Non-Linear Kalman Filtering Techniques for Term-Structure Models[*]

Jesper Lund[†]
Department of Finance
The Aarhus School of Business
Fuglesangs Alle 4
DK-8210 Aarhus V
Denmark

Phone: +45 8948-6362
Fax: +45 8615-1943
E-mail: jel@hha.dk

First draft: June 1997

[*] The most recent version of this paper is available at: `http://www.hha.dk/~jel`

[†] I thank Esben Høg for helpful comments.

Non-Linear Kalman Filtering Techniques for Term-Structure Models

Abstract

The state space form is a useful framework for estimating Markovian term-structure models with unobserved state variables. In this paper, we consider an econometric method which accommodates non-linearity in the measurement equation, for example when estimating exponential-affine models using prices of coupon bonds. The filtering algorithm is known as the iterative, extended Kalman filter (IEKF), and the model parameters are estimated by quasi maximum likelihood (QML), based on predictions errors obtained from the IEKF recursions. While, in general, the QML estimator is inconsistent, a Monte Carlo study demonstrates that the biases are very small, and economically insignificant, in sample configurations that are representative of real-world data.

The main contribution of the paper is a detailed account of an efficient computer implementation of the QML-IEKF technique. In this process, we calculate general expressions for the analytical derivatives of the log-likelihood function and the IEKF recursions, including the update step which is only defined implicitly as the solution to a non-linear GLS problem.

1 Introduction

A wide range of asset pricing models are based on the premise that all information about the economy is contained in a finite-dimensional vector of state variables whose dynamics are governed by a Markovian law of motion. Using arguments based on absence of arbitrage, or, alternatively, general equilibrium, asset prices are derived endogenously as functions of the state variables. The exact functional relationship depends on the stochastic process for the state variables and the associated risk premia, as well as the payoff characteristics of the asset, e.g. the time to maturity of a zero-coupon bond. Vasicek (1977) and Cox, Ingersoll and Ross (CIR) (1985b) introduced this framework in the term-structure literature, where the basic idea is a Markovian stochastic process for the instantaneous interest rate (short rate).[1] The latter feature forms a contrast to the Heath, Jarrow and Morton (1992) framework, where the short rate is only Markovian in certain special cases. See, e.g., Ritchken and Sankarasubramanian (1995) for further discussion of this issue.

The present paper deals with econometric techniques for Markovian term-structure models, like the CIR model. In most cases, the data set, containing either prices of coupon bonds, swap rates, or, perhaps, zero-coupon yields, has a "panel data" structure with a time dimension and a cross-sectional (maturity) dimension.[2] By construction, Markovian term-structure models impose joint restrictions on the dynamics and shape of the yield curve. Hence, for efficiency reasons the full data set should be exploited when estimating the unknown model parameters. Moreover, most Markovian term-structure models contain unobserved state variables, such as stochastic mean and volatility factors. These features make the state space setup a natural framework for estimation purposes.[3] If the data consist of zero-coupon yields, and the term-structure model under investigation belongs to the exponential-affine class, the model parameters can be estimated using the linear Kalman filter, see *inter alia* Pennacchi (1991), Jegadeesh and Pennacchi (1996), Chen and Scott (1995), Duan and Simonato (1995), and Lund (1997a).

In several practical applications, there is a non-linear relationship between the observed data and the unobserved state variables. The main examples involve prices of coupon bonds, and non-linear term-structure models, for example the SAINTS model proposed by Constantinides (1992). Estimating such models requires non-linear filtering, and since exact (optimal) filtering techniques tend to be computationally cumbersome, if not outright infeasible, due to a "curse of dimensionality" problem, we are often forced to use approximate methods. However, little is known about the statistical properties of these methods, whether pertaining to filtering of the unobserved state variables, or estimating the (constant) parameters of the model.

[1] General asset pricing models include Merton (1973), Breeden (1979) and CIR (1985a) in the continuous-time setting, and Lucas (1978) in the discrete-time setting.

[2] Note that, in practice, zero-coupon yields are not directly available (in the sense of being traded in the market), but they can be estimated from prices of coupon bonds or swap rates.

[3] Another possibility is the MLE "inversion" approach used by, e.g., Chen and Scott (1993) and Duffie and Singleton (1997), where the m latent variables are expressed as function of yields (or bond prices) for m maturities.

In this paper we consider one of the approximate filtering techniques, the iterative extended Kalman filter (IEKF) and provide two new contributions to the literature. First, we develop a computationally efficient implementation of the IEKF method, and as a key element of this part we calculate analytical derivatives of the (quasi) log-likelihood function. Second, in a Monte Carlo study, we investigate the finite sample properties of the quasi maximum likelihood (QML) estimator for two term-structure models.

The outline of the paper is as follows: in section 2 we introduce the state space form (model), and the associated statistical techniques, while section 3 describes the main examples of non-linear state space models in the term-structure setting. The main focus of the paper, the QML-IEKF method, is presented in section 4, along with a brief discussion of the asymptotic properties of the QML estimator of the model parameters. Sections 5 and 6 contain, respectively, a detailed discussion of the computational aspects of the QML-IEKF method, and the results from the Monte Carlo study. Finally, section 7 offers some concluding remarks.

2 A general framework for the state space form

The data consist of observations sampled at times t_1, t_2, ..., t_n that are not necessarily equally spaced. The observations at time t_k are collected in an $N_k \times 1$ vector, y_k, where the dimension N_k may depend on k. The data generating process (DGP) for y_k is specified in two steps. First, the *measurement* equation is given by:

$$y_k = Z(X_k, t_k;\, \psi) + \varepsilon_k, \tag{1}$$

where $E(\varepsilon_k) = 0$, and X_k is a $m \times 1$ vector of unobserved state variables. In general, we interpret ε_k as a measurement error term, so the function $Z(X_k, t_k;\, \psi)$ is the "theoretical" value of y_k for a given state vector, X_k. For a term-structure model, $Z(\cdot)$ is obtained from the bond-pricing equation, cf. the discussion in section 3.

Second, the dynamics of the unobserved state vector, X_k, are represented by the Markovian *transition* density,

$$p(X_k \,|\, X_{k-1};\, \psi). \tag{2}$$

With a further assumption about the distribution of the measurement errors, for example $\varepsilon_k \sim N(0, H_k(\psi))$ and independent over time, equations (1) and (2) completely specify the DGP for y_k.

2.1 Exact non-linear filtering

The econometric analysis of this model (DGP) can be divided into two separate, yet related, problems:

- Estimate the unobserved state variables, X_k, for $k = 1, 2, \ldots, n$. This is generally referred to as the *filtering* part.

- Estimate the model parameters in the vector ψ, preferably by maximum likelihood estimation (MLE).

The exact filtering recursions, described below, are the optimal solutions to these problems. To facilitate the discussion, let Y_k represent the information available at time t_k,

$$Y_k = (y_1, y_2, \ldots, y_k).$$

We begin by deriving the *prediction* density which is the distribution of X_k given Y_k:

$$\begin{aligned} p(X_k \mid Y_{k-1}) &= \int p(X_k, X_{k-1} \mid Y_{k-1}) dX_{k-1} \\ &= \int p(X_k \mid X_{k-1}) \, p(X_{k-1} \mid Y_{k-1}) dX_{k-1}, \end{aligned} \tag{3}$$

where (by definition) the integration is over the support of X_{k-1}. If the dimension of X is greater than one, the integral is implicitly understood to be a multi-dimensional integral. The optimal predictor of X_k, in a mean-squared-error (MSE) sense, is the conditional mean of X_k given Y_{k-1},

$$E[X_k \mid Y_{k-1}] = \int X_k p(X_k \mid Y_{k-1}) dX_k.$$

In the *update* step we use the additional information contained in y_k to obtain a better estimator of X_k. Since the state space model is non-linear, we must derive the full conditional distribution of X_k given Y_k,

$$\begin{aligned} p(X_k \mid Y_k) &= \frac{p(X_k, y_k \mid Y_{k-1})}{p(y_k \mid Y_{k-1})} \\ &= \frac{p(y_k \mid X_k, Y_{k-1}) \, p(X_k \mid Y_{k-1})}{p(y_k \mid Y_{k-1})} \\ &= \frac{p(y_k \mid X_k) \, p(X_k \mid Y_{k-1})}{p(y_k \mid Y_{k-1})}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} p(y_k \mid Y_{k-1}) &= \int p(y_k, X_k \mid Y_{k-1}) dX_k \\ &= \int p(y_k \mid X_k) \, p(X_k \mid Y_{k-1}) dX_k \end{aligned} \tag{5}$$

In going from the second to the third line of (4), note that once X_k is known, there is no further information about the distribution of y_k in the data history Y_{k-1}. This follows from the Markov property of the state space model (1) and (2). The same property is used in (5).

As in the prediction step, the optimal estimator of X_k is the conditional expectation, now given the larger information set Y_k,

$$E[X_k \mid Y_k] = \int X_k p(X_k \mid Y_k) dX_k.$$

Furthermore, when passing through the non-linear filtering recursions (3) and (4), for $k = 1, 2, \dots, n$, we compute the likelihood function of Y_n as a by-product. To see this, note that by the so-called prediction error decomposition,

$$\log L(y_1, \dots, y_n;\, \psi) = \sum_{k=1}^{n} \log p(y_k \,|\, Y_{k-1};\, \psi),$$

which is obtained directly from (5).

Unfortunately, except for the linear state space model discussed below, and a few other special cases, no closed-form solutions are known for the integrals in (3) and (5). Kitagawa (1987) suggests using numerical integration to compute the respective densities, but numerical integration is probably infeasible (in practice) if the dimension of X is greater than one (multi-factor models). Therefore, the main focus of the present paper is on approximate filtering techniques, especially the IEKF method.

2.2 Two special cases

2.2.1 The linear Gaussian state space model

The linear Gaussian state space model takes the following form:

$$\begin{aligned} y_k &= d_k(\psi) + Z_k(\psi)X_k + \varepsilon_k, \quad \varepsilon_k \sim N(0, H_k(\psi)) && (6) \\ X_k &= \Phi_{k0}(\psi) + \Phi_{k1}(\psi)X_{k-1} + u_k, \quad u_k \sim N(0, V_k(\psi)) && (7) \end{aligned}$$

Compared to the general state space model in the previous section, the measurement equation is linear in X_k, and the dynamics of the state vector are represented by a Gaussian VAR(1) process. The two error terms, ε_k and u_k, are assumed to be mutually independent, and serially uncorrelated. Finally, note that (by linearity) the system matrices $d_k(\psi)$, $Z_k(\psi)$, $H_k(\psi)$, $\Phi_{k0}(\psi)$, $\Phi_{k1}(\psi)$ and $V_k(\psi)$ are independent of the state vector X_k, but they may still vary deterministically over time, e.g. through unequally spaced observations.

Since all error terms in this state space model are normally distributed, the prediction and update densities, (3) and (4), can be shown to be normal (Gaussian) densities. Moreover, the conditional likelihood function (5) is also a Gaussian density. This means that the general filtering recursions for the conditional densities in (3) and (4) can be reduced to simpler recursions for the conditional means and covariance matrices, as the first and second moments completely characterize the normal distribution.

First, following, e.g., Harvey (1989), the prediction step can be represented by the mean recursion,[4]

$$\hat{X}_{k|k-1} = E\left[X_k \,|\, Y_{k-1}\right] = \Phi_{k0} + \Phi_{k1}\hat{X}_{k-1}$$

with mean square error (MSE) matrix

$$\Sigma_{k|k-1} = \Phi_{k1}\Sigma_{k-1}\Phi_{k1}' + V_k$$

[4] To simplify the notation in the following, we suppress the dependence of the system matrices on the parameter vector ψ.

