Using a Nonlinear Filter to Estimate a Multifactor Term Structure Model with Gaussian Mixture Innovations

Preliminary Version.
Comments welcome!

Wolfgang Lemke*

May 30, 2005

Abstract

This paper proposes a multifactor term structure model with factor innovations that have a Gaussian mixture distribution. The model allows for flexible modeling of the distribution of bond yields. Under the condition of no-arbitrage, yields are affine functions of factors. The model is estimated in a state space framework using a new nonlinear filtering algorithm. Estimation results for US data show that the mixture model is able to capture nonnormality in bond yield changes.

1 Introduction

The term structure of interest rates is a subject of interest in macroeconomics and finance alike. Learning about the nature of bond yield dynamics and its driving forces is important in different areas such as forecasting, monetary policy, debt policy, and derivative pricing.[1] Affine term structure models[2] simultaneously capture the dynamic and the cross-section properties of bond yields

*Deutsche Bundesbank, E-mail: wolfgang.lemke@bundesbank.de. This work is based on my Ph.D. thesis written at Bielefeld University. The paper represents the author's personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

[1] See Piazzesi (2005).

[2] See Duffie and Kan (1996) and Backus, Foresi, and Telmer (1998) for the discrete-time version.

while constraining the family of bond price processes to be arbitrage-free. The term 'affine' is due to the fact that bond yields are affine functions of a limited number of factors.

The stochastic properties of the factor process are inherited by bond yields. If, for instance, the factor process is a Gaussian VAR, bond yields of all maturities will be Gaussian as well. However, there is empirical evidence that bond yields and their first differences are not normally distributed. This paper provides an approach for a flexible modeling of the distribution of bond yields while staying within the class of affine models. Based on an idea in Backus et al. (1998) this is achieved by allowing factor innovations to be distributed as Gaussian mixtures. We derive an analytical formula for bond yields as a function of factors. The model allows the resulting distribution of yields and yield changes to assume a wide variety of shapes. In particular, it can account for non-vanishing skewness and excess kurtosis that varies with maturity.

For the estimation of multifactor term structure models using time series data of yields for different maturities, it has become common in the literature to translate the theoretical model into the statistical state space form. The measurement vector contains a set of bond yields for different maturities, the state vector represents the latent factors driving the term structure. The state space framework allows for adding measurement errors to the bond pricing equations. It permits to estimate unknown parameters and to filter out the unobservable factor processes. For affine multifactor Gaussian models, the corresponding state space model is linear and Gaussian. Hence, it can be estimated by maximum likelihood based on the Kalman filter. The literature contains numerous examples for this approach, e.g. De Jong (2000) and Cassola and Luis (2003).

For the class of term structure models considered in this paper, the corresponding state space model has a transition equation for which the innovation is distributed as a Gaussian mixture. As it is already shown by Sorenson and Alspach (1971), the exact filter for such a state space model is nonlinear in observations. Moreover, the exact filtering density at time t is a Gaussian mixture for which the number of components is exponentially growing with time. For instance, if the state innovation is distributed as a mixture of 2 normal distributions, the exact filtering density at time t contains 2^t component densities, rendering a practical application of the exact filter impossible. To deal with this problem we propose an approximate filter that preserves the nonlinearity of the exact solution but that restricts the number of components in the mixture distributions involved. The degree of complexity is controlled by a parameter k, so that in the example the true filtering density at time t is approximated by a mixture of 2^k densities only.

A two-factor term structure model with Gaussian mixture innovations is estimated with US data using the approximate nonlinear filter. The data set

contains time series of monthly yields for five different maturities. For comparison, we also estimate pure Gaussian models using the Kalman filter. Estimating the distribution of differenced yields from the data and comparing it to the distributions implied by the models, it turns out that the mixture model is superior compared to the Gaussian models.

The paper is organized as follows. Section 2 presents the data set and derives some stylized facts. In section 3, the term structure model is developed and the yield equation is derived. In section 4, the estimation approach is described which is employed for the empirical application in section 5. Section 6 concludes, the appendix provides details of the exact filter algorithm and our approximation.

2 Data and Stylized Facts

In this section we introduce the data set that will be used for the empirical application below. It also serves to derive some stylized facts that will motivate the term structure model in the following section. The data set is based on McCulloch and Kwon (1993) and Bliss (1997). It is the same set as used by Duffee (2002).[3] It consists of monthly observations of annual zero bond yields for the period of January 1962 to December 1998. The sample contains yields for maturities of 3, 6, 12, 24, 60 and 120 months. Thus, we have 6 time series of 444 observations each. Three of the six time series are graphed in figure 1, table 1 provides summary statistics of the data.

Mat	Mean	Std Dev	Skew	Kurt	Auto Corr
3	6.32	2.67	1.29	1.80	0.974
6	6.56	2.70	1.23	1.60	0.975
12	6.77	2.68	1.12	1.24	0.976
24	7.02	2.59	1.05	1.02	0.978
60	7.36	2.47	0.95	0.68	0.983
120	7.58	2.40	0.78	0.31	0.987

Table 1: Summary statistics of yields in levels. For each time to maturity (Mat) the columns contain mean, standard deviation, skewness, excess kurtosis, and autocorrelation at lag 1.

As table 1 shows, yields at all maturities are highly persistent. The mean increases with time to maturity. Ignoring the three-month yield, the standard deviation falls with maturity. For interpreting the coefficient of skewness and

[3] We obtained it from G. R. Duffee's website http://faculty.haas.berkeley.edu/duffee/affine.htm.

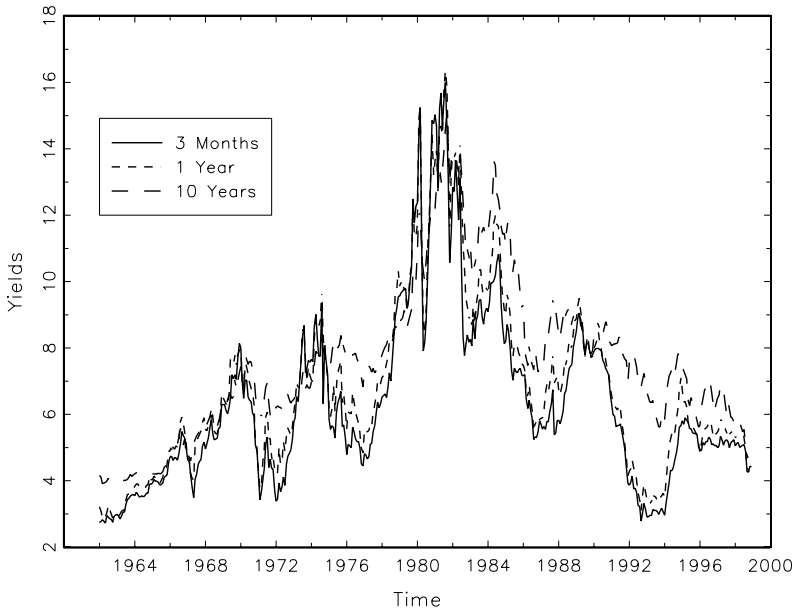

Figure 1: Yields from 01/1962 - 12/1998

Mat	3	6	12	24	60	120
3	1.000					
6	0.996	1.000				
12	0.986	0.995	1.000			
24	0.962	0.975	0.990	1.000		
60	0.909	0.924	0.950	0.982	1.000	
120	0.862	0.878	0.908	0.952	0.991	1.000

Table 2: Correlation of yields in levels

excess kurtosis, note that they should be close to zero if the data were normally distributed.

