleton, 2000; Chen and Joslin, 2012). The fact that $\beta_{11}^Q = \boldsymbol{\beta}_{II}^Q$ is a scalar in the $\mathbb{A}_1(3)$ model described in (6), allows to apply the computationally efficient

method proposed by Grasselli and Tebaldi (2008) to obtain an (almost) closed form solution for $\Phi(t, \mathbf{u})$ and $\mathbf{\Psi}(t, \mathbf{u})$. This methodology requires the matrix $\boldsymbol{\beta}_{II}^{Q}$ to be diagonal. Our Online-Appendix A-5 shows how Φ and $\mathbf{\Psi}$ could be derived for an $\mathbb{A}_m(d)$ model with diagonal $\boldsymbol{\beta}_{II}$ in a numerically parsimonious way.

Now we have to account for the fact that real world data cannot be observed on a continuous time scale, but only on a discrete grid $\Delta, 2\Delta, \ldots, \mathtt{t}\Delta, \ldots, T\Delta$, where T is the time series dimension and Δ is the step-width. As we use weekly data in empirical Section 4, we set $\Delta = 1/52$ and assume that $\mathbf{X}_{\mathtt{t}}$ stands for $\mathbf{X}(\mathtt{t}\Delta)$. Additionally, maturities τ available are given by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)'$, where M is the number of maturities observed. For model yields with a maturity $\tau_i \in \{\tau_1, \ldots, \tau_M\}$ observed at $t = \mathtt{t}\Delta$ we use the notation $y^0_{\mathtt{t}i}$, $i = 1, \ldots, M$. Since M yields cannot be matched exactly by d factors (usually $d < M$), we add the noise term $\varepsilon_{\mathtt{t}i}$ and arrive at the *yields observed*

$$y_{\mathtt{t}i} \;=\; y^0_{\mathtt{t}i} + \varepsilon_{\mathtt{t}i} = -\frac{1}{\tau_i}\left(\Phi(\tau_i, \mathbf{0}) + \mathbf{\Psi}(\tau_i, \mathbf{0})'\mathbf{X}_{\mathtt{t}}\right) + \varepsilon_{\mathtt{t}i}, \quad i = 1, \ldots, M, \quad \mathtt{t} = 1, \ldots, T.$$

With M maturities $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$ we define

$$\tilde{\mathbf{\Phi}} \;=\; \begin{pmatrix} -\Phi(\tau_1, \mathbf{0})/\tau_1 \\ \vdots \\ -\Phi(\tau_M, \mathbf{0})/\tau_M \end{pmatrix} \in \mathbb{R}^M, \quad \tilde{\mathbf{\Psi}} = \begin{pmatrix} -\mathbf{\Psi}(\tau_1, \mathbf{0})'/\tau_1 \\ \ldots \\ -\mathbf{\Psi}(\tau_M, \mathbf{0})'/\tau_M \end{pmatrix} \in \mathbb{R}^{M \times d} \quad \text{and} \quad \boldsymbol{\varepsilon}_{\mathtt{t}} = \begin{pmatrix} \varepsilon_{\mathtt{t}1} \\ \vdots \\ \varepsilon_{\mathtt{t}M} \end{pmatrix} \in \mathbb{R}^M,$$

such that the $M-$dimensional vector of yields, $\mathbf{y}_{\mathtt{t}} = (y_{\mathtt{t}1}, \ldots, y_{\mathtt{t}M})'$, is given by

$$\mathbf{y}_{\mathtt{t}} \;=\; \tilde{\mathbf{\Phi}} + \tilde{\mathbf{\Psi}}\mathbf{X}_{\mathtt{t}} + \boldsymbol{\varepsilon}_{\mathtt{t}} \in \mathbb{R}^M. \tag{9}$$

Based on (9) we observe that the moments of $y_{\mathtt{t}i}$ have to follow from the moments of $\mathbf{X}_{\mathtt{t}}$. For the noise term $\varepsilon_{\mathtt{t}i}$ we apply the following assumption.

Assumption 2. *Let $\varepsilon_{\mathtt{t}i}$, $\mathtt{t} = 1, \ldots, T$, $i = 1, \ldots, M$, be independent with zero mean, variance $0 < \sigma_i^2 < +\infty$ and $\mathbb{E}(\varepsilon_{\mathtt{t}i}^4) < +\infty$. In addition, $|\mathbb{E}(\varepsilon_{\mathtt{t}i}^p)| < +\infty$ for $i = 1, \ldots, M$ and $\mathbb{E}\left(\varepsilon_{\mathtt{t}i}^{2\iota-1}\right) = 0$ for $\iota = 1, \ldots, \lfloor p/2 \rfloor$, where $\lfloor p/2 \rfloor$ is the largest integer smaller or equal to $p/2$.*

Note that by Assumption 2, the yields of all maturities are assumed to be observed with noise. In addition,

$\mathbb{E}\left(\varepsilon_{\mathtt{t}i}\varepsilon_{\mathtt{t}j}\right) = 0$ for $i \neq j$, $i, j = 1, \ldots, M$ and $\mathbb{E}\left(\varepsilon_{\mathtt{t}i}^4\right) < +\infty$. By means of equation (9) and Assumption 2 we derive the moments of the empirical yields $\mathbb{E}\left(y_{\mathtt{t}i}^k y_{\mathtt{t}j}^l\right) = \mathbb{E}\left(\left([\tilde{\boldsymbol{\Phi}} + \tilde{\boldsymbol{\Psi}}\mathbf{X}_\mathtt{t} + \boldsymbol{\varepsilon}_\mathtt{t}]_i\right)^k \left([\tilde{\boldsymbol{\Phi}} + \tilde{\boldsymbol{\Psi}}\mathbf{X}_\mathtt{t} + \boldsymbol{\varepsilon}_\mathtt{t}]_j\right)^l\right)$, where $0 \leq k + l \leq p$ and $[\cdot]_i$ extracts the i-th element of a vector. Hence, we derive the first four moments of the yields observed, i.e. $\mathbb{E}\left(y_{\mathtt{t}i}^k\right)$, $k = 1, \ldots, 4$. In addition, applications in finance often take the auto-covariance of the yields, $\mathbb{E}(y_{\mathtt{t}i}y_{\mathtt{t}-1i})$, and the auto-covariance of the squared yields, $\mathbb{E}\left(y_{\mathtt{t}i}^2 y_{\mathtt{t}-1i}^2\right)$, into consideration (the auto-covariance of the squared yields is considered as an "indicator for volatility clustering" - see, e.g. the discussion in Piazzesi (2010)[p. 649]). Therefore also the terms $\mathbb{E}(y_{\mathtt{t}i}y_{\mathtt{t}-1i})$ and $\mathbb{E}\left(y_{\mathtt{t}i}^2 y_{\mathtt{t}-1i}^2\right)$ are calculated. Since this part is straightforward, but tedious algebraic manipulations were necessary to obtain all these moments, we present the results in the Online-Appendix A-4. We put the noise parameters necessary to obtain the moments of the observed yields into the parameter vector $\boldsymbol{\vartheta}_\sigma$. The dimension of $\boldsymbol{\vartheta}_\sigma$ depends on how σ_i^2, $i = 1, \ldots, M$, is specified and on the moments used in the estimation. If σ_i^2 is different for each maturity, we have M parameters for the second order moments of the noise. If, in addition, the fourth moments of the yields are calculated, the fourth moments of the noise enter into the calculations as well, i.e. we get another M parameters for the moments of the noise. In this case the dimension of $\boldsymbol{\vartheta}_\sigma$ is $2M$. Since the dimension of the model parameter $\boldsymbol{\vartheta}_{\mathbb{A}_{1(3)}}$ is already over twenty, we continue with a more parsimonious specification of the noise, where $\sigma_i^2 = \sigma^2$ and $\mathbb{E}\left(\varepsilon_{\mathtt{t}i}^4\right) = \tilde{\sigma}^4$ for all $i = 1, \ldots, M$. Hence, the dimension of $\boldsymbol{\vartheta}_\sigma$ is two if fourth moments are required in the calculation of the yields observed, otherwise it is one. This results in the model parameter vector $\boldsymbol{\vartheta} = \left(\boldsymbol{\vartheta}'_{\mathbb{A}_{1(3)}}, \boldsymbol{\vartheta}'_\sigma\right)'$ of dimension $\mathfrak{p}$, which is contained in the parameter space $\boldsymbol{\Theta} \in \mathbb{R}^\mathfrak{p}$. Note that due to parameter restrictions (see Appendix A-6) $\boldsymbol{\Theta}$ is proper subset of $\mathbb{R}^\mathfrak{p}$. The components of $\boldsymbol{\vartheta}$ are introduced by the first column of Table 1.

3 Parameter Estimation and Finite Sample Properties

In this section we describe the estimation procedure and its inference which allows parameter estimation based on observed yields, but without estimating the latent process $\mathbf{X}_\mathtt{t}$, $\mathtt{t} = 1, \ldots, T$. Section 3.1 applies GMM, where the parameter estimates are obtained by means of a multi-start random search procedure combined with a Quasi-Bayesian sampler, while Section 3.2 describes how to conduct inference. Here,

in addition, we describe how Quasi-Bayesian methods can be used to obtain the standard errors of our estimates. We shall observe that this is computationally costly.

By observing yields for maturities τ_i, $i = 1, \ldots, M$, in periods $\mathtt{t} = 1, \ldots, T$, we obtain $M-$dimensional vectors $\mathbf{y}_\mathtt{t} = (y_{\mathtt{t}1}, \ldots, y_{\mathtt{t}M})'$, $\mathtt{t} = 1, \ldots, T$, $MT-$dimensional vectors $\mathbf{y}_{1:T} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$, as well as $\tilde{\mathtt{q}}-$dimensional vectors $\tilde{\mathbf{m}}_{(\mathtt{t})}(\tilde{\mathbf{y}}_\mathtt{t}) = \left(y_{\mathtt{t}1}, \ldots, y_{\mathtt{t}M}^p, y_{\mathtt{t}1}y_{\mathtt{t}-1,1}, \ldots, y_{\mathtt{t}M}^2 y_{\mathtt{t}-1,M}^2\right)'$, where $\tilde{\mathbf{y}}_\mathtt{t} = \left(\mathbf{y}_\mathtt{t}', \mathbf{y}_{\mathtt{t}-1}'\right)'$, and $\tilde{\mathbf{m}}_T(\mathbf{y}_{1:T}) = \left(\frac{1}{T}\sum_{\mathtt{t}=1}^T y_{\mathtt{t}1}, \ldots, \frac{1}{T}\sum_{\mathtt{t}=1}^T y_{\mathtt{t}M}^p, \frac{1}{T-1}\sum_{\mathtt{t}=2}^T y_{\mathtt{t}1}y_{\mathtt{t}-1,1}, \ldots, \frac{1}{T-1}\sum_{\mathtt{t}=2}^T y_{\mathtt{t}M}^2 y_{\mathtt{t}-1,M}^2\right)'$. Let $\tilde{\boldsymbol{\mu}}(\boldsymbol{\vartheta}) = \left(\mathbb{E}(y_{\mathtt{t}1}), \ldots, \mathbb{E}\left(y_{\mathtt{t}M}^p\right), \mathbb{E}\left(y_{\mathtt{t}1}y_{\mathtt{t}-1,1}\right), \ldots, \mathbb{E}\left(y_{\mathtt{t}M}^2 y_{\mathtt{t}-1,M}^2\right)\right)'$ stands for the corresponding vector of moments of yields as a function of the unknown parameter vector $\boldsymbol{\vartheta} \in \boldsymbol{\Theta} \subset \mathbb{R}^\mathfrak{p}$. The components of the vector $\tilde{\boldsymbol{\mu}}(\boldsymbol{\vartheta})$ are provided in the Online-Appendix A-4, see equations (A-19), (A-23)-(A-26), (A-29) and (A-30).

