Bayesian Parameter Estimation and Identification of $\mathbb{A}_l(m)$-Affine Term Structure Models*

Leopold Sögner †

June 27, 2012

Abstract

This article investigates problems arising with near unit root behavior for affine term structure models. We show that with increasing serial correlation the Fisher information matrix approaches a singularity. We apply Markov Chain Monte Carlo simulation techniques in connection with regularized priors, as proposed in Schotman and van Dijk [1991], Jones [2003] and De Pooter *et al.* [2006] to simulate the joint posterior distribution of the model parameters. Sufficiently informative priors are necessary to obtain a well performing Bayesian sampler.

Keywords: Affine term-structure models, MCMC, near unit root behavior.

JEL: C01, C11, G12.

*The author appreciates helpful comments from Manfred Frühwirth, Robert Kunst, Jan Mutl, Paul Schneider, Volker Vonhoff, Martin Wagner and Marco Willner.

†Leopold Sögner, Department of Economics and Finance, Institute for Advanced Studies, Stumpergasse 56, 1060 Vienna, Austria, soegner@ihs.ac.at

1 Introduction

Term structure data usually exhibit a high degree of serial correlation. For these data, standard tests on a unit root often do not reject the null hypothesis of a unit root for usual significance levels. On the other side, from economic intuition and models used in mathematical finance, stationary time series for term structure data should be observed. This article sticks to the assumption of stationary interest rates and contributes to literature by highlighting pitfalls arising with parameter estimation for affine term structure models. The first innovative aspect of this paper is an analysis of the information matrix. Second, we observe that parameter estimation for a latent diffusion process becomes a difficult problem when the serial correlation is high (*near unit root behavior*). We observe that if the process approaches a unit root, the sampler produces a "wall", such that the posterior need not be integrable. That is to say we need not arrive at a *proper posterior distribution*. To enable Bayesian parameter estimation, we construct priors to account for this problem. We follow the works of Schotman and van Dijk [1991], Kleibergen and van Dijk [1994], Kleibergen and van Dijk [1998], Jones [2003], De Pooter *et al.* [2006], and De Pooter *et al.* [2008] to construct priors regularizing the posterior distribution of the parameters.

Affine term structure models describe yields by means of an affine function of an instantaneous affine vector diffusion process. The focus of this article is on the risk-free term structure, where only interest rate risk - and no other sources of risk like credit and liquidity risk - is investigated. The risk-free term structure is the basic building block of any *reduced form* credit risk model. Although other sources of risk can be described with more general models of the affine class, the mathematical structure of these settings is equal or similar to the structure investigated in this article (see e.g. Duffie and Singleton [1997], Collin-Dufresne and Goldstein [2001], Dai and Singleton [2002], Duffee [2002], Collin-Dufresne and Goldstein [2002], Collin-Dufresne *et al.* [2008], Ang *et al.* [2004], Collin-Dufresne *et al.* [2009], Filipović [2009], CDS pricing models, etc.). Therefore the results of the following analysis are also important for more general settings.

We already observe one of the core problems with the instantaneous yields, which is a singularity in

the Fisher information matrix when the yields are highly persistent. Further sections of this article will demonstrate that this problem neither disappears when non-instantaneous yields are considered nor arises from the use of Bayesian methods.

This paper combines results from different strands of econometrics literature: Near unit root behavior in discrete time models, continuous time financial econometrics and Bayesian techniques of prior selection to regularize the posterior distribution. For discrete time autoregressive models problems with near unit root behavior are known for a long time in classical econometrics: e.g. the bias in the OLS estimate of the persistence parameter (see Kendall [1954], Campbell *et al.* [1996][p. 273]), the Dickey Fuller example (see e.g. Davidson and MacKinnon [1993][p. 702], Greene [1997][p. 848] or Hamilton [1994][p. 486]), asymptotic distributions of the estimators (see e.g. Phillips [1998], Elliott and Stock [2001], Rothenberg and J. H.Stock [1997], Jansson and M. J. Moreira [2006]) and weak identification (see e.g. Blais [2009], Canova and Sala [2009], Dufour [1997], Dufour [2003], Ma and Nelson [2009]). For financial econometrics the reader is referred to Aït-Sahalia [2007], Piazzesi [2010] and Lewellen [2004]). Regarding information matrix issues, Aït-Sahalia and Jacod [2008] recently derived the behavior of the information matrix for Lévy processes; however, the processes investigated in their paper do not include mean reversion in the drift, as done in this article. Here, we shall observe that the speed of mean reversion will play a central role regarding weak identification issues

Econometric issues arising with parameter estimation for affine term structure models have been investigated in Ang and Piazzesi [2003], Chib and Ergashev [2009], Diebold *et al.* [2006], Duffee [2011], Aït-Sahalia and Kimmel [2009], Egorov *et al.* [2011], Hamilton and Wu [2010], Joslin *et al.* [2010], etc. Multi-factor Cox-Ingersoll-Ross models have been investigated in Frühwirth-Schnatter and Geyer [1996] and Sanford and Martin [2005]. Chib and Ergashev [2009] construct a Bayesian estimation procedure for an Ang *et al.* [2004] model with proposal densities based on mode and curvature of conditional distributions to improve the efficiency of the MCMC sampler. In our estimation procedure we also follow these ideas. Jones [2003] finds out that rather strong priors are necessary to estimate the parameters of the diffusion process. This paper also explains why this becomes necessary: When observing data with serial correlation

close to a unit root, at least some of the mean reversion parameters have to account for near unit root behavior. The closer to a unit root, the more we approach to a singularity in the information matrix. Almost recently Blais [2009] investigated identification issues and the specification of the market micro structure noise in a Bayesian setting. In some parts, it is closely related to the problems investigated in this paper. While Blais [2009] discusses the problem of label switching and different labeling subspaces in his paper, this paper sticks to one unique labeling subspace and contributes to literature by discussing weak identification issues arising from near unit root behavior.

This paper is organized as follows: Section 2 introduces affine settings. Section 3 investigates the likelihood of yields and the Fisher information matrix. In Section 4 we provide a Bayesian analysis for instantaneous yields, while Section 5 investigates parameter estimation for non-instantaneous yields. Section 6 applies our methodology to empirical data. Section 7 concludes.

2 Affine Term Structure Models

Assume a frictionless and arbitrage-free market in continuous time t and a filtered probability space, equipped with the empirical probability measure $\mathbb{P}$ and an equivalent martingale measure (risk-neutral measure) $\mathbb{Q}$. Throughout this paper we restrict to affine models of Dai and Singleton [2000] structure (i.e. the diffusion matrix can be *diagonalized*, for more details see Appendix A). I.e. we consider an affine process $(X(t))$ following the stochastic differential equation:

$$\begin{aligned} dX(t) &= \kappa^Q(\theta^Q - X(t))dt + \Sigma\sqrt{S(t)}dW^Q(t) \text{ where} \\ & \quad S_{ii}(t) = a_i + b_i^\top X(t) \text{ and } S_{ij} = 0 \text{ for } i,j = 1,\dots,m. \end{aligned} \tag{1}$$

$X(t) \in \mathbb{R}^m$ and $W^Q(t)$ is a m-dimensional Brownian motion under the equivalent martingale measure with independent components. κ^Q is a $m \times m$ matrix controlling the speed of mean reversion. Σ is a positive definite $m \times m$ matrix. $S(t)$ is a diagonal matrix including the components $S_{ii}(t) = a_i + b_i^\top X(t)$, a_i is a scalar and b_i a vector of dimension m. We get the vector $\mathcal{A} := (a_1, \dots, a_i, \dots, a_m)^\top$ and the matrix

$\mathcal{B}$ by horizontally stacking the vectors b_i, i.e. $\mathcal{B} = (b_1|\dots|b_m)$. Using this notation $\mathcal{B}_{ji} = b_{i(j)}$, with $i, j = 1, \dots, m$.

Market Prices of Risk and Dynamics in $\mathbb{P}$*:* We employ extended affine market prices of risk $\Lambda(t) = [\Sigma\sqrt{S(t)}]^{-1}\left(\mu^P(X_t) - \mu^Q(X_t)\right)$ (see Cheridito *et al.* [2007]). The drift term $\mu^Q(X_t) = \kappa^Q(\theta^Q - X(t))$ in (1) and the extended affine specification results in an affine drift term in the empirical measure $\mathbb{P}$, such that $\mu^P(X_t) = \kappa^P(\theta^P - X(t))$. Thus by construction, $(X(t))$ is an affine stochastic process with diagonal diffusion term also under $\mathbb{P}$, such that

$$dX(t) \quad = \quad \kappa^P(\theta^P - X(t))dt + \Sigma\sqrt{S(t)}dW^P(t), \tag{2}$$

where κ^P, θ^P and W^P have a structure analogous to κ^Q, θ^Q and W^Q with $dW^Q(t) = dW^P(t) - \Lambda(t)$. By estimating $\kappa^{\cdot}$ and $\theta^{\cdot}$ under both measures, $\mathbb{P}$ and $\mathbb{Q}$, the market price of risk parameters are estimated implicitly. This allows to study how the market compensates investors for bearing interest rate risk [see e.g. Driessen, 2005; Piazzesi, 2010].

Dai and Singleton [2000]-*canonical representation* and $\mathbb{A}_l(m)$ *Models*: Recent quantitative finance literature favors $\mathbb{A}_l(m)$ models (see e.g. Tang and Xia [2007]). They are described as follows:

Definition 1 ($\mathbb{A}_l(m)$-Term Structure Model)**.** Suppose that the risk-free term structure is driven by an affine process $(X(t))$ (under $\mathbb{Q}$) with diagonal diffusion matrix. $X(t)$ is a vector of dimension m which splits up into $X^B \in \mathbb{R}_+^l$ and $X^D \in \mathbb{R}^{m-l}$. The risk-free instantaneous discount rate $y(t,0) = \delta_0 + \delta^\top X(t)$, where δ is a m dimensional vector and $\delta_0 \in \mathbb{R}$. In an $\mathbb{A}_l(m)$ setting m is the number of Brownian motions and l is the number of different state variables that show up under the square root in (1); (see Dai and Singleton [2000]).

Regarding (1) it is worth noting that different parameter constellations can result in the same term structure, i.e. the model need not be *identified*. For example an unrestricted $\mathbb{A}_1(3)$ model has nineteen parameters (under $\mathbb{Q}$), while Dai and Singleton [2000] have shown that only fourteen parameters of this model can be identified. In addition the term under the square-root in $S(t)$ has to be positive

(*admissibility*). Therefore, the authors have provided *canonical representations* where the parameters are identified (under $\mathbb{Q}$) and the terms under the square root are positive. For more details see Dai and Singleton [2000] and Appendix A.[1]

Stationarity of $(X(t))$ under both measures requires positive definite matrices κ^P and κ^Q. For the square root components the modified Feller condition has to hold.[2] For independent square root components this reduces to $\kappa^Q_{ii}\theta^Q_i \geq \Sigma^2_{ii}/2$. Since we have assumed equal structures in $\mathbb{Q}$ and $\mathbb{P}$, all the requirements on the parameters under $\mathbb{Q}$ carry over to the parameters under $\mathbb{P}$. In this paper we assume:

Assumption 1. Consider a canonical representation of an $\mathbb{A}_l(m)$ model, where the admissibility and the Dai and Singleton [2000]-identification restrictions are fulfilled. The structures of the affine model in $\mathbb{P}$ and in $\mathbb{Q}$ are the same. $(X(t))$ is stationary under $\mathbb{Q}$ and $\mathbb{P}$.

Model Yields and Empirical Yields: Under the above assumptions, the time t yields $y(t, T-t)$ for a zero-coupon bond with maturity $\tau = T - t$ are given by

$$y(t,\tau) = -\frac{1}{\tau}\left(A(\tau) - B(\tau)^\top X(t)\right), \tag{3}$$

where $A(\tau) \in \mathbb{R}$ and $B(\tau) \in \mathbb{R}^m$ are functions of the parameters (under $\mathbb{Q}$). Generally, $A(\tau)$ and $B(\tau)$ can be found as solutions to ordinary differential equations of Riccati type [see Duffie and Kan, 1996]:

$$\begin{aligned}
\frac{dA(\tau)}{d\tau} &= -\theta^{Q\top}\kappa^{Q\top}B(\tau) + \frac{1}{2}\sum_{i=1}^{m}[\Sigma^\top B(\tau)]_i^2 a_i - \delta_0 \text{ with } A(0) = 0 \text{ and} \\
\frac{dB(\tau)}{d\tau} &= -\kappa^{Q\top}B(\tau) + \frac{1}{2}\sum_{i=1}^{m}[\Sigma^\top B(\tau)]_i^2 b_i + \delta \text{ with } B(0) = \mathbf{0}_{m\times 1}\,.
\end{aligned} \tag{4}$$

For extensions to jumps or more general transforms the reader is referred to Duffie *et al.* [2000], Chen and Joslin [2009] and Keller-Ressel and Mayerhofer [2011]. It is worth noting that the limits $\lim_{\tau\to 0} A(\tau)/\tau$

[1] It is also worth noting that different restrictions can be used to identify the parameters. Different opportunities to represent an affine term structure model follow from the transformations discussed in Dai and Singleton [2000][especially from Appendix A, C and E] and Filipović [2009][Chapter 10].

[2] See Duffie and Kan [1996][Condition A], Piazzesi [2010][also there denoted as Condition A] or Glasserman and Kim [2009].

and $\lim_{\tau\to 0} B(\tau)/\tau$ are $-\delta_0$ and δ, providing us with the short rate $y(t,0) = \delta_0 + \delta^\top X(t)$. The yields defined by (3) will be called *model yields*.

Although we have assumed a model in continuous time, the empirical/observed yields can only be measured on a discrete grid with step-width Δ. The corresponding model yields and instantaneous yields at $t = n\Delta$ and maturities τ_i, $i = 1, \ldots, k$, are $y_n(\tau_i)$. Consider the k-dimensional vector $\mathbf{A}$, with elements given by $A(\tau_i)/\tau_i$ and the $k \times m$ matrix $\mathbf{B}$, with rows derived by means of $B(\tau_i)^\top/\tau_i$, $i = 1, \ldots, k$. For the maturities $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_i, \ldots, \tau_k)^\top$ the k dimensional vector of model yields $\mathbf{y}_n$ is given by $\mathbf{y}_n = \mathbf{A} - \mathbf{B}X_n$. We assume the following relationship between the model yields $\mathbf{y}_n$ and the empirical yields $\mathbf{y}_n^{eps} = (y_n^{eps}(\tau_1), \ldots, y_n^{eps}(\tau_k))^\top$:

Assumption 2 (Empirical Yields)**.** The observed data $\mathbf{y}_n^{eps}$ and the model yields $\mathbf{y}_n$ are related by

$$y_n^{eps}(\tau_i) \quad = \quad y_n(\tau_i) + \sqrt{\sigma_{eps}^2(\tau_i)} e_{in} \ , \quad i = 1, \ldots, k. \tag{5}$$

e_{in} are *iid* standard normal variables for $i = 1, \ldots, k$. $k \geq m$.

Finance literature often motivates this noise term by market micro-structure noise arising from bid-ask bounces, discreteness of the pricing scale, trades on different markets, etc. (see Campbell *et al.* [1996] and Chen *et al.* [2007]). From an *econometric* point of view (5) is necessary to cope with the different dimensions of the latent process and the yields observed. A parsimonious model demands for $k > m$. Since it is hardly possible that empirical interest rate data exactly follow the model assumed by (3) for all t and maturities τ_i, the m factor setting cannot exactly match the corresponding yields $\mathbf{y}_n^{eps}$.

Remark 1. By means of (5) we have added noise to each maturity which eliminates this stochastic singularity problem. Alternatively, e_{in} can be stochastic for $i \in \{i | \tau_i \in \boldsymbol{\tau} \setminus \boldsymbol{\tau}^{fix}\}$ while $e_{in} = 0$ if $i \in \{i | \tau_i \in \boldsymbol{\tau}^{fix}\}$. $\boldsymbol{\tau}^{fix} \subset \boldsymbol{\tau}$ are the maturities observed without noise. We call this *particular* noise, while Assumption 2 describes *common* noise. Appendix C will demonstrate that the transformation between X_n and $\mathbf{y}_n$ can be ill conditioned. Therefore we stick to common noise.

Parameterization: In this article we stick to the parameterization $\Psi = \{\nu, \theta^P, \kappa^Q, \theta^Q, \delta_0, \Sigma^2, \boldsymbol{\sigma}_{eps}^2\}$;

where $\Sigma^2 = \Sigma\Sigma^\top$. $\delta = \mathbf{1}_{m\times 1}$, the non-zero elements of $\mathcal{A}$ are normalized to one. With independent square root components - as performed in the applied part - $\mathcal{B}_{ii} = 1$. $\nu := \exp(-\kappa^P\Delta)$ is the matrix exponential of the matrix $-\kappa^P\Delta$ (we can get κ^P from ν by means of the matrix logarithm (see Culver [1966], Horn and Johnson [1985])).[3] $\boldsymbol{\sigma}^2_{eps}$ is a $k \times k$ diagonal matrix with entries $\sigma^2_{eps}(\tau_i)$. Finally, $\mathbf{X} = (X_0, X_1, \ldots, X_N)$, $\mathbf{X}_{(1:N)} = (X_1, \ldots, X_N)$, $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ while $\mathbf{y}^{eps} = (\mathbf{y}^{eps}_1, \ldots, \mathbf{y}^{eps}_N)$.