Second, in the update step the additional information contained in y_k is used to obtain a more precise estimator of X_k, namely

$$\hat{X}_k = E\left(X_k \mid Y_k\right) = \hat{X}_{k|k-1} + \Sigma_{k|k-1} Z_k' F_k^{-1} v_k, \tag{8}$$

$$\Sigma_k = \left(\Sigma_{k|k-1}^{-1} + Z_k' H_k^{-1} Z_k\right)^{-1},$$

where

$$\begin{aligned} v_k &= y_k - E\left[y_k \,|\, Y_{k-1}\right] \;=\; y_k - \left(d_k + Z_k \hat{X}_{k|k-1}\right) \\ F_k &= \operatorname{Cov}(v_k) \;=\; Z_k \Sigma_{k|k-1} Z_k' + H_k. \end{aligned}$$

Finally, the log-likelihood function for the data is obtained directly as a by-product of the linear Kalman recursions,

$$\log L(y_1, y_2, \ldots, y_n;\ \psi) = \sum_{k=1}^{n} -\frac{N_k}{2}\log(2\pi) - \frac{1}{2}\log|F_k| - \frac{1}{2} v_k' F_k^{-1} v_k. \tag{9}$$

where $N_k = \dim(v_k)$. To start the Kalman recursions we need initial values of X_0 and Σ_0. If the state vector X_k is stationary, we can use the unconditional mean and covariance matrix of X_k, but another possibility is the diffuse prior approach, see Harvey (1989) for further discussion.

In the term-structure setting, we can use the linear Gaussian framework if

- The term-structure model is Gaussian, such as the one-factor Vasicek (1977) model, or the Beaglehole-Tenney (1991) "double-decay" model, and
- The data consist of zero-coupon yields which are assumed to be observed with measurement error, owing to, e.g., non-synchronous trading, rounding of prices, bid-ask spreads, or simply errors introduced by the particular method used to estimate the zero-coupon yields.

Gaussian models are estimated by the Kalman filter method in Pennacchi (1991), Duan and Simonato (1995) and Lund (1997a).[5]

[5] In the general exponential-affine model [Duffie and Kan (1996)] we obtain the same measurement equation as in (6), since the price of a zero-coupon bond is given by

$$P(t, t+\tau) = \exp\left[A(\tau) + B(\tau)' X_t\right],$$

but the transition dynamics are non-Gaussian, e.g. non-central χ^2 for the CIR model. Chen and Scott (1995) and Duan and Simonato (1995) suggest a QML approach based on the first and second (conditional) moments of the transition density. The resulting transition equation closely resembles (7), except that $V_k(\psi)$ depends linearly on the lagged state vector, X_{k-1}. However, as pointed out by Duan and Simonato (1995) this results in the QML estimator being inconsistent. See Lund (1997a) for further analysis, and a possible solution to this problem within the QML framework. In any case, the biases of the QML estimator appear to be small.

2.2.2 Non-linearity in the measurement equation only

In many cases, the non-linearity of the state space model is limited to the measurement equation:

$$\begin{aligned} y_k &= Z_k(X_k,\, \psi) + \varepsilon_k, \quad \varepsilon_k \sim N(0, H_k(\psi)) \qquad (10)\\ X_k &= \Phi_{k0}(\psi) + \Phi_{k1}(\psi)X_{k-1} + u_k, \quad u_k \sim N(0, V_k(\psi)) \qquad (11)\end{aligned}$$

Although the only difference compared to (6) and (7) is the non-linear transformation of X_k in the measurement equation, the exact filtering algorithm no longer simplifies to recursions for the first and second moments of X_k, and we are left with the general density recursions (3) and (4).

Nonetheless, there are two main reasons for separately discussing the state space model (10)–(11). First, several term-structure models, including the four examples in the next section, are all special cases of this model. Second, because the non-linearity is limited to the measurement equation, it is easier to develop good approximate filtering techniques. For example, the IEKF method is particularly effective in dealing with this type of non-linearity, as demonstrated by Jazwinski (1970).

Second, Frühwirth-Schnatter (1994) proposes a novel technique which is explicitly designed to exploit the structure of (10)–(11). The basic idea of her approach is to approximate the update density by a Gaussian density with the same mean and covariance matrix as the the exact update density (4). Of course, these moments need to be computed by numerical integration, but the dimension of the integration problem has been vastly reduced — in the one-factor case to 3n one-dimensional integrals that need to be computed by quadrature.[6] Evidence reported in Torous and Ball (1995) shows that the method is very effective when estimating a discrete-time log-normal stochastic volatility model.

Unfortunately, the Frühwirth-Schnatter (1994) approach still suffers from a "curse of dimensionality" problem since a m-factor model translates into numerical integration in m dimensions, and even with $m = 2$ this is rather impractical. Consequently, we use the IEKF method in the present paper.

3 Term-structure models in state space form

In this section we describe four examples of term-structure models cast in (non-linear) state space form. The common characteristics are a linear transition equation with Gaussian innovations, combined with a non-linear measurement equation, like the state space model (10)–(11). The first case is described in greatest detail since we use it in the Monte Carlo study in section 6.

[6] Another advantage of the Frühwirth-Schnatter (1994) approach is that we can use any (parametric) distribution for the measurement errors ε_k.

3.1 Estimation of exponential-affine term-structure models using prices of coupon bonds

Exponential-affine models are considerably easier to estimate if the data consist of zero-coupon yields, but such data are rarely available, except perhaps for short-term maturities. Therefore, most studies use zero-coupon yields that are estimated from prices of coupon bonds, for example the Fama-Bliss (1987) or McCulloch-Kwon (1993) data sets. Basically, there are two problems with this approach. First, the synthetic zero-coupon yield data contain less information than the original bond prices. Second, the method used to estimate the zero-coupon yields might introduce biases in the subsequent estimation results. We emphasize that the latter point is a conjecture as we are not aware of any studies relating to this question.

When applying the state space framework directly to bond prices, we get a non-linear measurement equation, where the i'th element is given by the expression:

$$P_i(t_k) \;=\; \sum_{j=1}^{M_i} c_{ij} \cdot \exp\left[A(T_{ij} - t_k;\, \psi) + B(T_{ij} - t_k;\, \psi)' X_k\right] \;+\; \varepsilon_{ik}, \tag{12}$$

where c_{ij} is the j'th payment of the i'th bond which is paid out at time T_{ij}.

The dynamics of the state variables (transition density) can be put in the linear VAR(1) form,

$$X_k = \Phi_{k0}(\psi) + \Phi_{k1}(\psi) X_{k-1} + u_k,$$

where $\Phi_{k0}(\psi)$, $\Phi_{k1}(\psi)$, and the distribution of the innovation u_k depend on the specific exponential-affine model. As already mentioned, we focus on Gaussian models in this paper which means that $u_k \sim N(0, V_k(\psi))$.[7] Langetieg (1980) derives general expressions for the system matrices in the transition equation.

3.2 Non-linear term-structure models

The vast majority of term-structure models with an analytical solution for bond prices belongs to the exponential-affine class. One of the relatively few exceptions is the SAINTS (Squared Autoregressive Instrumental Nominal Term Structure) model proposed by Constantinides (1992). In the SAINTS model, the state variables follow a Gaussian VAR(1) process, but the yield curve implied by the model is a linear-quadratic function of the state variables. This means that the measurement equation will always be non-linear in X_k, even when using zero-coupon yields to estimate the model.

[7] The mechanics of the IEKF method does not rely on specific distributional assumptions for either u_k or ε_k, so the main problem with non-Gaussian models is the fact that the covariance matrix of u_k, $V_k(\psi)$, is an affine function the lagged state vector, X_{k-1}, and that the support of X_k is restricted. These problems are addressed when presenting the IEKF method in section 4.

3.3 Models for pricing defaultable bonds

Claessens and Pennacchi (1996) and Cumby and Evans (1995) develop models for pricing credit risky bonds, in particular Brady bonds. The single state variable is a country "value index" whose technical role is triggering default when hitting zero. The authors show that prices of Brady bonds are a rather complicated non-linear function of this unobserved state variable (the value index). In both papers, QML combined with a non-linear Kalman filtering technique (known as the extended Kalman filter, or EKF) is used to estimate the model parameters.

3.4 A term-structure model with a monetary union (EMU)

One of the factors currently affecting long-term bond prices in Europe is the possible transition to a monetary union since this would eliminate yield spreads between member countries. The issue is quite complicated because of the prevailing uncertainty about the timing of EMU memberships, and possibly whether EMU will be formed in the first place. Lund (1997b) develops a term-structure model which explicitly takes into account the possibility of a monetary union. The model can be estimated using zero-coupon yield spreads to Germany (obtained from the swap market). All state variables in the EMU model are governed by Gaussian processes, but because of the EMU feature there is a non-linear relationship between yield spreads and the underlying state variables. We refer to Lund (1997b; section 3) for further details.

4 The iterative extended Kalman filter (IEKF)

As pointed out in section 2, exact filtering for non-linear state space models is generally considered to be computationally infeasible, except perhaps for one-factor models.[8] Therefore, we turn to approximate filtering techniques, although this move entails three major problems.

First, there is an efficiency loss for the estimator of the unobserved state variables (the filtering problem), and the filtered estimates may be biased as well. Second, the unknown parameters, ψ, of the state space model cannot be estimated by (exact) MLE, as this is inherently tied to the optimal (exact) filtering method. As a by-product of most filtering algorithms we construct a sequence of approximate prediction errors which can be used to form a quasi likelihood function of the Gaussian

[8] Statements along this line are quite prevalent in the econometrics literature, but they are based on the premise that exact filtering has to be done through numerical integration (quadrature). Recent advances in, especially, statistical computing and Markov Chain Monte Carlo (MCMC) methods have demonstrated that taking a Bayesian approach to analyzing non-linear state space models often reduces the computational burden considerably, without incurring the efficiency losses (and other problems) inherently associated with approximate filtering techniques, such as the IEKF method. The main (finance) applications of the MCMC approach are concerned with stochastic volatility models, see Jacquier et al. (1994) and Kim et al. (1996), but recently Frühwirth-Schnatter and Geyer (1996) have used the MCMC method to estimate multi-factor CIR models in the "panel data" framework. However, the Bayesian MCMC approach is outside the scope of the present paper.

form (9). However, as we discuss in section 4.2 below, it is generally not possible to prove that the resulting QML estimator is consistent.

Third, and finally, there are several approximate filtering techniques to choose from, and a priori it is difficult, if not impossible, to know which one is "optimal" for a given problem.[9] Arguably, this is a highly relevant concern for the IEKF method, but formally addressing the problem is outside the scope of the present paper.

4.1 A description of the IEKF algorithm

The state space model has the following form:

$$\begin{aligned} y_k &= Z_k(X_k, \psi) + \varepsilon_k, \quad \varepsilon_k \sim D(0, H_k(\psi)) && (13)\\ X_k &= \Phi_{k0}(\psi) + \Phi_{k1}(\psi) X_{k-1} + u_k, \quad u_k \sim D(0, V_k(\psi)), && (14) \end{aligned}$$

where $D(0, Q)$ refers to an arbitrary zero-mean distribution with covariance matrix Q. Since the IEKF method is based on linear projections, rather than conditional expectations, we do not need specific distributional assumptions for ε_k and u_k. For the present, though, we do assume that they are conditionally homoskedastic.