The generalized method of moments demands $\mathfrak{q}$ moments of yields to be selected such that $\tilde{\mathfrak{q}} \geq \mathfrak{q} \geq \mathfrak{p}$. By means of a $\mathfrak{q} \times \tilde{\mathfrak{q}}$ selector matrix $\mathcal{M}$, where $\mathcal{M}_{ij} = 1$ if the corresponding moment is used and $\mathcal{M}_{ij} = 0$ otherwise, we obtain $\boldsymbol{\mu}(\boldsymbol{\vartheta}) = \mathcal{M}\,\tilde{\boldsymbol{\mu}}(\boldsymbol{\vartheta}) \in \mathbb{R}^\mathfrak{q}$, $\mathbf{m}_{(\mathtt{t})}(\tilde{\mathbf{y}}_\mathtt{t}) = \mathcal{M}\,\tilde{\mathbf{m}}_{(\mathtt{t})}(\tilde{\mathbf{y}}_\mathtt{t}) \in \mathbb{R}^\mathfrak{q}$ and $\mathbf{m}_T(\mathbf{y}_{1:T}) = \mathcal{M}\,\tilde{\mathbf{m}}_T(\mathbf{y}_{1:T}) \in \mathbb{R}^\mathfrak{q}$. For the $\mathbb{A}_1(3)$ model considered in Section A-2, the dimension of the parameter vector $\boldsymbol{\vartheta}$ is 23 (i.e., $\mathfrak{p} = 23$; including fourth order moments of the yields results in $\mathfrak{p} = 24$). The number of maturities available is $M = 10$. For example, by using the moments $\mathbb{E}(y_{\mathtt{t}i})$, $\mathbb{E}\left(y_{\mathtt{t}i}^2\right)$ and $\mathbb{E}(y_{\mathtt{t}i}y_{\mathtt{t}-1,i})$, for $i = 1, \ldots, M$, we are already equipped with $3M$ moment conditions. Hence, for $M \geq 8$ the order condition $\mathfrak{q} \geq \mathfrak{p}$ is already met. Note that by using the moments or order $p \leq 4$ and the auto-covariances, the number of moments of yields $\tilde{\mathfrak{q}}$ is much larger than the number of parameters $\tilde{\mathfrak{p}}$, see equation (A-16) in the Online-Appendix.

To obtain parameter estimates by means of the GMM, a high-dimensional nonlinear minimization problem has to be solved and $\mathfrak{q}$ moment conditions have to be selected from the set of moments available. Here the question arises which moments help to identify the unknown model parameters $\boldsymbol{\vartheta}$ if only an $M-$dimensional vector of yields $\mathbf{y}_\mathtt{t}$, $\mathtt{t} = 1, \ldots, T$, is observed. Although we cannot assign particular moments to exactly one model parameter for the $\mathbb{A}_1(3)$ considered, we observe the following (based on a numerical analysis): If a short rate $r_\mathtt{t}$ were observed, the expected short rate would be determined by parameters γ_0 and θ^P. We observe that these parameters have a strong impact on the first order moments of the yields. This result can also be obtained in formal terms by using matrix $\mathbf{A}$, given by (7), and

results presented in Section 2.2. That is to say, γ_0 and θ^P determine the level of the yields. Parameters θ^Q and $\boldsymbol{\beta}^Q$ as well as γ_0, $\mathcal{B}^x_{12}$, $\mathcal{B}^x_{13}$ and the volatility parameters Σ_i $(i = 1, 2, 3)$, determine $\Phi(\mathtt{t}, \mathbf{u})$ and $\boldsymbol{\Psi}(\mathtt{t}, \mathbf{u})$ and therefore drive the slope and the curvature of the yield curve. Hence, these parameters are also associated to the first order moments. On the other hand it turns out that the first order moments of the yields are important to estimate γ_0, θ^P, θ^Q and $\boldsymbol{\beta}^Q$. By considering matrix $\mathbf{A}$ (see (7)) and the Riccati differential equations (given in (4)), which result in $\boldsymbol{\Phi}(t, \mathbf{u})$ and $\boldsymbol{\Psi}(t, \mathbf{u})$, we observe that without any further constraints all parameters (except the noise parameter σ^2_ε) have an impact on the second order moments of the model yields $y^0_{\mathtt{t}i}$. Hence, all parameters (including σ^2_ε) drive the volatility of the yields $y_{\mathtt{t}i}$ and thus the second order moments as well. The covariance structure of yields (also the off-diagonal elements of the covariance matrix of yields) is affected by $\boldsymbol{\beta}^P$. We observe that the second order moments of the yields are especially important to estimate Σ_i, $i = 1, 2, 3$, $\boldsymbol{\beta}^P$ and σ^2_ε. The auto-covariances are strongly connected to $\boldsymbol{\beta}^P$ (especially to the elements on the main diagonal). Thus, the auto-covariances of the yields help to identify $\boldsymbol{\beta}^P$ as well. Online-Appendix A-4 provides a lot of higher order moments. However, it turns out that the instability of the estimation routine we consider is amplified if higher order moments are added. Due to this instability, the Wald or the distance difference tests – to check for redundant moment conditions – provide us with very ambiguous results. Hence, the selection of these moments was performed by means of simulation experiments. Based on these simulation results, we work with $\mathfrak{q} = 3M = 30$ moment conditions, namely, $\mathbb{E}(y_{\mathtt{t}i})$, $\mathbb{E}\left(y^2_{\mathtt{t}i}\right)$ and $\mathbb{E}\left(y_{\mathtt{t}i}y_{\mathtt{t}-1i}\right)$, for $i = 1, \ldots, M$ and $\mathtt{t} = 2, \ldots, T$.

Next we define $\mathbf{h}_{(\mathtt{t})}\left(\boldsymbol{\vartheta}; \tilde{\mathbf{y}}_\mathtt{t}\right) = \mathbf{m}_{(\mathtt{t})}(\tilde{\mathbf{y}}_\mathtt{t}) - \boldsymbol{\mu}(\boldsymbol{\vartheta}) \in \mathbb{R}^\mathfrak{q}$ and $\mathbf{h}_T\left(\boldsymbol{\vartheta}; \mathbf{y}_{1:T}\right) = \mathbf{m}_T(\mathbf{y}_{1:T}) - \boldsymbol{\mu}(\boldsymbol{\vartheta}) \in \mathbb{R}^\mathfrak{q}$ as well as the GMM distance function

$$Q_T(\boldsymbol{\vartheta}; \mathbf{y}_{1:T}) = \mathbf{h}_T(\boldsymbol{\vartheta}; \mathbf{y}_{1:T})' \, \mathbf{C}_T \, \mathbf{h}_T(\boldsymbol{\vartheta}; \mathbf{y}_{1:T}). \tag{10}$$

The GMM estimate of $\boldsymbol{\vartheta}$ minimizes the distance function $Q_T(\cdot)$ in (10), where $\mathbf{C}_T$ is a $\mathfrak{q} \times \mathfrak{q}$ symmetric positive semi-definite weighting matrix (see, e.g. Ruud, 2000, Chapters 21-22). For regularity conditions and further issues on GMM estimation see, e.g. Hansen (1982); Newey and McFadden (1994); Altonji and Segal (1996); Pötscher and Prucha (1997); Windmeijer (2005); Guggenberger and Smith (2005); Newey

and Windmeijer (2009). In addition, the constraints imposed on the parameter space $\boldsymbol{\Theta}$ and described in Online-Appendix A-6 (following from admissibility, stationarity and estimation issues) will always be applied when $Q_T(\cdot)$ is minimized.

Let $\breve{\boldsymbol{\vartheta}}$ abbreviate a GMM estimator of $\boldsymbol{\vartheta}$. The asymptotic distribution of $\sqrt{T}\left(\breve{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}\right)$ is a normal distribution with mean vector $\mathbf{0}_{\mathfrak{p}}$ and the asymptotic covariance matrix $\mathbf{V}$, where $\mathbf{V}=\left(\mathbf{H}'\boldsymbol{\Lambda}^{-1}\mathbf{H}\right)^{-1}$, $\mathbf{H}=\mathbb{E}\left(\mathbf{D}_{\boldsymbol{\vartheta}}\mathbf{h}_{(\mathfrak{t})}\left(\boldsymbol{\vartheta};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)\right)\in\mathbb{R}^{\mathfrak{q}\times\mathfrak{p}}$, $\mathbf{D}_{\boldsymbol{\vartheta}}\mathbf{h}_{(\mathfrak{t})}\left(\boldsymbol{\vartheta};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)\in\mathbb{R}^{\mathfrak{q}\times\mathfrak{p}}$ is a matrix of partial derivatives of $\mathbf{h}_{(\mathfrak{t})}\left(\boldsymbol{\vartheta};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)$ and $\boldsymbol{\Lambda}\in\mathbb{R}^{\mathfrak{q}\times\mathfrak{q}}$ is the long run covariance matrix of $\mathbf{h}_{(\mathfrak{t})}\left(\boldsymbol{\vartheta};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)$ (see, e.g., Phillips and Hansen, 1990).