In addition, we assume that the term structure model is of *minimum dimension* (an $\mathbb{A}_0(m)$ setting exactly corresponds to a linear state space model as investigated in Brockwell and Davis [2006][p. 497]; for controllability in general see e.g. Meyn and Tweedie [2009][Chapter 7]). This implies that we cannot reconstruct the model yields with a latent process X_n of dimension smaller than m. In more details, following Karatzas and Shreve [1991][p. 354] and applying an Euler type approximation to the diffusion term provides a proxy of the solution of the affine stochastic differential equation (2):

$$\begin{aligned} X_n &= \nu X_{n-1} + (I_m - \nu)\theta^P + \Sigma \int_{(n-1)\Delta}^{n\Delta} \exp\left(-(n\Delta - u)\kappa^P\right) \sqrt{S(X(u))} dW(u) \\ &\approx \nu X_{n-1} + (I_m - \nu)\theta^P + \Sigma\sqrt{S(X_{n-1})}\sqrt{\Delta}\varepsilon_n \ . \end{aligned} \tag{6}$$

I_m is the identity matrix of dimension m and ε_n is a vector of dimension m with $iid\ \mathcal{N}(0,1)$ components. Since $S(X_{n-1})$ generally depends on X_{n-1} (e.g. some elements of $\mathcal{B}$ are non-zero), we only get a proxy by equation (6); in the following S_{n-1} abbreviates $S(X_{n-1})$. By means of (6) and Assumption 2 we get the following state space representation of the yields observed:

$$\begin{aligned} \mathbf{y}^{eps}_n &= \mathbf{y}_n + \sqrt{\boldsymbol{\sigma}^2_{eps}}\mathbf{e_n} = \mathbf{A} - \mathbf{B}X_n + \sqrt{\boldsymbol{\sigma}^2_{eps}}\mathbf{e_n} \ , \\ X_n &= \nu X_{n-1} + (I_m - \nu)\theta^P + \Sigma\sqrt{S_{n-1}}\sqrt{\Delta}\varepsilon_n \ . \end{aligned} \tag{7}$$

$\mathbf{e_n}$ is of dimension $k \times 1$. Appendix B shows that the model is of minimal dimension if Σ^2, ν and $\mathbf{B}$ have full rank m. Σ^2 and ν satisfy this property by Assumptons 1 and 2 while for $\mathbf{B}$ we impose

[3] Since Σ and $S(t)$ are diagonal matrices $\Sigma S(t)\Sigma^\top$ has to be equal to $\Sigma\Sigma^\top S(t)$; $\Sigma\Sigma^\top =: \Sigma^2$ can be derived by taking the squares of the individual components. If $S_{ii} = a_i + b_i^\top X$ and $a_i = \mathcal{A}_i$ and $b_i = \mathcal{B}_i^\top$ are both non-zero, then a_i can be normalized to one but $\mathcal{B}_i$ has to estimated. Appendix D.1 derives the information matrix with $\mathcal{B}$ as a free parameter.

Assumption 3. $rank(\mathbf{B}) = m$.

As the following example demonstrates, Dai and Singleton [2000] identification need not result in a model of minimal dimension.

Example 1. Consider an $A_0(2)$ model with two independent Ornstein-Uhlenbeck processes with the same parameters; i.e. $\kappa_{11} = \kappa_{22}$, $\kappa_{21} = 0$ by independence under $\mathbb{P}$ and $\mathbb{Q}$, $\theta^P = \theta^Q = 0$ and $\Sigma_{11}^2 = \Sigma_{22}^2$. $\tilde{X}(t) = X_1(t) + X_2(t)$ is an Ornstein-Uhlenbeck process with parameters $\tilde{\kappa}^P = \kappa_{11}^P = \kappa_{11}^P$, $\tilde{\theta}^P = 0$ and $\tilde{\Sigma}^2 = \Sigma_{11}^2 + \Sigma_{22}^2$. With $\tilde{\kappa}^Q = \kappa_{11}^Q$ we get $\mathbf{B}(X_1, X_2) = [\mathbf{B}]_{.,1}\tilde{X}$ ($[\mathbf{B}]_{..1}$ is the first column of $\mathbf{B}$; since $\tilde{\kappa}^Q = \kappa_{11}^Q$ the columns are the same). In addition for the Vasicek model $\mathbf{A}$ is linear in volatility parameter such that the sum of the components $\mathbf{A}_1$ and $\mathbf{A}_2$ for the initial two factor setting add up to $\tilde{\mathbf{A}}$. This implies that we can reduce this two-dimensional model to a one-dimensional one yielding the same term structure.

In the applied part where the parameters will be estimated by means of Bayesian methods we shall put a prior on the rank of $\mathbf{B}$. Since $rank(\mathbf{B}) = rank(\mathbf{B}^\top) = rank(\mathbf{B}^\top\mathbf{B})$ this can easily done by putting a prior on $det(\mathbf{B}^\top\mathbf{B})$.

3 Likelihood Analysis and the Information Matrix

Based on the model assumptions, we first derive the density of the latent process $f(\mathbf{X}; \Psi^P)$, where the corresponding parameters in the empirical measure $\Psi^P = \{\nu, \theta^P, \Sigma^2\}$. For the yields $\mathbf{y}^{eps}$ we already know that the model yields $\mathbf{y}$ are an affine transformations of $\mathbf{X}$ as described by (3). $A(\tau)$ and $B(\tau)$ are functions of the parameters under $\mathbb{Q}$, which are $\Psi^Q = \{\kappa^Q, \theta^Q, \Sigma^2\}$.[4] To derive the conditional distribution $f(\mathbf{y}|\mathbf{X}; \Psi)$, we have to consider the distribution due to market micro-structure noise. By (5) the relevant parameters are in the matrix $\boldsymbol{\sigma}_{eps}^2$. Then the joint density of $(\mathbf{y}, \mathbf{X})$ will be given by

$$f(\mathbf{y}^{eps}, \mathbf{X}; \Psi) = f(\mathbf{y}^{eps}|\mathbf{X}; \Psi) f(\mathbf{X}; \Psi) = f(\mathbf{y}^{eps}|\mathbf{X}; \Psi^Q, \boldsymbol{\sigma}_{eps}^2) f(\mathbf{X}; \Psi^P) \ . \tag{8}$$

[4]Note that $\Psi^P \cap \Psi^Q = \{\Sigma^2\} \cup \{\theta, \kappa | \theta_i^P = \theta_i^Q, \kappa_{ij}^P = \kappa_{ij}^Q\}$.

The joint log-likelihood $\ell(\Psi; \mathbf{y}^{eps}, \mathbf{X})$ is $\log f(\mathbf{y}^{eps}, \mathbf{X}; \Psi)$ evaluated at the data. Since an approximation of $f(\mathbf{X}; \Psi^P)$ will be used, we are going to derive a *quasi likelihood*.

In more details: By equation (6), $X_n|X_{n-1}$ is approximately multivariate normal with mean $\mu_{Xn} := \nu X_{n-1} + (I_m - \nu)\theta^P$ and covariance $\Sigma^2 S_{n-1}\Delta$. Using the fact the Σ^2 is diagonal yields

$$\log f(X_n|X_{n-1}; \Psi^P) \quad = \quad -\frac{m}{2} \cdot \log 2\pi - \frac{1}{2}\log(\prod_{i=1}^{m} \Sigma_{ii}^2 S_{n-1,ii}) - \frac{1}{2}\sum_{i=1}^{m} \frac{(X_{ni} - \mu_{Xni})^2}{\Sigma_{ii}^2 S_{n-1,ii}} \ . \tag{9}$$

With the $N+1$ observations $\mathbf{X}$ and the initial distribution $\pi(X_0; \Psi)$ we get the density of the latent process (X_n) by means of

$$\begin{aligned} f(\mathbf{X}_{(1:N)}|X_0, \Psi^P) &= \prod_{n=1}^{N} \log f(X_n|X_{n-1}; \Psi^P) \\ f(\mathbf{X}; \Psi^P) &= \left(\prod_{n=1}^{N} \log f(X_n|X_{n-1}; \Psi^P)\right) \cdot \pi(X_0; \Psi) = f(\mathbf{X}_{(1:N)}|X_0, \Psi^P)\pi(X_0; \Psi) \ . \end{aligned} \tag{10}$$

$\log f(\mathbf{X}; \Psi^P)$ evaluated at the data provides us with $\ell(\Psi^P; \mathbf{X})$. To get the density of the observed yields $\mathbf{y}^{eps}$, equation (3) tells us that $\mathbf{y}_n = \mathbf{A} - \mathbf{B}X_n$. Based on the model assumptions $\mathbf{y}_n^{eps}$ is normally distributed with mean $\mathbf{y}_n$ and a diagonal covariance matrix $\boldsymbol{\sigma}_{eps}^2$. I.e $f(\mathbf{y}_n^{eps}|\mathbf{X}_n; \Psi^Q, \boldsymbol{\sigma}_{eps}^2)$ is a normal density with mean vector $\mathbf{y}_n$ and covariance matrix $\boldsymbol{\sigma}_{eps}^2$. Since e_{in} is *iid* we get

$$f(\mathbf{y}^{eps}|\mathbf{X}; \Psi^Q, \boldsymbol{\sigma}_{eps}^2) = \Pi_{n=1}^{N} f(\mathbf{y}_n^{eps}|\mathbf{y}_n; \boldsymbol{\sigma}_{eps}^2) = \Pi_{n=1}^{N} f(\mathbf{y}_n^{eps}|\mathbf{X}_n; \Psi^Q, \boldsymbol{\sigma}_{eps}^2) \ . \tag{11}$$

$\log f(\mathbf{y}^{eps}|\mathbf{X}; \Psi^Q, \boldsymbol{\sigma}_{eps}^2)$ evaluated at the data yields $\ell(\Psi^Q, \boldsymbol{\sigma}_{eps}^2; \mathbf{y}^{eps}|\mathbf{X})$, such that the joint log-likelihood is given by

$$\ell(\Psi; \mathbf{y}^{eps}, \mathbf{X}) \quad = \quad \ell(\Psi^Q, \boldsymbol{\sigma}_{eps}^2; \mathbf{y}^{eps}|\mathbf{X}) + \ell(\Psi^P; \mathbf{X}) \ . \tag{12}$$

Information Matrix and Weak Identification: To investigate weak identification, we study the Fisher information matrix. A positive definite information matrix guarantees at least local identification of the

model parameters (see Bowden [1973]). Parameter estimation becomes difficult if the parameters Ψ result in an ill-conditioned information matrix. Following e.g. McLachlan and Krishnan [1997] the empirical information matrix of the full data $\mathbf{I}_c(\Psi, \mathbf{X}, \mathbf{y}) = -\frac{\partial^2 \ell(\Psi;\mathbf{X},\mathbf{y})}{\partial\Psi\partial\Psi^\top}$. The *Fisher information* of the complete data $\mathcal{I}_c(\theta) = \mathbb{E}\left(\mathbf{I}_c(\theta; \mathbf{X}, \mathbf{y})\right)$. Since the latent process $\mathbf{X}$ is not observed we have to consider the log-likelihood $\ell(\Psi; \mathbf{y}) = \log \int f(\mathbf{y}^{eps}|\mathbf{X}; \Psi) f(\mathbf{X}; \Psi) d\mathbf{X}$ and the restricted data information matrix $\mathbf{I}_r(\Psi, \mathbf{y}) = -\frac{\partial^2 \ell(\Psi;\mathbf{y})}{\partial\Psi\partial\Psi^\top}$ as well as the Fisher information $\mathcal{I}_r(\theta) = \mathbb{E}\left(\mathbf{I}_r(\theta, \mathbf{y})\right)$. From Orchard and Woodbury [1972] and Mislevy and Sheehan [1989] it is known that $\mathcal{I}_c(\Psi)$ and $\mathcal{I}_r(\Psi)$ are related as follows: $\mathcal{I}_c(\Psi) = \mathcal{I}_r(\Psi) + \mathcal{I}_m(\Psi)$, where the matrix $\mathcal{I}_m(\Psi)$ is positive semi-definite matrix, measuring the loss in information when the latent $\mathbf{X}$ is not observed. Since the matrix difference $\mathcal{I}_c(\Psi) - \mathcal{I}_r(\Psi)$ is positive semi-definite, the difference $\mathcal{I}_r(\Psi)^{-1} - \mathcal{I}_c(\Psi)^{-1}$ is positive semi-definite. That is to say, $\mathcal{I}_c(\Psi)^{-1}$ provides us with a lower bound of the Rao-Cramer lower bound when only $\mathbf{y}$ is observed. If this term becomes singular, then $\mathcal{I}_r(\Psi)^{-1}$ has to be singular as well. Since only parts of $\mathcal{I}_c(\theta)$ can be derived analytically we proceed as follows: First we obtain some analytical results, for the remaining parts we use numerical tools.

$\mathcal{I}_c(\Psi)$ will consist of three building blocks: The block regarding Ψ^P will be denoted $\mathcal{I}_c(\Psi^P)$. For the remaining parameters we get the blocks $\mathcal{I}_c(\Psi^Q)$ and $\mathcal{I}_c(\boldsymbol{\sigma}^2_{eps})$. With $\mathcal{I}_c(\Psi^P)$, we already observe the main problem: This block of the information matrix approaches a singularity if the speed of mean reversion implied by ν (or κ^P) becomes low.

Due to its analytical traceability we start with the Vasicek [1977] model, where $(X(t))$ follows an Ornstein-Uhlenbeck process, such that

$$X_n \quad = \quad \nu X_{n-1} + (1-\nu)\theta^P + \Sigma\sqrt{\Delta}\varepsilon_n \ . \tag{13}$$

Although, $\theta^P = 0$ by Assumptions 1-3 we treat θ^P as a free parameter in the following paragraph. The goal is to analytically demonstrate the problems arising with the parameter θ^P if the serial correlation becomes high. With $\kappa^P > 0$, the expected value $\mathbb{E}(X) = \theta^P$, the variance $\mathbb{V}(X) = \frac{\Sigma^2}{2\kappa^P} \approx \frac{\Sigma^2\Delta}{1-\nu^2}$. $\mathcal{I}_c(\Psi^P)$ is diagonal with the elements

$$\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \nu^2}\right) = \frac{1}{\Delta}\frac{N}{2\kappa^P},\ \mathbb{E}\left(-\frac{\partial^2 \ell(.)}{\partial(\theta^P)^2}\right) = \frac{1}{\Sigma^2\Delta}N(1-\nu)^2,\ \mathbb{E}\left(-\frac{\partial^2 \ell(.)}{\partial(\Sigma^2)^2}\right) = \frac{N}{2(\Sigma^2)^2}\ . \quad (14)$$

When the process approaches a unit root, i.e. $\nu \to 1$, the second term in (14) goes to zero. This implies that the data provides poor information on the parameter θ^P. Since the inverse of $\mathcal{I}_c(\Psi^P)$ is the Rao-Cramer lower bound for the estimator Ψ^P, the variance of the estimator of θ^P goes to infinity when we approach a unit root.[5]

Appendix D.1 approximates the information matrix $\mathcal{I}_c(\Psi^P)$ for an affine model with diagonal diffusion matrix. Here we observe that the expectation of the second derivatives with respect to θ^P, given by

$$N \cdot (I_m - \nu)^\top \mathbb{E}(\Sigma^2 S_{n-1}\Delta)^{-1}(I_m - \nu)\ , \quad (15)$$

becomes singular if $(I_m - \nu)$ is singular. This is the case if some eigenvalue of ν is equal to one. With eigenvalues of ν close to one we arrive at a weakly identified problem. For the purely Gaussian case this result has also been observed in Hamilton and Wu [2010]. A similar condition on the eigenvalues is also used in Chib and Ergashev [2009].

Next we investigate the $I(\Psi^Q)$ block: Some intuition can be obtained from the Gaussian settings:

Example 2. Assume that the term structure is described by the Vasicek model. Then $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q,.)}{\partial \kappa_Q^2}\right)$, goes to zero if $\Sigma^2 \to 0$. That is to say even in the simplest one factor setting there is a region of the parameter space where the model becomes weakly identified.

Example 3. Consider a two factor model with independent Ornstein-Uhlenbeck processes (e.g $A_0(2)$ model). Suppose that $\kappa^Q \to 0$, then $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q,.)}{\partial \kappa_{11}^Q \partial \kappa_{22}^Q}\right) = \frac{1}{64}\tau^4\sigma_1^2\Sigma_{22}^2$, $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q,.)}{\partial \kappa_{11}^Q \partial \kappa_{11}^Q}\right) = \frac{1}{64}\Sigma_{11}^2\left(\frac{8}{\kappa_{11}^P} + \tau^4\Sigma_{11}^2\right)$, $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q,.)}{\partial \kappa_{11}^Q \partial \Sigma_{11}^2}\right) = -\frac{1}{48}\Sigma_{11}^2\tau$ and $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q,.)}{\partial \kappa_{11}^Q \partial \Sigma_{22}^2}\right) = -\frac{1}{48}\Sigma_{22}^2\tau$. For the partial derivatives with respect to κ_{22}^Q we get equivalent expressions. If $\Sigma_{11}^2 = \Sigma_{22}^2$, then we observe that the

[5] Applying the reparametrization $\tilde{\gamma} := \theta(1-\nu)$, the singularity in the information matrix does not disappear. In addition in all the calculations we assume that the terms arising from $\pi(X_0;\Psi)$ can be neglected.

rows/columns of $\mathcal{I}_c(\Psi)$ corresponding to κ_{11} and κ_{22} become almost colinear if κ_{11}^P and κ_{22}^P become large.

By these examples we observe that the matrix $\mathcal{I}_c(\Psi^Q)$ can get close to a singularity. In general

$$\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|\mathbf{X})}{\partial \Psi^Q (\Psi^Q)^\top}\right) = \sum_{n=1}^{N} \mathbb{E}\left(\left[\frac{\partial(\mathbf{A}-\mathbf{B}X_n)}{\partial \Psi^Q}\right](\sigma^{\mathbf{2}}_{\mathbf{eps}})^{-1}\left[\frac{\partial(\mathbf{A}-\mathbf{B}X_n)}{\partial \Psi^Q}\right]^\top\right). \quad (16)$$

From (16) we expect weak identification issues if some terms in the gradient vector $\frac{\partial(\mathbf{A}-\mathbf{B}X_n)}{\partial \Psi^Q}$ get close to zero or if some rows or columns are almost the same. Since, in general, $\mathbf{A}$ and $\mathbf{B}$ are not available in closed form, we can only estimate $\mathcal{I}_c(\Psi^Q)$ by means of numerical tools.[6]

For the parameter $\delta_0 \in \Psi^Q$, we get $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|\mathbf{X})}{\partial \delta_0^2}\right) = N \sum_{i=1}^{k} \frac{1}{\sigma^2_{eps}(\tau_i)}$. The second partial derivative with respect to δ_0 neither depends on κ^Q nor on κ^P. I.e. in contrast to the mean parameters under $\mathbb{P}$ no problems should be expected with the estimation of this parameter.