The means of yields are graphed against the corresponding maturity in figure 2. Data are represented by filled circles. The connecting lines are drawn for optical convenience only. The picture shows that the mean yield curve has a concave shape: mean yields increase with maturity, but the increase becomes smaller as one moves along the abscissa. This is a typical shape for the *mean* yield curve. However, the shape of the yield curve observed from day to day can assume a variety of shapes. It may be inverted, i.e. monotonically decreasing, or contain 'humps'.

Finally, table 2 shows that yields exhibit a high contemporaneous correlation at all maturities. That is, interest rates of different maturities tend to move together.

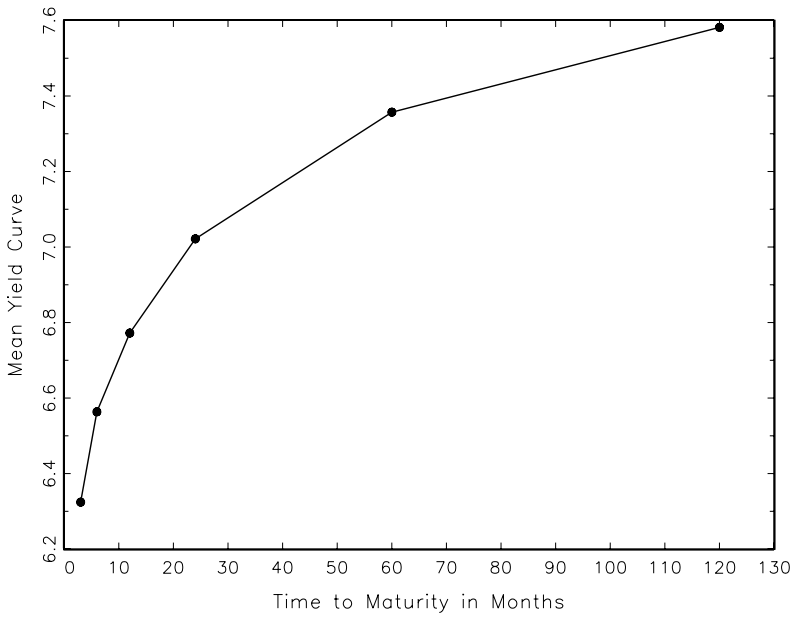

Figure 2: Mean yield curve

We now turn from levels to yields in first differences. That is, if $\{y_1^{n_i}, \ldots, y_T^{n_i}\}$ denotes an observed time series of the n_i-month yield in levels, we now consider the corresponding time series $\{\Delta y_2^{n_i}, \ldots, \Delta y_T^{n_i}\}$ with $\Delta y_t^{n_i} = y_t^{n_i} - y_{t-1}^{n_i}$.

Three of the six time series are graphed in figure 3. Table 3 shows summary statistics of yields in first differences. Again, the standard deviation falls with

Mat	Mean	Std Dev	Skew	Kurt	Auto Corr
3	0.0038	0.58	-1.80	14.32	0.115
6	0.0034	0.57	-1.66	15.76	0.155
12	0.0030	0.56	-0.77	12.31	0.158
24	0.0024	0.50	-0.36	10.35	0.146
60	0.0016	0.40	0.12	4.04	0.096
120	0.0015	0.33	-0.11	2.29	0.087

Table 3: Summary statistics of yields in first differences

time to maturity. The high autocorrelation that we have observed for yields in levels has vanished. Skewness is still moderate but excess kurtosis is vastly exceeding zero. Moreover, excess kurtosis differs with maturity having a general tendency to decrease with it. This leads to the interpretation that especially at the short end of the term structure, extreme observations occur much more often as being compatible with the assumption of a normal distribution.

The contemporaneous correlation of differenced yields is also high, as evident

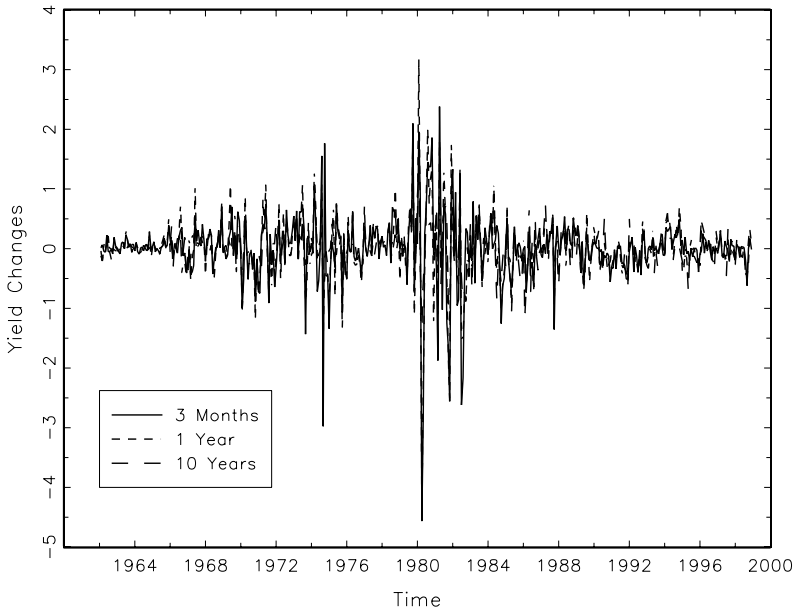

Figure 3: First difference of yields

from table 4. However, the correlations are consistently lower than for yields in levels.

Mat	3	6	12	24	60	120
3	1.000					
6	0.952	1.000				
12	0.867	0.957	1.000			
24	0.783	0.887	0.960	1.000		
60	0.645	0.762	0.859	0.936	1.000	
120	0.547	0.659	0.742	0.830	0.934	1.000

Table 4: Correlation of yields in first differences

3 The Model

We introduce a dynamic multifactor model in discrete time. Let P_t^n denote the price at time t of a zero-coupon bond that pays one unit of account at time $t+n$. The corresponding yield to maturity is given by

$$y_t^n = -\frac{\ln P_t^n}{n}. \tag{3.1}$$

A stochastic discount factor (SDF) or pricing kernel M_t prices bonds of all maturities, guaranteeing the absence of arbitrage opportunities. Thus, M_t is a strictly positive random variable with

$$E|M_t P_t^i| < \infty$$

and

$$P_t^n = E(M_{t+1} P_{t+1}^{n-1} | \mathcal{F}_t). \tag{3.2}$$

for all n and t, where $\mathcal{F}_t$ denotes the information set given at time t.[4] In the following we will use the short-hand notation $E_t(\cdot)$ for the conditional expectation $E(\cdot|\mathcal{F}_t)$.

In the framework of consumption-based asset pricing, the SDF represents the marginal rate of substitution between consumption in period t and period $t+1$.[5] As such, the specification of M_t depends on the specific utility function assumed. However, as consumption-based models fail to account for important features of asset prices, it has become common to specify the SDF as a more general function of explanatory variables. Moreover, in this paper we will treat the factors driving the SDF as latent variables as frequently done in the empirical term structure literature.