A first attempt to estimate the $\mathfrak{p}\times\mathfrak{p}$ covariance matrix $\mathbf{V}$ is the "standard GMM covariance matrix estimate" (see, e.g. Ruud, 2000, Chapters 21 and 22):

$$\begin{aligned}
&\widehat{\mathbf{V}}_T=\left(\widehat{\mathbf{H}}_T'\widehat{\boldsymbol{\Lambda}}_T^{-1}\widehat{\mathbf{H}}_T\right)^{-1}\in\mathbb{R}^{\mathfrak{p}\times\mathfrak{p}}\ ,\ \text{where}\\
&\widehat{\mathbf{H}}_T=\frac{1}{T-1}\sum_{\mathfrak{t}=2}^{T}\mathbf{D}_{\boldsymbol{\vartheta}}\mathbf{h}_{(\mathfrak{t})}\left(\breve{\boldsymbol{\vartheta}};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)\in\mathbb{R}^{\mathfrak{q}\times\mathfrak{p}}\ ,\ \widehat{\boldsymbol{\Lambda}}_T=\widehat{\boldsymbol{\Gamma}}_{T,0}+\sum_{j=1}^{T-1}\mathtt{k}\left(\frac{j}{B}\right)\left(\widehat{\boldsymbol{\Gamma}}_{T,j}+\widehat{\boldsymbol{\Gamma}}_{T,j}'\right)\\
&\text{and}\\
&\widehat{\boldsymbol{\Gamma}}_{T,j}=\frac{1}{T}\sum_{\mathfrak{t}=j+1}^{T}\mathbf{h}_{(\mathfrak{t})}\left(\breve{\boldsymbol{\vartheta}};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)\mathbf{h}_{(\mathfrak{t})}\left(\breve{\boldsymbol{\vartheta}};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)'\in\mathbb{R}^{\mathfrak{q}\times\mathfrak{q}},
\end{aligned}\tag{11}$$

where $\mathtt{k}\left(\frac{j}{B}\right)$ is a kernel function with bandwidth B. In our application we use the Bartlett-kernel and set B equal to the next smallest integer of $4\cdot\left(\frac{T}{100}\right)^{2/9}\approx 5.7$ for $T=500$ (see, e.g. Newey and West, 1987, 1994). Note that matrices of dimension $\mathfrak{p}\times\mathfrak{p}$ (with $\mathfrak{p}\geq 23$) have to be inverted in (11) and partial derivatives in matrix $\mathbf{D}_{\boldsymbol{\vartheta}}\mathbf{h}_{(\mathfrak{t})}\left(\breve{\boldsymbol{\vartheta}};\tilde{\mathbf{y}}_{\mathfrak{t}}\right)$ have to be derived numerically. Hence, estimating covariance matrix $\mathbf{V}$ by means of (11) can be numerically demanding. In addition, $\widehat{\mathbf{H}}_T$ as well as $\widehat{\boldsymbol{\Lambda}}_T$ also depend on $\mathbf{y}_{1:T}$, and therefore are subject to the variation of the finite samples.

To calculate the GMM distance function $Q_T(\cdot)$, see (10), the *continuous updating estimator* (*CUE*) of $\boldsymbol{\vartheta}$ is used. With the CUE estimator we run an iterative procedure, where in each iteration step $\mathtt{m}$, $\mathtt{m}=1,\ldots,\mathtt{M}$, we commute between (i) augmenting the "parameter-estimate" to $\breve{\boldsymbol{\vartheta}}^{(\mathtt{m})}$ with respect to $Q_T(\cdot)$ given $\mathbf{C}_T$ and (ii) updating $\mathbf{C}_T=\left(\widehat{\boldsymbol{\Lambda}}_T\right)^{-1}$, when $\breve{\boldsymbol{\vartheta}}$ (which occurs in $\widehat{\boldsymbol{\Lambda}}_T$) is replaced by the previous draw, $\breve{\boldsymbol{\vartheta}}^{(\mathtt{m})}$, in (11). When only one draw is considered, as with the random search (described later), then it is used to derive $\widehat{\boldsymbol{\Lambda}}_T$ in (11). This corresponds to standard efficient GMM. For the gradient based and

the Quasi-Bayesian estimator we also checked the impact on parameter estimation when using both the CUE estimator and standard efficient GMM estimator (where $\mathbf{C}_T$ follows from the first iteration step). Here, no significant differences were observed.

To investigate the properties of the GMM distance function and parameter estimation routines, we performed Monte Carlo experiments with simulated yields where $M = 10$ and $T = 500$. In particular, we used the (yearly) maturities $\boldsymbol{\tau} = \{1/12, 1/4, 1/2, 1, 2, 3, 5, 7, 10, 20\}$ that were used also in the empirical application presented in Section 4. In each Monte Carlo run, denoted by the subscript ℓ, we generate the yields $\{\mathbf{y}_{1:T}\}_\ell$ as follows: (i) Calculate $\mathbf{X}_{\mathtt{t},\ell}$ by means of the Euler scheme (see, e.g. Glasserman, 2003) where the initial point $\mathbf{X}_{\mathtt{0},\ell}$ is drawn from the stationary distribution of $\mathbf{X}_{\mathtt{t},\ell}$. (ii) Generate $\varepsilon_{i\mathtt{t},\ell}$ from a normal distribution with mean zero and variance $\sigma_\varepsilon^2 = 0.0067$, see Online-Appendix A-6 for description of the choice of $\sigma_\varepsilon^2 = 0.0067$. The true parameter $\boldsymbol{\vartheta}$ is provided either in the second column of Table 3 or in the second column of Table 4. In Table 3 the data are generated such that $\theta^Q = 10 \neq 1.5 = \theta^P$ and $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$, while in Table 4 they are generated such that $\theta^Q = \theta^P = 1.5$ and $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$.

Next, we investigate whether "undesired behavior" of the GMM distance function (e.g. multiple local minima) plays a major role when performing parameter estimation. To do this, we generate draws $\boldsymbol{\vartheta}^{(\mathtt{n})}$, $\mathtt{n} = 1, \ldots, \mathtt{N}$, as follows: If the support for coordinate j is the real axis then

$$
\begin{aligned}
[\boldsymbol{\vartheta}^{(\mathtt{n})}]_j &= [\boldsymbol{\vartheta}_c]_j + c_\vartheta [|\boldsymbol{\vartheta}_c|]_j \zeta_j^{(\mathtt{n})} \ , \text{ while} \\
[\boldsymbol{\vartheta}^{(\mathtt{n})}]_j &= \exp\left(\log[|\boldsymbol{\vartheta}_c|]_j + c_\vartheta \zeta_j^{(\mathtt{n})}\right) \operatorname{sgn}\left([\boldsymbol{\vartheta}_c]_j\right), \qquad (12)
\end{aligned}
$$

is used for the elements j living only on the non-positive or only on the non-negative part of the real axis. $\zeta_j^{(\mathtt{n})}$ is *iid* standard normal and c_ϑ is the distortion parameter. Here, $\boldsymbol{\vartheta}_c = \boldsymbol{\vartheta}$ and $c_\vartheta = 5$. We observed that the minima of the GMM distance function are relatively close to the true parameter values. However a larger $\mathtt{N}$, e.g. $\mathtt{N} = 5,000$, becomes necessary to obtain samples close to the true parameter value. In addition, given simulated samples with $T = 500$ periods and $M = 10$ maturities, we plotted the GMM distance function $Q_T(\boldsymbol{\vartheta}; \mathbf{y}_{1:T})$ against the i-th coordinate of $\boldsymbol{\vartheta}$ while keeping the other coordinates fixed at the true parameter values. For some parameters we observed non-convexities of the GMM distance function. A minimum is obtained at values close but not necessarily very close to the true parameter

value. E.g. with $\theta^Q = 1.5$, minima between approximately 1 and 2 were observed for various draws. With respect to parameters $\bar{\boldsymbol{\beta}}^Q$ and $\bar{\boldsymbol{\beta}}^P$ we observed that the GMM distance function is relatively flat in its coordinates. This effect is even stronger with the off-diagonal elements, where also non-convexities of $Q_T(\cdot)$ in β_{ij}^Q or β_{ij}^P, $i \neq j$, can show up. These observations explain the large range and a high dispersion of the estimates of $\bar{\boldsymbol{\beta}}^Q$ and $\bar{\boldsymbol{\beta}}^P$. In addition, we observe that the GMM distance function is flat in the parameters $\mathcal{B}_{12}^x$ and $\mathcal{B}_{13}^x$ (see *min*, *max* and *std* for β_{ij}^Q, β_{ij}^P, $\mathcal{B}_{12}^x$ and $\mathcal{B}_{13}^x$ in Tables 1 to 4). Although, we did not observe non-convexities for parameter $\vartheta_{23} = \sigma_\varepsilon^2$, its GMM distance function is quite flat. Since the variance of the yields is larger than σ_ε^2 by the model assumptions, we used the smallest sample variance of the observed yields as an upper bound for the parameter σ_ε^2 (this constraint is part of the constraints described in Online-Appendix A-6).

3.1 Parameter Estimation

In this subsection we investigate the properties of parameter estimation routines by means of Monte Carlo experiments. We used $M = 10$ yields of maturities $\boldsymbol{\tau}$ and the time series dimension of $T = 500$. The data are simulated as described in the above paragraphs. In each Monte Carlo run ℓ, $\ell = 1, \ldots, \mathtt{L}$, an estimation procedure is applied, where the true parameter $\boldsymbol{\vartheta}$ is either provided in the second column of Table 3 or in the second column of Table 4. In all Monte Carlo experiments an unrestricted model is estimated. That is, we obtain separate estimates for θ^Q and θ^P as well as for $\bar{\boldsymbol{\beta}}^Q$ and $\bar{\boldsymbol{\beta}}^P$, respectively.

"Standard GMM parameter" estimation: Suppose that an initial value, denoted by $\boldsymbol{\vartheta}^{(\mathtt{n})}$, is generated by means of (12) with the distortion parameter c_ϑ being set to 0, 0.1, 0.25, 0.5 and 1 and $\boldsymbol{\vartheta}_c = \boldsymbol{\vartheta}$. Then, $\boldsymbol{\vartheta}^{(\mathtt{n})}$ is used as the starting value of the MATLAB minimization routine `fminsearch` based on the Nelder-Mead algorithm (see `http://www.mathworks.de/de/help/matlab/ref/fminsearch.html`). We observe that the parameters can be estimated easily by means of this standard minimization tool when $c_\vartheta \leq 0.25$, i.e. when the optimization is started sufficiently close to the true parameter $\boldsymbol{\vartheta}$. However, the parameter estimation with $c_\vartheta = 0.5$ or $c_\vartheta = 1$ becomes a difficult problem.