The last block of the information matrix $\mathcal{I}_c(\boldsymbol{\sigma}^2_{eps})$ is obtained by taking the expectation of the second partial derivatives with respect to $\sigma^2_{eps}(\tau_i)$. While the off-diagonal elements are all zero, the diagonal elements of $\mathcal{I}_c(\boldsymbol{\sigma}^2_{eps})$ are

$$\mathbb{E}\left(-\frac{\partial^2}{\partial(\sigma^2_{eps}(\tau_i))^2}\ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|\mathbf{X})\right) = \frac{N}{2(\sigma^2_{eps}(\tau_i))^2}. \quad (17)$$

$\mathcal{I}_c(\Psi)$ is derived by putting together the corresponding blocks we have obtained above, the other elements of the matrix are zero. Note that the blocks for (Ψ^P, Ψ^Q) and $\boldsymbol{\sigma}^2_{eps}$ do not overlap. Σ^2 is an element of Ψ^P *and* Ψ^Q, depending on the market price of risk specifications further joint elements are possible.[7] Therefore the impact of these parameters on the information matrix is non-trivial.

In the following sections we perform parameter estimation for the following $\mathbb{A}_1(3)$ model. $N = 500$ observations are considered for the $k = 10$ maturities $\boldsymbol{\tau} = \{1/12, 1/4, 1/2, 1, 2, 3, 5, 7, 10, 20\}$.

[6] Since $\sigma^{\mathbf{2}}_{\mathbf{eps}}$ is a $k \times k$ diagonal matrix $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|X_n)}{\partial \Psi^Q (\Psi^Q)^\top}\right)$ can be derived in closed form given the partial derivatives of $\mathbf{A}$ and $\mathbf{B}$. The elements of this part of the information matrix are functions of these partial derivatives, $\mathbb{E}(X_n) = \theta^P$ and $\mathbb{E}\left(X_n X_n^\top\right)$ (see also Appendix D.1).

[7] With completely affine market prices of risk $\theta_i^P \neq \theta_i^Q$ and $\kappa_{ii}^P \neq \kappa_{ii}^Q$. Common elements of Ψ^P and Ψ^Q are Σ^2 and off diagonal elements of $\kappa^{\cdot}$. Otherwise Ψ^P and Ψ^Q overlap for some further θ_i and κ_{ii}.

Time is measured in years, the step width is $\Delta = 7/365$ accounting for weekly observations. The parameters $\sigma^2_{eps}(\tau_i)$ driving micro-structure noise are between 0.007 and 0.03.[8] We set $\mathcal{A}_1 = 0$, $\mathcal{A}_2 = \mathcal{A}_3 = 1$ and $\mathcal{B}_{11} = 1$, all other elements of $\mathcal{B}$ are zero. $\kappa^P = [0.7\ 0\ 0;\ 0\ 0.8\ 0;\ 0\ 0.1\ 0.9\ ]$, $\nu = [0.9867\ 0\ 0;\ 0\ 0.9848\ 0; 0\ -0.0019\ 0.9829\ ]$ is the corresponding matrix exponential, $\kappa^Q = [0.5\ 0\ 0;\ 0\ 0.7\ 0;\ 0\ 0.1\ 1\ ]$, $\theta^P = (1.5, 0, 0)^\top$, $\theta^Q = (2, 0, 0)^\top$, $\delta_0 = 2$ and $\Sigma = [0.25\ 0\ 0;\ 0\ 0.40\ 0;\ 0\ 0\ 0.50\ ]$ resulting in $\Sigma^2 = \Sigma\Sigma^\top = [0.0625\ 0\ 0;\ 0\ 0.16\ 0;\ 0\ 0\ 0.25\ ]$. This results in nine parameters under $\mathbb{Q}$. In addition we have four additional parameters under $\mathbb{P}$ (θ^P_1 and κ^P_{ii}, $i = 1, \ldots, 3$) and the three micro-structure noise parameters $\boldsymbol{\sigma}^2_{eps}$. $\kappa^Q_{23} = \kappa^P_{23}$ is assumed. Summing up, this results in 16 parameters. This setting allows for a closed form solution of $A(\tau)$ and $B(\tau)$ and satisfies the stationarity, admissibility and the Feller condition. With these parameters we derived $\mathcal{I}_c$ by means of the above calculations. We observe: (i) A high standard deviation for the parameter θ^P as expected from the above calculations. A modest degree of serial correlation sharply decreases the corresponding elements of $\mathcal{I}_c^{-1}$. (ii) When κ^Q and ν remain fixed as above but Σ^2 decreases then the diagonal elements of $\mathcal{I}_c^{-1}$ corresponding to ν and κ^Q increase; for small Σ^2 these elements become large. (iii) Given high values of κ^Q the derivatives with respect to κ^Q become small. This results in larger values for the elements corresponding to κ^Q in $\mathcal{I}_c^{-1}$. Intuitively, with larger κ^Q the paths of $\mathbf{B}$ rapidly move from values close to -1 to values close to zero where the partial derivatives with respect to κ^Q become small. (iv) As can be expected from (16) an increase in σ^2_{eps} also raises the terms in $\mathcal{I}_c^{-1}$ corresponding to the parameters under $\mathbb{Q}$.

4 Analysis of Instantaneous Yields

This paper applies the Bayesian approach to estimate the model parameters. As already discussed and demonstrated in Chib and Ergashev [2009], the Bayesian approach can be motivated by the complex and

[8] In a prior version we work with the noise specification $\sigma^2_{eps}(\tau_i) = \exp(a_{0eps} + a_{1eps}\tau_i + a_{2eps}\tau_i^2)$ (e.g. motivated by Brandt and He [2002]). The $\sigma^2_{eps}(\tau_i)$ used in this version are obtained by means of $\exp(a_{0eps} + a_{1eps}\tau_i + a_{2eps}\tau_i^2)$ setting $a_{0eps} = -5$ $a_{1eps} = 0.25$ and $a_{2eps} = -0.04$. When working with a_{0eps}, a_{1eps} and a_{2eps} we observed that these parameters are difficult to estimate. When sticking to the Bayesian approach $\sigma^2_{eps}(\tau_i)$ can be sampled by means of the Gibbs sampler when assuming a conjugate (truncated) inverse Gamma prior. In addition by estimating $\sigma^2_{eps}(\tau_i)$ for each maturity separately, we directly observe - with simulated data - for which maturities the variance of the noise terms is difficult to estimate and - for empirical data - how different maturities are affected by noise.

possibly multi-modal structure of the log-likelihood function in multivariate settings.

4.1 Instantaneous Yields in the Vasicek and the CIR Setting

Let us we start with Vasicek [1977] and the Cox *et al.* [1985] (*CIR*) model. For the latter the diffusion term in (13) has to be replaced by $\Sigma\sqrt{X_{n-1}}\Delta^{0.5}\varepsilon_n$. These models are non-linear in the parameters due to the term $\theta^P(1-\nu)$. A standard three step Gibbs sampler can be constructed (see Appendix F), where natural conjugate priors are used for θ^P and Σ^2 which is a normal prior for θ^P, $\theta^P \sim \mathcal{N}(a_{\theta,0}, A_{\theta,0})$, and an inverse gamma prior for Σ^2, $\Sigma^2 \sim \mathcal{IG}(n_0, S_0)$. Since ν should fulfill $\nu \in [0,1]$, we use a uniform prior for this parameter.[9]

We draw 50,000 MCMC samples, including 20,000 burn-in steps, from simulated Ornstein-Uhlenbeck and CIR paths. We set $\theta^P = 3$ and $\Sigma^2 = 1.2^2$ and $\Sigma^2 = 0.7^2$, for the Vasicek [1977] and the Cox *et al.* [1985] model, respectively. $\Delta = 7/365$. ν we set to 0.76, 0.9 and 0.99. The parameters of the priors are $n_0 = 1$, $S_0 = 1$, $a_{\theta,0} = 0$ and $A_{\theta,0} = 1000$. Figure 1 presents representative MCMC output for these two settings (Vasicek [1977] - left sub-figures, Cox *et al.* [1985] - right sub-figures), for different ν, starting with $\nu = 0.76$ in the first row to $\nu = 0.99$ in the third row. For low ν samples are well behaved, with $\nu = 0.9$ this is still the case but the standard deviation of the parameter θ^P starts to increase (take a look on the scale of the horizontal axis). With high serial correlation the sampler produces a "wall". The standard deviation for the Vasicek model is higher than the standard deviation for the CIR model. Nevertheless the standard deviations are very high in both models. This corresponds to our analytical results with the information matrix, where the standard deviation of the parameter θ^P increases drastically when we approach a unit root.

For the model considered above the *conditional densities* of ν and θ^P are the conditional densities used in the Gibbs sampling steps. Based on De Pooter *et al.* [2006] or De Pooter *et al.* [2008] Appendix E derives the *marginal distribution* of θ^P, ν for the Vasicek model. Here we observe that the joint distribution $\pi(\theta^P, \nu|\mathbf{X})$ becomes improper with $\nu = 1$. With ν close to one it becomes almost flat.

[9] Also in the Vasicek, we consider θ^P as a free parameter to demonstrate the impact of near unit root behavior on parameter estimation.

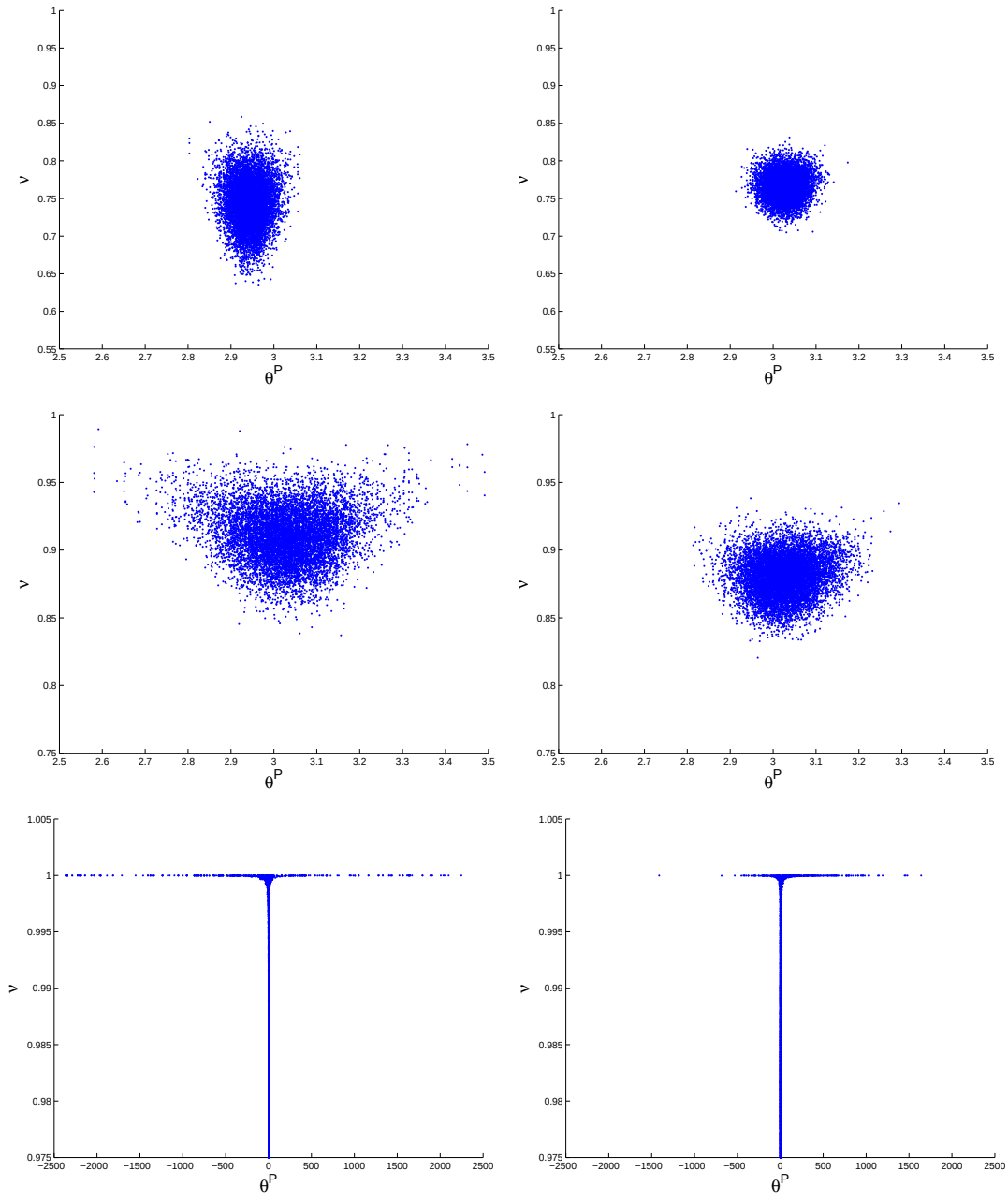

Figure 1: Vasicek and CIR Model: Samples of the joint posterior of θ^P and ν, 50,000 MCMC steps, 20,000 burn-in. Vasicek model left sub-figures, CIR model right sub-figures. First row $\nu = 0.76$ (moderate persistence), second row $\nu = 0.9$; third row $\nu = 0.99$ (near unit root); $\theta^P = 3$. Uniform prior on ν, conjugate normal prior with parameters $a_{\theta,0} = 0$ and $A_{\theta,0} = 1000$ on θ^P. Note that the range of the horizontal axis is different for different ν.

4.2 Near Unit Root Behavior and Sufficiently Informative Priors

In a Bayesian setting we are able to compensate for the term $(1-\nu)^{-1}$ in the priors. De Pooter *et al.* [2006] propose either to put zero measure on ν larger than $1-\varepsilon$ (this approach has been applied in Jones [2003]), construct a Kleibergen and van Dijk [1998] prior or to choose a Schotman and van Dijk [1991] prior. In the third approach, Schotman and van Dijk [1991] use a conjugate normal prior for θ^P with a variance term proportional to $\Sigma^2\frac{1}{(1-\nu)^2}$. The higher the degree of persistence of the stochastic process the lower the prior information on the parameter θ^P. We augment this idea to the m dimensional setting, such that

$$\pi_{SD}(\theta^P|\nu,\Sigma^2) = \mathcal{N}\left(a_{\theta,0}, \tilde{A}_{\theta,0}\cdot((I_m-\nu)^{-1})\Sigma^2((I_m-\nu)^{-1})^\top\right) . \tag{18}$$

The *Schotman and Van Dijk* prior (18) will be used in all further parts of this paper; $\tilde{A}_{\theta,0}$ is set to 1000, $a_{\theta,0}$ remains $0_{m\times 1}$. With (18) the integral of $\pi(\theta^P,\nu|\mathbf{X})$ for the Vasicek model becomes finite. However, we observe in simulation experiments that the wall does not disappear. Therefore, we have to construct a *sufficiently informative prior*.

In Section 3 we observed that the information matrix becomes ill-conditioned, if some eigenvalues of ν were close to one. Therefore, in addition to the prior (18), we use a prior punishing ν with eigenvalues $\boldsymbol{\lambda}_\nu = (\lambda_{\nu 1},\ldots,\lambda_{\nu m})^\top$ close to one. We choose a function $g:\mathbb{R}\to\mathbb{R}$ putting equal probability weight to any $\lambda_{\nu i}\in[\lambda_*,\lambda^*]$, with $0\le\lambda_*\le\lambda^*\le 1$. To the left and to the right of this interval we assign smaller prior probabilities. The degree of punishment will be controlled by the hyper-parameters γ_* and γ^* fulfilling $\gamma_* = \gamma^*\frac{\log(1-\lambda^*)}{\log(\lambda_*)}$. Then with $c=(1-\lambda^*)^{\gamma^*}$ we get

$$g(\lambda_{\nu i}) \;=\; \left(\lambda_*^{\gamma_*}\mathbf{1}_{\lambda_{\nu i}<\lambda_*} + c\mathbf{1}_{\lambda_*\le\lambda_{\nu i}\le\lambda^*} + (1-\lambda_{\nu i})^{\gamma^*}\mathbf{1}_{\lambda_{\nu i}>\lambda^*}\right)\mathbf{1}_{\lambda_{\nu i}\in[0,1]} \text{ and } \pi_{SI}(\nu)\propto\prod_{i=1}^m g(\lambda_{\nu i}) \; . \tag{19}$$

In the limit $\gamma^*\to\infty$, (19) corresponds to a *shrinkage prior*, where no prior mass is put on $\lambda_{\nu i}>\lambda^*$. $\lambda_*^{\gamma_*}=0$ for the remaining part of this article.[10]

[10](i) If ν is lower triangular $\lambda_{\nu i}=\nu_{ii}$. (ii) To sample ν, the Metropolis Hasting algorithm has to be used. Although

Results for the Ornstein-Uhlenbeck and the CIR Setting: We generate $\mathtt{M} = 500$ paths of (X_n). For each $\mathtt{m} = 1, \ldots, \mathtt{M}$ we obtain the estimates $\hat{\Psi}^P_{\mathtt{m}}$ by using MCMC (50,000 MCMC steps and 20,000 burn in-steps). The parameters are $\theta^P = 3$, $\nu = 0.99$ and $\Sigma^2 = 1.2^2$ for the Vasicek and $\Sigma^2 = 0.7^2$ for the CIR setting. The estimates $\hat{\Psi}^P_{\mathtt{m}}$ are given by the sample means from the MCMC steps following the burn-in phase. From these $\mathtt{M}$ estimates we calculated the the mean, the median, the standard deviation SD, the minimum, the maximum, and the 2.5% and 97.5% quantiles. This is done for the Vasicek and the CIR model. Table 1 and 2 present the results from this Monte-Carlo study for the prior (19) and the shrinkage prior.[11] $\gamma^* = 2$ with (19), the different λ^* are shown in Table 1 and 2. We observe that with (19) the variation of the estimates still remains substantial. By increasing γ^* we can obtain more reliable estimates. To avoid time consuming fine tuning, we propose to stick to a shrinking prior with $\lambda^* = 0.995$. This is sufficient for the simulated data where the true ν is known. For the empirical data such a strong assumption seems to fit as well (see also Jones [2003]) when λ^* is "sufficiently larger" than the true ν but "sufficiently smaller than 1". The application of the shrinking prior is not completely free of cost. First, of course, the true parameter has to be within the interval $[0, \lambda^*)$. Additionally, if the parameter $\kappa^P = -\log(\nu)/\Delta$ is of our main interest, we observe that although the impact of the prior on ν seems to be reasonably small, the impact on $\hat{\kappa}^P$ can be quite substantial.