We denote by X_t the d-dimensional vector of factors. Its dynamic evolution is specified as a VAR(1) process, i.e.

$$X_t = a + \mathcal{K} X_{t-1} + u_t, \tag{3.3}$$

where a is a $d \times 1$ vector of constants, and $\mathcal{K}$ is a $d \times d$ matrix. The eigenvalues of $\mathcal{K}$ are assumed to lie inside the unit circle, which guarantees stationarity of the process $\{X_t\}$.

The pricing kernel is affine in the vector of factors and its innovations,

$$-\ln M_{t+1} = \delta + \gamma' X_t + \lambda' u_{t+1}, \tag{3.4}$$

[4] For a proof of the equivalence between the existence of an SDF and the absence of arbitrage opportunities see Irle (1998).

[5] See, e.g., Campbell, Lo, and MacKinlay (1997) or Cochrane (2001).

where δ is a scalar and γ and λ are both $d \times 1$ vectors. The components of the vector λ will be referred to as the market price of risk parameters.

A solution of the model is a family of functions $g_n(\cdot)$ that map the factor vector X_t into the corresponding arbitrage-free yield y_t^n for each n,

$$y_t^n = g_n(X_t).$$

Thus, the whole term structure of interest rates at time t is determined by the realization of X_t. The dynamics of any yield y_t^n and its stationary distribution depend on the dynamics of X_t.

If for the model (3.3) - (3.4) the factor innovation is Gaussian, the solution function is affine, i.e. we have[6]

$$y_t^n = \frac{A_n}{n} + \frac{1}{n} B_n' X_t, \tag{3.5}$$

where A_n and B_n are a scalar and a $d \times 1$ vector, respectively, that depend on the model parameters and time to maturity n but not on t.

The yield equation (3.5) implies that if X_t is a stationary Gaussian VAR, yields of all maturities will be normally distributed. The same holds for all linear combination of yields, e.g. term spreads, and for yields in first differences. However, normality of yield changes is at odds with the stylized facts as illustrated above.

One approach to generate non-normal yields with an arbitrage-free model is to replace the simple normal distribution by a mixture of normal distributions. This is done by Backus et al. (1998) for a one-factor model. The model can capture excess kurtosis in yield changes. However, since there is only one source of randomness in the model, yields of all maturities share the same kurtosis. This paper generalizes the idea of Backus et al. to the multivariate case. That is, for the multifactor model (3.3) - (3.4) above it is assumed that u_t has a multivariate Gaussian mixture distribution. We write

$$u_t \sim i.i.d. \sum_{b=1}^{B} \omega_b N(\mu_b, V_b), \quad \sum_{b=1}^{B} \omega_b = 1, \ \sum_{b=1}^{B} \omega_b \mu_b = 0, \tag{3.6}$$

to denote that the density of u_t is given by

$$p(x) = \sum_{b=1}^{B} \omega_b \frac{1}{\sqrt{(2\pi)^g |V_b|}} \exp\left(-\frac{1}{2} (x - \mu_b)' V_b^{-1} (x - \mu_b) \right).$$

This formulation allows high flexibility in modeling the shape of the distribution. For instance, the density of u_t may be asymmetric, fat-tailed or bimodal.[7]

[6] See, e.g., Backus et al. (1998) or Campbell et al. (1997).

[7] For a treatment of finite mixture models and their properties see McLachlan and Peel (2000) or Titterington, Smith, and Makov (1985).

Fortunately, by going from a simple normal to a normal mixture, the structure of the solution for bond prices is retained as the following proposition shows.

Proposition 3.1 (Yields in the linear multifactor Gaussian mixture model). *For the multifactor model (3.3), (3.4), (3.6), zero bond yields are given as*

$$y_t^n = \frac{A_n}{n} + \frac{1}{n} B_n' X_t \tag{3.7}$$

with[8]

$$B_n = (I - \mathcal{K}^n)(I - \mathcal{K})^{-1}\gamma \tag{3.8}$$

$$A_n = \sum_{i=0}^{n-1} G(B_i) \tag{3.9}$$

where

$$G(B_i) = \delta + B_i' a - \ln \left[\sum_{b=1}^{B} \omega_b \cdot e^{-(\lambda + B_i)'\mu_b + \frac{1}{2}(\lambda + B_i)' V_b (\lambda + B_i)} \right].$$

Proof. We start with the guess that bond prices are affine in factors.

$$-\ln P_t^n = A_n + B_n' X_t$$

For computing the functional forms of the scalar A_n and the d-dimensional vector B_n we use the fundamental pricing equation (3.2) in logs

$$-\ln P_t^{n+1} = -\ln E_t(M_{t+1} P_{t+1}^n). \tag{3.10}$$

The logarithm of the product on the right-hand side is given by

$$\begin{aligned} & \ln M_{t+1} + \ln P_{t+1}^n \\ = \;& -\delta - A_n - B_n' a - (\gamma' + B_n' \mathcal{K}) X_t - (\lambda' + B_n') u_{t+1} \\ := \;& V_{t+1}. \end{aligned}$$

The conditional distribution of V_{t+1} is not normal but a d-variate normal mixture with B components. For the right-hand side or (3.10) we have to compute

$$E_t \left(e^{\ln M_{t+1} + \ln P_{t+1}^n} \right)$$

which has the form

$$E_t \left(e^{c_0 + c_1' u_{t+1}} \right)$$

with $c_0 = -\delta - A_n - B_n' a - (\gamma' + B_n' \mathcal{K}) X_t$, $c_1 = -(\lambda + B_n)$.

[8] Empty sums are evaluated as zero.

Following from a result in Lemke (2005), we have

$$\begin{aligned} & E_t\left(e^{c_0+c_1' u_{t+1}}\right) \\ = \; & e^{c_0}\left(\sum_{b=1}^{B}\omega_b\, e^{c_1'\mu_b+\frac{1}{2}c_1' V_b c_1}\right). \end{aligned}$$

Plugging back in the original variables we thus obtain

$$\begin{aligned} & \ln E_t\left(e^{\ln M_{t+1}+\ln P_{t+1}^n}\right) \\ = \; & -\delta - A_n - B_n' a - (\gamma' + B_n'\mathcal{K})X_t \\ & + \ln\left[\sum_{b=1}^{B}\omega_b \cdot e^{-(\lambda+B_n)'\mu_b+\frac{1}{2}(\lambda+B_n)' V_b(\lambda+B_n)}\right]. \end{aligned}$$

For the fundamental pricing equation (3.10) to hold, the coefficient functions A_n and B_n have to satisfy the following set of difference equations

$$B_{n+1} = \gamma + \mathcal{K}' B_n \tag{3.11}$$

$$\begin{aligned} A_{n+1} = \; & \delta + A_n + B_n' a \\ & - \ln\left[\sum_{b=1}^{B}\omega_b \cdot e^{-(\lambda+B_n)'\mu_b+\frac{1}{2}(\lambda+B_n)' V_b(\lambda+B_n)}\right]. \end{aligned} \tag{3.12}$$

with initial conditions $A_0 = 0$ and $B_0 = 0$. The vector difference equation for B_n is the same as in the Gaussian multifactor model, so again

$$B_n = (I + \mathcal{K}' + \mathcal{K}'^2 + \ldots + \mathcal{K}'^{n-1})\gamma = (I - \mathcal{K}'^n)(I - \mathcal{K}')^{-1}\gamma.$$

The solution of the difference equation for A_n leads to (3.9). □

The mixture model nests the linear Gaussian multifactor model as a special case. In Lemke (2005) the properties of the model are analyzed in more detail. For instance, it turns out that the model can exhibit excess kurtosis that varies with time to maturity. As regards the stationary distribution of yields, the pure Gaussian model implies a Gaussian distribution for yields. For the mixture model, the unconditional distribution of yields is not straightforward to derive. We leave this as a topic for future research. For a given set of model parameters, however, the distribution of yields can be approximated by using Monte Carlo methods as done in section 5 below.