Random Search (part I): To cope with this problem we apply a multi-start random search method (see, e.g. Törn and Zilinskas, 1989). That is, parameter estimation is started with the random draws

$\boldsymbol{\vartheta}^{(\mathtt{n})}$, where $\mathtt{n} = 1, \ldots, \mathtt{N} = 1,000$. Each $\boldsymbol{\vartheta}^{(\mathtt{n})}$ is from $\boldsymbol{\Theta}$ and the draws outside $\boldsymbol{\Theta}$ are skipped. The samples $\boldsymbol{\vartheta}^{(\mathtt{n})}$ are generated in the same way as in (12) with distortion parameter $c_\vartheta = 1$ and $\boldsymbol{\vartheta}_c = \boldsymbol{\vartheta}$. Our parameter estimate $\widehat{\boldsymbol{\vartheta}}$ is provided by the $\boldsymbol{\vartheta}^{(\mathtt{n})}$ resulting in the smallest GMM distance function (10). We also force our multi-start random search routine to generate samples such that $\left(\theta^Q\right)^{(\mathtt{n})} = \left(\theta^P\right)^{(\mathtt{n})}$ as well as $\left(\theta^Q\right)^{(\mathtt{n})} \neq \left(\theta^P\right)^{(\mathtt{n})}$ for both experiments presented in Tables 3 and 4, respectively. The same procedure is also applied to $\left(\bar{\boldsymbol{\beta}}^Q\right)^{(\mathtt{n})}$ and $\left(\bar{\boldsymbol{\beta}}^P\right)^{(\mathtt{n})}$. A further alternative to obtain an estimate of $\boldsymbol{\vartheta}$ is to follow the suggestions of Chernozhukov and Hong (2003) and use the draws from an ergodic Markov Chain, $\boldsymbol{\vartheta}^{(\mathtt{m})}$, $\mathtt{m} = 1, 2, \ldots, \mathtt{M}$. In particular, *adaptive MCMC* was applied (see, e.g., Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009) as described in Online-Appendix A-7 in more detail. We denote the estimate obtained by the adaptive MCMC procedure by $\tilde{\boldsymbol{\vartheta}}$. Note, however, that the main advantage of adaptive MCMC seems to be for inference, namely when estimating covariance matrices for the Wald type test. For example, see Table A-10 in Online-Appendix A-7, where the Wald test statistics were calculated using "standard" estimates of asymptotic covariance matrix and Table 5 where the Wald test statistics were obtained using adaptive MCMC approach.

Our results suggest (see Tables A-2 to A-9 in the Online-Appendix) that the numerical minimization routine based on the Nelder-Mead algorithm, where $\boldsymbol{\vartheta}^{(\mathtt{n})}$ with the smallest GMM distance function from multi-start random search method is used as the starting value, does not improve the properties of the estimation routine. The best results are obtained either with the estimator $\tilde{\boldsymbol{\vartheta}}$, where first multi-start random search and then adaptive MCMC methods are used, or with the estimator $\widehat{\boldsymbol{\vartheta}}$, where only multi-start random search is applied.

A further alternative to this approach is to start the Bayesian sampler at some initial value described by equation (12) and then check whether the draws obtained by the sampler cluster around the true parameter $\boldsymbol{\vartheta}$. Here we observe that the convergence of the chain is slow and thus we do not get close to $\boldsymbol{\vartheta}$ in reasonable time. The exception is a sampler started sufficiently close to the true parameter. Hence, we first apply random search methods to obtain $\widehat{\boldsymbol{\vartheta}}$ and then run – if necessary – the Bayesian sampler started at $\widehat{\boldsymbol{\vartheta}}$ (e.g. to obtain standard errors as demonstrated in the following Section 3.2). Thus, the two step optimization approach as used in Monfort et al. (2015) (where they maximize the likelihood) is very

similar to ours, where we minimize the GMM distance function.

To compare the performance of estimators $\widehat{\boldsymbol{\vartheta}}$ and $\tilde{\boldsymbol{\vartheta}}$, we calculate estimates of (the absolute value of) their bias and of the root means squared error ($RMSE$), which are presented in columns 3 to 6 of Tables 1 and 2. Here the bias, $|\mathbb{E}\left(\widehat{\vartheta}_i - \vartheta_i\right)|$, is estimated with $|\widehat{\vartheta}_i - \vartheta_i| = \left|\frac{1}{\mathtt{L}}\sum_{\ell=1}^{\mathtt{L}} \widehat{\vartheta}_{\ell i} - \vartheta_i\right|$ and $RMSE$, $\sqrt{\mathbb{E}\left(\widehat{\vartheta}_i - \vartheta_i\right)^2}$, is estimated with $\sqrt{sdt\left(\widehat{\vartheta}_i - \vartheta_i\right)^2 + |\widehat{\vartheta}_i - \vartheta_i|^2}$, where std denotes the sample standard deviation. For those coordinates i where $\tilde{\vartheta}_i$ is not better than $\widehat{\vartheta}_i$, the performance of $\tilde{\vartheta}_i$ is not much worse than for its competitor $\widehat{\vartheta}_i$ (here, "better" means that bias and $RMSE$ are smaller for more coordinates of $\tilde{\boldsymbol{\vartheta}}$ than for $\widehat{\boldsymbol{\vartheta}}$). From the estimates $\tilde{\boldsymbol{\vartheta}}_\ell$, $\ell = 1, \ldots, \mathtt{L} = 250$, we obtain the sample $mean$, minimum (min), maximum (max), standard deviation (std), skewness ($skew$) and excess-kurtosis ($kurt$). These descriptive statistics are reported in columns 3 to 8 in Tables 3 and 4.

Random Search (part II): The above results also indicate that either a lot of draws, $\mathtt{N}$, or a "good guess" about the true parameter value are necessary when performing parameter estimation. To investigate this claim we set $\boldsymbol{\vartheta}_c = \boldsymbol{\vartheta}$ and use 5,000 draws where $c_\vartheta = 5$ or $c_\vartheta = 10$ and for a small number of draws (namely 10 draws) we apply $c_\vartheta = 0.01$. We observe that the smallest GMM distance functions was obtained for draws with $c_\vartheta = 0.01$. However, by excluding the draws based on $c_\vartheta = 0.01$ and only working with relatively large c_ϑ, good parameter estimates require a high number of draws, in particular, 5,000 and more. Hence, from this analysis we conclude that parameter estimation either demands for a "(very) good guess" of the location of parameter $\boldsymbol{\vartheta}$ or a high number of search steps. We call a scenario where the econometrician knows the location of true $\boldsymbol{\vartheta}$ with a very high probability as *the strong prior information scenario* (i.e., $\boldsymbol{\vartheta}_c = \boldsymbol{\vartheta}$ and $c_\vartheta \leq 1$). On the other hand, the scenario, where a set $\boldsymbol{\Theta}_0 \subset \mathbb{R}^{\mathtt{p}}$ is sufficiently large (i.e., much larger than the set effectively covered by the procedure using pseudo-random numbers described in part I) and $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}_0$, is referred to as *the weak prior information scenario*. Appendix A-7 provides more details on our choice of $\boldsymbol{\Theta}_0$ and $\boldsymbol{\vartheta}_c$. The implementation of the weak prior information scenario is as follows: Let $\boldsymbol{\vartheta}^{(\circ)}$ denote some element $\in \boldsymbol{\Theta}_0$. A grid with e.g. 100 intervals per parameter on $\boldsymbol{\Theta}_0$, results in 100^{23} values $\boldsymbol{\vartheta}^{(\circ)}$ where the GMM distance function has to be evaluated. This is definitely too costly from a computational point of view. By contrast, we observed surprisingly good results when generating $\mathtt{N}_0 = 200,000$ uniformly distributed draws $\boldsymbol{\vartheta}^{(\circ)}$ from $\boldsymbol{\Theta}_0$, and then taking the mean of the 100

draws with the smallest GMM distance function $\widehat{\boldsymbol{\vartheta}}_{\texttt{Step 0}}$ (note that, $\texttt{N}_0$ is sufficiently large such that all of these 100 draws of $\boldsymbol{\vartheta}^{(\texttt{o})}$ satisfy the constraints imposed in Appendix A-6, i.e. these draws are also contained in $\boldsymbol{\Theta}$). Next we proceed with $\texttt{N} = 2,000$ normally distributed samples using (12), where $\boldsymbol{\vartheta}_c = \widehat{\boldsymbol{\theta}}_{\texttt{Step 0}}$ and $c_\vartheta = 1$. Finally the Quasi-Bayesian sampler is applied. Estimates based on $\texttt{N}_0 + \texttt{N}$ random search steps are abbreviated by $\widehat{\boldsymbol{\vartheta}}$, while estimates based on $\texttt{N}_0 + \texttt{N}$ random search steps and $\texttt{M}$ quasi-Bayesian steps are denoted $\tilde{\boldsymbol{\vartheta}}$. Estimates and further descriptive statistics based on the weak prior information are presented in the columns 7 to 10 in Tables 1 and 2 as well as in the last four columns of Tables 3 and 4. First, we observe that the performance of the Bayesian estimate $\tilde{\boldsymbol{\vartheta}}$ is slightly better than $\widehat{\boldsymbol{\vartheta}}$. However, these differences are small. By comparing the estimates with strong prior information to the scenario with weak prior information, we observe that the biases and the $RMSE$s increase sharply. The effects are more pronounces when $\theta^Q \neq \theta^P$ and $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$. We observe that especially the off-diagonal parameters of $\bar{\boldsymbol{\beta}}^Q$ and $\bar{\boldsymbol{\beta}}^P$ as well as the parameters $\mathcal{B}^x_{12}$ and $\mathcal{B}^x_{13}$ are difficult to estimate. These results are hardly a surprise given the analysis of the GMM distance function already provided in this section, where we observe that the GMM distance function is flat in β^Q_{ij} or β^P_{ij}, $i \neq j$, as well as in the parameters $\mathcal{B}^x_{12}$ and $\mathcal{B}^x_{13}$.