In addition we also estimated ν, θ^P, σ^2 by means of maximum likelihood. Without any restrictions in the optimization routine we obtained results comparable to the results at the beginning of Section 4.1. For most $\mathtt{m}$ we observe that the maximization routine provides us with very small or very large estimates of the parameters (also values larger than $\pm 10^{50}$ are observed), the highest variation is observed with the

already investigated in Hoogerheide *et al.* [2007], it is worth noting that sampling θ^P by means of the Metropolis Hastings needs some tuning if ν is close to one. With Gibbs sampling the variance of the conditional posterior $p(\theta^P|\mathbf{X}, \nu, \Sigma^2)$ becomes automatically large with ν close to one, while in the MH scheme efficient sampling requires that this effect is included in the proposal density. (iii) Alternatively we can also use an *informative normal prior* with $a_{\nu,0}$ equal to the (highest) first order autocorrelation $\widehat{ACF}_1(y_n^{eps}(\tau_i))$, $i = 1, \ldots, k$, and $A_{\nu,0} = \tilde{A}_{\nu,0}\Delta/T$. $\tilde{A}_{\nu,0}$ is set to 5, 10 or 1000, where the Gibbs sampler can also be applied. For a univariate affine term structure model, the first order autocorrelation of the yields fulfills $ACF_1(y_n^{eps}(\tau_i)) = \frac{\nu \mathbb{V}(y_n(\tau_i))}{\mathbb{V}(y_n^{eps}(\tau_i))} \leq \nu$. The less or equal to is caused by market micro-structure noise. For $m \geq 2$ a prior of this kind is a much stronger a-priori assumption on the eigenvalues of ν. This prior will not be applied further in this article. (iv) On the other side $\max\{\widehat{ACF}_1(y_n^{eps}(\tau_i))\}$ can be used as a lower bound for λ^* with the shrinkage prior. Smaller cut-offs should not be used due to the relationship obtained above.

[11] Note that Table 1 and 2 present means of the parameter estimates $\hat{\Psi}^P_{\mathtt{m}}$. Further tables present parameter estimates from one MCMC chain.

parameter θ^P as with MCMC output.

4.3 Instantaneous Yields with an $\mathbb{A}_1(3)$ Setting

We continue to work with the parameters used at the end of Section 3. As regards (3), this specification has the advantage that closed form solutions for $A(\tau)$ and $B(\tau)$ are available.[12] Therefore we avoid problems that might arise with the numerical solution of ordinary differential equations and increase computing speed. Satisfying the Feller condition for the square root component requires $\kappa_{11}^P \theta_1^P \geq \Sigma_{11}^2/2$. For Σ_{ii}^2 we stick to an inverse Gamma prior with parameters n_0 and S_0. This yields:

$$\pi(\Psi^P) \quad \propto \quad \pi_{SI}(\nu) \cdot \pi_{SD}(\theta|\Sigma^2, \nu) \cdot \mathbf{1}_{(\kappa_{11}^P \theta_1^P \geq \Sigma_{11}^2/2)} \cdot \pi(\Sigma^2) \; . \tag{20}$$

By the prior (19), either with γ^* finite or with the shrinkage prior as a limit, we automatically fulfill the restriction required for eigenvalues of ν smaller than one. Performing Bayesian parameter estimation with this model confirms the results obtained with the Vasicek and the CIR model. With a shrinkage prior, where $\lambda^* = 0.995$, we have a prior which is easy to implement with relatively good sampling properties. Therefore, we continue to work with a shrinkage prior on the eigenvalues of ν also in this $\mathbb{A}_1(3)$ model.

5 Yields observed with Common Micro-Structure Noise

We proceed with the $\mathbb{A}_1(3)$ setting already investigated in the Sections 3 and 4.3. To perform Bayesian parameter estimation we augment the set of parameters (see Tanner and Wong [1987]) by the latent process $\mathbf{X}$. While the density of $X_1, \ldots, X_N$ is determined by the model assumptions (see (10)), we have to specify the prior $\pi(X_0, \Psi)$. In addition, we have to specify the priors for κ^Q, θ^Q, $\boldsymbol{\sigma}_{eps}^2$ and δ_0. For the diagonal components of κ^Q, X_{01} and θ^Q we use a gamma prior with parameters $n_{0Q} = 1$ and $S_{0Q} = 1$, while for δ_0, $X_{0,2}$ and $X_{0,3}$ - all living on $\mathbb{R}$ - we use a normal prior with mean parameter zero and variance 1000. Since $\kappa_{32}^P = \kappa_{32}^Q$ was assumed, the prior for this parameter is already specified. To derive

[12] Here the *Mathematica* package has been used.

a stationary process under $\mathbb{Q}$ the Feller condition for the square root component requires $\kappa_{11}^Q \theta_1^Q \geq \Sigma_{11}^2/2$.

To derive a $\mathbf{B}$ of rank m, we apply the prior $\pi(det(\mathbf{B}^\top \mathbf{B}))$ on the determinant of $(\mathbf{B}^\top \mathbf{B})$. Here we assume that $\pi(det(\mathbf{B}^\top \mathbf{B})) \propto |det(\mathbf{B}^\top \mathbf{B})|^{d_1} \mathbf{1}_{|det(\mathbf{B}^\top \mathbf{B})|<d_0}$; where d_1 was set to 1 and $d_0 = 10^{-10}$. For the data considered in this article we observe that the impact of this prior can be neglected.

For the micro structure noise parameters $\pi(\boldsymbol{\sigma}_{eps}^2) = \prod_{i=k} \pi(\sigma_{eps}^2(\tau_i))$ is assumed. $\pi(\sigma_{eps}^2(\tau_i))$ is truncated inverse gamma $\mathcal{IG}_T(n_{0eps}, S_{0eps})$ with $n_{0eps} = 1$ and $S_{0eps} = 1$. The truncation is such that $\pi(\sigma_{eps}^2(\tau_i)) > 0$ for $0 \leq \sigma_{eps}^2(\tau_i) \leq \mathbb{V}(y_{eps}(\tau_i))$. $\mathbb{V}(y_{eps}(\tau_i))$ can be estimated from prior data, or if not available – by being less clean – from the actual data. This truncation was necessary to improve the properties of the Bayesian sampler. In more details: For the updates of the latent process $\mathbf{X}$ we mix between random walks proposals and proposals based on running the Kalman filter as introduced in Frühwirth-Schnatter and Geyer [1996]. For the second opportunity to work $\boldsymbol{\sigma}_{eps}^2$ should not be too large. If the sampler is started with some $\mathbf{X}$ not sufficiently close to the true $\mathbf{X}$, the sampler generates $\boldsymbol{\sigma}_{eps}^2$ much larger than the true $\boldsymbol{\sigma}_{eps}^2$. A lot of these samples are even larger than $\mathbb{V}(y_{eps}(\tau_i))$, while $\sigma_{eps}^2(\tau_i) < \mathbb{V}(y_{eps}(\tau_i))$ by the model assumptions. Without a-priori restrictions on $\sigma_{eps}^2(\tau_i)$ proposing from the Kalman filter turned out to be inefficient. This is the reason why we impose a truncated inverse gamma prior on the variance of the noise terms. For the sampling of Ψ see Appendix F.

The parameters are sampled by means of a MCMC sampler. We set $\lambda^* = 0.995$ and work with a shrinkage prior; working with prior (19), $\gamma^* = 2$ and $\lambda^* = 0.99$ does not improve the estimation results. Table 3 presents typical MCMC output for simulated data. Starting the sampler at different initial values results in very similar estimates. As already observed with instantaneous yields the estimates of ν are close to their true parameter values, while the non-linear transformations κ^P show an upward bias. A modest upward bias is also observed for most of the estimates of the volatility terms Σ^2. For the estimates of θ^Q quite a large variation is observed. Finally we have to point out that the parameters $\sigma_{eps}^2(\tau_i)$ are difficult to estimate for the smaller maturities, the upward bias can be substantial. We try to explain why the noise for smaller maturities is so difficult as follows: When we consider (7) we observe that for smaller maturities the absolute values of the elements of $B(\tau_i)$ are larger while $|A(\tau_i)|$ is small and vice

versa. That is to say the larger τ_i the smaller the impact of $B(\tau_i)X_n$; i.e. the variance of the model yields decreases. For the small maturities very precise estimates of $\mathbf{X}$ are necessary to obtain precise estimates of $\sigma^2_{eps}(\tau_i)$. However, when starting the sampler with the true parameters and the true $\mathbf{X}$ it is not very difficult to estimate the noise parameters for the different maturities.

6 Parameter Estimation with Empirical Data

This section applies the econometric tools developed in the former sections to empirical term structure data. From the Federal Reserve (http://federalreserve.gov/releases/h15/data.htm) we downloaded yields for the time span March 8, 2003 to June 26, 2009. A full panel of maturities from one month to thirty years is available for these periods. Since the thirty year maturity time series exhibits a lot of missing values this maturity has been excluded. This gives $\boldsymbol{\tau} = \{1/12, 1/4, 1/2, 1, 2, 3, 5, 7, 10, 20\}$, $k = 10$ and $N = 413$ observations per yield time series. These discrete time yields were also translated to continuous compounding. Although this H-15 data set can only be seen as a proxy for the risk-free term structure, we also use it since it is often used in recent literature (e.g. Chib and Ergashev [2009]). Standard tests on a unit root only reject the zero of a unit root for the long maturity.

In addition we derived a risk-free term structure data from USD LIBOR (maturities of 1, 3, 6, 9 and 12 months from Bloomberg) and USD swap rates (middle rates, for maturities 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25 and 30 years from Datastream). Similar to Filipović [2009][Chapter 2] we derived continuously compounded spot rates by means of "bootstrapping". Here we worked with $k = 11$ maturities, $\tau = \{1/12, 1/4, 1/2, 1, 2, 5, 7, 10, 15, 20, 30\}$, and $N = 500$ observations. The time span considered was July 1, 2002 to June 2004.

Tables 4 and 5 present the parameter estimates for the empirical data. We want to point out that some differences in the parameter estimates can be observed when comparing the estimates for the two data sets. With both data sets the estimates of the parameters ν_{ii} are all larger than 0.95, the standard deviations are low as observed with simulated data. We also know that the impacts of small changes in ν have relatively

large impacts in κ^P. In $\hat{\kappa}^P$, especially in $\hat{\kappa}^P_{32}$ and $\hat{\kappa}^P_{33}$ differences are observed. While $\hat{\kappa}^Q_{11}$ and $\hat{\kappa}^Q_{22}$ are quite similar we have once again differences in the third term, a similar effect arises with the estimates $\hat{\Sigma}^2$. The estimates $\hat{\delta}_0$ and $\hat{\theta}^Q$ are different. We also have to point out that due to higher variance of the H-15 yields (range between 0.24 for the highest maturity to 2.5 for the short term maturities) compared to the European yield data (range 0.057 to 0.19, highest variance for medium maturities) the estimates of some Σ^2 should be higher; in our estimates this is the case with $\hat{\Sigma}^2_{33}$. In addition the estimates of the micro-structure noise parameters $\hat{\sigma}^2_{eps}(\tau_i)$ are higher for the H-15 data set. Based on our estimates of the noise terms $\hat{\sigma}^2_{eps}(\tau_i)$, $\frac{\hat{\sigma}^2_{eps}(\tau_i)}{\hat{\mathbb{V}}(y_{eps}(\tau_i))}$ estimates the proportion of the micro-structure noise in terms of the variance of the yields observed. For the European yield data we observe that between 32% and 99% of the variance is due to market-micro structure noise, while for the H-15 data the numbers vary between 12% and 98%. This impact can be considered to be substantial. The very high percentages are observed with the longest maturites where the impact of $\mathbf{X}$ on the yields becomes very small and the model yields are mainly determined by $\mathbf{A}$. This explains why almost all the variation with the large maturites is considered to be micro-structure noise. In addition we have to remark that based on our expierence with similuated data, especially the estimates for $\sigma^2_{eps}(\tau_i)$ for the shorter maturities have to be interpreted with care. Last but not least the inefficiency factors are high but also remain in the range reported in Chib and Ergashev [2009].

Finance literature often compares the parameter estimates of κ^P to κ^Q and θ^P to θ^Q to infer risk premia. Suppose that we stick to the following rule of thumb: parameters are said to be significantly different if the intervals $[\hat{\kappa}^Q_{ii} \pm sd(\kappa^Q_{ii})]$ and $[\hat{\kappa}^P_{ii} \pm sd(\kappa^P_{ii})]$ do not overlap. Based on this rule only for $\hat{\kappa}^P_{33}$ in the H-15 data set a significant effect can be observed, for all other mean reversion parameters no significant risk premium can be observed. In contrast to the mean reversion parameters $\hat{\theta}^Q$ is significantly larger than $\hat{\theta}^P$ in both data sets. However, we once again have to point out that the standard deviations of the samples of θ^P strongly depend on the prior used. Since we already know that the standard deviation of the estimates of θ^P are strongly influenced by the choice of the prior, the results of the above comparison should be handled with care.

7 Conclusions

In this article we investigated the impacts arising from near unit root behavior on parameter estimation with affine term structure models. We showed that the information matrix approaches a singularity when serial correlation increases. To cope with this problem in a Bayesian framework, we constructed priors regularizing the marginal distribution and allowing for stable parameter estimation. More precisely, we applied a multivariate version of the Schotman and van Dijk [1991] prior to the level parameters. Since this is not sufficient to get reliable parameter estimates, an informative prior punishing parameter values where weak identification occurs is compared to a more simpler shrinkage prior. Due to its simplicity and the fact that the more complicated prior does not really improve the estimation results this article recommends to work with a shrinkage prior on the mean reversion parameters, which is in line with Jones [2003]. By means for this prior, eigenvalues of this matrix close to one have zero prior probability mass. That is to say sufficiently strong priors are necessary to get reliable parameter estimates.

This article provides also important insights for a finance audience. The first point is that the level parameter of the risk-free term structure is difficult to estimate due to a high degree of serial correlation. This has important implications: When using affine term structure models, this implies that this parameter can only be estimated with a low precision. Second interpreting differences in the level parameters as risk premia, should also be handled with care.

Last but not least we have to raise the question why affine term structure models have become so popular although there are so many problems from an econometric point of view. Regarding this issue, affine term structure models provide a mathematically elegant and consistent way to describe the whole term structure by a parsimonious model. The principle of no-arbitrage is fulfilled for all yields. In addition this class of models offers a natural way to include other sources of risk such as credit and liquidity risk, and can therefore be used for bond, corporate default swap and option pricing issues [among a plenty of literature see e.g. Lando, 1998; Duffie and Singleton, 1999; Driessen, 2005; Feldhütter and Lando, 2008; Pan and Singleton, 2008]. Thus, we conclude that if we continue to stick to this class of models, we have

to be careful as regards parameter estimation.

A The Canonical Representation of an $\mathbb{A}_l(m)$ Model

Given the notation of Section 2 an affine stochastic process is defined as follows:

Definition 2 (Affine Process)**.** $(Y(t))$ follows an affine stochastic process $dY(t) = \beta(Y(t))dt + \varrho(Y(t))dW(t)$ if the (positive definite) diffusion matrix $\varrho(Y(t))\varrho(Y(t))^\top$ and the drift term $\beta(Y(t))$ are affine functions in $Y(t)$ (see Filipović [2009][Definition 10.1 and Theorem 10.1]).

Diagonal Diffusion Term: T_A is called an *affine transformation* if $T_A Y(t) = LY(t) + \delta$. L is a $m \times m$ non-singular matrix and δ is a vector of dimension m (see Dai and Singleton [2000][Appendix A]). Equipped with T_A we get:

Definition 3 (Affine Process with Diagonal Diffusion Term)**.** An affine stochastic process is said to have a diagonal diffusion term if there exists an affine transformation $T_A Y$, such that $\tilde{\beta}(X(t))$ is affine in $X(t)$ and $[\varrho(X(t))\varrho(X(t))^\top]_{ii} = \tilde{a}_i + \tilde{b}_i^\top X(t)$ while $[\varrho(X(t))\varrho(X(t))^\top]_{ij} = 0$, for $i, j = 1, \ldots, m$; $\tilde{a}_i > 0$ and $\tilde{b}_i$ is a vector of dimension m. See Cheridito *et al.* [2008] and Dai and Singleton [2000].