The filtering recursions of the IEKF method can be divided into a prediction and update step. Both steps provide an estimator of the unobserved state vector and an associated MSE matrix. They are denoted by, respectively, $\hat{X}_{k|k-1}$ and $\Sigma_{k|k-1}$ for the prediction step, and $\hat{X}_k$ and Σ_k for the update step. Since the transition equation is linear, we use the same prediction step as in section 2.2.1,

$$\hat{X}_{k|k-1} = \Phi_{k0} + \Phi_{k1} \hat{X}_{k-1},$$

with MSE matrix

$$\Sigma_{k|k-1} = \Phi_{k1} \Sigma_{k-1} \Phi'_{k1} + V_k.$$

The update step is less straightforward because of the non-linear measurement equation in (13), and the different approximate filtering techniques can primarily be distinguished according to their implementation of the update step. To understand the intuition behind the update step of the IEKF method (below), it is useful to consider an alternative interpretation of the update step for the linear Gaussian state space model. Specifically, Duncan and Horn (1972) show that calculating (8) is equivalent to solving the generalized least squares problem:

$$\begin{aligned} F_L(X) &= \left(X - \hat{X}_{k|k-1}\right)' \Sigma^{-1}_{k|k-1} \left(X - \hat{X}_{k|k-1}\right) + \\ &\quad (y_k - d_k - Z_k X)' H_k^{-1} (y_k - d_k - Z_k X). && (15) \end{aligned}$$

In other words, the update step (8) can be interpreted as a linear projection, whereas in section 2.2.1 it is stated as the conditional expectation of X_k, given Y_k.

[9] See Tanizaki (1996) for an extensive account of non-linear filtering techniques for economic models.

With (15) as the main motivation, the update step for the IEKF method is represented by the non-linear GLS problem:

$$\hat{X}_k = \text{argmin}_X \;\; F_{NL}(X),$$

where

$$\begin{aligned} F_{NL}(X) \;\; &= \;\; \left(X - \hat{X}_{k|k-1}\right)' \Sigma_{k|k-1}^{-1} \left(X - \hat{X}_{k|k-1}\right) \; + \\ & \quad\;\; \left(y_k - Z_k(X)\right)' H_k^{-1} \left(y_k - Z_k(X)\right). \end{aligned} \tag{16}$$

We further define the MSE matrix for $\hat{X}_k$ as,

$$\Sigma_k = \left(\Sigma_{k|k-1}^{-1} + \frac{\partial Z_k(\hat{X}_k)'}{\partial X} H_k^{-1} \frac{\partial Z_k(\hat{X}_k)}{\partial X'}\right)^{-1}, \tag{17}$$

which may be recognized as the (asymptotic) covariance matrix if $\hat{X}_k$ is viewed as a standard parameter estimator in a non-linear GLS setting.

Since (16) needs to be minimized at each time series observation, and for each candidate parameter value ψ, when maximizing the likelihood function, it is extremely important that we use an efficient algorithm. In our experience, the Gauss-Newton algorithm with analytical derivatives is an overall efficient choice. Its iteration scheme is given by:

$$\begin{aligned} \hat{X}_k^{j+1} \;\; &= \;\; \hat{X}_k^j \;-\; \theta^{j+1} \left\{\Sigma_{k|k-1}^{-1} + \frac{\partial Z_k(\hat{X}_k^j)'}{\partial X} H_k^{-1} \frac{\partial Z_k(\hat{X}_k^j)}{\partial X'}\right\}^{-1} \times \\ & \quad\; \left\{\Sigma_{k|k-1}^{-1}(\hat{X}_k^j - \hat{X}_{k|k-1}) \;-\; \frac{\partial Z_k(\hat{X}_k^j)'}{\partial X} H_k^{-1} \left(y_k - Z_k(\hat{X}_k^j)\right)\right\}, \end{aligned} \tag{18}$$

where θ^j is a step length, chosen at the j'th iteration to ensure a decrease in the criterion function $F_{NL}(X)$. As starting value for the Gauss-Newton iterations, we use the previous estimate from the prediction step, $\hat{X}_{k|k-1}$.

With the extended Kalman filter (EKF), see e.g. Harvey (1989), the update step is obtained by linearizing the measurement and transition equations and applying the standard (linear) Kalman filter to the linearized model. For the state space model (13)–(14), the EKF procedure corresponds to just one iteration of (18), starting from $X = \hat{X}_{k|k-1}$. Jazwinski (1970) compares the properties of the IEKF and EKF methods, and concludes that the IEKF method is more effective in dealing with non-linearities in the measurement equation.

What remains to be done is devising a method for estimating the unknown model parameters, ψ. As we have already pointed out, MLE is not an option, and the estimation method for ψ could be completely separated from the non-linear filtering algorithm, at least in principle. Note, however, that since the data are often non-stationary — for example prices of coupon bonds whose stochastic properties change over time due to maturity shortening — we cannot use methods that rely on convergence of unconditional moments (such as GMM).

Instead, we estimate the model parameters by the quasi maximum likelihood (QML) principle. The quasi log-likelihood function is given by:

$$\log L(y_1, \dots, y_n;\ \psi) \;=\; \sum_{k=1}^{n} -\frac{N_k}{2}\log(2\pi) - \frac{1}{2}\log|F_k| - \frac{1}{2}v_k' F_k^{-1} v_k, \tag{19}$$

where

$$v_k \;=\; y_k - Z_k(\hat{X}_{k|k-1}) \tag{20}$$

$$F_k \;=\; \frac{\partial Z_k(\hat{X}_{k|k-1})}{\partial X'}\Sigma_{k|k-1}\frac{\partial Z_k(\hat{X}_{k|k-1})'}{\partial X} + H_k \tag{21}$$

Using the prediction error (20) and its covariance matrix (21) corresponds to linearizing the measurement equation (13) around $X = \hat{X}_{k|k-1}$.

4.2 IEKF with non-Gaussian transition equations

In the following, we make a brief digression and discuss possible generalizations of the IEKF method to conditionally heteroskedastic transition equations, that is state space models where $V_k(\psi)$ depends on the unobserved state vector X_{k-1}.

Duan and Simonato (1995) show that all exponential-affine models can be put in VAR(1) form, like (14), and that the conditional covariance matrix of the innovation, u_k, is an affine function of the (lagged) state vector X_{k-1}. Furthermore, term-structure models such as the CIR model also restrict the support of the state variables, typically to the non-negative part of the real line, and without imposing this restriction there is no guarantee that the covariance matrix of u_k remains positive definite. However, the mechanics of the IEKF update step does not automatically ensure that $\hat{X}_k$ satisfies these restrictions.

There are several modifications of the basic IEKF method that would make it possible to estimate exponential-affine models (in addition to Gaussian models):

- The update step can be modified to minimize (16) subject to the requisite non-negativity conditions. Conceptually, this is probably the best solution. However, the minimization problem in the update step becomes much more complex (and time-consuming), and the same caveat applies to calculating analytical derivatives (cf. section 5) that are often critical to successfully maximizing the quasi log-likelihood function over ψ.

- Duan and Simonato (1995) and Chen and Scott (1995) estimate multi-factor CIR models (with a linear measurement equation as their data consist of zero-coupon yields), and they propose a simpler solution which involves replacing negative values by zero.

- Finally, we can simply ignore the non-negativity restrictions, thus avoiding any new complications in the update step. Of course, we need some modification to keep V_k, and hence F_k, positive definite. One possibility is using the absolute value of the state variables when calculating V_k. Alternative, Lund (1997a) suggests that we use the unconditional covariance matrix of u_k. With a linear measurement equation, the latter suggestion actually ensures that QML is consistent, see Lund (1997a) for further details.

Analyzing the pros and cons of the different approaches is outside the scope of the present paper, and so we concentrate on Gaussian term-structure models.

4.3 Asymptotic properties of the IEKF-QML estimator

It is well known that maximum likelihood (ML) estimators are consistent and asymptotically normally distributed under quite general conditions. However, the results do not apply to the IEKF method since the prediction error v_k entering (19) does not have a conditional normal distribution with mean zero and covariance matrix F_k. In other words, the likelihood function (19) is misspecified.

Fortunately, there is a well-developed statistical theory for misspecified models, known as quasi maximum likelihood (QML) theory, which can be used in the IEKF context. We briefly review the main QML results below, and refer to White (1982), Gallant and White (1988) and White (1994) for an in-depth exposition.

The QML estimator for n observations, $\hat{\psi}_n$, is obtained by maximizing the quasi log-likelihood function:

$$Q_n(\psi) = \frac{1}{n}\sum_{k=1}^{n} l_k(\psi) = \frac{1}{n}\sum_{k=1}^{n} \log L_k(\psi), \tag{22}$$

where $\log L_k(\psi)$ is defined in (19). Following Gallant and White (1988; ch. 3), we define ψ_n^* as the global maximizer of the non-stochastic function

$$\bar{Q}_n(\psi) = \frac{1}{n}\sum_{k=1}^{n} E\left[l_k(\psi)\right] = \frac{1}{n}\sum_{k=1}^{n} \int l_k(\psi)\, dG_k, \tag{23}$$

where G_k is the (true) distribution of the k'th contribution to the likelihood function. This distribution may depend on k, thus allowing for non-stationary data generating processes (DGPs).

Under certain regularity conditions, see Gallant and White (1988), a version of the uniform law of large numbers (ULLN) can be used to show that

$$Q_n(\psi) - \bar{Q}_n(\psi) \;\to\; 0 \quad \text{a.s.} \tag{24}$$

and uniformly in the parameter space Ψ. As a direct consequence of (24), it follows that

$$\hat{\psi}_n - \psi_n^* \;\to\; 0 \quad \text{a.s.} \tag{25}$$

To summarize, (25) establishes that the limiting behavior of the QML estimator $\hat{\psi}_n$ is well-defined, but apart from that the result is somewhat abstract and of limited practical use as the non-stochastic sequence $\{\psi_n^*\}$ is unknown. Moreover, consistency of the QML estimator (in the normal sense) further requires that $\psi_n^* \to \psi_0$, where ψ_0

is the true, but unknown, value of the parameter vector. Bollerslev and Wooldridge (1992) show that a Gaussian QML estimator is consistent if

$$\begin{aligned} E\left[v_k \,|\, Y_{k-1}\right] &= 0 & (26)\\ E\left[v_k v_k' \,|\, Y_{k-1}\right] &= F_k, & (27)\end{aligned}$$

which means that the first and second conditional moments of y_k are correctly specified.[10]

In the QML-IEKF framework, the prediction errors, v_k, correspond to a linearized model, and we cannot expect (26) and (27) to hold because of the approximation error. Therefore, we are unable to prove that the QML estimator is consistent. However, there does not seem to exist a consistent estimation method for the non-linear state space model (13)–(14) which, at the same time, is computationally tractable. For example, the extended Kalman filter (EKF), used in e.g. Claessens and Pennacchi (1996) and Cumby and Evans (1995), suffers from exactly the same problems since the only difference between the two methods is the update step. In any case, we should base our choice of estimation technique on the magnitude of the small sample bias, and this issue is explored with the Monte Carlo study in section 6.

Gallant and White (1988; ch. 5) also derive the asymptotic distribution of the QML estimator. There are two main conditions for proving asymptotic normality of the QML estimator:

- There exists a non-stochastic $O(1)$ (i.e. bounded) sequence of positive definite matrices, $\{B_n^*\}$, such that

  $$B_n^{*\,-1/2} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\partial}{\partial \psi} \log L_k\left(\psi_n^*\right) \;\Rightarrow\; N(0, I), \tag{28}$$

  where $\Rightarrow$ denotes convergence in distribution. Equation (28) says that B_n^* is the asymptotic covariance matrix of the average (normalized) score.