3.2 Inference

To test for parameter restrictions, let $\breve{\boldsymbol{\vartheta}}$ and $\breve{\mathbf{V}}_T$ denote GMM-based estimates of $\boldsymbol{\vartheta}$ and $\mathbf{V}$ and assume that the null hypothesis consists of $\mathfrak{r}_\mathfrak{p}$ restrictions. Suppose that these restrictions are described by a twice continuously differential function $\mathbf{r}(\boldsymbol{\vartheta}) : \mathbb{R}^\mathfrak{p} \to \mathbb{R}^{\mathfrak{r}_\mathfrak{p}}$ and the $\mathfrak{r}_\mathfrak{p} \times \mathfrak{p}$ matrix of partial derivatives

$$\mathbf{R} = \mathbf{D}_{\boldsymbol{\vartheta}}\mathbf{r}(\breve{\boldsymbol{\vartheta}}) = \begin{pmatrix} \frac{\partial r_1(\breve{\boldsymbol{\vartheta}})}{\partial \vartheta_1} & \dots & \frac{\partial r_1(\breve{\boldsymbol{\vartheta}})}{\partial \vartheta_\mathfrak{p}} \\ \dots & \dots & \dots \\ \frac{\partial r_{\mathfrak{r}_\mathfrak{p}}(\breve{\boldsymbol{\vartheta}})}{\partial \vartheta_1} & \dots & \frac{\partial r_{\mathfrak{r}_\mathfrak{p}}(\breve{\boldsymbol{\vartheta}})}{\partial \vartheta_\mathfrak{p}} \end{pmatrix}, \tag{13}$$

which has rank $\mathfrak{r}_\mathfrak{p}$. Under the null hypothesis we have $\mathbf{r}(\boldsymbol{\vartheta}) = \mathbf{0}_{\mathfrak{r}_\mathfrak{p}}$ and thus the Wald-type statistic becomes

$$\mathscr{W}_T \;=\; T\mathbf{r}(\breve{\boldsymbol{\vartheta}})' \left(\mathbf{R}\breve{\mathbf{V}}_T\mathbf{R}'\right)^{-1} \mathbf{r}(\breve{\boldsymbol{\vartheta}}) \;, \tag{14}$$

which weakly converges to $\mathscr{W}$, where $\mathscr{W}$ follows a χ^2-distribution with $\mathfrak{r}_\mathfrak{p}$ degrees of freedom. The null hypothesis is rejected if $\mathscr{W}_T > \chi^2_{\mathfrak{r}_\mathfrak{p},1-\alpha_S}$, where α_S is the significance level and $\chi^2_{\mathfrak{r}_\mathfrak{p},1-\alpha_S}$ is the $1-\alpha_S$ percentile of a χ^2-distribution with $\mathfrak{r}_\mathfrak{p}$ degrees of freedom.

As our test statistics rely on asymptotic results, we have to investigate the finite sample properties of our tests. Since a lot of parameters are considered and various restrictions can be constructed, we focus now on the restrictions $\theta^Q = \theta^P$ as well as $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ which are often discussed in finance literature. In particular, if the goal is to test the null hypothesis $\theta^Q = \theta^P$ against the alternative $\theta^Q \neq \theta^P$, then $\mathfrak{r}_\mathfrak{p} = 1$, $\mathbf{r}(\boldsymbol{\vartheta}) = (1, -1, 0, \ldots, 0)\,\boldsymbol{\vartheta} = \theta^Q - \theta^P$ and $\mathbf{R} = (1, -1, 0, \ldots, 0)$. To test $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ against $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$, we get $\mathbf{R} = \big(\mathbf{0}_{(7\times 2)}, \mathbf{I}_7, -\mathbf{I}_7, \mathbf{0}_{(7\times 7)}\big)$.

Online-Appendix A-7 (see Table A-10) demonstrates that the performance of the Wald-type test implemented in a standard way (as well as the distance difference test) is poor (for further details see Online-Appendix A-7). To cope with these problems, we follow the suggestions of Chernozhukov and Hong (2003); Andrieu and Thoms (2008); Roberts and Rosenthal (2009) and use the draws from an ergodic Markov Chain, $\big(\boldsymbol{\vartheta}^{(\mathfrak{m})}\big)$, to estimate the covariance matrix $\mathbf{V}$ (on more details see Online-Appendix A-7). A quasi-Bayesian estimate of $\mathbf{V}$, denoted by $\tilde{\mathbf{V}}_T$, and the parameter estimate $\tilde{\boldsymbol{\vartheta}}$ are used to obtain the Wald-type statistic $\mathscr{W}_T$.

Since finance literature distinguishes between affine market prices of risk, where $\theta^Q \neq \theta^P$ but $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ and extended affine market price of risk specifications, where $\theta^Q \neq \theta^P$ and $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$ is allowed (as proposed and mathematically investigated in Cheridito et al. (2007); Le et al. (2010)), we perform tests for the null-hypothesis $\theta^Q = \theta^P$ and for the null-hypothesis $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ separately. Table 5 presents simulation results (based on the Bayesian sampler) when testing the null hypothesis $\theta^Q = \theta^P$ against the alternative hypothesis $\theta^Q \neq \theta^P$ as well as $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ against $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$. The data generating process follows from the affine term structure model in equation (9), where the noise terms $\varepsilon_{\mathfrak{t}i}$ are *iid* normal with mean zero and variance $\sigma^2_\varepsilon = 0.0067$ (as presented in the second column of Tables 3 or 4). To investigate the *size* of the Wald-type tests, we generate the data by using the (true) parameter vector presented in the second column of Table 4. When testing the true null-hypothesis $\theta^Q = \theta^P$ against the alternative $\theta^Q \neq \theta^P$ as well as $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ against the alternative $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$, we observe that the rejection rates are

above the theoretical significance level of $\alpha_S = 5\%$ for strong prior information. In more detail, when testing the (true) null-hypothesis $\theta^Q = \theta^P$ oversizing becomes relatively high for both prior information scenarios. When testing the (true) null-hypothesis $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$, oversizing becomes relatively high for strong prior information, while undersizing is observed for weak prior information. We claim that undersizing is mainly caused by the difficulty to estimate the off-diagonal parameters of $\bar{\boldsymbol{\beta}}^Q$ and $\bar{\boldsymbol{\beta}}^P$. On the other hand, to analyze the *power* of the Wald-type test, we generated yields by using the (true) parameters presented in the second column of Table 3, where $\theta^Q \neq \theta^P$ as well as $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$, and perform Wald-type tests of the false null-hypothesis $\theta^Q = \theta^P$ against $\theta^Q \neq \theta^P$ as well as $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ against $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$. Here Table 5 shows that the (false) null hypothesis is rejected in almost all simulation runs.

4 Parameter Estimation in Empirical Data

This section applies the estimator developed in the previous sections to empirical data. We use H-15 interest rate data of "Treasury constant maturity" yields on weekly frequency (measured every Friday) from the Federal Reserve (see `http://federalreserve.gov/releases/h15/data.htm`). The time period considered is August 3, 2001 to August 30, 2013. An almost full panel of maturities from one month to twenty years is available for these periods. Thus, we have $M = 10$ maturities such that $\boldsymbol{\tau} = \{1/12, 1/4, 1/2, 1, 2, 3, 5, 7, 10, 20\}$, where τ_i is measured in years, and $T = 631$ observations per yield. Note that as in our simulations the dimension is in the same ballpark, namely $M = 10$ and $T = 500$ and as performance of the Wald-type tests on parameter restrictions using the Bayesian sampler was 'reasonable', we use this Wald-type test also on our empirical data. Although the H-15 data set can only be seen as a proxy for the risk-free term structure, we follow the related literature (see, e.g. Chib and Ergashev, 2009) and work with this dataset.

Since we a-priori don't know the location of the parameter in the empirical data we apply the estimation routine designed for weak prior information. That is, we generated $\mathtt{N}_0 = 200,000$ uniformly distributed draws $\boldsymbol{\vartheta}^{(\mathtt{o})}$ from $\boldsymbol{\Theta}_0$, followed by $\mathtt{N} = 2,000$ normally distributed samples using (12) with $c_\theta = 1$ and $\boldsymbol{\vartheta}_c = \widehat{\boldsymbol{\vartheta}}_{\mathtt{Step\ 0}}$. Finally, adaptive MCMC is applied.

To check for the stability of our estimation routine in the empirical data, we obtained $\mathtt{L} = 5$ estimates,

each based on $\mathtt{N}_0 = 200,000$ uniformly distributed draws $\boldsymbol{\vartheta}^{(\mathtt{o})}$, $\mathtt{N} = 2,000$ normally distributed samples and $\mathtt{M} = 6,000$ adaptive MCMC steps (but of course using a different seed). By doing this, we observe that in all simulation runs, $\ell = 1, \ldots, \mathtt{L} = 5$, the intervals $\left[\tilde{\boldsymbol{\vartheta}}_\ell\right]_\iota \pm SE\left(\left[\tilde{\boldsymbol{\vartheta}}_\ell\right]_\iota\right)$ overlap for $\iota = 1, \ldots, \mathfrak{p} = 23$. The term $SE\left(\left[\tilde{\boldsymbol{\vartheta}}_\ell\right]_\iota\right)$ denotes the standard error of the ι-coordinate of $\tilde{\boldsymbol{\vartheta}}_\ell$. The standard error follows from $SE\left(\left[\tilde{\boldsymbol{\vartheta}}_\ell\right]_\iota\right) = \frac{1}{T^{0.5}}\tilde{\mathbf{V}}_T\left(\left[\tilde{\boldsymbol{\vartheta}}_\ell\right]_{\iota\iota}\right)^{0.5}$.

To obtain parameter estimates, the draws of the Bayesian sampler $\boldsymbol{\vartheta}^{(\mathtt{m})}$, $\mathtt{m} = 1001, \ldots, 6000$, are used to obtain the estimate $\tilde{\boldsymbol{\vartheta}}$, which comprises $\tilde{\theta}^Q = 5.6017$, $\tilde{\theta}^P = 0.4532$, $\tilde{\beta}^Q_{11} = -0.1890$, $\tilde{\beta}^Q_{21} = 0.2062$, $\tilde{\beta}^Q_{31} = 0.0543$, $\tilde{\beta}^Q_{22} = -2.088$, $\tilde{\beta}^Q_{32} = -0.4168$, $\tilde{\beta}^Q_{23} = -0.1250$, $\tilde{\beta}^Q_{33} = -1.6590$, $\tilde{\beta}^P_{11} = -0.7769$, $\tilde{\beta}^P_{21} = 0.2793$, $\tilde{\beta}^P_{31} = 0.2412$, $\tilde{\beta}^P_{22} = -0.6783$, $\tilde{\beta}^P_{32} = 0.0844$, $\tilde{\beta}^P_{23} = -0.0617$, $\tilde{\beta}^P_{33} = -0.9461$, $\tilde{\mathcal{B}}^x_{12} = 0.4690$, $\tilde{\mathcal{B}}^x_{13} = 0.6358$, $\tilde{\gamma}_0 = 1.2030$, $\tilde{\Sigma}_1 = 0.3640$, $\tilde{\Sigma}_2 = 0.7797$, $\tilde{\Sigma}_3 = 1.1451$, $\tilde{\sigma}^2_\varepsilon = 0.0714$. By means of $\tilde{\mathbf{V}}_T$ we obtain standard errors: $SE\left(\tilde{\theta}^Q\right) = 0.5902$, $SE\left(\tilde{\theta}^P\right) = 0.0993$, $SE\left(\tilde{\beta}^Q_{11}\right) = 0.0462$, $SE\left(\tilde{\beta}^Q_{21}\right) = 0.1173$, $SE\left(\tilde{\beta}^Q_{31}\right) = 0.04890$, $SE\left(\tilde{\beta}^Q_{22}\right) = 0.7100$, $SE\left(\tilde{\beta}^Q_{32}\right) = 0.2932$, $SE\left(\tilde{\beta}^Q_{23}\right) = 0.1968$, $SE\left(\tilde{\beta}^Q_{33}\right) = 0.2338$, $SE\left(\tilde{\beta}^P_{11}\right) = 0.4452$, $SE\left(\tilde{\beta}^P_{21}\right) = 0.1374$, $SE\left(\tilde{\beta}^P_{31}\right) = 0.1535$, $SE\left(\tilde{\beta}^P_{22}\right) = 0.3565$, $SE\left(\tilde{\beta}^P_{32}\right) = 0.2822$, $SE\left(\tilde{\beta}^P_{23}\right) = 0.3161$, $SE\left(\tilde{\beta}^P_{33}\right) = 0.5253$, $SE\left(\tilde{\mathcal{B}}^x_{12}\right) = 0.3182$, $SE\left(\tilde{\mathcal{B}}^x_{13}\right) = 0.2809$, $SE\left(\tilde{\gamma}_0\right) = 0.0993$, $SE\left(\tilde{\Sigma}_1\right) = 0.1164$, $SE\left(\tilde{\Sigma}_2\right) = 0.4723$, $SE\left(\tilde{\Sigma}_3\right) = 0.2321$, $SE\left(\tilde{\sigma}^2_\varepsilon\right) = 0.03562$.