By considering (1) and Dai and Singleton [2000], the process $(X(t))$ can be transformed by means of T_A such that Σ becomes diagonal. That is to say (1) is a (maybe transformed) representation of an affine process with diagonal diffusion matrix.[13]

Dai and Singleton [2000]-*canonical representation*: Let us partition the matrices κ^Q and $\mathcal{B}$ as follows:

$$\kappa^Q = \begin{pmatrix} \kappa^{BB}_{l\times l} & \kappa^{BD}_{l\times(m-l)} \\ \kappa^{DB}_{(m-l)\times l} & \kappa^{DD}_{(m-l)\times(m-l)} \end{pmatrix} \; ; \quad \mathcal{B} = \begin{pmatrix} I_{l\times l} & {\mathcal{B}^{BD}}_{l\times(m-l)} \\ \mathbf{0}_{(m-l)\times l} & \mathbf{0}_{(m-l)\times(m-l)} \end{pmatrix} . \tag{21}$$

θ is partitioned into θ^B and θ^D, where the first term is of dimension l while the second is of dimension $m - l$. The same slip-up has already been applied to $X(t)$. The diffusion matrix is diagonal, such that

[13] Regarding the question, whether any affine process can be transformed to the structure given by (1), Cheridito *et al.* [2008] have shown that this need not be the case. Cheridito *et al.* [2008][Theorem 2.1] provide a condition when such a transformation of the affine model in Definition 2 to the structure given by (1) is possible; counterexamples when their condition is not met are provided as well. For $m \leq 3$ such a transformation exists. Even if higher dimensional processes are used (e.g. Duffee [2011] with term structure data, when credit risk is added as in Feldhütter and Lando [2008] or CDS spreads are priced Schneider *et al.* [2010]), models with diagonal/diagonalizable diffusion matrix are mainly applied in financial models.

$\Sigma\sqrt{S(X(t))}(\Sigma\sqrt{S(X(t))})^\top = \Sigma^2\sqrt{S(X(t))}$. Dai and Singleton [2000] have demonstrated that under the following conditions the model is admissible and identified.

Definition 4 (Dai and Singleton [2000]-*canonical representation* of an $\mathbb{A}_l(m)$ Model)**.** Consider (1) with diagonal diffusion matrix. Admissibility and identification require: (i.a) If $l > 0$ then κ^Q is of the structure (21). First, $\kappa^{BD} = \mathbf{0}_{l\times m-l}$, where $\mathbf{0}_{l\times m-l}$ is a $l \times m-l$ matrix of zeros. $\kappa_{ij} \leq 0$ for $1 \leq j \leq l$ and $i \neq j$ and $\sum_{j=1}^{l} \kappa_{ij}\theta_j > 0$ for $i = 1, \ldots, l$ (which specifies the κ^{BB} and κ^{DB} blocks). (i.b) If $l = 0$, then κ^Q is a lower triangular matrix. (ii) θ_i^Q satisfies $\theta_i^Q \geq 0$ for $i = 1, \ldots, l$. $\theta_i^Q = 0$ for $i = l+1, \ldots, m$. δ_0 and θ_i, $i = 1, \ldots, l$ are free. δ is a free parameter, with $\delta_i \geq 0$ for $i > l+1$. (iii) Regarding $\mathcal{B}$, $I_{l\times l}$ is the identity of dimension l. The elements of the submatrix $\mathcal{B}^{BD}$ fulfill $\mathcal{B}_{ij} \geq 0$ for $1 \leq i \leq l$ and $l+1 \leq j \leq m$. $\mathcal{B}_{ii} = 1$ for $i = 1, \ldots, l$. This results in the matrix $\mathcal{B}$ as described in (21). $\alpha_i = 0$ for $i = 1, \ldots, l$ and $\alpha_i = 1$ for $i = l+1, \ldots, m$. (iv) The elements of the main diagonal of Σ are equal to 1; $\Sigma_{ij} = 0$ for all $i, j = 1, \ldots, m$, $i \neq j$, by the assumption of a diagonal diffusion matrix.

B The Minimal Model

Let us consider the state space model (7):

$$\begin{aligned} \mathbf{y}_n^{eps} &= \mathbf{A} - \mathbf{B}X_n + \sqrt{\boldsymbol{\sigma}_{eps}^2}\mathbf{e_n} \\ X_n &= \nu X_{n-1} + (I_m - \nu)\theta^P + \Sigma\sqrt{S_{n-1}}\sqrt{\Delta}\varepsilon_n \ , \end{aligned}$$

with the non-singular $k \times k$ matrix $\boldsymbol{\sigma}_{eps}^2$; where $0 < \sigma_{eps}^2(\tau_i) < \infty$. In addition we assume that the $k \times m$ matrix $\mathbf{B}$ has rank m, which is the dimension of X_n.

Definition 5 (Minimal Dimension)**.** A state space model is called *controllable* if for any two vectors x_a and $x_b \in \mathbb{R}^m$, there exists an integer v and noise terms ε_n such that $X_v = x_b$ if $X_0 = x_a$. A state space model is called *observable* if and only if X_0 is completely determined by $\mathbf{y}_n$, $n \geq 0$, given $\mathbf{e_n} = 0$. The model is called minimal if it is controllable and observable. (see Brockwell and Davis [2006][Chapter 12.4]).

Regrading controllability, we get $X_0 = x_a$, $X_1 = \nu x_a + (I_m - \nu)\theta^P + H_1\varepsilon_1$, with $H_1 = \Sigma\sqrt{S_{n-1}(x_a)}\sqrt{\Delta}$, $X_2 = \nu^2 x_a + (I_m + \nu)(I_m - \nu)\theta^P + \nu H_1\varepsilon_1 + H_2\varepsilon_2$, $\ldots$, $X_v = \nu^v x_a + (I_m + \nu)\theta^P \sum_{i=0}^{v-1}\nu^i + \sum_{i=0}^{v-1}\nu^i H_{i+1}\varepsilon_{i+1}$; where $H_i(x_{i-1}) = \Sigma\sqrt{S_{i-1}(x_{i-1})}\sqrt{\Delta}$. Then x_b can be derived from x_a if $C_v = (H_v, \nu H_{v-1}, \nu^2 H_{v-2}, \ldots, \nu^{v-1}H_1)$ has rank m, which is the dimension of X_n (see Brockwell and Davis [2006][p. 490]). H_i is diagonal, positive definite and of full rank by the model assumptions. ν is of full-rank by the stationarity assumption. Therefore all the terms in C have rank m, such that C has rank m. That is to say the model in controllable. Observability follows directly from Brockwell and Davis [2006][Theorem 12.4.4]: Consider the $m \times jk$ matrix $O_j = (\mathbf{B}^\top, \nu^\top\mathbf{B}^\top, \ldots, \nu^{\top,j-1}\mathbf{B}^\top)$. If O_m has rank m, then the system is observable. In our case O_m is of rank m since $\mathbf{B}$ has rank m by Assumption 3. This can be summarized as follows:

Lemma 1. Suppose that Assumptions 1-3 hold, then the system (7) is of minimal dimension.

Example 1 already demonstrated that we can replace the processes X_{1n} and X_{2n} by $\tilde{X}_{1n} = X_{1n} + X_{2n}$ and $\tilde{X}_{2n} = 0$ and get the same term structure. In terms of this section we get:

Example 4 (Counterexample)**.** Consider a two factor Vasicek model investigated in Example 1. Since $\kappa^Q_{11} = \kappa^Q_{22}$ we get $B_1(\tau_i) = B_2(\tau_i) = \frac{1}{\kappa^Q\tau_i}(1 - \exp(-\kappa^Q\tau_i))$ for all τ_i. In this case $\mathbf{B}$ has rank 1. ν is diagonal with elements $\nu_{ii} = \exp(-\kappa^P_{ii}\Delta)$. The elements of O_m are given by

$$\nu^j\mathbf{B}^\top = \begin{pmatrix} \nu^j_{11} & 0 \\ 0 & \nu^j_{22} \end{pmatrix}\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{21} & \cdots & \mathbf{B}_{k1} \\ \mathbf{B}_{12} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{k2} \end{pmatrix} = \begin{pmatrix} \nu^j_{11}\mathbf{B}_{11} & \nu^j_{11}\mathbf{B}_{21} & \cdots & \nu^j_{11}\mathbf{B}_{k1} \\ \nu^j_{22}\mathbf{B}_{12} & \nu^j_{22}\mathbf{B}_{22} & \cdots & \nu^j_{22}\mathbf{B}_{k2} \end{pmatrix}$$

such that if $\nu^{\top,j}\mathbf{B}^\top$ is a linear combination of the rows of $\mathbf{B}$. This results in $rank(O_m) = 1$.

What remains to discuss is when $\mathbf{B}$ has full rank: For the general systems of ODEs (4) we still assume that $\mathbf{B}$ has full rank. For independent Gaussian terms different κ^Q_{ii} are sufficient.

C A Note on Particular Noise

Let us now assume that m maturities are observed without noise, that is to say in contrast to Assumption 2, $e_{in} = 0$ for m maturities. For these m yields we arrive at the log-likelihood

$$\ell(\Psi; \mathbf{y}^{fix,eps}) = -N \log |det \mathbf{B}^{fix}| + \ell(\Psi^P; \mathbf{X}) \ , \tag{22}$$

while for the remaining $k-m$ yields observed with noise we get $\ell(\Psi; \mathbf{y}^{eps,nf}, \mathbf{X})$. $\mathbf{B}^{fix}$ stands for the sub-matrix of $\mathbf{B}$ corresponding to the maturities observed without noise. $\mathbf{y}^{fix,eps} \in \mathbb{R}^m$ and $\mathbf{y}^{eps,nf} \in \mathbb{R}^{k-m}$ are the yields observed without and with noise. In the one dimensional case, where $|det\mathbf{B}^{fix}| = |\mathbf{B}^{fix}|$ no problems arise if κ^Q is sufficiently larger than zero. With $m > 1$ a further important problem arises. Although, the matrix $\mathbf{B}$ has to be of full rank by the minimality requirement, the matrix $\mathbf{B}^{fix}$ can be ill conditioned. The fraction of the largest over the smallest eigenvalue of $\mathbf{B}^{fix}$ can become quite large if the rows of $\mathbf{B}^{fix}$ are close to colinearity. If this is the case $|det\mathbf{B}^{fix}|$ becomes a dominating term in the likelihood (22). Due to the high condition number, the impact of a small change in some component of Ψ^Q can be tremendous. On the other had with common noise $\mathbf{B}$ is only used to transform X_n into model yields. Since financial applications favor multi-factor term structure models, this analysis provides us with the important insight that the assumption of particular noise should not be applied. This problem goes back to a standard problem of numerical linear algebra.[14] In other words, with particular noise $X_n = (\mathbf{B}^{fix})^{-1}(\mathbf{y}_n^{fix,eps} - \mathbf{A}^{fix})$. Small changes in $\mathbf{B}^{fix}$ result in large changes in its inverse. X_n is strongly affected by small changes in Ψ^Q. The following examples should shed some light on this problem:

Example 5. Consider a two factor Vasicek model with independent factors. Here, with $\tilde{B}_j(\tau_i) = \frac{1}{\kappa^Q_{jj}\tau_i}(1-$

[14] In a former version of this paper we estimated a Vasicek model with no market price of risk. This setting can be transformed to the structure of the instantaneous yield model. After applying the priors used in Section 4, we observed good sampling behavior with particular micro-structure noise for the Vasicek model with zero market price of risk. When insisting on particular noise with $\mathbb{A}_1(3)$ models, the likelihood approximations of Aït-Sahalia and Kimmel [2009] can be used.

$\exp(-\kappa_{jj}^Q \tau))$, we get

$$\mathbf{B}^{fix} = \begin{pmatrix} \tilde{B}_1(\tau_1) & \tilde{B}_2(\tau_1) \\ \tilde{B}_1(\tau_2) & \tilde{B}_2(\tau_2) \end{pmatrix}.$$

The eigenvalues are $\frac{1}{2}(\tilde{B}_1(\tau_1)+\tilde{B}_2(\tau_2)) \pm \sqrt{(\tilde{B}_1(\tau_1)+\tilde{B}_2(\tau_2))^2/4 - \left[\tilde{B}_1(\tau_1)\tilde{B}_2(\tau_2) - \tilde{B}_1(\tau_2)\tilde{B}_2(\tau_1)\right]}$. If the determinant $\tilde{B}_1(\tau_1)\tilde{B}_2(\tau_2) - \tilde{B}_1(\tau_2)\tilde{B}_2(\tau_1) = 0$, the eigenvalues are $\tilde{B}_1(\tau_1)\tilde{B}_2(\tau_2)$ and 0, such that the condition number goes to infinity. A singular $\mathbf{B}^{fix}$ is derived with $\tau_1 = \tau_2$ or $\kappa_{11}^Q = \kappa_{22}^Q$. Therefore also for $\tau_1 \approx \tau_2$ and $\kappa_{11}^Q \approx \kappa_{22}^Q$ the condition numbers can remain large.

Example 6. Consider the $\mathbb{A}_1(3)$ model investigated in Section 5. Assume $\tau^{fix} = \{2, 5, 10\}$, then the eigenvalues of $\mathbf{B}^{fix}$ are 3.8607, -0.0958 and 0.0010, such that the fraction of largest over the smallest eigenvalue in absolute terms is 3697.9. When using different τ^{fix} the smallest fraction of eigenvalues still remains above 1000.

D The Information Matrix

D.1 Information Matrix for the Instantaneous Process of an $\mathbb{A}_l(m)$ Setting

Let us consider (6): For a stationary (X_n) the covariance matrix is $Cov(X_n) = \mathbb{E}(X_n - \theta)^2 = \mathbb{E}(X_n X_n^\top) - \theta^P \theta^{P,\top}$. This matrix is positive definite. By the Feller condition we get $(X_n^B) > 0$ (a.s.). This results in $S(X_{n-1}) > 0$ by the assumptions of Section 2. For the following analysis it is sufficient to know that the $m \times m$ matrix $Cov(X_n)$ is positive definite.[15] It can be derived in closed form - up to an evaluation of a matrix exponential - from Cuchiero *et al.* [2010]. Alternatively, if $S(X_{n-1})$ is constant or approximated by $S(\theta^P)$, we get from Hamilton [1994][p. 265]: $vec(Cov(X_n) = (I_{(m^m)} - \mathcal{C})^{-1} vec(\Sigma^2 S_{n-1}(\theta^P)\Delta)$ where $\mathcal{C} = (\nu \bigotimes \nu)$. $\bigotimes$ stands for the Kronecker product.

Let us consider the log-likelihood (10). ν' are the non-restricted parameters of the $m \times m$ matrix ν. $0 \leq m'_\nu \leq m^2$ is the number of free parameters in ν. By means of matrix calculus (see e.g. Poirier

[15] Existence of $2k$ moments is treated in Filipović [2009][Chapter 10, Lemma 10.7], for general stationarity conditions the reader is referred to Kim and Glasserman [2008].

[1995][Appendix B]) we get:

$$\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial vec(\nu')vec(\nu')^\top} = -\sum_{n=1}^{N}\left[\frac{\partial(X_n-\nu X_{n-1}-(I_m-\nu)\theta^P)}{\partial vec(\nu')}\right]^\top \tilde{\Sigma}_n^{-1}\left[\frac{\partial(X_n-\nu X_{n-1}-(I_m-\nu)\theta^P)}{\partial vec(\nu')}\right] . \quad (23)$$

The matrix $\frac{\partial}{\partial vec(\nu')}(X_n-\nu X_{n-1}-(I_m-\nu)\theta^P)$ is of dimension $m\times m'_\nu$. The elements of this matrix are given by $\mathcal{P}_{\nu'}[(-X_{1n}+\theta_1^P),(-X_{2n}+\theta_2^P),\ldots,(-X_{mn}+\theta_m^P)]$, where $\mathcal{P}_{\nu'}$ projects on the m' columns of the $m\times m^2$ matrix $[(-X_{1n}+\theta_1^P),\ldots,(-X_{mn}+\theta_m^P)]$.[16]

Using the diagonal structure of the diffusion matrix we get

$$\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \nu_{ij}\partial \nu_{vw}} = -\sum_{n=1}^{N}\mathbf{1}_{(i=v)}\frac{(-X_{jn}+\theta_j^P)\mathbf{1}_{\nu_{ij}\neq 0}(-X_{wn}+\theta_w^P)\mathbf{1}_{\nu_{vw}\neq 0}}{\Sigma_{ii}^2 S_{ii}(X_{n-1})\Delta} , \quad (24)$$

for $i,j,v,w=1,\ldots,m$. In (24) X_n enters into the numerator while X_{n-1} enters in the denominator. Consider this fraction as a function $g(X_n,X_{n-1})$ such that we can approximate the expectation of the fractions (24) by the first order approximations (see Paolella [2007][Chapter 2.3]). This yields

$$\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \nu_{ij}\partial \nu_{vw}}\right) \approx N\cdot\frac{[Cov(X_n)]_{jw}}{\Sigma_{ii}^2 S_{ii}(\theta^P)\Delta}\mathbf{1}_{\nu_{ij}\neq 0}\mathbf{1}_{\nu_{vw}\neq 0}\mathbf{1}_{(i=v)} . \quad (25)$$

In $S_{ii}(\theta^P)$ the parameter θ^P is plugged in for X_{n-1}. $[Cov(X_n)]_{jw}$ is the element (j,w) of the covariance matrix. In the same way as described above we can derive all blocks of the proxy of the expected values of the $m'_\nu\times m'_\nu$ Hessian. Since $Cov(X_n)$ is positive definite, $\mathbb{E}\left(-\frac{\partial^2\ell(\Psi^P;\mathbf{X})}{\partial\nu\partial\nu^\top}\right)$ has to be positive definite. If only a $m'_{\nu_i}\times m'_{\nu_i}$ submatrix of the $m\times m$ is considered, this submatrix has to be positive definite by the principal minors criterion.

Based on (25) $\mathbb{E}\left(-\frac{\partial^2\ell(\Psi^P;\mathbf{X})}{\partial vec(\nu')vec(\nu')^\top}\right)$ becomes close to a singularity if $\Sigma^2 S_{ii,n-1}(\theta^P)\Delta$ becomes large. Alternatively the determinant of the covariance matrix can be small as well. From estimates reported in literature neither the former nor the latter case can be expected.