- There exists a non-stochastic sequence of matrices, $\{A_n^*(\psi)\}$, such that

  $$\frac{\partial^2 Q_n(\psi)}{\partial \psi \partial \psi'} - A_n(\psi) \;\rightarrow\; 0 \quad \text{a.s. and uniformly in } \Psi. \tag{29}$$

  This means that there is a well-defined limit (a.s.) for the Hessian of (22). If the DGP is stationary, we further have that $A_n(\psi_n^*) \rightarrow A(\psi^*)$. The asymptotic distribution theory below applies to either case, though.

Under conditions (28) and (29), Gallant and White (1988) show that

$$B_n^{*\,-1/2} A_n^* \sqrt{n} \left(\hat{\psi}_n - \psi_n^*\right) \;\Rightarrow\; N(0, I), \tag{30}$$

[10] Specifically, Bollerslev and Wooldridge (1992) show that ψ_0 is the global minimizer of (23) if the conditions (26) and (27) hold.

where $A_n^* \equiv A_n(\psi_n^*)$. The upshot of (30) is that the covariance matrix of the QML estimator $\hat{\psi}_n$ can be estimated by the formula:

$$\mathrm{Cov}(\hat{\psi}_n) = \frac{1}{n} A_n^{-1}(\hat{\psi}_n) B_n(\hat{\psi}_n) A_n^{-1}(\hat{\psi}_n),$$

where $A_n(\hat{\psi}_n)$ is the Hessian of the log likelihood function,

$$A_n(\hat{\psi}_n) = \frac{1}{n} \sum_{k=1}^{n} \frac{\partial^2}{\partial\psi\partial\psi'} \log L_k\left(\hat{\psi}_n\right),$$

and B_n is a consistent estimator of the covariance matrix of the average (normalized) QML score, cf. (28). In general, the score, $s_k = \partial \log L_k / \partial\psi$, is serially correlated, so we cannot estimate B_n^* by the outer product of the gradient (OPG) formula. Instead, we may use the Newey-West (1987) estimator,

$$B_n(\hat{\psi}_n) = \frac{1}{n} \left\{ \sum_{k=1}^{n} s_k s_k' + \sum_{h=1}^{L} \sum_{k=h+1}^{n} \left(1 - \frac{h}{L+1}\right) \left(s_k s_{k-h}' + s_{k-h} s_k'\right) \right\},$$

or another autocorrelation and heteroskedasticity consistent covariance matrix estimator. See Andrews (1991), Andrews and Monahan (1992), Gallant and White (1988; ch. 6), and Newey and West (1994) for further details.

5 Implementation of the QML-IEKF technique

Prior to computing the quasi log-likelihood function (19) for a candidate value of the parameter vector ψ, we must solve n non-linear GLS problems as the prediction errors v_k entering (19) depend on the updated state vector, $\hat{X}_k$. Consequently, computing the likelihood function is a time-consuming exercise. Moreover, the most effective optimization algorithms require at least first-order derivatives (the gradient) as input, and sometimes we also need the Hessian, i.e. second-order derivatives.

If we compute the gradient by finite differences (numerical derivatives), we have to repeat the n GLS problems each time we perturb the parameter vector. Hence, with p parameters in the vector ψ, the workload increases by a factor of p or $2p$, depending on whether we use single-sided or double-sided derivatives. In addition, the Gauss-Newton iterations for each of the n GLS problems are terminated when some convergence criteria are satisfied, for example when the norm of the gradient of (16) is less then some small value, say 10^{-7}. A small change in the parameter vector ψ could change the number of iterations needed for convergence at observation k, and this would introduce an artificial discontinuity in $\hat{X}_k(\psi)$ which carries over to the quasi log-likelihood function. In situations like this, an optimizer expecting a smooth criterion function could easily get stuck, as Gill et al. (1981) point out.

If we use analytical derivatives for the gradient, we eliminate the above-mentioned problems with artificial discontinuities, and we only have to perform the n GLS minimizations when computing the likelihood function (19), and not when computing the gradient (at the same value of ψ). Both factors should contribute considerably to speeding up the maximization of the likelihood function.

For the linear Kalman filter, cf. section 2.2.1, Harvey (1989) provides expressions for analytical derivatives of the log-likelihood function. Since the IEKF quasi likelihood function depends on $\hat{X}_k$, which is the outcome of a non-linear minimization problem, calculating analytical derivatives seems impossible at first, but in the following we develop a solution to the problem. To our knowledge, this has not been done before. As in Harvey (1989) we set up recursions for the analytical derivatives that run alongside the regular IEKF recursions. Thus, the derivative recursions (below) have prediction and update steps, as well as a part dealing with the k'th contribution to the likelihood function (19). Furthermore, we discuss an optimization algorithm based on either the scoring or the Newton-Raphson algorithm. In the latter case, the Hessian is computed by numerical differentiation of the analytical gradient.

Our new method for calculating analytical derivatives applies to the general non-linear state space model (13)–(14). However, in the remaining part of the paper we focus on the case where the individual measurement errors in the vector ε_k are cross-sectionally independent, and distributed with a common variance, that is

$$H_k = \mathrm{Cov}(\varepsilon_k) = \sigma_\varepsilon^2 I_{N_k} \tag{31}$$

In many applications the dimension N_k of the observation vector y_k is "large", say in excess of 15–20, and N_k varies over time. This makes is difficult to use more elaborate specifications of H_k than (31), especially because we want to keep the dimension of the parameter space at a manageable level. Moreover, there is a significant computational advantage associated with (31) as several key expressions simplify. Basically, the complexity of computing the likelihood function reduces from an $O(nN_k^2)$ to an $O(nN_k)$ operation.

5.1 Derivative recursions for the prediction step

These expressions are completely analogous to the linear Kalman filter, and so we simply restate the results from Harvey (1989, p. 143). The derivatives of $\hat{X}_{k|k-1}$ and $\Sigma_{k|k-1}$ with respect to ψ_i are given by

$$\frac{\partial \hat{X}_{k|k-1}}{\partial \psi_i} = \frac{\partial \Phi_{k0}}{\partial \psi_i} + \frac{\partial \Phi_{k1}}{\partial \psi_i}\hat{X}_{k-1} + \Phi_{k1}\frac{\partial \hat{X}_{k-1}}{\partial \psi_i}, \tag{32}$$

and

$$\frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} = \frac{\partial \Phi_{k1}}{\partial \psi_i}\Sigma_{k-1}\Phi_{k1}' + \Phi_{k1}\frac{\partial \Sigma_{k-1}}{\partial \psi_i}\Phi_{k1}' + \Phi_{k1}\Sigma_{k-1}\frac{\partial \Phi_{k1}}{\partial \psi_i}' + \frac{\partial V_k}{\partial \psi_i}, \tag{33}$$

respectively.

5.2 Derivative recursions for the update step

The derivative recursions for the prediction step involve $\partial\hat{X}_k/\partial\psi_i$ and $\partial\Sigma_k/\partial\psi_i$ from the previous update step $(k-1)$. The difficult part is clearly the first derivative since the functional relationship between $\hat{X}_k$ and ψ is not defined explicitly.

We begin by noting that $\hat{X}_k$ is the minimizer of the function (16) which implies that $\hat{X}_k$ is implicitly defined by:

$$\frac{\partial F_{NL}}{\partial X}(\hat{X}_k) \;=\; 0. \tag{34}$$

In the present case, and because of (31), equation (34) may be reformulated as

$$0 \;=\; \sigma_\varepsilon^2 \Sigma_{k|k-1}^{-1}\left(\hat{X}_k - \hat{X}_{k-1}\right) - \frac{\partial Z_k(\hat{X}_k,\psi)'}{\partial X}\left\{y_k - Z_k(\hat{X}_k,\psi)\right\}. \tag{35}$$

Since (35) holds for any ψ, we can differentiate with respect to ψ_i on both sides of the equation, and solve for $\partial\hat{X}_k/\partial\psi_i$. In this connection, note that the left hand side of (35) is zero for any value of ψ. Furthermore, it is important to recognize that the $N_k \times 1$ vector

$$Z(\hat{X}_k,\psi), \tag{36}$$

and the $N_k \times m$ matrix

$$\frac{\partial Z(\hat{X}_k,\psi)}{\partial X'} \tag{37}$$

depend on ψ in two ways. First, there is the direct dependence through the function argument ψ. Second, the vector $\hat{X}_k$ is itself an implicit function of ψ. Therefore, by the chain rule, the total derivatives of the j'th element/row of (36) and (37) are given by:[11]

$$\frac{\partial Z_{kj}(\hat{X}_k(\psi),\psi)}{\partial\psi_i} \;=\; \frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial\psi_i} + \frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial X'}\frac{\partial\hat{X}_k}{\partial\psi_i} \tag{38}$$

$$\frac{\partial^2 Z_{kj}(\hat{X}_k(\psi),\psi)}{\partial X\partial\psi_i} \;=\; \frac{\partial^2 Z_{kj}(\hat{X}_k,\psi)}{\partial X\partial\psi_i} + \frac{\partial^2 Z_{kj}(\hat{X}_k,\psi)}{\partial X\partial X'}\frac{\partial\hat{X}_k}{\partial\psi_i} \tag{39}$$

[11] When writing

$$\frac{\partial Z_{kj}(\hat{X}_k(\psi),\psi)}{\partial\psi_i}$$

in (38) we mean the total derivative with respect to ψ_i, whereas the notation

$$\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial\psi_i}$$

denotes the derivative with respect to ψ_i for a fixed value of the first function argument $\hat{X}_k$. The same principle applies in other cases, including (39).

After taking derivatives with respect to ψ_i on both sides of (35), and using equations (38) and (39), we arrive at:

$$
\begin{aligned}
0_{m\times 1} \;=\; & \left(\frac{\partial\sigma_\varepsilon^2}{\partial\psi_i}\Sigma_{k|k-1}^{-1} - \sigma_\varepsilon^2\Sigma_{k|k-1}^{-1}\frac{\partial\Sigma_{k|k-1}}{\partial\psi_i}\Sigma_{k|k-1}^{-1}\right)\left(\hat{X}_k - \hat{X}_{k|k-1}\right) \\
& - \sigma_\varepsilon^2\Sigma_{k|k-1}^{-1}\frac{\partial\hat{X}_{k|k-1}}{\partial\psi_i} + \sigma_\varepsilon^2\Sigma_{k|k-1}^{-1}\frac{\partial\hat{X}_k}{\partial\psi_i} \\
& - \sum_{j=1}^{N_k}\left\{\frac{\partial^2 Z_{kj}(\hat{X}_k,\psi)}{\partial X\partial\psi_i} + \frac{\partial^2 Z_{kj}(\hat{X}_k,\psi)}{\partial X\partial X'}\frac{\partial\hat{X}_k}{\partial\psi_i}\right\}\left[y_{kj} - Z_{kj}(\hat{X}_k,\psi)\right] \\
& + \sum_{j=1}^{N_k}\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial X}\left\{\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial\psi_i} + \frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial X'}\frac{\partial\hat{X}_k}{\partial\psi_i}\right\} \qquad (40)
\end{aligned}
$$

The next step is to isolate all terms in (40) involving $\partial\hat{X}_k/\partial\psi_i$, and solve the resulting system of equations, which yields

$$
\frac{\partial\hat{X}_k}{\partial\psi_i} \;=\; C^{-1}(\hat{X}_k,\psi)\,b(\hat{X}_k,\psi), \qquad (41)
$$

where

$$
\begin{aligned}
C(\hat{X}_k,\psi) \;=\; & \sigma_\varepsilon^2\Sigma_{k|k-1}^{-1} \;+\; \sum_{j=1}^{N_k}\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial X}\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial X'} \\
& - \sum_{j=1}^{N_k}\frac{\partial^2 Z_{kj}(\hat{X}_k,\psi)}{\partial X\partial X'}\left[y_{kj} - Z_{kj}(\hat{X}_k,\psi)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
b(\hat{X}_k,\psi) \;=\; & \left(\sigma_\varepsilon^2\Sigma_{k|k-1}^{-1}\frac{\partial\Sigma_{k|k-1}}{\partial\psi_i}\Sigma_{k|k-1}^{-1} - \frac{\partial\sigma_\varepsilon^2}{\partial\psi_i}\Sigma_{k|k-1}^{-1}\right)\left(\hat{X}_k - \hat{X}_{k|k-1}\right) \\
& + \sigma_\varepsilon^2\Sigma_{k|k-1}^{-1}\frac{\partial\hat{X}_{k|k-1}}{\partial\psi_i} \;+\; \sum_{j=1}^{N_k}\frac{\partial^2 Z_{kj}(\hat{X}_k,\psi)}{\partial X\partial\psi_i}\left[y_{kj} - Z_{kj}(\hat{X}_k,\psi)\right] \\
& - \sum_{j=1}^{N_k}\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial X}\frac{\partial Z_{kj}(\hat{X}_k,\psi)}{\partial\psi_i}.
\end{aligned}
$$

Apart from a scaling factor, the matrix $C(\hat{X}_k,\psi)$ can be recognized as the Hessian for the function (16). Since $\hat{X}_k$ is the minimizer of (16), the matrix $C(\hat{X}_k,\psi)$ should be positive definite, and hence invertible, which ensures a well-defined solution in (41).