Following mathematical finance literature (see, e.g. Cheridito et al., 2007; Cochrane, 2005), a usual way to investigate how the market demands for a compensation for the risk generated by $\mathbf{W}^P(t)$ (*risk premium*), is to consider the market price of risk process $(\boldsymbol{\phi}(\mathbf{X}(t)))_{t\geq 0}$. This process depends on the model parameters $\boldsymbol{\vartheta}$. If $\mathbf{b}^Q = \mathbf{b}^P$ and $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$, then $\boldsymbol{\phi}(\mathbf{X}(t)) = \mathbf{0}_d$. In terms of the parametrization used in this article, $\boldsymbol{\phi}(\mathbf{X}(t)) = \mathbf{0}_d$ if $\theta^Q = \theta^P$ and $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$, while if $\theta^Q \neq \theta^P$ or $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$, then $\boldsymbol{\phi}(\mathbf{X}(t)) \neq \mathbf{0}_d$ (almost surely). In the following we test whether this is the case.

By considering the estimates $\tilde{\theta}^Q$ and $\tilde{\theta}^P$ and their estimated standard errors $SE\left(\tilde{\theta}^Q\right)$ and $SE\left(\tilde{\theta}^P\right)$, respectively, we observe that the difference in the parameter estimates is relatively large, compared to their estimated standard deviations. We obtained the Wald statistic $\mathscr{W}_T$ with p-value being approximately < 0.0001. Based on this, the null hypothesis $\theta^Q = \theta^P$ is rejected at the $\alpha_S = 0.01$ significance level for this empirical dataset.

Next, we test the null hypothesis $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ against the alternative hypothesis $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$, where $\bar{\boldsymbol{\beta}}^{\cdot}$

contains seven parameters. By estimating $\bar{\boldsymbol{\beta}}^Q - \bar{\boldsymbol{\beta}}^P$ and its covariance matrix from Monte Carlo output, we obtain the Wald statistic $\mathscr{W}_T$ with a corresponding p-value of < 0.01. That is, here the null hypothesis $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ is rejected on usual significance levels. Summing up, by these results the market price of risk process is significantly different from zero.

5 Conclusions

In this article we developed a new method allowing for parameter estimation based on the exact moments of the yields for affine term structure models. By applying the results of Cuchiero et al. (2012) on $p-$polynomial processes the conditional moments are derived. By assuming a stationary process, we obtain the exact moments of the yields as well as the first order auto-covariance of the yields and the squared yields. By means of these moments, the model parameters can be estimated by the generalized method of moments.

Since the number of parameters is relatively large and the moments are non-linear in the model parameters, the implementation of the generalized method of moments becomes a non-trivial problem. We observe that standard minimization routines perform poorly. To cope with this problem, we use multi-start random search methods combined with Quasi-Bayesian methods, as proposed in Chernozhukov and Hong (2003), to estimate the model parameters as well as the asymptotic covariance matrix and to perform inference.

Another main contribution of this article is a rigorous investigation of the testing problem, whether parameters controlling for the mean of the latent affine process in the empirical and in the equivalent martingale measure are different. We observe substantial undersizing, when implementing a Wald-type test based on standard estimates of the covariance matrix of the unknown parameter. However, by applying Quasi-Bayesian methods to obtain the standard errors of the corresponding components of the parameter vector we observe then that the rejection rates of the true null hypothesis are close to theoretically correct levels.

In a final step, our estimation methodology is applied to empirical term structure data. By applying the testing procedure proposed in this article, the null hypothesis of equal parameters controlling for the

mean of the latent affine process, in the empirical as well as in the equivalent martingale measure, is rejected. Our estimates thus support the presence of a significant market price of risk.

Acknowledgments

The authors thank Eberhard Mayerhofer, Robert Kunst, Paul Schneider and Chen Zhou as well as the participants of the *CFE 2012, 2013* conferences, the *GPSD 2014* conference, the *COMPSTAT 2014* conference and the *SFG 2017* conference for interesting discussions and comments. Financial support from the Austrian Central Bank under Anniversary Grant Nr. 14678 is gratefully acknowledged. Moreover, we are grateful to two anonymous referees for helpful comments.

Supplementary material

Supplementary material associated with this article can be found as annexes in the electronic version of the manuscript.

		"Strong Prior Information"				"Weak Prior Information"			
		Bias		$RMSE$		Bias		$RMSE$	
	$\boldsymbol{\vartheta}$	$\lvert\widehat{\vartheta}_i - \vartheta_i\rvert$	$\lvert\tilde{\vartheta}_i - \vartheta_i\rvert$	$\widehat{RMSE}_i$	$\widetilde{RMSE}_i$	$\lvert\widehat{\vartheta}_i - \vartheta_i\rvert$	$\lvert\tilde{\vartheta}_i - \vartheta_i\rvert$	$\widehat{RMSE}_i$	$\widetilde{RMSE}_i$
θ^Q	10	0.1105	0.1627	0.9514	0.1845	0.0211	0.3163	1.0371	1.0970
θ^P	1.5	0.0080	0.0474	0.0805	0.2644	0.1496	0.2545	0.6153	0.5832
β^Q_{11}	-5	0.0020	0.0650	0.4758	0.3509	0.3238	0.1243	2.6409	1.8208
β^Q_{21}	2	0.0206	0.0160	0.2002	0.1095	0.1165	0.2741	0.9522	0.9750
β^Q_{31}	1	0.0060	0.0119	0.1008	0.1027	1.1174	0.9503	1.1889	1.0715
β^Q_{22}	-5	0.0204	0.0490	0.4374	0.1967	0.6856	1.1203	2.2239	2.1613
β^Q_{32}	2	0.0290	0.0399	0.1968	0.1107	1.8284	1.6069	1.6884	1.5597
β^Q_{23}	4	0.0391	0.0338	0.4059	0.2148	3.6090	3.3463	1.7212	1.7389
β^Q_{33}	-5	0.0378	0.0957	0.4473	0.2087	0.5098	1.0133	2.0426	2.1011
β^P_{11}	-0.8	0.0017	0.3418	0.1615	0.4725	0.5066	0.4081	0.5279	0.4168
β^P_{21}	0.02	0.0000	0.3049	0.0042	0.3699	0.2351	0.2167	0.2551	0.0836
β^P_{31}	0.01	0.0000	0.2024	0.0021	0.2494	0.2505	0.2346	0.2721	0.1156
β^P_{22}	-0.7	0.0034	0.3739	0.1414	0.5491	0.3178	0.3426	0.4883	0.4101
β^P_{32}	0.01	0.0001	0.2818	0.0020	0.5096	0.0005	0.0140	0.1838	0.2118
β^P_{23}	0	0.0001	0.1526	0.0021	0.4069	0.0130	0.0058	0.1797	0.2068
β^P_{33}	-0.7	0.0005	0.3825	0.1438	0.5546	0.3487	0.3326	0.5022	0.4240
$\mathcal{B}^x_{12}$	0.05	0.0001	0.2460	0.0103	0.3210	0.4575	0.4472	0.4744	0.2148
$\mathcal{B}^x_{13}$	0.1	0.0004	0.1702	0.0206	0.2394	0.3929	0.3540	0.4120	0.2316
γ_0	2	0.0669	0.0192	0.4199	0.1929	0.6067	0.4819	0.7790	0.7035
Σ_1	0.7	0.0045	0.0140	0.0343	0.0890	0.0833	0.0660	0.2377	0.2863
Σ_2	1	0.0012	0.1605	0.0682	0.2362	0.1172	0.1889	0.2555	0.3078
Σ_3	0.8	0.0022	0.0522	0.0505	0.1573	0.0541	0.0034	0.2477	0.3035
σ^2_ε	0.0067	0.0002	0.0008	0.0004	0.0006	0.0024	0.0026	0.0011	0.0005

Table 1: Comparison of parameter estimates for the $\mathbb{A}_1(3)$ model: (i) multi-start random search only, $\widehat{\boldsymbol{\vartheta}}$, and (ii) multi-start random search and MCMC, $\tilde{\boldsymbol{\vartheta}}$. Data are simulated with $M = 10$, $T = 500$, $\theta^Q \neq \theta^P$ and $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$. $\mathtt{N}_0 = 200,000$ initial search steps for weak prior information. $\mathtt{N} = 1,000$ with strong prior information and $\mathtt{N} = 2,000$ with weak prior information, $c_\vartheta = 1$. $\mathtt{M} = 6,000$ Quasi-Bayesian steps; 1,000 burn in steps. $\widehat{RMSE}$ denotes estimates of the $RMSE$ based on (only) the multi-start random search parameter estimates and $\widetilde{RMSE}$ denotes estimates of the $RMSE$ based on multi-start random search and MCMC parameter estimates. Statistics for $\widehat{\boldsymbol{\vartheta}}$ and $\tilde{\boldsymbol{\vartheta}}$ are obtained from $\mathtt{L} = 250$ simulation runs.