4 Estimation Approach

In the literature, the state space approach has often been adopted for the estimation of term structure models.[9] The statistical state space model is a representation of the joint dynamic evolution of an observable random vector y_t

[9] This has mostly been done for continuous-time models, as, for instance, by Babbs and Nowman (1999), Babbs and Nowman (1998), Ball and Torous (1996), de Jong (2000), Duan

and a generally unobservable state vector α_t.[10] The state space model contains a measurement equation and a transition equation. The transition equation governs the evolution of the state vector,

$$\alpha_t = T\alpha_{t-1} + c + \eta_t. \tag{4.1}$$

The measurement equation specifies how the state interacts with the vector of observations,

$$y_t = M\alpha_t + d + \epsilon_t. \tag{4.2}$$

The quantities d, c, M, T, H, Q are vectors and matrices of appropriate dimension. η_t is the innovation of the state process, ϵ_t is referred to as the measurement error. The model is completed by specifying the distribution of the initial state vector α_0 and the joint evolution of η_t and ϵ_t.

Once the term structure model of the preceding section is cast into state space form, the statistical inference associated with state space models can be conducted to estimate unknown model parameters, to estimate the latent factor process driving the term structure, and to make one- or multistep-predictions. Moreover, goodness-of-fit criteria developed for state space models can be employed to judge the adequacy of the term structure model specification under consideration.

For estimating our term structure model in state space form, we first transform the factor evolution to the form of a state space model's transition equation. This is straightforward as the factor evolution in (3.3) is already of the form (4.1). That is, we have $c = a$, $T = \mathcal{K}$, and $\eta_t = u_t$.

The measurement equation arises by choosing observed interest rates as left-hand-side variables, whereas the right-hand-side is the sum of the theoretical solution implied by the term structure model and a measurement error. Recall that bond yields are given by

$$y_t^n = \frac{A_n}{n} + \frac{1}{n} B_n' X_t, \tag{4.3}$$

where A_n and B_n depend on the parameters of the factor process and on market price of risk parameters collected in a vector λ. Let the measurement vector at time t contain observed yields of k different maturities, say $n_1, \ldots, n_k$. Then the theoretical model implies that

$$\begin{pmatrix} y_t^{n_1} \\ \vdots \\ y_t^{n_k} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1} A_{n_1} \\ \vdots \\ \frac{1}{n_k} A_{n_k} \end{pmatrix} + \begin{pmatrix} \frac{1}{n_1} B_{n_1}' \\ \vdots \\ \frac{1}{n_k} B_{n_k}' \end{pmatrix} \alpha_t. \tag{4.4}$$

and Simonato (1999), Geyer and Pichler (1999) and Schwaar (1999). Cassola and Luis (2003) is an example for estimating a discrete-time Gaussian model.

[10] See, e.g., Brockwell and Davis (1996), Durbin and Koopman (2001) or Hamilton (1994).

Adding a vector of measurement errors $\epsilon_t = (\epsilon_t^{n_1}, \ldots, \epsilon_t^{n_k})'$ leads to a linear measurement equation of the form (4.2),

$$y_t = d + M\alpha_t + \epsilon_t, \tag{4.5}$$

with obvious definitions of the vector d and the matrix M.

Denote by $\mathcal{Y}_s = \{y_0, y_1, \ldots, y_s\}$ a sequence of observations of the measurement vector. If η_t and ϵ_t are both Gaussian, the filtering densities $p(\alpha_t|\mathcal{Y}_t)$ as well as the prediction densities $p(\alpha_t|\mathcal{Y}_{t-1})$ and $p(y_t|\mathcal{Y}_{t-1})$ are Gaussian. They can be computed by the Kalman filter. Unknown model parameters can be estimated by maximum likelihood. The log-likelihood is given by

$$l(\psi; \mathcal{Y}_T) = \sum_{i=1}^{T} p(y_t|\mathcal{Y}_{t-1}) \tag{4.6}$$

where ψ contains the unknown model parameters.

For the model considered here, we assume that the measurement error is in fact normally distributed, i.e.

$$\epsilon_t \sim N(0, H). \tag{4.7}$$

However, the distribution of the state innovation η_t is not a simple normal but a mixture of normals, given by (3.6),

$$\eta_t \sim \sum_{b=1}^{B} \omega_b N(\mu_b, V_b).$$

The implication of this deviation from the linear purely Gaussian state space model is that now the filtering and prediction densities are not Gaussian any more as shown by Sorenson and Alspach (1971). They are rather mixtures of normals. The filtering density at time t is given by

$$p(\alpha_t|\mathcal{Y}_t) = \sum_{i=1}^{l_t} \omega_{i\,t|t} \cdot \phi(\alpha_t;\, a_{i\,t|t},\, \Sigma_{i\,t|t}), \tag{4.8}$$

where $a_{i\,t|t}$ and $\Sigma_{i\,t|t}$ are the means and variance-covariance matrices of the component densities, respectively. These as well as the weights $\omega_{i\,t|t}$ are nonlinear functions of the observations.

The conditional expectation and its variance-covariance matrix can be computed as

$$E(\alpha_t|\mathcal{Y}_t) \;=\; \sum_{i=1}^{l_t} \omega_{i,\,t|t}\, a_{i,\,t|t} =: a_{t|t},$$

and

$$\begin{aligned} & Var(\alpha_t|\mathcal{Y}_t) \\ = & \sum_{i=1}^{l_t} \omega_{i,t|t} \left(\Sigma_{i,t|t} + (a_{i,t|t} - a_{t|t})(a_{i,t|t} - a_{t|t})'\right) =: \Sigma_{t|t} \end{aligned}$$

respectively.

The one-step prediction densities for the state and observation-vector, $p(\alpha_t|\mathcal{Y}_{t-1})$ and $p(y_t|\mathcal{Y}_{t-1})$, have a similar structure. Accordingly, the log-likelihood is a sum of mixture distributions. Details are given in the appendix.

The filtering and prediction densities can be computed in an iterative fashion. The algorithm is described in the appendix and can be interpreted as a bunch of Kalman filters working parallel. The main problem, however, is that the number of components is growing exponentially with time: at time t the exact filtering density given above has

$$l_t = B^t$$

components. That is, if our model has $B = 2$ components in the mixture distribution, the filtering density at time $t = 10$ is a mixture of 1024 normals. Hence, for time series of length typically encountered in practice, computing the exact filter becomes impossible. This is why we use an approximate filter, the structure of which will be sketched in the following. The appendix contains a more detailed description.

For our proposed approximation scheme, the maximum number of components appearing in the employed mixture distributions is governed by a parameter $k < T$. After an initial phase, the exact filtering and prediction densities – mixtures with B^t components – are approximated by mixtures with B^k components only. This approximating density results from applying the exact filter to the most recent k observations only. A suitable initialization of the filter takes the first $t - k$ observations into account in a condensed form. We abbreviate the approximation scheme as AMF(k), standing for 'approximate mixture filter of degree k'. Next, we describe verbally how the approximation works.[11]

First, the exact filter is run up to time $t = k$ yielding the exact filtering densities $p(\alpha_t|\mathcal{Y}_t)$ for $t = 1, \ldots, k$. The last of these densities, $p(\alpha_k|\mathcal{Y}_k)$ is a mixture of B^k normals.