This completes the difficult part of obtaining analytical derivatives for the update step, and we turn to $\partial\Sigma_k/\partial\psi_i$. With $H_k = \mathrm{Cov}(\varepsilon_k)$ specified as in (31), the MSE matrix Σ_k in (17) can be written as

$$
\Sigma_k \;=\; \sigma_\varepsilon^2\left(\sigma_\varepsilon^2\Sigma_{k|k-1}^{-1} + \frac{Z_k(\hat{X}_k,\psi)'}{\partial X}\frac{Z_k(\hat{X}_k,\psi)}{\partial X'}\right)^{-1} \;\equiv\; \sigma_\varepsilon^2 D_k^{-1}(\hat{X}_k,\psi), \qquad (42)
$$

and the derivative with respect to ψ_i is given by:

$$\frac{\partial \Sigma_k}{\partial \psi_i} = \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} D_k^{-1}(\hat{X}_k, \psi) \; - \; \sigma_\varepsilon^2 D_k^{-1}(\hat{X}_k, \psi) \frac{\partial D_k(\hat{X}_k(\psi), \psi)}{\partial \psi_i} D_k^{-1}(\hat{X}_k, \psi), \tag{43}$$

where

$$\begin{aligned} \frac{\partial D_k(\hat{X}_k(\psi), \psi)}{\partial \psi_i} &= \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} \Sigma_{k|k-1}^{-1} - \sigma_\varepsilon^2 \Sigma_{k|k-1}^{-1} \frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} \Sigma_{k|k-1}^{-1} + \\ & \frac{\partial^2 Z_k(\hat{X}_k(\psi), \psi)'}{\partial X \partial \psi_i} \frac{\partial Z_k(\hat{X}_k, \psi)}{\partial X'} + \frac{\partial Z_k(\hat{X}_k, \psi)'}{\partial X} \frac{\partial^2 Z_k(\hat{X}_k(\psi), \psi)}{\partial X' \partial \psi_i} \end{aligned} \tag{44}$$

The last matrix in (44), with the dimension $N_k \times m$, can be obtained directly from (39).

5.3 Computation of the likelihood function[12]

The k'th contribution to the log-likelihood function is given by:

$$\log L_k(\psi) = -\frac{N_k}{2} \log(2\pi) - \frac{1}{2} \log |F_k| - \frac{1}{2} v_k' F_k^{-1} v_k, \tag{45}$$

where

$$v_k = y_k - Z(\hat{X}_{k|k-1}, \psi) \tag{46}$$

$$\begin{aligned} F_k &= \frac{\partial Z(\hat{X}_{k|k-1}, \psi)}{\partial X'} \Sigma_{k|k-1} \frac{\partial Z'(\hat{X}_{k|k-1}, \psi)}{\partial X} + H_k \\ &\equiv Z_k^* \Sigma_{k|k-1} Z_k^{*'} + H_k \end{aligned} \tag{47}$$

If we substitute $H_k = \sigma_\varepsilon^2 I$ into (47), and use a matrix inversion lemma known as Woodbury's formula, see Harvey (1989), F_k^{-1} simplifies to

$$\begin{aligned} F_k^{-1} &= H_k^{-1} - H_k^{-1} Z_k^* \left(\Sigma_{k|k-1}^{-1} + Z_k^{*'} H_k^{-1} Z_k^* \right)^{-1} Z_k^{*'} H_k^{-1} \\ &= \sigma_\varepsilon^{-2} \left[I - \sigma_\varepsilon^{-2} Z_k^* \left(\Sigma_{k|k-1}^{-1} + \sigma_\varepsilon^{-2} Z_k^{*'} Z_k^* \right)^{-1} Z_k^{*'} \right] \\ &= \sigma_\varepsilon^{-2} \left[I - Z_k^* \left(\sigma_\varepsilon^2 \Sigma_{k|k-1}^{-1} + Z_k^{*'} Z_k^* \right)^{-1} Z_k^{*'} \right], \end{aligned} \tag{48}$$

and the determinant of F_k can be written as

$$\begin{aligned} |F_k| &= |H_k| \cdot |\Sigma_{k|k-1}| \cdot |\Sigma_{k|k-1}^{-1} + Z_k^{*'} H_k^{-1} Z_k^*| \\ &= \sigma_\varepsilon^{2(N_k - m)} \cdot |\Sigma_{k|k-1}| \cdot |\sigma_\varepsilon^2 \Sigma_{k|k-1}^{-1} + Z_k^{*'} Z_k^*| \end{aligned} \tag{49}$$

The next step is to substitute (48) and (49) into (45), which gives

$$v_k' F_k^{-1} v_k = \sigma_\varepsilon^{-2} \left[v_k' v_k - \left(Z_k^{*'} v_k \right)' D_k^{-1}(\hat{X}_{k|k-1}, \psi) \left(Z_k^{*'} v_k \right) \right] \tag{50}$$

$$\log |F_k| = (N_k - m) \log(\sigma_\varepsilon^2) + \log |\Sigma_{k|k-1}| + \log |D_k(\hat{X}_{k|k-1}, \psi)|, \tag{51}$$

[12] The computational approach outlined in sections 5.3–5.4 can also be applied (with advantage) in the linear setting when N_k is much greater than m, and H_k is given by (31).

where

$$D_k(\hat{X}_{k|k-1},\psi) \;=\; \sigma_\varepsilon^2 \Sigma_{k|k-1}^{-1} + Z_k^{*'} Z_k^*$$

is defined as in (42).

In (50) and (51) we only compute inverses and determinants of $m \times m$ matrices, and we totally avoid inverting any $N_k \times N_k$ matrices. Roughly speaking, the number of operations has been reduced from $O(N_k^2)$ to $O(N_k)$.

5.4 Derivatives of the likelihood function

The derivative of the k'th contribution to the log-likelihood function is given by:

$$\frac{\partial \log L_k(\psi)}{\partial \psi_i} = -\frac{1}{2}\frac{\partial}{\partial \psi_i}\log|F_k| - \frac{1}{2}\frac{\partial}{\partial \psi_i} v_k' F_k^{-1} v_k. \tag{52}$$

In the following we provide computationally efficient formulae for each of the two terms in (52).

5.4.1 Derivatives of the first term in (52)

First, note that

$$\frac{\partial \log|A|}{\partial z} = \mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial z}\right), \tag{53}$$

cf. Harvey (1989, p. 140). By applying (53) to the right hand side of (51), we obtain the following:

$$\begin{aligned}\frac{\partial \log|F_k|}{\partial \psi_i} \;&=\; \frac{N_k - m}{\sigma_\varepsilon^2}\frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} \;+\; \mathrm{Tr}\left(\Sigma_{k|k-1}^{-1}\frac{\partial \Sigma_{k|k-1}}{\partial \psi_i}\right) \;+\\ &\quad \mathrm{Tr}\left(D_k^{-1}(\hat{X}_{k|k-1},\psi)\,\frac{\partial D_k(\hat{X}_{k|k-1}(\psi),\psi)}{\partial \psi_i}\right)\end{aligned}$$

The derivative of $\Sigma_{k|k-1}$ has already been calculated in the prediction step, see equation (33) above. We leave out the details regarding the derivative of the $m\times m$ matrix $D(\hat{X}_{k|k-1},\psi)$, as the requisite expression is completely analogous to (44), except that it is evaluated at $X = \hat{X}_{k|k-1}$ instead of $X = \hat{X}_k$.

5.4.2 Derivatives of the second term in (52)

Straightforward calculations give

$$\begin{aligned}-\frac{1}{2}\frac{\partial\, v_k' F_k^{-1} v_k}{\partial \psi_i} \;&=\; -v_k' F_k^{-1}\frac{\partial v_k}{\partial \psi_i} + \frac{1}{2} v_k' F_k^{-1}\frac{\partial F_k}{\partial \psi_i} F_k^{-1} v_k\\ &=\; -w_k'\frac{\partial v_k}{\partial \psi_i} + \frac{1}{2} w_k'\frac{\partial F_k}{\partial \psi_i} w_k,\end{aligned} \tag{54}$$

where

$$w_k \;=\; F_k^{-1} v_k \;=\; \sigma_\varepsilon^{-2} \left[v_k - Z_k^* D_k^{-1}(\hat{X}_{k|k-1}, \psi) \left(Z_k^{*'} v_k \right) \right],$$

and

$$\begin{aligned} \frac{\partial v_k}{\partial \psi_i} \;&=\; -\frac{\partial Z_k(\hat{X}_{k|k-1}(\psi), \psi)}{\partial \psi_i} \\ &=\; -\frac{\partial Z_k(\hat{X}_{k|k-1}, \psi)}{\partial \psi_i} \;-\; \frac{\partial Z_k(\hat{X}_{k|k-1}, \psi)}{\partial X'} \frac{\partial \hat{X}_{k|k-1}}{\partial \psi_i}. \end{aligned} \tag{55}$$

The derivative of the $N_k \times N_k$ matrix F_k is given by:

$$\frac{\partial F_k}{\partial \psi_i} \;=\; \frac{\partial Z_k^*}{\partial \psi_i} \Sigma_{k|k-1} Z_k^{*'} + Z_k^* \frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} Z_k^{*'} + Z_k^* \Sigma_{k|k-1} \frac{\partial Z_k^{*'}}{\partial \psi_i} + \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} I, \tag{56}$$

where

$$\begin{aligned} \frac{\partial Z_k^*}{\partial \psi_i} \;&=\; \frac{\partial^2 Z_k(\hat{X}_{k|k-1}(\psi), \psi)}{\partial X' \partial \psi_i} \\ &=\; \frac{\partial^2 Z_k(\hat{X}_{k|k-1}, \psi)}{\partial X' \partial \psi_i} \;+\; \sum_{j=1}^{m} \frac{\partial^2 Z_k(\hat{X}_{k|k-1}, \psi)}{\partial X' \partial X_j} \frac{\partial \hat{X}_{k|k-1,j}}{\partial \psi_i}. \end{aligned}$$

Finally, if we substitute (56) into the second term in (54), we get

$$\begin{aligned} w_k' \frac{\partial F_k}{\partial \psi_i} w_k \;&=\; \left(\frac{\partial Z_k^{*'}}{\partial \psi_i} w_k \right)' \Sigma_{k|k-1} \left(Z_k^{*'} w_k \right) \;+\; \left(Z_k^{*'} w_k \right)' \frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} \left(Z_k^{*'} w_k \right) \\ &\quad + \; \left(Z_k^{*'} w_k \right)' \Sigma_{k|k-1} \left(\frac{\partial Z_k^{*'}}{\partial \psi_i} w_k \right) \;+\; \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} w_k' w_k, \end{aligned}$$

which only involves $O(N_k)$ operations since we avoid directly computing (56).