		"Strong Prior Information"				"Weak Prior Information"			
		Bias		$RMSE$		Bias		$RMSE$	
	$\boldsymbol{\vartheta}$	$\lvert\widehat{\vartheta}_i - \vartheta_i\rvert$	$\lvert\tilde{\vartheta}_i - \vartheta_i\rvert$	$\widehat{RMSE}_i$	$\widetilde{RMSE}_i$	$\lvert\widehat{\vartheta}_i - \vartheta_i\rvert$	$\lvert\tilde{\vartheta}_i - \vartheta_i\rvert$	$\widehat{RMSE}_i$	$\widetilde{RMSE}_i$
θ^Q	1.5	0.0076	0.0494	0.2557	0.3066	0.1901	0.1541	0.6617	0.6122
θ^P	1.5	0.0057	0.0334	0.2256	0.3816	0.1629	0.1278	0.6449	0.6489
β^Q_{11}	-0.8	0.0908	0.0806	0.5344	0.4756	0.1321	0.1490	0.6558	0.6705
β^Q_{21}	0.02	0.0012	0.0471	0.0185	0.1172	0.2321	0.2779	0.2780	0.2133
β^Q_{31}	0.01	0.0001	0.0479	0.0038	0.1423	0.2450	0.2909	0.3222	0.2712
β^Q_{22}	-0.7	0.0212	0.0599	0.1357	0.2606	0.3343	0.3783	0.6658	0.6448
β^Q_{32}	0.01	0.0002	0.0088	0.0029	0.1074	0.0390	0.0280	0.3267	0.3590
β^Q_{23}	0	0.0003	0.0041	0.0030	0.0657	0.0414	0.0501	0.3680	0.3837
β^Q_{33}	-0.7	0.0212	0.0367	0.1354	0.1909	0.3682	0.4003	0.6221	0.6175
β^P_{11}	-0.8	0.0728	0.3100	0.5140	0.6540	0.1242	0.3261	0.5033	0.4423
β^P_{21}	0.02	0.0014	0.2125	0.0153	0.3222	0.2298	0.2195	0.2726	0.0852
β^P_{31}	0.01	0.0016	0.2089	0.0165	0.3169	0.2269	0.2211	0.2677	0.0825
β^P_{22}	-0.7	0.0069	0.2450	0.1915	0.4709	0.3110	0.4494	0.4778	0.4125
β^P_{32}	0.01	0.0002	0.1630	0.0029	0.4081	0.0270	0.0431	0.2704	0.2123
β^P_{23}	0	0.0002	0.0769	0.0026	0.3435	0.0239	0.0237	0.2699	0.2089
β^P_{33}	-0.7	0.0014	0.2164	0.1840	0.4389	0.3658	0.4924	0.5153	0.4140
$\mathcal{B}^x_{12}$	0.05	0.0063	0.1661	0.0698	0.2583	0.4249	0.3894	0.5086	0.2128
$\mathcal{B}^x_{13}$	0.1	0.0010	0.1565	0.0287	0.2682	0.3989	0.3469	0.4874	0.2173
γ_0	2	0.0253	0.0033	0.2864	0.4574	0.1135	0.1418	0.7657	0.7510
Σ_1	0.7	0.0101	0.0533	0.0891	0.1620	0.0313	0.0400	0.2934	0.2670
Σ_2	1	0.0133	0.0279	0.1287	0.2189	0.1522	0.1392	0.3768	0.3399
Σ_3	0.8	0.0129	0.0076	0.0878	0.1531	0.0993	0.0831	0.3533	0.3255
σ^2_ε	0.0067	0.0001	0.0008	0.0011	0.0015	0.0002	0.0002	0.0029	0.0016

Table 2: Comparison of parameter estimates for the $\mathbb{A}_1(3)$ model: (i) multi-start random search only, $\widehat{\boldsymbol{\vartheta}}$, and (ii) multi-start random search and MCMC, $\tilde{\boldsymbol{\vartheta}}$. Data are simulated with $M = 10$, $T = 500$, $\theta^Q = \theta^P$ and $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$. $\mathtt{N}_0 = 200,000$ initial search steps for weak prior information. $\mathtt{N} = 1,000$ with strong prior information and $\mathtt{N} = 2,000$ with weak prior information, $c_\vartheta = 1$. $\mathtt{M} = 6,000$ Quasi-Bayesian steps; 1,000 burn in steps. $\widehat{RMSE}$ denotes estimates of the $RMSE$ based on (only) the multi-start random search parameter estimates and $\widetilde{RMSE}$ denotes estimates of the $RMSE$ based on multi-start random search and MCMC parameter estimates. Statistics for $\widehat{\boldsymbol{\vartheta}}$ and $\tilde{\boldsymbol{\vartheta}}$ are obtained from $\mathtt{L} = 250$ simulation runs.

		"Strong Prior Information"						"Weak Prior Information"					
	$\boldsymbol{\vartheta}$	*mean* $\tilde{\boldsymbol{\vartheta}}$	*min*	*max*	*sdt*	*skew*	*kurt*	*mean* $\tilde{\boldsymbol{\vartheta}}$	*min*	*max*	*sdt*	*skew*	*kurt*
θ^Q	10	9.837	9.711	10.380	0.177	0.485	-0.628	9.684	7.205	12.320	1.057	-0.257	-0.415
θ^P	1.5	1.453	0.752	1.961	0.259	-0.986	1.109	1.754	0.282	3.452	0.574	0.122	-0.107
β^Q_{11}	-5	-4.935	-5.853	-4.488	0.344	-0.765	0.130	-5.124	-9.808	-2.617	1.810	-1.082	0.383
β^Q_{21}	2	1.984	1.700	2.207	0.109	-0.158	-0.388	1.726	0.084	4.774	0.962	1.333	1.927
β^Q_{31}	1	0.988	0.775	1.256	0.103	-0.121	-0.203	1.950	0.053	4.746	1.058	0.878	0.030
β^Q_{22}	-5	-4.951	-5.490	-4.503	0.195	0.212	0.008	-3.880	-9.876	-0.219	2.117	-1.359	1.295
β^Q_{32}	2	1.960	1.747	2.216	0.110	-0.010	-0.012	0.393	-4.737	4.923	1.544	0.226	1.654
β^Q_{23}	4	3.966	3.579	4.501	0.215	-0.078	-0.129	0.654	-4.766	9.310	1.719	0.906	5.904
β^Q_{33}	-5	-4.904	-5.522	-4.501	0.200	-0.080	0.305	-3.987	-9.875	-0.194	2.040	-1.111	0.546
β^P_{11}	-0.8	-1.142	-1.837	-0.489	0.328	-0.299	-0.563	-1.208	-1.921	-0.291	0.405	0.560	-0.569
β^P_{21}	0.02	0.325	0.022	0.932	0.209	1.073	0.793	0.237	0.039	0.441	0.082	-0.030	-0.581
β^P_{31}	0.01	0.212	0.013	0.965	0.146	2.485	9.632	0.245	0.005	1.172	0.115	4.178	32.937
β^P_{22}	-0.7	-1.074	-1.954	-0.416	0.405	-0.194	-0.972	-1.043	-1.862	-0.046	0.409	0.116	-0.540
β^P_{32}	0.01	-0.272	-1.093	1.097	0.425	0.786	0.915	0.024	-0.394	0.432	0.211	-0.027	-0.988
β^P_{23}	0	-0.153	-0.902	0.769	0.377	-0.084	-0.198	-0.006	-0.421	0.393	0.207	-0.082	-0.803
β^P_{33}	-0.7	-1.083	-2.830	-0.278	0.401	-0.911	4.061	-1.033	-2.698	-0.018	0.424	-0.252	1.211
$\mathcal{B}^x_{12}$	0.05	0.296	0.053	0.951	0.206	1.622	2.424	0.497	0.071	0.972	0.215	0.153	-0.791
$\mathcal{B}^x_{13}$	0.1	0.270	0.061	0.797	0.169	1.088	0.316	0.454	0.037	0.928	0.228	0.102	-0.951
γ_0	2	1.981	1.616	2.401	0.187	-0.268	-0.462	1.518	0.507	3.815	0.692	0.867	0.234
Σ_1	0.7	0.686	0.481	0.867	0.088	-0.434	0.528	0.634	0.262	1.447	0.244	0.546	0.665
Σ_2	1	0.839	0.608	1.190	0.174	0.417	-1.119	0.811	0.300	1.477	0.299	0.077	-0.913
Σ_3	0.8	0.748	0.557	1.109	0.149	0.728	-0.318	0.803	0.147	1.458	0.299	0.124	-0.783
σ^2_ε	0.0067	0.006	0.006	0.007	0.000	0.294	0.010	0.004	0.002	0.006	0.000	-0.454	3.464

Table 3: Parameter estimates $\tilde{\boldsymbol{\vartheta}}$ for the $\mathbb{A}_1(3)$ based on multi-start random search and $\mathtt{M} = 6,000$ Quasi-Bayesian steps. Data are simulated with $M = 10$, $T = 500$, $\theta^Q \neq \theta^P$ and $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$. $\mathtt{N} = 1,000$ with strong prior information and $\mathtt{N} = 2,000$ with weak prior information, $c_\vartheta = 1$. $\mathtt{N}_0 = 200,000$ initial search steps for weak prior information. Statistics are obtained from $\mathtt{L} = 250$ simulation runs. *mean*, *min*, *max*, *std*, *skew* and *kurt* stand for the sample mean, minimum, maximum, standard deviation, skewness and excess-kurtosis of the point estimates $\tilde{\boldsymbol{\vartheta}}_\ell$, $\ell = 1, \ldots, \mathtt{L}$. The true parameter values $\boldsymbol{\vartheta}$ are reported in the second column.