Next we consider the parameter θ^P, with the non-restricted components $\theta^{P'}$, its number is m'_θ. By

[16]The symbol $\mathcal{P}_\cdot$ is used as a projection device either to project on the non-restricted elements of a vector or to project on the elements of a matrix which are non-zero due to restrictions on the parameters.

matrix calculus we get

$$\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial\theta^{P\prime}(\theta^{P\prime})^\top} \;=\; -\sum_{n=1}^{N} \mathcal{P}_{\theta^{P\prime}} \left[(I_m-\nu)^\top(\Sigma^2 S_{n-1}\Delta)^{-1}(I_m-\nu)\right] \;. \tag{26}$$

Assume that all elements are free parameters, then the expectation of (26) becomes

$$N\cdot(I_m-\nu)^\top(\Sigma^2)^{-1}\Delta\mathbb{E}(S_{n-1}^{-1})(I_m-\nu)\;. \tag{27}$$

By the diagonal structure, $(\Sigma^2)^{-1}$ is given by its reciprocals. $\mathbb{E}(S_{n-1}^{-1})$ can be approximated by a (first order) approximation of the expectation of $1/S_{ii,n-1}$ (e.g. $1/S_{ii,n-1} \approx 1/S_{ii}(\theta^P)$). $\bar{\Sigma}^{-1}$ is a proxy of $(\Sigma^2)^{-1}\Delta\mathbb{E}(S_{n-1}^{-1})$. Here it is sufficient to know that $\mathbb{E}(S_{n-1}^{-1}) > 0$, which is implied by $S_{n-1} > 0$ (a.s.). Assume that all parameters of θ are free. To get a regular matrix, the quadratic form $(I_m - \nu)^\top\bar{\Sigma}^{-1}(I_m-\nu)$ has to be positive definite which is the case if its determinant is larger than zero. By the properties of determinants we get $det\left[(I_m-\nu)^\top\bar{\Sigma}^{-1}(I_m-\nu)\right] = det\left[(I_m-\nu)\right] det\left[\bar{\Sigma}^{-1}\right] det\left[(I_m-\nu)\right]$. Since $det\left[\bar{\Sigma}^{-1}\right] > 0$, (27) is positive definite if $det\left[(I_m-\nu)\right] > 0$. This condition is satisfied if the eigenvalues of $I_m - \nu$ are strictly positive, which is the case if the eigenvalues of ν are smaller than one. These eigenvalues have been abbreviated by $\boldsymbol{\lambda}_\nu = (\lambda_{\nu 1},\ldots,\lambda_{\nu m})^\top$. If some parameters of θ^P are restricted, then we get the corresponding part of the information matrix by means of

$$\mathbb{E}\left(-\frac{\partial^2\ell(\Psi^P;\mathbf{X})}{\partial\theta^{P\prime}(\theta^{P\prime})^\top}\right) = N\cdot\mathcal{P}_{\theta^{P\prime}}\left[(I_m-\nu)^\top(\Sigma^2)^{-1}\Delta\mathbb{E}(S_{n-1}^{-1})(I_m-\nu)\right] \;. \tag{28}$$

If one eigenvalue of ν is equal to one, this does not automatically imply that (28) is singular. The matrix is regular if the projection on the $\mathcal{P}_{\theta^{P\prime}}$ on the rows of $(I_m-\nu)$ has rank m'_θ. Since for an arbitrary matrix A, $rank(A) = rank(A^\top) = rank(AA^\top) = rank(A^\top A)$, the rank of $\mathcal{P}_{\theta^{P\prime}}((I_m-\nu)(I_m-\nu)^\top)$ has to be less or equal to m'_θ. The projection is a submatrix of $(I_m-\nu)(\Sigma^2)^{-1}\Delta\mathbb{E}(S_{n-1}^{-1})(I_m-\nu)^\top$. Full rank of $(I_m-\nu)(\Sigma^2)^{-1}\Delta\mathbb{E}(S_{n-1}^{-1})(I_m-\nu)^\top$ implies a rank of m'_θ for this submatrix.

For the matrix Σ^2 we get:

$$\frac{\partial^2 \ell(\Psi^P; \mathbf{X})}{\partial \Sigma_{ii}^2} = -\sum_{n=1}^{N} \frac{\frac{1}{2}\Sigma_{ii}^2 - \frac{\zeta_i^2}{S_{ii,n-1}\Delta}}{(\Sigma_{ii}^2)^3} \text{ with } \zeta_i = X_{in} - \sum_j \nu_{ij} X_{i,n-1} = \Sigma_{ii}^2 S_{ii,n-1} \Delta \varepsilon_i \ . \tag{29}$$

All the ij terms are zero be the diagonal structure of Σ^2. (29) can be transformed to

$$\frac{\partial^2 \ell(\Psi^P; \mathbf{X})}{\partial \Sigma_{ii}^2} = -\sum_{n=1}^{N} \frac{\frac{1}{2}\Sigma_{ii}^2 - \frac{\Sigma_{ii}^2 S \Delta \varepsilon_i}{S_{ii,n-1}\Delta}}{(\Sigma_{ii}^2)^3} = -\sum_{n=1}^{N} \frac{\frac{1}{2} - \varepsilon_i}{(\Sigma_{ii}^2)^2} \ . \tag{30}$$

Since X_{n-1} and ε_i are independent, taking expectations yields:

$$\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^P; \mathbf{X})}{\partial \Sigma_{ii}^2}\right) = \frac{N}{2(\Sigma_{ii}^2)^2} \ . \tag{31}$$

If also some diagonal elements would be fixed to some positive values a-priori then we could proceed as with projections as applied to the other parameters. Last but not least, if some $\mathcal{B}_{ij}$ are free parameters, we get

$$\begin{aligned} \frac{\partial^2 \ell(\Psi^P; \mathbf{X})}{\partial \mathcal{B}_{ij} \partial \mathcal{B}_{vw}} &= -\sum_{n=1}^{N} X_{j,n-1} X_{w,n-1} \frac{-\frac{1}{2}\Sigma_{ii}^2 S_{ii,n-1}\Delta + \zeta_i^2}{\Sigma^2 \Delta S_{ii,n-1}^3} \\ &= -\sum_{n=1}^{N} X_{j,n-1} X_{w,n-1} \frac{-\frac{1}{2}\Sigma_{ii}^2 S_{ii,n-1}\Delta + \Sigma_{ii}^2 S_{ii,n-1}\Delta \varepsilon_{in}^2}{(\Sigma^2\Delta) S_{ii,n-1}^3} \ . \end{aligned} \tag{32}$$

These terms can be non-zero for $i = v$ only. The conditional expectation of these terms are

$$\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^P; X_n)}{\partial \mathcal{B}_{ij} \partial \mathcal{B}_{vw}} | X_{n-1}\right) = \mathbb{E}\left(X_{j,n-1} X_{w,n-1} \frac{\frac{1}{2}\Sigma_{ii}^2 S_{ii,n-1}\Delta}{\Sigma^2 \Delta S_{ii,n-1}^3} | X_{n-1}\right) = \frac{X_{j,n-1} X_{w,n-1}}{2 S_{ii,n-1}^2} \ . \tag{33}$$

Similar to what we did with ν, where X_{n-1} shows up in the numerator and the denominator, we can do a first order approximation (see Paolella [2007][Chapter 2.3]), where each element $[Cov(X_n) + \theta^P \theta^{P\top}]_{jw}$ of the $m'_{\mathcal{B}i} \times m'_{\mathcal{B}i}$ block matrix for each $i = 1, \ldots, m$ is divided by $2\ \mathbb{E}(S_{ii,n-1}^2)$; $\mathcal{B}'$ are the free elements of $\mathcal{B}$, m'_B is the number of free elements - $m'_{\mathcal{B}i}$ is the number of free elements of the first column of $\mathcal{B}$. Since

the matrix $Cov(X_n)$ is positive definite, each of these blocks is positive definite, such that $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \mathcal{B}_{ij}\partial \mathcal{B}_{vw}}\right)$ has to be positive definite. This matrix can become ill-condition matrix with large $S_{ii,n-1}$ and an ill-conditioned covariance matrix of X_n. Projecting on the non-restricted elements provides us with the block regarding $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \mathcal{B}' \partial \mathcal{B}'^{\top}}\right)$.

Insofar we have calculated the blocks of the information matrix $I(\Psi^P)$, which are located along the main diagonal of this $m'_P \times m'_P$ matrix; m'_P was the number non-restricted parameters under $\mathbb{P}$ which is equivalent to the number of free elements in Ψ^P. In the general setting the expectations of the mixed partial derivatives $\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \nu_{ij}\partial B_{vw}}$, $\frac{\partial^2 \ell(\Psi^P;\mathbf{X})}{\partial \nu_{ij}\partial \theta_i}$, etc. need not be zero as in the Vasicek setting. However, to show that $I(\Psi^P)$ is (strictly) positive definite, we can consider the principal minors $\mathcal{M}_{\cdot\cdot}$ of this matrix ($m''_P \times m''_P$ submatrices with $m''_P \leq m'_P$). For a positive definite matrix all principal minors along the main diagonal of $\mathcal{M}_{1:m''_P}$ have to be positive (see Mas-Colell *et al.* [1995][Theorem M.D.2]). For the underlying setting let us rearrange the matrix $I(\Psi^P)$ such that is start with the block for $\theta^{P'}$. Here we have observed that this block is singular if its rank is less than m'_{θ^P}. If this block has not full rank at least one of the m'_{θ^P} principal minors has to be zero. In this case the information matrix becomes singular. If some of these principal minors are small, then the information matrix becomes ill-conditioned. This is the case of some eigenvalues of ν are close to one. If necessary, we can proceed in the same way for the other parameters.

D.2 The $\mathbb{A}_l(m)$ Affine Setting and Micro-Structure Noise

Generally, Ψ^Q affects $\mathbf{A}$ and $\mathbf{B}$, which allows us - given X_n - to derive the model yields $\mathbf{y}_n$. We consider $\ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|\mathbf{X})$ as described by (12). The non-restricted elements of Ψ^Q are denoted by Ψ^Q_-. For those parameters in Ψ^Q_- also entering in Ψ^P we get additional information from $\mathbf{y}^{eps}$. Here the corresponding expectations of partial derivatives have to be put to be added to elements of $I(\Psi^P)$. Such elements are the non-restricted elements of Σ^2 and maybe elements of κ^Q and θ^Q which are equal in both measures. For fixed $\mathbf{X}$ we get

$$\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|\mathbf{X})}{\partial vec(\Psi^Q_-) vec(\Psi^Q_-)^{\top}}|\mathbf{X}\right) = \sum_{n=1}^{N}\left[\frac{\partial(\mathbf{A}-\mathbf{B}X_n)}{\partial vec(\Psi^Q_-)}\right]^{\top}(\boldsymbol{\sigma}^2_{eps})^{-1}\left[\frac{\partial(\mathbf{A}-\mathbf{B}X_n)}{\partial vec(\Psi^Q_-)}\right]. \quad (34)$$

$(\boldsymbol{\sigma}^2_{eps})^{-1}$ consists of the reciprocals of the noise volatility terms described by (5). $\left[\frac{\partial(\mathbf{A}-\mathbf{B}X_n)}{\partial vec(\Psi^Q_-)}\right]^\top$ is of dimension k times the number of free elements m'_Q. The existence of these partial derivatives already follows from Gronwall [1919]. From (4) we know that $\frac{\partial(\mathbf{B}X_n)}{\partial vec(\Psi^Q_-)}$ only depends on κ^Q and Σ; for Gaussian processes it depends on κ^Q only. Although $\mathbf{A}$ and $\mathbf{B}$ and its corresponding derivatives are not available in closed form, we can derive the terms of this symmetric $m'_Q \times m'_Q$ matrix $\mathbb{E}\left(-\frac{\partial^2 \ell(\Psi^Q, \boldsymbol{\sigma}^2_{eps}; \mathbf{y}^{eps}|\mathbf{X})}{\partial vec(\Psi^Q_-) vec(\Psi^Q_-)^\top}|\mathbf{X}\right)$ by means of

$$\mathbb{E}\left[\sum_{l=1}^{k'} \sigma^2_{eps}(\tau_l)^{-1} \frac{\partial(\mathbf{A}(\tau_l)-\mathbf{B}(\tau_l)X_n)}{\partial \Psi^Q_{i-}} \frac{\partial(\mathbf{A}(\tau_l)-\mathbf{B}(\tau_l)X_n)}{\partial \Psi^Q_{j-}}\right], \tag{35}$$

for all $\Psi^Q_{i-}, \Psi^Q_{j-} \in \Psi^Q_- \times \Psi^Q_-$. $\frac{\partial \mathbf{A}(\tau_l)}{\partial \Psi^Q_{i-}}$ and $\frac{\partial \mathbf{B}(\tau_l)}{\partial \Psi^Q_{i-}}$ have to be derived numerically. In addition (35) demands for $\mathbb{E}(X_n) = \theta^P$ and $\mathbb{E}(X_n X_n^\top) = Cov(X_n) + \theta^P \theta^{P,\top}$, which are available in closed form.

E The Marginal Distribution of θ^P and ν in the Vasicek Model

By means of De Pooter *et al.* [2006] or De Pooter *et al.* [2008] we get the *marginal distribution* of the parameters θ^P, ν for the Vasicek model. With $X := (X_1, \ldots, X_N)$, $X_{-1} := (X_0, X_1, \ldots, X_{N-1})$, $M_{A.} := I_N - A.(A.^\top A.)^{-1} A.^\top$, I_N is the N-dimensional identity matrix, $A_{\theta^P} = X_{-1} - \theta^P$, $A_\nu = (1-\nu)X$, $c(\Psi^P) := F_{\mathcal{N}}(\frac{1-\nu}{\sigma_\nu}) - F_{\mathcal{N}}(\frac{-\nu}{\sigma_\nu})$, $F_{\mathcal{N}}(.)$ is the probability distribution function of a standard normal random variable, and $\sigma^2_\nu = \Sigma^2 \left((X - \nu X_{-1})^\top (X - \nu X_{-1})\right)^{-1}$ the marginals are given by:

$$\pi(\theta^P|\mathbf{X}, \nu, \Sigma^2) \propto \left((X-\theta^P)^\top M_{A_{\theta^P}} (X-\theta^P)\right)^{-\frac{N-1}{2}} \left((X_{-1}-\theta^P)^\top (X_{-1}-\theta^P)^\top\right)^{-\frac{1}{2}} c(\Psi^P) \tag{36}$$

$$\pi(\nu|\mathbf{X}) \propto \left((X-\nu X_{-1})^\top M_{A_\nu} (X-\nu X_{-1})\right)^{-\frac{N-1}{2}} N^{-\frac{1}{2}} (1-\nu)^{-1} \mathbf{1}_{\{\nu \in [0,1]\}}\,. \tag{37}$$

While the density (36) cannot be attributed to a density currently known in literature, we observe that the marginal density (37) factorizes into a student-t kernel, the term $(1-\nu)^{-1}$ and the indicator function $\mathbf{1}_{\{\nu\in[0,1]\}}$. Thus, when $\nu \to 1$ then $(1-\nu)^{-1} \to \infty$. Considering (36), $\nu \to 1$ results in $\sigma^2_\nu \to \infty$ and $c(\Psi^P) \to 0$. Therefore the joint distribution $\pi(\theta^P, \nu|\Sigma^2, \mathbf{X})$ becomes improper with $\nu = 1$. With ν close to one it becomes almost flat.

F Bayesian Sampling of the Parameters

Let (X_n) follow an affine diffusion with diagonal diffusion matrix. We get the parameters by means of Gibbs sampling and/or the Metropolis Hastings (MH) algorithm as follows:

`Step 1`: sample θ^P from $p(\theta^P|\mathbf{X}, \nu, \Sigma^2)$

`Step 2`: sample ν from $p(\nu|\mathbf{X}, \theta^P, \Sigma^2)$

`Step 3`: sample Σ^2 from $p(\Sigma^2|\mathbf{X}, \theta^P, \nu)$

Ad `Step 1`: With ν fixed, $Y_{n,\theta} = X_n - \nu X_{n-1} \in \mathbb{R}^m$ and $Z_{\theta,n} = (I_m - \nu) \in \mathbb{R}^{m\times m}$. We can write $Y_{n,\theta} = Z_{\theta,n}\theta + \sqrt{\tilde{\Sigma}}\varepsilon_n$; here $\tilde{\Sigma}_n = \Sigma^2 S_{n-1}\Delta$. I.e. we get a regression model with heterogeneous innovations (see e.g. Frühwirth-Schnatter [2006]). If some components of θ^P are restricted, the analysis can be performed in an equivalent way. With conjugate priors, θ^P can be sampled from a normal distribution with parameters $a_{\theta,p} = A_{\theta,p} \cdot \left([\sum_{n=1}^N (Z_{n,\theta}^\top \tilde{\Sigma}_n^{-1} Y_\theta)] + A_{\theta,0}^{-1} a_{\theta,0}\right)$ and $A_{\theta,p} = \left([\sum_{n=1}^N (Z_{n,\theta}^\top \tilde{\Sigma}_n^{-1} Z_{n,\theta})] + A_{\theta,0}^{-1}\right)^{-1}$. $A_{\theta,0}$ is the prior variance and $a_{\theta,0}$ is the prior mean. Applying the prior π_{SD} results in $A_{\theta,0} = \tilde{A}_{\theta,0} \cdot (I_m - \nu)^{-1}\Sigma^2(I_m - \nu)^{-1\top}$. For a non-conjugate prior the MH algorithm has to be applied. This normal conditional density can also be used as a proposal density $q(\theta^P)$ in a MH step (see also Chib and Ergashev [2009], where similar tailored proposal densities are used). For $m > 1$ we follow this approach, while for the Vasicek and the CIR model the Gibbs sampler was applied.

Remark 2. Especially in the Vasicek setting sampling from $p(\theta^P|\mathbf{X}, \nu, \Sigma^2)$ in `Step 1` can be performed by using the Gibbs sampler. Alternatively, the Metropolis-Hastings algorithm can be used. With the MH algorithm it is important to note that sampling of θ^P requires a careful choice of the proposal densities $q(.)$. By using a normal random walk proposal $\theta_i^{P,prop} = \theta_i^P + c_{\theta_i^P}\varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$ and $c_{\theta_i^P}$ constant (and "as usual not too large to get sufficiently high acceptance probabilities"), we hardly get large deviations from θ_i^P as done by Gibbs sampler when ν_{ii}^P is close to one. To tackle this problem we propose from a normal random walk proposal with $c_{\theta_i^P} \propto (1 - \nu_{ii}^P)^{-1}$. For the Vasicek model we compared the posterior samples and observed minor differences when the MH algorithm is used instead of Gibbs sampler. Here

the reader is also referred to Hoogerheide *et al.* [2007] and Hoogerheide and van Dijk [2008].