Continuing with the exact filter would deliver the exact density for time $t = k + 1$ as a mixture with B^{k+1} components. However, we want to constrain the number of components to B^k. The idea is now to apply the exact

[11] We will refer to the filtering densities only. The idea is the same for the prediction densities. In the summary of the approximation algorithm below, it will be documented how they are computed.

filter algorithm, but only to the last k observations of $\mathcal{Y}_{k+1}$, i.e. to the subsequence $\{y_2, \ldots, y_{k+1}\}$. The filter is initialized by the univariate normal with mean $a_{1|1}$ and variance $\Sigma_{1|1}$, the latter being the mean and the variance of the B-component mixture $p(\alpha_1|\mathcal{Y}_1)$. Thus, the initial condition contains information about y_1 in a condensed form, the exact density $p(\alpha_1|\mathcal{Y}_1)$ is replaced by a simple normal. Applying the exact filter in this fashion to the most recent k observations yields a mixture with B^k components, denoted by $\tilde{p}(\alpha_{k+1}|\mathcal{Y}_{k+1})$, that approximates the exact filtering density at time $k+1$.

A similar procedure is applied for approximating each of the filtering densities from $t = k+1$ to $t = 2k$. For obtaining an approximation of the density $p(\alpha_t|\mathcal{Y}_t)$, the exact filter is applied to the k most recent observations only. The first $t-k$ observations $\{y_1, \ldots, y_{t-k}\}$, however, are not ignored. They enter the estimation process through the initial condition. The exact filter is initialized by a simple normal, and the mean of that normal is $a_{t-k|t-k}$, the optimal estimate of the state at $t-k$, given the observations from 1 to $t-k$. Since the algorithm is iteratively applied, the estimate $a_{t-k|t-k}$ and its variance-covariance matrix $\Sigma_{t-k|t-k}$ are already available.

In this fashion approximate densities $\tilde{p}(\alpha_t|\mathcal{Y}_t)$ for $t = k+1, \ldots, 2k$ are obtained. Each of them is a mixture of B^k components.

Analog operations can be conducted for approximating the filtering densities for $t = 2k+1, \ldots, T$. At time $t \geq 2k+1$ the approximate density is generated by an application of the exact filter to $\{y_{t-k+1}, \ldots, y_t\}$. For computing the initial condition at time $t-k$, one would again collapse the mixture density $p(\alpha_{t-k}|\mathcal{Y}_{t-k})$ to a simple normal. However, since we are beyond $t = 2k$, we do not have the exact filtering density $p(\alpha_{t-k}|\mathcal{Y}_{t-k})$ for time $t-k$ available. We only have $\tilde{p}(\alpha_{t-k}|\mathcal{Y}_{t-k})$ available, a mixture of B^k components that approximates $p(\alpha_{t-k}|\mathcal{Y}_{t-k})$. Nevertheless, we can proceed as usual and collapse this density into a simple normal.

Similar to the filtering densities, the prediction densities are also approximated by mixtures with B^k components. With the sequence of approximate prediction densities at hand, an approximate log-likelihood can be constructed by replacing the exact densities $p(y_t|\mathcal{Y}_{t-1})$ in (4.6) by their approximating counterparts $\tilde{p}(y_t|\mathcal{Y}_{t-1})$.

In Lemke (2005) Monte Carlo simulations have been carried out to assess the properties of the AMF(k) . It turns out that for the data generating processes considered there, the approximate filtering densities generated by the AMF(k) are good approximations to the exact ones (which have been computed for time series of length $T = 10$), even for small k such as $k = 1, 2, 3$. Moreover, it turns out that for $B = 2$, increasing k beyond 3 does not yield any substantial changes of results. In most cases, $k = 1$ does already lead to quite good approximations of the exact filter. Finally, results from the AMF(k) have been compared to

results from the Kalman filter which is still the best *linear* filter for the linear state space model with mixture innovations. The AMF(k) performs consistently better than the Kalman filter, the degree of improvement being dependent on the model parameterization.

5 Empirical Application

With the estimation methodology at hand we now conduct an empirical study in which we estimate three discrete-time term structure models. We use the data set of US treasury yields that has been presented in section 2. It is not claimed that the models that we use in our study are in some sense optimal specifications for our data set. Rather, the main purpose of this section is to show the methodology at work. Moreover, we want to point out what difference it can make to use a mixture model as opposed to a Gaussian model with the same number of factors.

5.1 Models and Parameterization

We estimate three specifications of the model described in section 3: a Gaussian two-factor model, a two-factor model with a two-component mixture, and a three-factor Gaussian model. Recall that the mixture model from section 3 nests a purely Gaussian model as a special case.[12]

The models are characterized by a vector-valued factor process

$$X_t = \mathcal{K} X_{t-1} + u_t \tag{5.1}$$

and a specification of the stochastic discount factor (SDF), that is of the form

$$-\ln M_{t+1} = \delta + \iota' X_t + \lambda' u_{t+1}. \tag{5.2}$$

Note that we have set the intercept in the factor process equal to zero. As described in more detail in Lemke (2005), the model in its original specification is overparameterized, so dropping the intercept is innocuous.

For the Gaussian models, the factor innovation satisfies

$$u_t \sim N(0, V), \tag{5.3}$$

whereas for the mixture model

$$u_t \sim \sum_{b=1}^{B} \omega_b N(\mu_b, V_b), \quad \sum_{b=1}^{B} \omega_b = 1, \ \sum_{b=1}^{B} \omega_b \mu_b = 0. \tag{5.4}$$

[12] Strictly speaking we should refer to the model outlined in section 3 as 'a class of models'.

Going from the general to the specific, the factor process of the two-factor Gaussian model is given by

$$\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{pmatrix} \begin{pmatrix} X_{1\,t-1} \\ X_{2\,t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{5.5}$$

where the distribution of the factor innovation is

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_1^2 & 0 \\ 0 & v_2^2 \end{pmatrix} \right). \tag{5.6}$$

The SDF satisfies

$$-\ln M_{t+1} = \delta + X_{1t} + X_{2t} + \lambda_1 u_{1\,t+1} + \lambda_2 u_{2\,t+1}. \tag{5.7}$$

Interchanging the two factors will not alter the implied term structure. For the Gaussian two-factor model and the mixture model that will be described hereafter, we will sort the factors by their persistence. That is, they are arranged such that $\kappa_1 > \kappa_2$.