5.5 An efficient optimization algorithm

In our experience, the scoring or Newton-Raphson algorithm, combined with a trust region strategy (instead of the usual line search), provides the best overall performance. The iteration scheme is given by:

$$\psi^{i+i} = \psi^i + \left[\mathcal{H}(\psi^i) + \lambda_i I \right]^{-1} \frac{\partial \log L(\psi^i)}{\partial \psi}, \tag{57}$$

where ψ^i is the value of the parameter vector after the i'th iteration, and λ_i is chosen adaptively by the trust region algorithm, see Dennis and Schnabel (1983, 1989) for a detailed discussion. Compared to the line search method, the trust region method is particularly effective in dealing with cases where $\mathcal{H}(\psi)$ is not necessarily positive definite, such as the Newton-Raphson algorithm [Goldfeld et al. (1966)].

The matrix $\mathcal{H}(\psi)$ in (57) is either the Hessian (with the opposite sign),

$$-\frac{\partial^2 \log L(\psi)}{\partial \psi \partial \psi}, \tag{58}$$

or the expected value of this matrix. These cases correspond to, respectively, the Newton-Raphson and scoring algorithms.

The exact Hessian (58) is computed by numerical differentiation of the analytical score (52), whereas the "expected" Hessian is obtained from the approximation[13]

$$-E\left[\frac{\partial^2 \log L(\psi)}{\partial\psi_i\partial\psi_i}\right] \approx \sum_{k=1}^{n} \frac{\partial {v_k}'}{\partial\psi_i} F_k^{-1}\frac{\partial v_k}{\partial\psi_j} + \frac{1}{2}\mathrm{Tr}\left(F_k^{-1}\frac{\partial F_k}{\partial\psi_i}F_k^{-1}\frac{\partial F_k}{\partial\psi_j}\right). \tag{59}$$

Note that the right hand side of (59) only involves first order derivatives of v_k and F_k which are already calculated when computing the gradient of the likelihood function. Furthermore, as we show in the appendix, we can avoid direct computations of the $N_k \times N_k$ matrices in (59), i.e. F_k^{-1} and $\partial F_k/\partial\psi_i$. This means that (59) can be completed in just $O(N_k)$ operations.

Our normal strategy for maximizing the quasi likelihood function is to start with $\mathcal{H}(\psi)$ equal to the expected Hessian, i.e. the scoring algorithm. If the algorithm has not converged after a prespecified number of iterations, we switch to the Newton-Raphson scheme and start computing the Hessian with finite differences of the analytical score. The basic idea is avoiding the expensive Hessian calculations until we are close to the maximum.[14] In this context, it is worth emphasizing that the theoretical advantage of the Newton-Raphson algorithm, namely quadratic convergence, only applies in a small neighborhood of the maximum.

5.6 Analytical derivatives — a worthwhile effort?

Admittedly, deriving analytical expressions for the first-order derivatives (gradient) and the expected Hessian (cf. the appendix) is a time-consuming process, as is the next step of implementing the requisite formulae in a computer program. However, it is important to realize that most of the work is a one-off investment. For example, when estimating exponential-affine models (with prices of coupon bonds as data), only the following parts of the computer program depend on the specific model under investigation:

- The system matrices in the transition equation, Φ_{k0}, Φ_{k1}, and V_k, and the derivatives of these matrices with respect to ψ_i.

- The functions $A(\tau)$ and $B(\tau)$ in the measurement equation (12), as well as derivatives of these functions with respect to ψ_i.

[13] Contrary to the linear case, the right hand side in (59) is only an approximation to the expected Hessian since $E(v_k \mid Y_{k-1}) \neq 0$ because of the linearization error. Of course, this problem does not rule out that (59) is a good candidate for $\mathcal{H}(\psi)$ in the trust region algorithm (57). Specifically, note that (59) is positive definite by construction.

[14] A more elaborate rule for switching between the scoring and Newton-Raphson algorithms could be based on some "estimate" of the distance to the maximum. For example, we could use the reduction in the norm of the gradient at the present iteration relative to an average of the previous iterations. However, the simple rule outlined in the text works quite well, and in many cases convergence is obtained prior to switching to the Newton-Raphson algorithm.

The model-specific derivatives could even be computed by finite differences without any significant loss of speed or accuracy.

Furthermore, there are really two major advantages of using analytical derivatives when maximizing the likelihood function. We have already discussed the first in our introduction to section 5, namely speed and accuracy. It is significantly faster to compute the gradient with analytical derivatives. Second, the optimization algorithm discussed above is only effective when combined with analytical derivatives. Computing the Hessian without an analytical gradient is a very slow process, and although the expected Hessian only involves first-order derivatives of v_k and F_k, we can no longer avoid computing (and multiplying) the $N_k \times N_k$ matrices in (59). Therefore, we are probably better off with optimization algorithms that only require first-order derivatives of $\log L_k(\psi)$, such the BHHH or BFGS (quasi-Newton) methods. However, in our experience, these algorithms tend to require more iterations to achieve convergence than the scoring / Newton-Raphson algorithm.

6 Monte Carlo study of the QML-IEKF method

As discussed in section 4.3, the QML estimator derived from the IEKF technique is not consistent. On the other hand, consistent alternatives seem to require exact filtering, either via numerical integration or MCMC methods, both of which are considerably more time-consuming than the IEKF method. Therefore, if we can demonstrate that the QML estimator performs well in finite samples, including that the biases are sufficiently small and economically insignificant, the IEKF method should still be regarded as useful, the lack of consistency notwithstanding.[15] We investigate the issue in this section, focusing on the case where Gaussian term-structure models are estimated using prices of coupon bonds.

Throughout, the simulated data consist of 1000 time-series observation, each containing 10 bond prices for bullets with maturities of 1–5, 7, 10, 15, 20 and 30 years. In most cases, the sampling frequency is weekly, corresponding to a sample period of about 20 years. The coupon rates are 6% for the 1 and 2 year bonds, 7% for the 3–7 year bonds, and 8% for the remaining bonds. With the parameter values used below, this data specification ensures that the bonds, on average, trade around the par value of 100. The measurement errors, ε_{ik} are independently, normally distributed, $N(0, \sigma_\varepsilon^2)$, where, unless otherwise noted, the standard deviation σ_ε is 0.3, or 30 basis points. For each model and parameter configuration, we use 500 Monte Carlo replications.

[15] One could argue that this pertains to *all* econometric estimators, whether consistency has been demonstrated or not. Clearly, the asymptotic analysis does not apply to the finite sample properties unless the sample is sufficiently large, but what constitutes a "sufficiently large" sample varies from case to case. Sometimes a few hundred observations, or even less, are sufficient, whereas in other cases, even 5000 observations may not be enough. Pritsker (1996) presents an interesting example of the latter case. In general, though, asymptotic properties are a useful starting point that, whenever possible, should be supplemented by Monte Carlo studies.

6.1 Vasicek model

First, we investigate the properties of QML for the one-factor Vasicek model,

$$dr_t = \kappa(\mu - r_t)dt + \sigma dW_t, \tag{60}$$

with a constant market price of risk λ. As shown by Vasicek (1977), the price of a zero-coupon bond is given by

$$P(t, t+\tau) = \exp\left[A(\tau) + B(\tau)r_t\right],$$

where

$$\begin{aligned} B(\tau) &= \frac{e^{-\kappa\tau} - 1}{\kappa}, \\ A(\tau) &= -R(\infty)\left(\tau + B(\tau)\right) - \frac{\sigma^2}{4\kappa}B^2(\tau), \end{aligned}$$

and

$$R(\infty) = \mu - \frac{\lambda\sigma}{\kappa} - \frac{1}{2}\left(\frac{\sigma}{\kappa}\right)^2$$

is the asymptotic interest rate, $\lim_{\tau\to\infty} -\log P(t, t+\tau)/\tau$. Prices of coupon bonds, including bullets, follow in straightforward fashion from (12).

In Table 1, we consider six parameter configurations (cases) for the Vasicek model. The starting point, case 1, has

$$(\kappa, \mu, \sigma, \lambda, \sigma_\varepsilon) = (1.0000,\ 0.0650,\ 0.0300,\ -0.5000,\ 0.3000),$$

which implies that $R(\infty) = 0.0796$. We also explore the possible effect of a higher sampling frequency (daily data) in case 2, and the magnitude of the measurement errors, with σ_ε equal to 10 and 100 basis points (cases 3 and 4, respectively). Finally, in cases 5 and 6, we vary the speed of mean reversion, letting $\kappa = 0.25$ and $\kappa = 2.0$. Here, σ and λ are recalibrated to ensure roughly the same unconditional variance of r_t and asymptotic interest rate, $R(\infty)$, as in case 1.

The results of the Monte Carlo study are displayed in Table 1 where we report the sample mean and standard deviations for 500 parameter estimates (replications). Uniformly across all cases, the average estimates are very close to the true values. In fact, when taking the standard errors and the number of replications into account, there does not appear to be any discernible biases. The market price of risk parameter λ is estimated with least precision, but this is largely due to its correlation with $\hat{\mu}$, and $R(\infty)$ is estimated very precisely.

If we estimate an AR(1) process, like (60), from a univariate time series of r_t, the maximum likelihood estimate of κ tends to be biased upwards in small samples. As in Ball and Torous (1996), there is no such bias when the term-structure model is estimated with a panel data approach. This suggests that most of the information in the data about κ are associated with the cross-sectional properties of the model, that is the shape of the yield curve.

Judging from case 2, the sampling frequency only affects the standard errors for μ and λ. This is to be expected, though, since 4 years of daily data contain less information about the unconditional (long-run) distribution that 20 years of weekly data. On the other hand, the linearization error for v_k, which is really the main cause of inconsistency for the QML estimator, cf. section 4.3, should be smaller for daily data, but in the present case there are no biases in the first place.

The standard deviation of the measurement error, σ_ε, mainly affects the precision of $\hat{\kappa}$. To explain this, note that by increasing σ_ε, the observed yield curve becomes more erratic (less smooth), and hence less informative about κ. Finally, a comparison across cases 1, 5 and 6 shows that the relative precision of $\hat{\kappa}$ is greater for smaller values of κ, whereas exactly the opposite effect occurs for $\hat{\mu}$ and $\hat{\lambda}$. The latter is explained by the fact that weaker mean reversion is equivalent to less time-series information about long-run properties, such as the unconditional mean μ. At the same time, for lower values of κ, a change in r_t has a larger effect on long-maturity bond prices, and this should increase the cross-sectional information content about κ.