Ad `Step 2`: With θ^P fixed, we define $Y_{n,\nu} = X_n - \theta^P \in \mathbb{R}^m$ and $Z_{n,\nu} = X_{n-1} - \theta^P \in \mathbb{R}^{m \times m'_\nu}$; m'_ν is the number of non-zero parameters in ν. From the $m \times m$ matrix ν, we get the vector $\beta(\nu)$ by deletion of the elements equal to zero of $vec(\nu)$. $\beta(\nu)$ has dimension m'_ν. Now we get $Y_{n,\nu} = Z_{n,\nu}\beta(\nu) + \sqrt{\tilde{\Sigma}}\varepsilon_n$. ν can be sampled from a normal distribution with parameters $a_{\nu,p} = A_{\nu,p} \cdot \left([\sum_{n=1}^N (Z_{n,\nu}^\top \tilde{\Sigma}_n^{-1} Y_{n,\nu})]\right)$ and $A_{\nu,p} = \left([\sum_{n=1}^N (Z_{n,\nu}^\top \tilde{\Sigma}_n^{-1} Z_{n,\nu})]\right)^{-1}$. Alternatively a conjugate normal prior with parameters $a_{\nu,0}$ and $A_{\nu,0}$ can by applied, such that $a_{\nu,p}$ and $A_{\nu,p}$ become $a_{\nu,p} = A_{\nu,p} \cdot \left([\sum_{n=1}^N Z_{n,\nu}^\top \tilde{\Sigma}_n^{-1} Y_{n,\nu}] + A_{\nu,0}^{-1} a_{\nu,0}\right)$ and $A_{\nu,p} = \left([\sum_{n=1}^N Z_{n,\nu}^\top \tilde{\Sigma}_n^{-1} Z_{n,\nu}] + A_{\nu,0}^{-1}\right)^{-1}$, respectively (see e.g. Cameron and Trivedi [2005]). A conjugate truncated normal could also be applied by using $a_{\nu,0}$ and $A_{\nu,0}$ as above and $\nu \in (0,1)$. Here, ν is sampled from a normal distribution with parameters $a_{\nu,p}$ and $A_{\nu,p}$. The sample is accepted if $\nu \in (0,1)$. For $m > 1$ the eigenvalues of ν have to be in this interval. With the prior $\pi_{SI}(.)$, the MH algorithm has to be used but the above conditional density can be used as a proposal density.

For the random walk proposals we use $\nu_{ij}^{P,prop} = \nu_{ij}^P + c_{\kappa^P}\varepsilon$; we set $c_{\kappa^P} = 0.1$ and $\varepsilon \sim \mathcal{N}(0,1)$. While for the Vasicek and the CIR model the Gibbs sampler has been applied (if possible due to the prior), we sample ν for the multivariate setting by means of the MH algorithm with random walk proposals (in contrast to Chib and Ergashev [2009]).

Ad Step 3: If Σ^2 has diagonal structure, we sample Σ_{ii}^2 from an inverse gamma distribution with parameters n_{ip} (*degrees of freedom parameter*) and S_{ip} (*scale parameter*) based on the assumption of the conjugate inverse Gamma prior. The parameters are given by $n_{ip} = n_0 + N/2$ and $S_{ip} = S_0 + \frac{1}{2}\sum_{n=1}^N (X_n - \nu X_{n-1} - \theta^P(1-\nu))/[S_{ii,n-1}\Delta^{0.5}])^2$. In the Vasicek in the CIR model we apply a Gibbs sampler with this conjugate prior. For the yields observed where Σ^2 enters into $\mathbf{A}$ and $\mathbf{B}$, the Metropolis Hastings algorithm has to be applied. Here we mix between proposals from these densities and random walk proposals.

For the parameter δ_0 we applied the MH algorithm. We propose from a normal density which is derived in a similar way as the conditional density in Step 1. Given $\mathbf{X}$, the other parameters and the fact the δ_0 enters into $\mathbf{A}$ in a linear way allows us to write $\mathbf{y}_n^{eps} = \delta_0 + \mathbf{A}^- - \mathbf{B}X_n + \mathbf{e}_n$, where $\mathbf{A}^-$ is $\mathbf{A}$ without the δ_0 component. In this case sampling δ_0 corresponds to Bayesian sampling of a sample mean

with heterogeneous innovations. For the updates of θ^Q and κ^Q, random walk updates have been applied. For $\boldsymbol{\sigma}^2_{eps}$ we applied the Gibbs sampler as in Step 3.

References

Yacine Aït-Sahalia and Jean Jacod. Fisher's information for discretely sampled lévy processes. *Econometrica*, 76:727–761, 2008.

Yacine Aït-Sahalia and Robert L. Kimmel. Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions. *SSRN eLibrary*, 2009.

Yacine Aït-Sahalia. *Estimating Continuous-Time Models Using Discretely Sampled Data.* in Richard Blundell, Torsten Persson, Whitney K. Newey: *Advances in Economics and Econometrics, Theory and Applications*, chapter 9, Cambridge University Press, 2007.

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

Andrew Ang, Sen Dong, and Monika Piazzesi. No-Arbitrage Taylor Rules. *SSRN eLibrary*, 2004.

Sebastien Blais. Bayesian analysis of affine term structure models. *Working Paper, Bank of Canada*, 2009.

Roger Bowden. The theory of parametric identification. *Econometrica*, 41(6):1069–1074, 1973.

Michael W. Brandt and Ping He. Simulated likelihood estimation of affine term structure models from panel data. Working paper, University of Pennsylvania, 2002.

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods.* Springer Series and Statistics. Springer, New York, 2 edition, 2006.

A. Colin Cameron and Pravin K. Trivedi. *Microeconometrics.* Cambridge University Press, Cambridge, MA, 2005.

John W. Campbell, Andrew W. Lo, and Craig MacKinlay. *The Econometrics of Financial Markets.* Princeton University Press, Oxford, UK, 1996.

Fabio Canova and Luca Sala. Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics*, 56(4):431–449, May 2009.

Hui Chen and Scott Joslin. Generalized transform analysis of affine processes and asset pricing applications. *Working Paper, MIT*, 2009.

Long Chen, David A. Lesmond, and Jason Wei. Corporate yield spreads and bond liquidity. *The Journal of Finance*, 62(1):119–149, 2007.

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

Patrick Cheridito, Damir Filipovic, and Robert L. Kimmel. A Note on the Dai-Singleton Canonical Representation of Affine Term Structure Models. *SSRN eLibrary*, 2008.

Siddhartha Chib and Bakhodir Ergashev. Analysis of multi-factor affine yield curve models. *Journal of the American Statistical Association*, 104(488):1324–1336, 2009.

Siddhartha Chib. Markov chain monte carlo methods: computation and inference. In J.J. Heckman and E.E. Leamer, editors, *Handbook of Econometrics*, volume 5 of *Handbook of Econometrics*, chapter 57, pages 3569–3649. Elsevier, 2001.

Pierre Collin-Dufresne and Robert Goldstein. Do credit spreads reflect stationary leverage ratios? *Journal of Finance*, 56(5):1929–1957, 2001.

Pierre Collin-Dufresne and Robert Goldstein. Do bonds span the fixed income markets? Theory and evidence for unspanned stochastic volatility. *Journal of Finance*, 57(4):1685–1730, 2002.

Pierre Collin-Dufresne, Robert S. Goldstein, and Christopher S. Jones. Identification of maximal affine term structure models. *Journal of Finance*, 63(2):743–795, 2008.

Pierre Collin-Dufresne, Robert S. Goldstein, and Christopher S. Jones. Can interest rate volatility be extracted from the cross section of bond yields? *Journal of Financial Economics*, 2009. Forthcoming.

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

Christa Cuchiero, Josef Teichmann, and Martin Keller-Ressel. Polynomial processes and their application to mathematical finance. *Finance & Stochastics*, 2010. forthcoming.

Walter J. Culver. On the existence and uniqueness of the real logarithm of a matrix. *Proceeding of the American Mathematical Society*, 17:1146–1151, 1966.

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

Qiang Dai and Kenneth J. Singleton. Expectation puzzles, time-varying risk premia, and affine models of the term structure. *Journal of Financial Economics*, 63(3):415–441, 2002.

James Davidson and Russel G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, New York, 1993.

Michiel D. De Pooter, Rene Segers, and Herman K. van Dijk. On the practice of bayesian inference in basic economic time series models using gibbs sampling. Working paper, Tinbergen Institute, TI 2006-076/04, Erasmus University Rotterdam, 2006.

Michiel D. De Pooter, Francesco Ravazzolo, Rene Segers, and Herman K. van Dijk. Bayesian near-boundary analysis in basic macroeconomic time series models. Working paper, Tinbergen Institute, Erasmus University Rotterdam, 2008.

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

Joost Driessen. Is default event risk priced in corporate bonds? *Review of Financial Studies*, 18(1):165–195, 2005.

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

Gregory R. Duffee. Information in (and not in) the term structure. Technical report, forthcoming in Review of Financial Studies, 2011.

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

Darrell Duffie and Kenneth Singleton. An econometric model of the term structure of interest rate swap yields. *Journal of Finance*, 52(4):1287–1323, 1997.

Darrell Duffie and Kenneth Singleton. Modeling term structures of defaultable bonds. *Review of Financial Studies*, 12(4):687–720, 1999.

Darrell Duffie, Jun Pan, and Kenneth Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376, 2000.

Jean-Marie Dufour. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65:1365–1388, 1997.

Jean-Marie Dufour. Identification, weak instruments and statistical inference in econometrics. Cahiers de recherche 10-2003, Centre interuniversitaire de recherche en ï¿½conomie quantitative, CIREQ, 2003.

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

G. Elliott and J. H. Stock. Confidence intervals for autoregressive coefficients near one. *Journal of Econometrics*, 103:155–181, 2001.

Peter Feldhütter and David Lando. Decomposing swap spreads. *Journal of Financial Economics*, 88(2):375–405, May 2008.

Damir Filipović. *Term-Structure Models: A Graduate Course*. Springer, Berlin, 2009.

Sylvia Frühwirth-Schnatter and Alois Geyer. 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 BA, 1996.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, New York, 2006.

Paul Glasserman and Kyoung-Kuk Kim. Moment Explosions and Stationary Distributions in Affine Diffusion Models. *Mathematical Finance, Forthcoming*, 2009.

William H. Greene. *Econometric Analysis*. Perentice Hall, New Jersey, 3 edition, 1997.

Thomas H. Gronwall. Note on the derivative with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

J.D. Hamilton and C. Wu. Identification and estimation of gaussian affine term structure models. Technical report, Working paper, University of California, San Diego, 2010.

James Douglas Hamilton. *Time Series Analysis*. Princeton University Press, New York, 1994.

Lennart Hoogerheide and Herman K. van Dijk. Possibly ill-behaved posteriors in econometric models. *Tinbergen Institute Discussion Papers*, (08-036/4), April 2008.

Lennart F. Hoogerheide, Johan F. Kaashoek, and Herman K. van Dijk. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. *Journal of Econometrics*, 139(1):154 – 180, 2007.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

M. Jansson and M. J. M. J. Moreira. Optimal inference in regression models with nearly integrated regressors. *Econometrica*, 74(3):681–714, 2006.

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

S. Joslin, K. Singleton, and H. Zhu. A new perspective on gaussian dynamic term structure models. *Review of Financial Studies*, 24:926–970, 2010.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, 2 edition, 1991.

Matrin Keller-Ressel and Eberhard Mayerhofer. On exponential moments of affine processes. Technical report, Working paper, Deutsche Bundesbank, Frankfurt, 2011.

M. Kendall. Note on bias in the estimation of autocorrelation. *Biometrika*, 41(2):403–404, 1954.

K.-K. Kim and P. Glasserman. Moment explosion and stationary distributions in affine diffusion models. *Mathematical Finance*, 2008. to appear.

F. Kleibergen and H.K. van Dijk. On the shape of the likelihood/posterior in cointegration models. *Econometric Theory*, 10(3/4):514–551, 1994.

F. Kleibergen and H.K. van Dijk. A bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14:701–743, 1998.

David Lando. On Cox processes and credit risky securities. *Review of Derivatives Research*, 2:99–120, 1998.

J. Lewellen. Predicting returns with financial ratios. *Journal of Financial Economics*, 74:209–235, 2004.

Jun Ma and Charles R. Nelson. Valid inference for a class of models where standard inference performs poorly; including nonlinear regression, arma, garch, and unobserved components. *Working paper, University of Washington*, 2009.

Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory.* Oxford University Press, New York, 1995.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Wiley, New York, 1997.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability.* Cambridge University Press (Cambridge Mathematical Library), New York, 2 edition, 2009.

Robert J. Mislevy and Kathleen M. Sheehan. Information matrices in latent-variable models. *Journal of Educational and Behavioral Statistics*, 14(4):335–350, 1989.

Terence Orchard and Max A. Woodbury. A missing information principle: Theory and applications. *Proceeding of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (University of California Press)*, 1:697–715, 1972.

Jun Pan and Kenneth J. Singleton. Default and recovery implicit in the term structure of sovereign CDS spreads. *Journal of Finance*, 63(5):2345–2384, 2008.

Marc Paolella. *Intermediate Probability - A Computational Approach.* Wiley, 2007.

P. Phillips. Regression theory for near-integrated time series. *Econometrica*, 56(5):1021–1043, 1998.

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

Dale J. Poirier. *Intermediate Statistics and Econometrics: A Comparative Approach.* MIT Press, Cambridge, Mass., 1995.

T. J. Rothenberg and J. H. J. H.Stock. Inference in a nearly integrated autoregressive model with non-normal innovations. *Journal of Econometrics*, 80:269–286, 1997.

Andrew D. Sanford and Gael M. Martin. Simulation-based Bayesian estimation of an affine term structure model. *Computational Statistics and Data Analysis*, 49:527–554, 2005.

Paul Schneider, Leopold Sögner, and Tanja Veža. The economic role of jumps and recovery rates in the market for corporate default risk. *Journal of Financial and Quantitative Analysis*, 45:1517–1547, 2010.

P. Schotman and H.K. van Dijk. A bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics*, 49:195–238, 1991.

Huarong Tang and Yihong Xia. An international examination of affine term structure models and the expectations hypothesis. *Journal of Financial and Quantitative Analysis*, 42(1):41–80, 2007.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

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

	$mean(\hat{\Psi}^P)$	$sd(\hat{\Psi}^P)$	$min(\hat{\Psi}^P)$	$max(\hat{\Psi}^P)$	$q(\hat{\Psi}^P, 0.025)$	$median(\hat{\Psi}^P)$	$q(\hat{\Psi}^P, 0.975)$
Parameter Estimates with the Prior (19)							
Parameter Estimates for the OU model with $\gamma^* = 2$, $\lambda^* = 0.9999$							
ν	0.9882	0.0085	0.9498	1.0000	0.9700	0.9893	0.9997
θ^P	4.8286	127.3978	-1829.6188	1906.1439	-187.0803	3.1611	210.2170
σ^2	1.4439	0.0917	1.1148	1.9084	1.2752	1.4401	1.6346
κ^P	0.6251	0.4531	0.0010	2.6948	0.0179	0.5625	1.5939
Parameter Estimates for the CIR model with $\gamma^* = 2$, $\lambda^* = 0.9999$							
ν	0.9871	0.0080	0.9519	1.0000	0.9703	0.9878	0.9988
θ^P	28.0817	126.0652	-1436.7104	2320.7493	-52.5715	3.9422	324.4285
σ^2	0.4919	0.0312	0.3800	0.6488	0.4344	0.4906	0.5568
κ^P	0.6843	0.4265	0.0018	2.5807	0.0609	0.6410	1.5747
Parameter Estimates for the OU model with $\gamma^* = 2$, $\lambda^* = 0.995$							
ν	0.9882	0.0086	0.9495	1.0000	0.9698	0.9893	0.9995
θ^P	2.4114	127.9933	-1880.2639	1864.3128	-200.6264	3.0177	198.4717
σ^2	1.4480	0.0920	1.1178	1.9111	1.2788	1.4441	1.6393
κ^P	0.6261	0.4557	0.0005	2.7117	0.0241	0.5627	1.6016
Parameter Estimates for the CIR model with $\gamma^* = 2$, $\lambda^* = 0.995$							
ν	0.9864	0.0082	0.9507	0.9999	0.9695	0.9872	0.9985
θ^P	26.0766	121.2544	-1419.3769	2263.3184	-51.2701	3.5733	309.4119
σ^2	0.4862	0.0309	0.3750	0.6414	0.4295	0.4849	0.5504
κ^P	0.7184	0.4334	0.0039	2.6426	0.0785	0.6741	1.6223
Parameter Estimates for the OU model with $\gamma^* = 2$, $\lambda^* = 0.99$							
ν	0.9871	0.0080	0.9519	1.0000	0.9703	0.9878	0.9988
θ^P	28.0817	126.0652	-1436.7104	2320.7493	-52.5715	3.9422	324.4285
σ^2	1.4919	0.0312	1.3800	1.6488	1.4344	1.4906	1.5568
κ^P	0.6843	0.4265	0.0018	2.5807	0.0609	0.6410	1.5747
Parameter Estimates for the CIR model with $\gamma^* = 2$, $\lambda^* = 0.99$							
ν	0.9872	0.0081	0.9520	1.0000	0.9704	0.9880	0.9988
θ^P	26.6796	125.1440	-1472.8413	2303.1051	-54.7794	3.6674	320.7548
σ^2	0.4905	0.0312	0.3787	0.6479	0.4332	0.4892	0.5553
κ^P	0.6768	0.4273	0.0009	2.5724	0.0606	0.6318	1.5704

Table 1: Parameter estimates for the Vasicek [1977] and the Cox *et al.* [1985] model. Data simulated with $\theta^P = 3$, $\nu = 0.99$ and $\Sigma^2 = 1.2^2$ for the Vasicek and $\Sigma^2 = 0.7^2$ for the CIR setting. Statistics obtained from $\mathtt{M} = 500$ simulation runs. Prior (19) applied to ν.