		"Strong Prior Information"						"Weak Prior Information"					
	$\boldsymbol{\vartheta}$	*mean* $\tilde{\boldsymbol{\vartheta}}$	*min*	*max*	*sdt*	*skew*	*kurt*	*mean* $\tilde{\boldsymbol{\vartheta}}$	*min*	*max*	*sdt*	*skew*	*kurt*
θ^Q	1.5	1.451	0.403	3.684	0.304	0.995	15.250	1.346	0.364	4.469	0.611	1.129	2.317
θ^P	1.5	1.467	0.277	3.684	0.381	1.033	7.747	1.372	0.255	3.140	0.648	0.500	-0.266
β_{11}^Q	-0.8	-0.881	-3.805	-0.205	0.476	-3.946	17.470	-0.949	-8.173	-0.098	0.670	-5.289	53.276
β_{21}^Q	0.02	0.067	0.002	0.770	0.108	4.079	19.435	0.298	0.012	1.670	0.208	2.077	8.849
β_{31}^Q	0.01	0.058	0.001	1.437	0.134	7.909	73.170	0.301	0.008	2.105	0.267	3.208	16.716
β_{22}^Q	-0.7	-0.760	-2.540	-0.241	0.258	-4.134	21.712	-1.078	-6.613	-0.155	0.643	-3.119	21.438
β_{32}^Q	0.01	0.001	-0.901	0.264	0.107	-5.480	42.218	-0.018	-2.876	1.266	0.359	-1.942	15.415
β_{23}^Q	0	-0.004	-0.378	0.266	0.066	-1.150	7.583	0.050	-0.938	3.935	0.384	4.008	41.255
β_{33}^Q	-0.7	-0.737	-2.188	-0.242	0.190	-4.021	24.057	-1.100	-6.042	-0.212	0.617	-2.458	15.419
β_{11}^P	-0.8	-1.110	-5.404	-0.181	0.609	-2.986	13.686	-1.126	-2.319	-0.362	0.393	-0.019	-0.617
β_{21}^P	0.02	0.233	0.002	1.303	0.243	1.455	2.420	0.239	0.013	0.471	0.085	0.149	0.579
β_{31}^P	0.01	0.219	0.001	1.221	0.240	1.500	2.238	0.231	0.010	0.482	0.082	0.100	0.910
β_{22}^P	-0.7	-0.945	-3.119	-0.211	0.398	-1.896	5.786	-1.149	-1.956	-0.346	0.389	0.107	-1.054
β_{32}^P	0.01	-0.153	-1.898	0.874	0.374	-1.372	4.139	-0.033	-0.445	0.466	0.212	0.342	-0.507
β_{23}^P	0	-0.077	-1.020	1.010	0.335	0.097	1.252	-0.024	-0.468	0.468	0.203	0.275	-0.535
β_{33}^P	-0.7	-0.916	-2.203	-0.173	0.383	-1.138	0.867	-1.192	-1.955	-0.208	0.394	0.230	-0.754
$\mathcal{B}_{12}^x$	0.05	0.216	0.017	1.262	0.203	1.965	5.315	0.439	0.021	0.965	0.210	0.130	-0.582
$\mathcal{B}_{13}^x$	0.1	0.256	0.011	1.141	0.219	1.765	2.758	0.447	0.031	1.021	0.211	0.167	-0.310
γ_0	2	2.003	-1.285	3.787	0.457	-1.297	13.969	2.142	0.583	3.964	0.750	0.061	-0.503
Σ_1	0.7	0.647	0.090	1.176	0.156	-0.730	1.513	0.660	0.253	1.471	0.267	0.948	0.378
Σ_2	1	0.972	0.170	3.280	0.218	5.604	58.796	0.861	0.258	1.478	0.340	0.093	-1.227
Σ_3	0.8	0.808	0.361	2.311	0.153	4.510	38.857	0.883	0.270	1.489	0.325	0.041	-0.998
σ_ε^2	0.0067	0.008	0.005	0.013	0.001	1.418	2.734	0.007	0.002	0.015	0.002	1.330	4.028

Table 4: Parameter estimates $\tilde{\boldsymbol{\vartheta}}$ for the $\mathbb{A}_1(3)$ based on multi-start random search and $\mathtt{M} = 6000$ Quasi-Bayesian steps. Data are simulated with $M = 10$, $T = 500$, $\theta^Q = \theta^P$ and $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$. $\mathtt{N} = 1,000$ with strong prior information and $\mathtt{N} = 2,000$ with weak prior information, $c_\vartheta = 1$. $\mathtt{N}_0 = 200,000$ initial search steps for weak prior information Statistics are obtained from $\mathtt{L} = 250$ simulation runs. *mean*, *min*, *max*, *std*, *skew* and *kurt* stand for the sample mean, minimum, maximum, standard deviation, skewness and excess-kurtosis of the point estimates $\tilde{\boldsymbol{\vartheta}}_\ell$, $\ell = 1, \ldots, \mathtt{L}$. The true parameter values $\boldsymbol{\vartheta}$ are reported in the second column.

H_0:	$\theta^Q = \theta^P$				$\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$			
DGP	$\theta^Q \neq \theta^P$ $\boldsymbol{\beta}^Q \neq \boldsymbol{\beta}^P$		$\theta^Q = \theta^P$ $\boldsymbol{\beta}^Q = \boldsymbol{\beta}^P$		$\theta^Q \neq \theta^P$ $\boldsymbol{\beta}^Q \neq \boldsymbol{\beta}^P$		$\theta^Q = \theta^P$ $\boldsymbol{\beta}^Q = \boldsymbol{\beta}^P$	
Prior Information	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak
Rejection Rate	1.000	1.000	0.140	0.068	1.000	1.000	0.080	0.024

Table 5: Parameter tests based on the Wald-type test (14): $\mathscr{W}_T$ obtained by means the estimate $\tilde{\boldsymbol{\vartheta}}$ and the Quasi-Bayesian estimate $\tilde{\mathbf{V}}_T$. The quantities presented are rejection rates of the null hypothesis presented in the first row given significance level $\alpha_S = 5\%$. The data generating process (DGP) is simulated with $M = 10$, $T = 500$ and $c_\theta = 1$. The 'true' parameters are provided in the second column of Table 3 for the $\theta^Q \neq \theta^P$ and $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$ case, while for the $\theta^Q = \theta^P$ and $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ case the 'true' parameters are provided in the second column of Table 4. Tests on the null hypothesis $\theta^Q = \theta^P$ against the two sided alternative hypothesis $\theta^Q \neq \theta^P$ are provided in columns 2-3 (power) and 4-5 (size) and tests on the null hypothesis $\bar{\boldsymbol{\beta}}^Q = \bar{\boldsymbol{\beta}}^P$ against the two sided alternative hypothesis $\bar{\boldsymbol{\beta}}^Q \neq \bar{\boldsymbol{\beta}}^P$ are provided in columns 6-7 (power) and 8-9 (size). Statistics are obtained from $\mathtt{L} = 250$ simulation runs.

References

Aït-Sahalia, Y. (1996a). Nonparametric pricing of interest rate derivative securities. *Econometrica*, 64:527–560.

Aït-Sahalia, Y. (1996b). Testing continuous-time models of the spot interest rate. *The Review of Financial Studies*, 9(2):385–426.

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely-sampled diffusions: A closed-form approximation approach. *Econometrica*, 70:223–262.

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

Altonji, J. G. and Segal, L. M. (1996). Small-Sample Bias in GMM Estimation of Covariance Structures. *Journal of Business & Economic Statistics*, 14(3):353–66.

Andersen, T. G., Chung, H.-J., and Sorensen, B. E. (1999). Efficient method of moments estimation of a stochastic volatility model: A monte carlo study. *Journal of Econometrics*, 91(1):61–87.

Andrieu, C. and Thoms, J. (2008). A tutorial of adaptive MCMC. *J. Stat Comput*, 18:343–373.

Chen, H. and Joslin, S. (2012). Generalized transform analysis of affine processes and applications in finance. *The Review of Financial Studies*, 25(7):2225–2256.

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

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346.

Chib, S. and Ergashev, B. (2009). Analysis of multifactor affine yield curve models. *Journal of the American Statistical Association*, 104(488):1324–1337.

Cochrane, J. (2005). *Asset Pricing*. Princeton University Press, revised edition.

Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407.

Creal, D. D. and Wu, J. C. (2015). Estimation of affine term structure models with spanned or unspanned stochastic volatility. *Journal of Econometrics*, 185(1):60–81.

Cuchiero, C., Teichmann, J., and Keller-Ressel, M. (2012). Polynomial processes and their application to mathematical finance. *Finance & Stochastics*, 16(4):711–740.

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

Dai, Q. and Singleton, K. J. (2003). Term structure dynamics in theory and reality. *The Review of Financial Studies*, 16(3):631–678.

de Los Rios, A. D. (2015). A new linear estimator for gaussian dynamic term structure models. *Journal of Business & Economic Statistics*, 33(2):282–295.

Diebold, F. X., Rudebusch, G. D., and Aruoba, S. B. (2006). The macroeconomy and the yield curve: A dynamic latent factor approach. *Journal of Econometrics*, 131:309–338.

Duffee, G. R. (2011). Information in (and not in) the term structure. *The Review of Financial Studies*, 24(9):2895–2934.

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

Egorov, A. V., Li, H., and Ng, D. (2011). A tale of two yield curves: Modeling the joint term structure of dollar and euro interest rates. *Journal of Econometrics*, 162(1):55–70.

Filipović, D. (2009). *Term-Structure Models: A Graduate Course*. Springer, Berlin.

Frühwirth-Schnatter, S. and Geyer, A. (1996). Bayesian estimation of econometric multi-factor cox-ingersoll-ross-models of the term structure of interest rates via MCMC methods. Working paper, Vienna University of Economics and Business.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Stochastic Modelling and Applied Probability, No. 53, Springer, New York.

Glasserman, P. and Kim, K.-K. (2010). Moment Explosions and Stationary Distributions in Affine Diffusion Models. *Mathematical Finance*, 20(1):1–33.

Grasselli, M. and Tebaldi, C. (2008). Solvable affine term structure models. *Mathematical Finance*, 18(1):135–153.

Guggenberger, P. and Smith, R. J. (2005). Generalized empirical likelihood estimators and tests under partial, weak, and strong identification. *Econometric Theory*, null:667–709.

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

Hamilton, J. D. and Wu, J. C. (2014). Testable implications of affine term structure models. *Journal of Econometrics*, 178:231 – 242. Recent Advances in Time Series Econometrics.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Jones, C. S. (2003). Nonlinear mean reversion in the short-term interest rate. *The Review of Financial Studies*, 16(3):793–843.

Joslin, S., Le, A., and Singleton, K. J. (2013). Why gaussian macro-finance term structure models are (nearly) unconstrained factor-vars. *Journal of Financial Economics*, 109(3):604 – 622.

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

Le, A., Singleton, K. J., and Dai, Q. (2010). Discrete-time affineq term structure models with generalized market prices of risk. *Review of Financial Studies*, 23(5):2184–2227.

Monfort, A., Renne, J.-P., and Roussellet, G. (2015). A Quadratic Kalman Filter. *Journal of Econometrics*, 187(1):43–56.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):pp. 703–708.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):pp. 631–653.

Newey, W. K. and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719.

Phillips, P. C. B. and Hansen, B. E. (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies*, 57(1):99–125.

Piazzesi, M. (2010). *Affine Term Structure Models*. In Y. Aït-Sahalia and L. Hansen (Eds.), Handbook of Financial Econometrics, North-Holland, Amsterdam.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models, Asymptotic Theory*. Springer, New York.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. Oxford University Press, New York.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance*, 52(5):1973–2002.

Törn, A. and Zilinskas, A. (1989). *Global Optimization*. Lecture Notes in Computer Science 350. Springer.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5:177–188.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, 126(1):25–51.

Zhou, H. (2001). Finite sampler properties of EMM, GMM, QMLE, and MLE for a square-root interest rate diffusion model. *Journal of Computational Finance*, 5:89–122.

Zhou, H. (2003). Itô conditional moment generator and the estimation of short-rate process. *Journal of Financial Econometrics*, 1(2):250–271.