6.2 Beaglehole-Tenney "double decay" model

Next, we turn to a two-factor model, originally proposed by Beaglehole and Tenney (1991),

$$\begin{aligned} dr_t &= \kappa_1(\mu_t - r_t)\,dt + \sigma_1 dW_{1t} \\ d\mu_t &= \kappa_2(\theta - \mu_t)\,dt + \sigma_2 dW_{2t}, \end{aligned}$$

where the Brownian motions W_{1t} and W_{2t} are correlated, with ρ denoting the correlation coefficient. The market prices of risk are constant, λ_1 and λ_2. The price of a zero-coupon bond is given by:

$$P(t, t+\tau) = \exp\left[A(\tau) + B(\tau)r_t + C(\tau)\mu_t\right].$$

As the closed-form expressions for $A(\tau)$, $B(\tau)$ and $C(\tau)$ are rather lengthy, we refer to Sørensen (1994) and Jegadeesh and Pennacchi (1996) for the requisite formulae.

Furthermore, since there are now nine parameters and two state variables, performing the IEKF filtering recursions, i.e. solving the non-linear GLS problem (16) for each k, and computing the gradient of the log-likelihood function takes considerably more time than in the one-factor Vasicek case. Therefore, we only consider two parameter configurations in this part of the Monte Carlo study, both assuming weekly data and $\sigma_\varepsilon = 0.30$. The true parameter values are given in Table 2, along with the sample mean and standard deviations of the QML estimates over 500 Monte Carlo replications.

As in Table 1, the results are encouraging for the QML estimator although small biases are noticeable in case I, especially for $\hat{\kappa}_1$ and $\hat{\sigma}_1$. Similar biases do not show up in case II where κ_1 and κ_2 are smaller, corresponding to less mean reversion. Moreover, as argued above, less mean reversion implies that long-term bonds contain more information about the mean reversion parameters via the cross-sectional properties of the term-structure model. Thus, the upward bias for $\hat{\kappa}_1$ in case I most likely reflects the usual small-sample biases that occur when estimating autoregressive parameters, and not the QML-IEKF method as such.

Apart from this minor and economically insignificant problem, there is a close resemblance between Tables 1 and 2 with respect to the performance of the QML estimator. For example, the market prices of risk, λ_1 and λ_2, are estimated somewhat imprecisely in Table 2, but this is clearly caused by a "multicollinearity" problem since the asymptotic interest rate, $R(\infty)$, is estimated very precisely.

7 Concluding remarks

The Monte Carlo evidence presented in section 6 is strongly supportive of the QML-IEKF method as finite sample biases are virtually non-existent, and key model parameters are estimated quite precisely (the risk premia being the usual exception). The positive results should inspire further work in the area, especially in the following directions. First, it would be interesting to study the properties of the QML-IEKF technique for general exponential-affine models, such as multi-factor CIR models. As discussed in section 4.2, non-Gaussian models present additional complications, and the best solution to these problems is, by no means, obvious.

Second, while the present Monte Carlo study has focused on the properties of the QML estimator for the (constant) model parameters, an equally important issue in many applications is the performance of the filtering algorithm. Thus, we should compare the mean squared error (MSE) of the IEKF method to other filtering methods, including (preferably) the optimal filter. If the issue is discussed separately from parameter estimation, a comparison of IEKF and the integration-based optimal filter is clearly computationally feasible, perhaps even for a two-factor model. Another possibility, of course, is adapting the MCMC analysis of Frühwirth-Schnatter and Geyer (1996) to the state space model (10)–(11). In any case, it is something that we leave for future research.

Appendix

In the appendix we show how the expected Hessian (59) can be computed in $O(N_k)$ operations. The first term, involving $\partial v_k / \partial \psi_i$, can be rewritten as

$$\begin{aligned}
\frac{\partial v_k}{\partial \psi_i}' F_k^{-1} \frac{\partial v_k}{\partial \psi_j} &= \sigma_\varepsilon^{-2} \frac{\partial v_k}{\partial \psi_i}' \left[I - Z_k^* \left(\sigma_\varepsilon^2 \Sigma_{k|k-1} + Z_k^{*'} Z_k^* \right)^{-1} Z_k^{*'} \right] \frac{\partial v_k}{\partial \psi_j}' \\
&= \sigma_\varepsilon^{-2} \left\{ \frac{\partial v_k}{\partial \psi_i}' \frac{\partial v_k}{\partial \psi_j} - \left(Z_k^{*'} \frac{\partial v_k}{\partial \psi_i} \right)' \left(\sigma_\varepsilon^2 \Sigma_{k|k-1} + Z_k^{*'} Z_k^* \right)^{-1} \left(Z_k^{*'} \frac{\partial v_k}{\partial \psi_j} \right) \right\}. \qquad (61)
\end{aligned}$$

Computing (61) for all elements of the expected Hessian matrix requires p matrix multiplications

$$Z_k^{*'} \frac{\partial v_k}{\partial \psi_i}, \quad \text{for } i = 1, 2, \ldots, p,$$

where p is the number of parameters, and $p(p+1)/2$ inner products of the form

$$\frac{\partial v_k}{\partial \psi_i}' \frac{\partial v_k}{\partial \psi_i}, \quad \text{for } i = 1, 2, \ldots, p, \ j = i, \ldots, p$$

Note that the vector $\partial v_k / \partial \psi_i$ follows from the gradient calculation in section 5.4, cf. equation (55).

Speeding up the computation of the second term,

$$\operatorname{Tr} \left(F_k^{-1} \frac{\partial F_k}{\partial \psi_i} F_k^{-1} \frac{\partial F_k}{\partial \psi_j} \right), \qquad (62)$$

is somewhat more involved. After rather lengthy calculations (expanding and collecting terms) we obtain the following intermediate result:

$$\begin{aligned}
F_k^{-1} \frac{\partial F_k}{\partial \psi_i} &= \sigma_\varepsilon^{-2} \left[I - Z_k^* \left(\sigma_\varepsilon^2 \Sigma_{k|k-1} + Z_k^{*'} Z_k^* \right)^{-1} Z_k^{*'} \right] \times \\
&\quad \left\{ Z_k^* \frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} Z_k^{*'} + \frac{\partial Z_k^*}{\partial \psi_i} \Sigma_{k|k-1} Z_k^{*'} + Z_k^* \Sigma_{k|k-1} \frac{\partial Z_k^{*'}}{\partial \psi_i} + \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} I \right\} \\
&= \sigma_\varepsilon^{-2} \left\{ \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} I + Z_k^* A_{i,1k} Z_k^{*'} + Z_k^* A_{i,2k} \frac{\partial Z_k^{*'}}{\partial \psi_i} + \frac{\partial Z_k^*}{\partial \psi_i} \Sigma_{k|k-1} Z_k^{*'} \right\}, \qquad (63)
\end{aligned}$$

where

$$\begin{aligned}
A_{i,1k} &= \frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} - \left(\sigma_\varepsilon^2 \Sigma_{k|k-1} + Z_k^{*'} Z_k^* \right)^{-1} \times \\
&\quad \left\{ Z_k^{*'} Z_k^* \frac{\partial \Sigma_{k|k-1}}{\partial \psi_i} + Z_k^{*'} \frac{\partial Z_k^*}{\partial \psi_i} \Sigma_{k|k-1} + \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} I_m \right\},
\end{aligned}$$

and

$$A_{i,2k} = \Sigma_{k|k-1} - \left(\sigma_\varepsilon^2 \Sigma_{k|k-1} + Z_k^{*'} Z_k^* \right)^{-1} Z_k^{*'} Z_k^* \Sigma_{k|k-1}$$

Next, we multiply (63) for i and j, respectively, and after rearranging the result in the same form as (63), we get

$$\begin{aligned}
F_k^{-1}\frac{\partial F_k}{\partial \psi_i}F_k^{-1}\frac{\partial F_k}{\partial \psi_j} &= \sigma_\varepsilon^{-4} \times \\
&\left\{ \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i}\frac{\partial \sigma_\varepsilon^2}{\partial \psi_j} I_{N_k} + Z_k^* C_{ij,1k} Z_k^{*'} + \frac{\partial Z_k^*}{\partial \psi_j} C_{i,2k} Z_k^{*'} + Z_k^* C_{ij,3k} \frac{\partial Z_k^{*'}}{\partial \psi_j} \right. \\
&\left. + \; Z_k^* C_{ij,4k} \frac{\partial Z_k^{*'}}{\partial \psi_i} + \frac{\partial Z_k^*}{\partial \psi_i} C_{j,5k} Z_k^{*'} + \frac{\partial Z_k^*}{\partial \psi_i} C_{j,6k} \frac{\partial Z_k^{*'}}{\partial \psi_j} \right\}, \qquad (64)
\end{aligned}$$

where

$$\begin{aligned}
C_{ij,1k} &= \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} A_{j,1k} + \frac{\partial \sigma_\varepsilon^2}{\partial \psi_j} A_{i,1k} + A_{i,1k} Z_k^{*'} Z_k^* A_{j,1k} + A_{i,1k} Z_k^{*'} \frac{\partial Z_k^*}{\partial \psi_j} \Sigma_{k|k-1} \\
&+ \; A_{i,2k} \frac{\partial Z_k^{*'}}{\partial \psi_i} Z_k^* A_{j,1k} + A_{i,2k} \frac{\partial Z_k^{*'}}{\partial \psi_i} \frac{\partial Z_k^*}{\partial \psi_j} \Sigma_{k|k-1} \\
C_{i,2k} &= \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} \Sigma_{k|k-1} \\
C_{ij,3k} &= \frac{\partial \sigma_\varepsilon^2}{\partial \psi_i} A_{j,2k} + A_{i,1k} Z_k^{*'} Z_k^* A_{j,2k} + A_{i,2k} \frac{\partial Z_k^{*'}}{\partial \psi_i} Z_k^* A_{j,2k} \\
C_{ij,4k} &= \frac{\partial \sigma_\varepsilon^2}{\partial \psi_j} A_{i,2k} \\
C_{j,5k} &= \frac{\partial \sigma_\varepsilon^2}{\partial \psi_j} \Sigma_{k|k-1} + \Sigma_{k|k-1} Z_k^{*'} Z_k^* A_{j,1k} + \Sigma_{k|k-1} Z_k^{*'} \frac{\partial Z_k^*}{\partial \psi_j} \Sigma_{k|k-1} \\
C_{j,6k} &= \Sigma_{k|k-1} Z_k^{*'} Z_k^* A_{j,2k}
\end{aligned}$$

The last six elements (matrices) in the sum within the braces in (64) have a common structure. Specifically, each element is a $N_k \times N_k$ matrix of the form

$$Z_1 C Z_2' \qquad (65)$$

where Z_1 and Z_2 are $N_k \times m$ matrices, and the dimension of the middle matrix, C, is $m \times m$. The trace of (65) is given by

$$\mathrm{Tr}\,[Z_1 C Z_2'] = \mathrm{Tr}\,[C\,(Z_2' Z_1)]\,,$$

where the equality is obtained from a property of the matrix trace operator. In the last expression we apply the trace operator to the product of C and $Z_2' Z_1$, each of which are $m \times m$ matrices. When N_k is much larger than m, this is considerably faster than computing the trace of (65) directly.