Parameter Estimates with Shrinkage Prior							
	$mean(\hat{\Psi}^P)$	$sd(\hat{\Psi}^P)$	$min(\hat{\Psi}^P)$	$max(\hat{\Psi}^P)$	$q(\hat{\Psi}^P, 0.025)$	$median(\hat{\Psi}^P)$	$q(\hat{\Psi}^P, 0.975)$
Parameter Estimates for the OU model, shrinkage prior, $\lambda^* = 0.9999$							
ν	0.9862	0.0086	0.9478	0.9999	0.9682	0.9869	0.9992
θ^P	2.9612	13.7410	-186.1015	190.1522	-20.6280	3.0207	25.8990
σ^2	1.4337	0.0911	1.1063	1.8921	1.2662	1.4299	1.6229
κ^P	0.7291	0.4590	0.0053	2.8061	0.0443	0.6888	1.6915
Parameter Estimates for the CIR model, shrinkage prior, $\lambda^* = 0.9999$							
ν	0.9864	0.0087	0.9475	0.9999	0.9682	0.9871	0.9992
θ^P	6.2284	14.7442	-156.2127	242.4959	-6.7318	3.3476	42.6995
σ^2	0.4883	0.0310	0.3770	0.6442	0.4313	0.4870	0.5528
κ^P	0.7208	0.4618	0.0052	2.8210	0.0417	0.6784	1.6921
Parameter Estimates for the OU model, shrinkage prior, $\lambda^* = 0.999$							
ν	0.9848	0.0081	0.9480	0.9990	0.9677	0.9854	0.9978
θ^P	3.0093	1.8791	-16.8882	22.9827	-0.7151	3.0004	6.7872
σ^2	1.4493	0.0920	1.1176	1.9145	1.2800	1.4454	1.6406
κ^P	0.8044	0.4318	0.0523	2.7981	0.1163	0.7706	1.7162
Parameter Estimates for the CIR model, shrinkage prior, $\lambda^* = 0.999$							
ν	0.9846	0.0082	0.9476	0.9990	0.9676	0.9852	0.9977
θ^P	3.0209	1.8547	-16.8976	22.6495	-0.7079	3.0304	6.6702
σ^2	1.4327	0.0910	1.1055	1.8904	1.2654	1.4289	1.6218
κ^P	0.8112	0.4340	0.0530	2.8200	0.1179	0.7789	1.7244
Parameter Estimates for the OU model, shrinkage prior, $\lambda^* = 0.995$							
ν	0.9827	0.0071	0.9482	0.9950	0.9673	0.9834	0.9938
θ^P	3.0622	0.6808	-1.4025	7.5643	1.6617	3.0615	4.4677
σ^2	1.4405	0.0914	1.1117	1.9014	1.2724	1.4366	1.6306
κ^P	0.9117	0.3797	0.2615	2.7828	0.3223	0.8727	1.7378
Parameter Estimates for the CIR model, shrinkage prior, $\lambda^* = 0.995$							
ν	0.9827	0.0072	0.9473	0.9950	0.9670	0.9834	0.9938
θ^P	3.0710	0.6693	-1.0711	8.1534	1.8808	3.0103	4.6143
σ^2	0.4912	0.0312	0.3791	0.6494	0.4339	0.4899	0.5561
κ^P	0.9160	0.3839	0.2615	2.8306	0.3224	0.8762	1.7524

Table 2: Parameter estimates for the Vasicek [1977] and the Cox *et al.* [1985] model. Data simulated with $\theta^P = 3$, $\nu = 0.99$ and $\Sigma^2 = 1.2^2$ for the Vasicek and $\Sigma^2 = 0.7^2$ for the CIR setting. Statistics obtained from $\mathtt{M} = 500$ simulation runs. Shrinkage prior applied to ν.

Parameter Estimates for $\mathbb{A}_1(3)$ Setting									
	true	mean	sd	min	max	q(0.025)	median	q(0.975)	IEF
ν_{11}	0.9867	0.9823	0.0065	0.9605	0.9950	0.9684	0.9826	0.9931	82.5969
ν_{22}	0.9848	0.9938	0.0011	0.9883	0.9950	0.9911	0.9941	0.9950	12.3765
ν_{32}	-0.0019	0.0038	0.0017	-0.0013	0.0081	0.0001	0.0038	0.0072	130.0790
ν_{33}	0.9829	0.9464	0.0134	0.9157	0.9768	0.9215	0.9477	0.9704	155.4359
κ_{11}^P	0.7000	1.1913	0.4846	0.2618	2.8435	0.3689	1.1567	2.2269	
κ_{22}^P	0.8000	0.4763	0.3682	0.2614	2.9426	0.2639	0.3383	1.6589	
κ_{32}^P	0.1000	-0.0966	0.1098	-0.4479	0.3572	-0.3016	-0.0987	0.1128	
κ_{33}^P	0.9000	2.4559	0.7983	0.2618	4.9630	0.8550	2.4795	4.0057	
κ_{11}^Q	0.5000	0.4610	0.0671	0.3123	0.6089	0.3459	0.4480	0.5892	198.1873
κ_{22}^Q	0.7000	0.7634	0.0498	0.6282	0.8684	0.6620	0.7720	0.8424	197.0050
κ_{33}^Q	1.0000	1.0923	0.0772	0.9277	1.2707	0.9497	1.1089	1.2380	198.9622
θ_1^P	1.5000	1.6014	0.1151	0.6656	2.8691	1.3789	1.5871	1.8949	4.9220
θ_1^Q	2.0000	3.1638	0.3816	1.8168	3.8732	2.1398	3.2499	3.6659	193.3803
δ_0	1.0000	0.7357	0.3195	0.2270	1.9946	0.4071	0.6012	1.5206	195.7308
Σ_{11}^2	0.0625	0.0938	0.0199	0.0491	0.1841	0.0606	0.0937	0.1385	181.5149
Σ_{22}^2	0.1600	0.2365	0.0504	0.1384	0.4424	0.1715	0.2202	0.3568	181.2781
Σ_{33}^2	0.2500	0.2480	0.0311	0.1542	0.3818	0.1936	0.2458	0.3142	147.8921
Σ_{11}	0.2500	0.3045	0.0323	0.2216	0.4291	0.2461	0.3061	0.3721	
Σ_{22}	0.4000	0.4837	0.0500	0.3721	0.6652	0.4142	0.4692	0.5973	
Σ_{33}	0.5000	0.4970	0.0311	0.3927	0.6179	0.4400	0.4958	0.5605	
$\sigma_{eps}^2(1/12)$	0.0069	0.0477	0.0576	0.0112	0.1947	0.0133	0.0185	0.1947	63.9079
$\sigma_{eps}^2(1/4)$	0.0072	0.0429	0.0525	0.0100	0.1773	0.0122	0.0167	0.1773	59.3977
$\sigma_{eps}^2(1/2)$	0.0076	0.0368	0.0435	0.0094	0.1472	0.0112	0.0149	0.1472	61.1095
$\sigma_{eps}^2(1)$	0.0086	0.0305	0.0315	0.0103	0.1101	0.0118	0.0151	0.1101	60.5161
$\sigma_{eps}^2(2)$	0.0107	0.0248	0.0157	0.0123	0.0641	0.0144	0.0176	0.0641	57.8162
$\sigma_{eps}^2(3)$	0.0130	0.0235	0.0090	0.0142	0.0451	0.0165	0.0197	0.0451	56.6530
$\sigma_{eps}^2(5)$	0.0183	0.0257	0.0035	0.0173	0.0321	0.0209	0.0246	0.0321	58.5145
$\sigma_{eps}^2(7)$	0.0238	0.0277	0.0011	0.0216	0.0284	0.0247	0.0284	0.0284	14.4051
$\sigma_{eps}^2(10)$	0.0302	0.0293	0.0009	0.0230	0.0298	0.0266	0.0298	0.0298	9.4167
$\sigma_{eps}^2(20)$	0.0183	0.0190	3.6E-5	0.0167	0.0190	0.0190	0.0190	0.0190	1.1070

Table 3: $\mathbb{A}_1(3)$ **Model, Simulated Data:** MCMC estimates of parameters Ψ; shrinkage prior with $\lambda^* = 0.995$. $N = 500$ observations, $k = 10$ maturities $\boldsymbol{\tau} = \{1/12, 1/4, 1/2, 1, 2, 3, 5, 7, 10, 20\}$. The columns provide the true parameter values and the descriptive statistics sample mean, standard deviation (sd), minimum, maximum, 0.025% quantile, median, the 0.975 quantile and the Chib [2001] inefficiency factor. 50,000 MCMC steps, 20,000 burn in.

Parameter Estimates for $\mathbb{A}_1(3)$ Setting - Empirical US Data								
	mean	sd	min	max	q(0.025)	median	q(0.975)	IEF
ν_{11}	0.9859	0.0035	0.9811	0.9950	0.9813	0.9852	0.9936	13.9133
ν_{22}	0.9902	0.0038	0.9811	0.9950	0.9819	0.9912	0.9948	51.2329
ν_{32}	0.0912	0.0536	0.0108	0.2419	0.0225	0.0805	0.2046	198.4277
ν_{33}	0.9871	0.0041	0.9811	0.9950	0.9813	0.9866	0.9945	17.7597
κ_{11}^P	0.7392	0.1861	0.2616	0.9950	0.3335	0.7757	0.9841	
κ_{22}^P	0.5141	0.2007	0.2614	0.9949	0.2693	0.4610	0.9530	
κ_{32}^P	-4.8105	2.8287	-12.8454	-0.5684	-10.7918	-4.2542	-1.1907	
κ_{33}^P	0.6794	0.2165	0.2614	0.9950	0.2881	0.7019	0.9858	
κ_{11}^Q	0.1472	0.0569	0.0639	0.3957	0.0721	0.1459	0.2547	199.0140
κ_{22}^Q	0.7934	0.0900	0.5310	0.9750	0.5691	0.8111	0.9329	198.3452
κ_{33}^Q	2.4098	0.3720	1.0959	2.7352	1.2617	2.5574	2.7004	199.9012
θ_1^P	1.9354	0.1465	0.7277	3.5848	1.6842	1.9011	2.3229	3.0251
θ_1^Q	15.9040	3.7424	6.2730	24.6725	8.2071	16.5237	22.1365	199.5094
δ_0	0.0372	0.0869	-0.3222	0.3270	-0.1303	0.0194	0.1874	179.5633
Σ_{11}^2	0.0905	0.0163	0.0405	0.1445	0.0511	0.0934	0.1162	171.0889
Σ_{22}^2	0.3433	0.1846	0.1003	0.9085	0.1227	0.2893	0.7091	195.3897
Σ_{33}^2	1.3265	0.5252	0.2089	2.4453	0.2563	1.4723	2.0368	193.2727
Σ_{11}	0.2994	0.0286	0.2012	0.3801	0.2260	0.3057	0.3409	
Σ_{22}	0.5648	0.1557	0.3167	0.9532	0.3503	0.5379	0.8421	
Σ_{33}	1.1199	0.2690	0.4570	1.5638	0.5062	1.2134	1.4272	
$\sigma_{eps}^2(1/12)$	0.7684	0.8143	0.0492	2.4694	0.0604	0.4370	2.4694	197.3384
$\sigma_{eps}^2(1/4)$	0.4959	0.6196	0.0195	2.4992	0.0244	0.2781	2.1706	198.0231
$\sigma_{eps}^2(1/2)$	0.4020	0.5901	0.0110	2.4926	0.0154	0.1814	2.1160	198.1032
$\sigma_{eps}^2(1)$	0.2631	0.4706	0.0103	2.1548	0.0139	0.0628	1.7133	198.1821
$\sigma_{eps}^2(2)$	0.2021	0.3321	0.0109	1.6277	0.0151	0.0797	1.2502	197.9106
$\sigma_{eps}^2(3)$	0.2117	0.2493	0.0295	1.2600	0.0413	0.1392	1.0056	197.2610
$\sigma_{eps}^2(5)$	0.2290	0.1258	0.0789	0.7748	0.1169	0.2021	0.6371	193.3388
$\sigma_{eps}^2(7)$	0.2423	0.0702	0.1193	0.5379	0.1607	0.2304	0.4570	176.3249
$\sigma_{eps}^2(10)$	0.2184	0.0417	0.1188	0.3545	0.1596	0.2123	0.3280	117.7855
$\sigma_{eps}^2(20)$	0.2448	0.0104	0.1769	0.2493	0.2114	0.2493	0.2493	23.8948

Table 4: $\mathbb{A}_1(3)$ **Model, H-15 Data:** MCMC estimates of parameters Ψ and the Chib [2001] inefficiency factor IEF: The columns provide descriptive statistics calculated from the posterior. The estimates are based on 50,000 MCMC steps, 20,000 burn in. Shrinkage prior with $\lambda^* = 0.995$. $N = 413$ observations, $k = 10$ maturities $\boldsymbol{\tau} = \{1/12, 1/4, 1/2, 1, 2, 3, 5, 7, 10, 20\}$.

Parameter Estimates for $\mathbb{A}_1(3)$ Setting - Empirical European Data								
	mean	sd	min	max	q(0.025)	median	q(0.975)	IEF
ν_{11}	0.9887	0.0040	0.9722	0.9950	0.9804	0.9889	0.9947	53.2176
ν_{22}	0.9918	0.0024	0.9824	0.9950	0.9863	0.9923	0.9949	20.4088
ν_{32}	0.0182	0.0035	0.0057	0.0265	0.0114	0.0183	0.0242	112.3791
ν_{33}	0.9591	0.0119	0.9272	0.9882	0.9352	0.9596	0.9808	161.3796
κ_{11}^P	0.9862	0.5449	0.2614	3.3264	0.3001	0.8544	2.3163	
κ_{22}^P	0.5468	0.2558	0.2614	1.9257	0.2708	0.4759	1.2771	
κ_{32}^P	-0.3721	0.4825	-1.4198	0.7810	-1.1056	-0.4682	0.6002	
κ_{33}^P	2.5286	0.9804	0.2703	6.4312	0.8207	2.4644	4.3904	
κ_{11}^Q	0.2455	0.0923	0.1083	0.5857	0.1281	0.2222	0.5218	199.5059
κ_{22}^Q	0.7603	0.1622	0.4984	1.0261	0.5203	0.7460	1.0004	199.7201
κ_{33}^Q	1.1244	0.0424	1.0428	1.2512	1.0534	1.1161	1.2175	196.2825
θ_1^P	2.0134	0.0841	0.8178	2.8744	1.7721	2.0244	2.1274	32.9397
θ_1^Q	9.1917	0.9575	7.6658	12.7288	7.8739	9.0578	11.9971	198.0933
δ_0	-0.8778	0.1989	-1.8218	-0.2427	-1.4092	-8E-01	-0.6436	187.7038
Σ_{11}^2	0.0395	0.0151	0.0195	0.0944	0.0234	0.0341	0.0774	194.0363
Σ_{22}^2	0.1030	0.0485	0.0454	0.2807	0.0536	0.0859	0.2239	195.4531
Σ_{33}^2	0.1189	0.0306	0.0696	0.2529	0.0830	0.1099	0.1997	188.3213
Σ_{11}	0.1957	0.0351	0.1397	0.3073	0.1528	0.1847	0.2782	
Σ_{22}	0.3133	0.0700	0.2131	0.5298	0.2315	0.2930	0.4732	
Σ_{33}	0.3422	0.0420	0.2637	0.5029	0.2880	0.3314	0.4469	
$\sigma_{eps}^2(1/12)$	0.0310	0.0166	0.0117	0.0647	0.0151	0.0238	0.0647	87.4174
$\sigma_{eps}^2(1/4)$	0.0293	0.0170	0.0068	0.0581	0.0109	0.0215	0.0581	100.2041
$\sigma_{eps}^2(1/2)$	0.0186	0.0181	0.0049	0.0570	0.0061	0.0095	0.0570	86.5584
$\sigma_{eps}^2(1)$	0.0645	0.0189	0.0197	0.0851	0.0296	0.0662	0.0851	119.2183
$\sigma_{eps}^2(2)$	0.1366	0.0293	0.0665	0.1636	0.0801	0.1469	0.1636	145.6772
$\sigma_{eps}^2(5)$	0.1700	0.0180	0.1159	0.1899	0.1352	0.1704	0.1899	105.7929
$\sigma_{eps}^2(7)$	0.1577	0.0135	0.1096	0.1713	0.1299	0.1596	0.1713	84.7395
$\sigma_{eps}^2(10)$	0.1370	0.0096	0.0977	0.1459	0.1160	0.1394	0.1459	57.3023
$\sigma_{eps}^2(15)$	0.1085	0.0051	0.0740	0.1113	0.0938	0.1113	0.1113	35.1514
$\sigma_{eps}^2(20)$	0.0917	0.0037	0.0665	0.0933	0.0801	0.0933	0.0933	46.1553
$\sigma_{eps}^2(30)$	0.0811	0.0018	0.0631	0.0816	0.0747	0.0816	0.0816	12.3991

Table 5: $\mathbb{A}_1(3)$ **Model, Yields from LIBOR and Swap Rates:** MCMC estimates of parameters Ψ and the Chib [2001] inefficiency factor IEF: The columns provide descriptive statistics calculated from the posterior. The estimates are based on 50,000 MCMC steps, 20,000 burn in. Shrinkage prior with $\lambda^* = 0.995$. $N = 500$ observations, $k = 11$ maturities $\tau = \{1/12, 1/4, 1/2, 1, 2, 5, 7, 10, 15, 20, 30\}$.