Concerning the two-factor mixture model, the factor process and the SDF equation are of the same form as for the two-factor Gaussian model. The distribution of the factor innovation is specified as a Gaussian mixture with two components,

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \sim \omega N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_{11}^2 & 0 \\ 0 & v_{21}^2 \end{pmatrix} \right) + (1-\omega) N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_{12}^2 & 0 \\ 0 & v_{22}^2 \end{pmatrix} \right) \tag{5.8}$$

We have tried three different specifications, one with $v_{11} \neq v_{12}$ and $v_{21} \neq v_{22}$, another with $v_{11} \neq v_{12}$ and $v_{21} = v_{22}$, and a third with $v_{11} = v_{12}$ and $v_{21} \neq v_{22}$. It turned out that the third one performed best and we only report the results of this specification. In order to identify the two components we assume that $v_{21} \geq v_{22}$. This assumption is embedded into the specification by parameterizing the first component variance as a multiple of the second. Summing up, we will assume that

$$v_{11} = v_{12} =: v_1, \quad \text{and} \quad v_{21}^2 = c_{22} v_{22}^2, \quad c_{22} \geq 1. \tag{5.9}$$

Finally, the three-factor Gaussian model consists of the factor process

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \begin{pmatrix} \kappa_1 & 0 & 0 \\ 0 & \kappa_2 & 0 \\ 0 & 0 & \kappa_3 \end{pmatrix} \begin{pmatrix} X_{1\,t-1} \\ X_{2\,t-1} \\ X_{3\,t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix} \tag{5.10}$$

with

$$\begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_1^2 & 0 & 0 \\ 0 & v_2^2 & 0 \\ 0 & 0 & v_3^2 \end{pmatrix} \right) \tag{5.11}$$

The pricing kernel is given by

$$-\ln M_{t+1} = \delta + X_{1t} + X_{2t} + X_{3t} + \lambda_1 u_{1\,t+1} + \lambda_2 u_{2\,t+1} + \lambda_3 u_{3\,t+1}. \qquad (5.12)$$

Similarly as for the two-factor models, we assume that $\kappa_1 > \kappa_2 > \kappa_3$.

All three models have the property that both the matrix $\mathcal{K}$ and the (component) variance-covariance matrices are diagonal. For all three models, this implies that the factors are independent from each other. Of course, this is a restrictive assumption whose validity could be tested for. For the two-factor models, correlation of the factors could be induced by introducing an additional free parameter for the (2,1)-element of $\mathcal{K}$. The hypothesis of uncorrelated factors would then correspond to this parameter being zero. Such a test, however, will not be conducted here and we will stick to the more simple specification.

Each model is cast into its corresponding state space form and the parameters are estimated by maximum likelihood. For the Gaussian models, the state space model is linear and Gaussian, and the exact likelihood can be constructed using the Kalman filter. For the two-factor mixture model, the state space model is linear but the state innovations are distributed as a Gaussian mixture. For this model, we construct an approximate likelihood based on the AMF(1) filter.[13] We will now explain some details of the estimation process and turn to the results in the next section.

From the data set presented in section 2, we use time series of yields for maturities of 3, 6, 12, 24, 60, and 120 months. The yields are annualized, the models, however, hold for monthly yields. The models imply that for some arbitrary n, the joint evolution of factor and yield are given by (5.1) and (4.3). Then the annualized yield $\tilde{y}_t^n := 1200 \cdot y_t^n$ satisfies[14]

$$\begin{aligned} \tilde{y}_t^n &= \frac{A_n^*}{n} + \frac{1}{n} B_n^{*\prime} X_t \\ X_t &= \mathcal{K} X_{t-1} + u_t \end{aligned}$$

with $A_n^* = 1200 \cdot A_n$ and $B_n^* = 1200 \cdot B_n$. It is this kind of representation that we use in the empirical study. This implies that the parameters that we obtain are those that correspond to the original monthly yields. Accordingly, they can be compared in size with parameters from the literature that have been obtained for other samples using possibly different statistical techniques. The reason for using annualized yields (as opposed to monthly yields) lies in the fact that for monthly yields the measurement error in the corresponding state space model would have a have a very low standard deviation (of around 7e-6). This would possibly lead to numerical difficulties.

[13] Using the AMF(2) filter delivered nearly the same results.

[14] We have to multiply by 1200 (and not by 12 only) since yields in the data set are expressed in percentages.

We do not want to carry on with the tilde on top of our annualized yields, so we drop it from here on and understand each y_t^n as an annualized yield.

For the state space models associated with our theoretical term structure models, the measurement vector y_t is five-dimensional,

$$y_t = (y_t^{n_1}, y_t^{n_2}, \ldots, y_t^{n_5})', \quad (n_1, n_2, \ldots, n_5)' = (3, 6, 12, 60, 120)'.$$

For each term structure model we identify the factor vector with the state vector, i.e. $\alpha_t = X_t$. The measurement equation has the form

$$\begin{pmatrix} y_t^{n_1} \\ \vdots \\ y_t^{n_5} \end{pmatrix} = \begin{pmatrix} 1200 \cdot \frac{1}{n_1} A_{n_1} \\ \vdots \\ 1200 \cdot \frac{1}{n_5} A_{n_5} \end{pmatrix} + \begin{pmatrix} 1200 \cdot \frac{1}{n_1} B'_{n_1} \\ \vdots \\ 1200 \cdot \frac{1}{n_5} B'_{n_5} \end{pmatrix} \alpha_t + \begin{pmatrix} \epsilon_{1t} \\ \vdots \\ \epsilon_{5t}. \end{pmatrix} \tag{5.13}$$

where the functional forms of the A_{n_i} and the B_{n_i} differ across models, of course. Written more compact in the familiar notation of a state space model,

$$y_t = d + M\alpha_t + \epsilon_t. \tag{5.14}$$

For the measurement error we use the simple specification

$$\epsilon_t \sim N(0,\, h^2 I_5). \tag{5.15}$$

This is not an innocuous assumption since it implies that the difference between theoretical and observed yields has the same variance for all maturities. We also tried a specification in which the variances were allowed to be pairwise different. However, it turned out that the other parameter estimates have not been affected much by this change of specification.

For the two-factor Gaussian model, the unknown model parameters to be estimated are κ_1, v_1^2, λ_1, κ_2, v_2^2, λ_2, δ, and h^2. The parameters κ_1, v_1^2, κ_2, v_2^2 of the theoretical model appear in both, the transition equation and the measurement equation, whereas the parameters λ_1, λ_2, and δ appear in the intercept vector d of the measurement equation only.

Concerning the four parameters v_1^2, λ_1, v_2^2 and λ_2, the model may be equivalently parameterized in v_1, $\lambda_1 v_1$, v_2, and $\lambda_2 v_2$.[15] This can be seen as follows. The only places in which the parameters λ_1 and λ_2 appear are the functions A_n. For a Gaussian model, A_n is computed as

$$A_n \;=\; \sum_{i=0}^{n-1} G(B_i) \tag{5.16}$$

where

$$G(B_i) = \delta + B_i' a - \frac{1}{2}(\lambda + B_i)' V (\lambda + B_i).$$

[15] This is also done by Cassola and Luis (2003).

With a diagonal V matrix, expanding the expression $(\lambda + B_i)'V(\lambda + B_i)$ yields

$$(\lambda + B_i)'V(\lambda + B_i) = \sum_{j=1}^{2} \lambda_j^2 v_j^2 + 2B_{ij}\lambda_j v_j^2 + B_{ij}^2 v_j^2 \quad (5.17)$$

$$= \sum_{j=1}^{2} (\lambda_j v_j)^2 + 2B_{ij}(\lambda_j v_j) \cdot v_j + B_{ij}^2 v_j^2 \quad (5.18)$$

where B_{ij}, $j = 1, 2$ denotes the jth component of B_i. Thus, λ_j only shows up as a multiplier of v_j.