Finally, by applying this idea in (64), we get

$$\begin{aligned}
\mathrm{Tr}\left(F_k^{-1}\frac{\partial F_k}{\partial\psi_i}F_k^{-1}\frac{\partial F_k}{\partial\psi_j}\right) \;&=\; \sigma_\varepsilon^{-4}\times \\
&\left\{N_k\frac{\partial\sigma_\varepsilon^2}{\partial\psi_i}\frac{\partial\sigma_\varepsilon^2}{\partial\psi_j} + \mathrm{Tr}\left[C_{ij,1k}Z_k^{*'}Z_k^*\right] + \mathrm{Tr}\left[C_{i,2k}Z_k^{*'}\frac{\partial Z_k^*}{\partial\psi_j}\right] + \mathrm{Tr}\left[C_{ij,3k}\frac{\partial Z_k^{*'}}{\partial\psi_j}Z_k^*\right]\right. \\
&\left. + \;\mathrm{Tr}\left[C_{ij,4k}\frac{\partial Z_k^{*'}}{\partial\psi_i}Z_k^*\right] + \mathrm{Tr}\left[C_{j,5k}Z_k^{*'}\frac{\partial Z_k^*}{\partial\psi_i}\right] + \mathrm{Tr}\left[C_{j,6k}\frac{\partial Z_k^{*'}}{\partial\psi_j}\frac{\partial Z_k^*}{\partial\psi_i}\right]\right\}
\end{aligned}$$

In summary, we have simplified (62), which is given by the trace of a product four $N_k\times N_k$ matrices, to applying the trace operator to a series of $m\times m$ matrices, each of which can be computed in $O(N_k)$ operations.

References

Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, D.W.K. and J.C. Monahan (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–967.

Ball, C.A. and W.N. Torous (1996), "Unit Roots and the Estimation of Interest Rate Dynamics," *Journal of Empirical Finance*, 3, 215–238.

Beaglehole, D.R. and M.S. Tenney (1991), "General Solutions of Some Interest Rate-Contingent Claim Pricing Equations," *Journal of Fixed Income*, 1, Sept., 69–83.

Bollerslev, T. and J.M. Wooldridge (1992), "Quasi-Maximum Likelihood Estimation of Dynamic Models with Time-Varying Covariances," *Econometric Reviews*, 11, 143–172.

Breeden, D.T. (1979), "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, 7, 265–296.

Chen, R.R. and L. Scott (1993), "Maximum Likelihood Estimation for a Multifactor General Equilibrium Model of the Term Structure of Interest Rates," *Journal of Fixed Income*, 3, December, 14–31.

Chen, R.R. and L. Scott (1995), "Multi-Factor Cox-Ingersoll-Ross Models of the Term Structure: Estimates and Tests from a Kalman Filter Model," Manuscript, University of Georgia.

Claessens, S. and G.G. Pennacchi (1996), "Estimating the Likelihood of Mexican Default from Market Prices of Brady Bonds," *Journal of Financial and Quantitative Analysis*, 31, 109–126.

Constantinides, G.M. (1992), "A Theory of the Nominal Term Structure of Interest Rates," *Review of Financial Studies*, 5, 531–552.

Cox, J.C., J.E. Ingersoll and S.A. Ross (1985a), "An Intertemporal General Equilibrium Model of Asset Prices," *Econometrica*, 53, 363–384.

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

Cumby, R.E. and M.D.D. Evans (1995), "The Term Structure of Credit Risk: Estimates and Specification Tests," Manuscript, Georgetown University.

Dennis, J.E., Jr., and R.B. Schnabel (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ.

Dennis, J.E., Jr., and R.B. Schnabel (1989), "A View of Unconstrained Optimization," in Nemhauser, G.L.. A.H.G. Rinnooy Kan, and M.J. Todd, editors, *Handbook of Operations Research and Management Science, Volume I*, North Holland, Amsterdam.

Duan, J.C. and J.G. Simonato (1995), "Estimating and Testing Exponential-Affine Term Structure Models by Kalman Filter," Manuscript, McGill University.

Duffie, D. and R. Kan (1996), "A Yield Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

Duffie, D. and K.J. Singleton (1997), "An Econometric Model of the Term Structure of Interest Rate Swap Yields," Manuscript, Graduate School of Business, Stanford University. Forthcoming, *Journal of Finance*, September 1997.

Duncan, D.B. and S.D. Horn (1972), "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," *Journal of the American Statistical Association*, 67, 815–821.

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

Frühwirth-Schnatter, S. (1994), "Applied State Space Modelling of Non-Gaussian Time Series Using Integration-Based Kalman Filtering," *Statistics and Computing*, 4, 259–269.

Frühwirth-Schnatter, S. and A.L.J. Geyer (1996), "Bayesian Estimation of Econometric Multi-Factor Cox-Ingersoll-Ross Models of the Term Structure of Interest Rates via MCMC Methods," Manuscript, Vienna University of Economics and Business Administration.

Gallant, A.R. and H. White (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell, Oxford.

Gill, P.E., W. Murray and M.H. Wright (1981), *Practical Optimization*, Academic Press, New York.

Goldfeld, S., R. Quandt and H. Trotter (1966), "Maximization by Quadratic Hill Climbing," *Econometrica*, 34, 541–551.

Harvey, A.C. (1989), *Forecasting, Structural Models and the Kalman Filter*, Cambridge University Press, New York.

Heath, D., R. Jarrow and A. Morton (1992), "Bond Pricing and the Term Structure of Interest Rates," *Econometrica*, 60, 77–105.

Jacquier, E., N.G. Polson and P.E. Rossi (1994), "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics*, 12, 371–389.

Jazwinski, A.H. (1970), *Stochastic Processes and Filtering Theory*, Academic Press, New York.

Jegadeesh, N. and G.G. Pennacchi (1996), “The Behavior of Interest Rates Implied by the Term Structure of EuroDollar Futures,” *Journal of Money, Credit and Banking*, 28, 426–446.

Kim, S., N. Shephard and S. Chib (1996), “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models,” Manuscript, Nuffield College, Oxford.

Kitagawa, G. (1987), “Non-Gaussian State Space Modeling of Nonstationary Time Series,” *Journal of the American Statistical Association*, 82, 1032–1063.

Langetieg, T.C. (1980), “A Multivariate Model of the Term Structure,” *Journal of Finance*, 35, 71–97.

Lucas, R.E. (1978), “Asset Prices in an Exchange Economy,” *Econometrica*, 46, 1429–1445.

Lund, J. (1997a), “Econometric Analysis of Continuous-Time Arbitrage-Free Models of the Term Structure of Interest Rates,” Manuscript, Department of Finance, Aarhus School of Business.

Lund, J. (1997b), “A Model for Studying the Effect of EMU on European Yield Curves,” Manuscript, Department of Finance, Aarhus School of Business.

McCulloch, J.H. and H.C. Kwon (1993), “U.S. Term Structure Data 1947–1991,” Manuscript, Department of Economics, Ohio State University.

Merton, R.C. (1973), “An Intertemporal Capital Asset Pricing Model,” *Econometrica*, 41, 867–887.

Newey, W. and K.D. West (1987), “A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, 55, 703–708.

Newey, W. and K.D. West (1994), “Automatic Lag Selection in Covariance Matrix Estimation,” *Review of Economic Studies*, 61, 631–653.

Pennacchi, G.G. (1991), “Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data,” *Review of Financial Studies*, 4, 53–86.

Pritsker, M. (1996), “Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models,” Manuscript, Federal Reserve Board, Washington, DC.

Ritchken, P. and L. Sankarasubramanian (1995), “Volatility Structure of Forward Rates and the Dynamics of the Term Structure,” *Mathematical Finance*, 5, 55–72.

Sørensen, C. (1994), "Option Pricing in a Gaussian Two-Factor Model of the Term Structure of Interest Rates," Manuscript, Institute of Finance, Copenhagen Business School.

Tanizaki, H. (1996), *Nonlinear Filters: Estimation and Applications, 2nd Edition*, Springer Verlag, New York.

Torous, W.N. and C.A. Ball (1995), "Regime Shifts in Short-Term Riskless Interest Rates", Manuscript, Owen Graduate School of Business, Vanderbilt University.

Vasicek, O.A. (1977), "An Equilibrium Characterization of the Term Structure of Interest Rates," *Journal of Financial Economics*, 5, 177–188.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, New York, NY.

Table 1:
Results of Monte Carlo study
Vasicek one-factor model

Case No.		Parameter					
(Freq.)		σ_ε	κ	μ	σ	λ	$R(\infty)$
1	True	0.3000	1.0000	0.0650	0.0300	-0.5000	0.0796
(50)	Mean	0.2999	1.0004	0.0650	0.0300	-0.5016	0.0795
	S.e.	[0.0022]	[0.0120]	[0.0055]	[0.0008]	[0.1863]	[0.0000]
2	True	0.3000	1.0000	0.0650	0.0300	-0.5000	0.0796
(250)	Mean	0.2999	1.0006	0.0655	0.0300	-0.4848	0.0795
	S.e.	[0.0021]	[0.0142]	[0.0107]	[0.0010]	[0.3583]	[0.0000]
3	True	0.1000	1.0000	0.0650	0.0300	-0.5000	0.0796
(50)	Mean	0.1000	1.0000	0.0649	0.0300	-0.5032	0.0795
	S.e.	[0.0007]	[0.0040]	[0.0055]	[0.0006]	[0.1861]	[0.0000]
4	True	1.0000	1.0000	0.0650	0.0300	-0.5000	0.0796
(50)	Mean	0.9997	1.0037	0.0653	0.0300	-0.4928	0.0795
	S.e.	[0.0071]	[0.0402]	[0.0056]	[0.0016]	[0.1879]	[0.0000]
5	True	0.3000	0.2500	0.0650	0.0150	-0.3000	0.0812
(50)	Mean	0.2999	0.2500	0.0643	0.0150	-0.3123	0.0812
	S.e.	[0.3000]	[0.0010]	[0.0109]	[0.0003]	[0.1829]	[0.0000]
6	True	0.3000	2.0000	0.0650	0.0400	-0.8000	0.0808
(50)	Mean	0.3000	2.0069	0.0651	0.0401	-0.7992	0.0808
	S.e.	[0.3000]	[0.0038]	[0.0038]	[0.0017]	[0.1954]	[0.0000]

Notes: The simulated samples consist of 1000 time series observations, each containing 10 bond prices, all bullets with maturities of 1–5, 7, 10, 15, 20 and 30 years.

The number in parenthesis below the Case No. corresponds to the sampling frequency of the data (time series observations per year).

For each case/parameter we report three numbers. The true value of the parameter is displayed in the first line. The second and third lines contain, respectively, the sample mean and standard error (in brackets) of the QML estimates from 500 Monte Carlo replications.

Table 2:
Results of Monte Carlo study
Double-Decay two-factor model

	Case I		Case II	
Parameter	True value	Mean [Std.Err]	True value	Mean [Std.Err]
σ_ε	0.3000	0.2998 [0.0023]	0.3000	0.2997 [0.0023]
κ_1	2.0000	2.0549 [0.2419]	1.2500	1.2525 [0.0464]
κ_2	0.3000	0.2999 [0.0037]	0.1000	0.1000 [0.0008]
θ	0.0700	0.0691 [0.0085]	0.070	0.0697 [0.0167]
σ_1	0.0300	0.0304 [0.0029]	0.0200	0.0199 [0.0011]
σ_2	0.0100	0.0100 [0.0005]	0.0100	0.0100 [0.0003]
ρ	0.5000	0.4986 [0.0671]	-0.2500	-0.2483 [0.0475]
λ_1	-0.4000	-0.4216 [0.2234]	-0.3000	-0.2921 [0.2172]
λ_2	-0.1000	-0.1179 [0.1947]	-0.1000	-0.1041 [0.1734]
$R(\infty)$	0.0784	0.0784 [0.0000]	0.0801	0.0801 [0.0001]

Notes: The simulated samples consist of 1000 time series observations (at the weekly frequency), each containing 10 bond prices, all bullets with maturities of 1–5, 7, 10, 15, 20 and 30 years.

The sample mean and standard errors (in brackets) are computed over 500 Monte Carlo replications.