The same argument goes through for the three-factor model, which will be parameterized in v_1, $\lambda_1 v_1$, v_2, $\lambda_2 v_2$, v_3, $\lambda_3 v_3$. A similar reasoning holds for the two-factor mixture model. For each mixture component b one can expand the exponent $(\lambda + B_i)'V_b(\lambda + B_i)$ in (3.9) in the same fashion as just shown for the Gaussian case.[16] Thus, our two-factor model is parameterized in v_1, $\lambda_1 v_1$, v_{22}, $v_{21} = (\sqrt{c_{22}} v_{22})$, and $\lambda_2 v_{22}$.

Estimating the model, it turned out that δ and the market price of risk parameters $\lambda_1 v_1$ and $\lambda_2 v_2$ cannot be estimated very accurately. Moreover, the estimated covariance matrix shows that they are highly correlated.[17] In particular, the parameter $\lambda_1 v_1$ has been individually insignificant, so we dropped it from the model.

Summing up, the following parameters will be estimated. For the Gaussian two-factor model,

$$\kappa_1,\ v_1,\ \kappa_2,\ v_2,\ \lambda_2 v_2,\ \delta,\ h^2,$$

for the Gaussian two-factor mixture model,

$$\kappa_1,\ v_1,\ \kappa_2,\ v_{22},\ \lambda_2 v_{22},\ c_{22},\ \omega,\ \delta,\ h^2$$

and for the Gaussian three-factor model,

$$\kappa_1,\ v_1,\ \kappa_2,\ v_2,\ \lambda_2 v_2,\ \kappa_3,\ v_3,\ \lambda_3 v_3,\ \delta,\ h^2.$$

Note that some of the parameters have to satisfy certain restrictions. We have:

$-1 \le \kappa_i \le 1,\ i = 1, 2, 3$	(stationarity of the factor process)
$v_i \ge 0,\ i = 1, 2, 3,$ and $v_{22} \ge 0$	(v_i and v_{22} are standard deviations)
$c_{22} \ge 1$	(by our assumption above)
$0 < \omega < 1$	(ω is a component weight)
$h^2 \ge 0$	(h^2 is a variance)

[16] The parameterization that we use would not be possible if $\mu_b \neq 0$ as can be seen from (3.9).

[17] All of these three parameters only enter the intercept vector d and do not show up elsewhere in the model. However, there is no identification problem as one might suspect. All of these parameters are individually identified, since we use five yields in the measurement vector.

These constraints have been taken care of by reparameterizing the model parameters accordingly (for example by squaring to ensure nonnegativity).

5.2 Estimation Results

Table 5 contains the maximum likelihood estimates of the parameters. Estimated standard errors are given in parentheses.The dimension and sign of the estimates are reasonable for all parameters.

The first factor is highly persistent as the estimate of κ_1 is nearly one for all models. Estimated standard errors may be interpreted with some caution since the estimate is very close to the boundary of the parameter space. For future studies we suggest using the bootstrap in order to obtain reliable confidence intervals. The standard deviation v_1 of the first factor is estimated with satisfiable precision and it does not differ much across models.

The second factor exhibits lower autocorrelation (κ_2) than the first factor, but it is still very high. The innovation of the second factor is the place in which the Gaussian models differ from the mixture model. For the latter model, the marginal distribution of the factor innovation is a mixture of two normals,

$$u_{2t} \sim \omega N(0,\, v_{21}^2) + (1-\omega)N(0, v_{22}^2), \quad \text{with} \quad v_{21}^2 = c_{22} \cdot v_{22}^2.$$

Judging on the basis of a standard t-test, the estimate of the weight ω is significantly different from zero and the estimate of the variance ratio c_{22} is different from unity.[18] So the results suggest that for the sample at hand the density for the second factor innovation is in fact a 'true' mixture of normals. It can be interpreted in such a way that in 86.2 percent of the time the innovation is drawn from a normal with standard deviation $v_{22} = 0.00023$, and in 13.8 percent it is drawn from a normal whose standard deviation is $5.11 (= \sqrt{26.1})$ times bigger.

For the mixture model, the estimates of v_{22}, ω and c_{22} imply that the estimate of the standard deviation of the second factor innovation is given by

$$\hat{v}_2 := \left(\hat{\omega} \cdot \hat{c}_{22} \cdot \hat{v}_{22}^2 + (1-\hat{\omega}) \cdot \hat{v}_{22}^2\right)^{0.5} = 0.000486.$$

This does not deviate much from the estimated standard deviation of the second factor innovation for the Gaussian two-factor model.

In the mixture model, the parameter estimates imply for the excess kurtosis of u_{2t},

$$\widehat{kurt}(u_{2t}) = \frac{3\left[\hat{\omega} \cdot \left(\hat{c}_{22} \cdot \hat{v}_{22}^2\right)^2 + (1-\hat{\omega}) \cdot \hat{v}_{22}^4\right]}{\hat{v}_2^4} - 3 = 11.284,$$

[18] In face of the fact that we use the approximate likelihood generated by the AMF, the estimated standard deviations should be used with caution.

Recall that the excess kurtosis is zero (by definition) for the Gaussian models.

The two panels in figure 4 show the marginal densities of the factor innovations that are implied by the parameter estimates. The left panel contains the estimated densities of u_{1t}, the innovation of the first factor. The solid line corresponds to the Gaussian two-factor model, the dashed line corresponds to the mixture model. Recall that both densities are normal. They differ from each other due to the fact that they have slightly different variances. The right panel shows a more substantial difference. The solid line depicts the density of u_{2t} for the Gaussian model. The dashed line represents the density of u_{2t} for the mixture model. The density is a Gaussian mixture with two components. It is remarkably different compared to its Gaussian counterpart although it implies nearly the same variance.

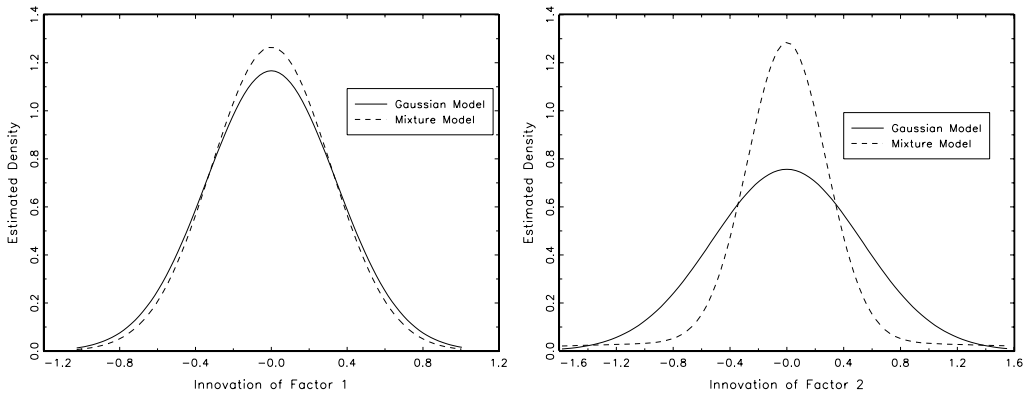

Figure 4: For the two-factor models: Estimated densities of the innovation of the first factor (left panel) and the second factor (right panel)

For all three models, the market price of risk parameters, $\lambda_2 v_2$ and $\lambda_2 v_{22}$ have the expected negative sign which corresponds to a positive term premium. These parameters are estimated with lower relative precision compared to the other parameters discussed so far. For the three-factor model, $\lambda_2 v_2$ is not even significantly different from zero. The parameter δ that governs the average level of the yield curve is individually estimated quite precisely. However, the estimated autocorrelation matrix of estimates (not reported here) shows that for all models considered, the correlation of the market price of risk parameters and δ is high.

Heuristically, these properties may be explained by the fact that the market price of risk parameters and δ only show up in the intercept vector d of the measurement equation. Since the factors have mean zero, it is easy to see from equation (5.14) that the vector d contains the individual means of yields included in y_t. Now, since all yields are highly autocorrelated, their means - and in turn the parameters that parameterize them - cannot be estimated very precisely.

For the three-factor model, the parameters κ_3, v_3, and $\lambda_3 v_3$ of the additional factor process had to be estimated. The estimate of the autocorrelation parameter κ_3 is remarkably smaller than those of the first two factors. The estimated innovation variance ${v_3}^2$ is similar in size to that of the second factor. Unlike for the second factor, the market price of risk parameter $\lambda_3 v_3$ is individually significantly different from zero.

The estimated variance $\hat{h}^2$ of the measurement error has the same size for both two-factor models. Recall that the measurement error captures the difference between observed annualized yields and the theoretical yields implied by the respective model under consideration. The estimates for the two-factor models imply that this error has a standard deviation of $0.186 (= \sqrt{0.0346})$ percentage points. The standard deviation implied by the three-factor model is half as large, it amounts to 0.092 percentage points.

The bottom of table 5 contains the values of the log-likelihood at maximum for the three models. We also provide the value of Akaike's information criterion, defined as

$$AIC = -2 \ln \mathcal{L}(\hat{\psi}) + 2w$$

where w is the number of unknown parameters. The AIC decreases in the value of the likelihood and increases in the number of parameters that have to be estimated. Using the AIC as a model selection criterion, the model with the smallest value of the AIC is chosen. Employing this measure for selecting one of our three models, the three-factor model would be preferred. Comparing between the two two-factor models only, the mixture model would beat the pure Gaussian model. A worthwhile exercise for future research would consist of choosing a mixture distribution for the innovations of the three-factor model and checking if this enhanced three-factor model beats the pure Gaussian one considered here.

Figure 5 displays the average observed yield curve together with the average estimated yield curves for the three models.[19] For convenience the points of the average observed yield curve are connected in the picture. For $(n_1, n_2, \ldots, n_6)' = (3, 6, 12, 24, 60, 120)'$, the observed average yield curve consists of the points $(n_i, \bar{y}_t^{n_i})$, where

$$\bar{y}_t^{n_i} = \frac{1}{T} \sum_{t=1}^{T} y_t^{n_i}, \quad i = 1, \ldots, 6,$$

is the average of the annualized n_i-month yields over the 444 observations. Note that the 24-month yield, that has not been used for the estimation, is also included. The average estimated yield curve is given by the points $(n_i, \hat{y}_t^{n_i})$

[19] The points representing the two-factor Gaussian model and those representing the two-factor mixture model nearly coincide and are hard to distinguish from each other.

where

$$\hat{y}_t^{n_i} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{A}_{n_i}}{n_i} + \frac{1}{n_i}\hat{B}'_{n_i}a_{t|t}\right). \tag{5.19}$$

Here $\hat{A}_{n_i}$ and $\hat{B}_{n_i}$ are the coefficient functions implied by the models where the parameters are replaced by their maximum likelihood estimates. The $a_{t|t}$ are the filtered states at time t. Thus, for a given time t, $a_{t|t}$ is an estimate of the factor vector X_t, which is constructed using all information up to this point in time. The figure shows that the mean yield curve is matched well by all models

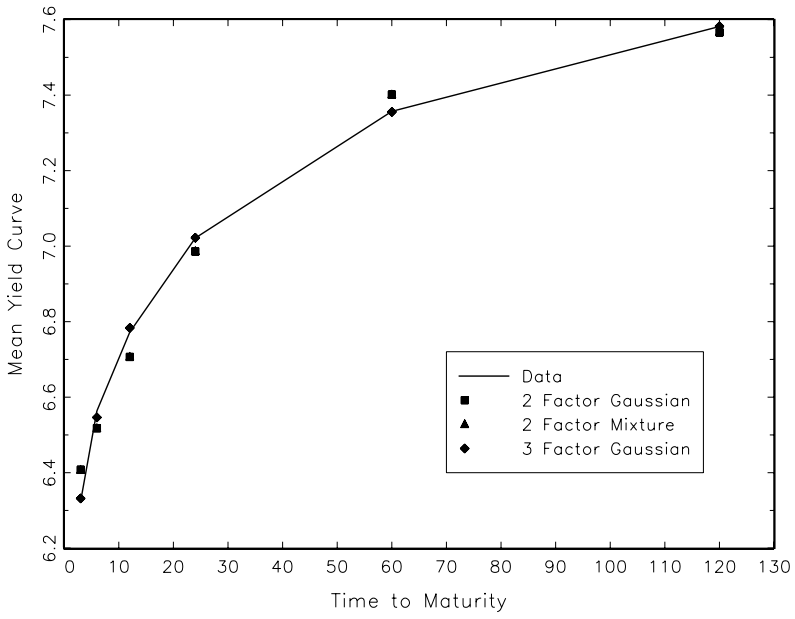

Figure 5: Mean yield curve

whereas the three-factor model seems to have a slight edge over the other two models.

In univariate time series analysis, diagnostic tests for fitted models are often based on residuals. In particular, residuals should be uncorrelated over time. Tests for the correlation of residuals are based on the autocorrelations of the estimated residuals. In multivariate time series analysis there is more than one autocorrelation for a given lag. Let $\{v_t\}$ be a vector valued series of residuals, where the v_t are of dimension $N \times 1$ each. Then for a given lag k there are N^2 possibly different autocorrelations, namely between v_{it} and $v_{j,t-k}$ for all pairs (i,j), $i = 1,\ldots,N$, $j = 1,\ldots,N$.[20] In the literature that deals with the estimation of term structure models in a state space framework, the analysis is

[20] Note that in general the autocorrelation between $v_{i\,t}$ and $v_{j,\,t-k}$ is different from that between $v_{j\,t}$ and $v_{i,\,t-k}$.

generally restricted to to univariate autocorrelations, i.e those between $v_{i\,t}$ and $v_{i\,t-k}$.

For our models we want to provide two measures for the autocorrelation of residuals. First, we will show the five univariate autocorrelation functions. Second, we provide a measure that tries to capture multivariate autocorrelation in a condensed form. We do not seek to formally test on autocorrelation of residuals by, for instance, using a multivariate portmenteau statistic. This is partly due to the fact that we do not know how such a statistic would behave for our model with Gaussian mixture innovations.

The residual vector v_t at time t is given by

$$v_t = y_t - y_{t|t-1},$$

where $y_{t|t-1}$ is the one-step forecast of y_t based on observations up to time $t-1$. The (i,j)-element of the autocorrelation matrix of v_t for lag k, $\Gamma(k)$, is given by[21]

$$\Gamma(k)_{ij} = \frac{\sum_{t=15+k}^{T}(v_{i\,t} - \bar{v}_i)(v_{j,\,t-k} - \bar{v}_j)}{\sqrt{\sum_{t=15+k}^{T}(v_{i\,t} - \bar{v}_i)^2} \cdot \sqrt{\sum_{t=15+k}^{T}(v_{j\,t} - \bar{v}_j)^2}}, \tag{5.20}$$

where

$$\bar{v}_l = \frac{1}{T-15}\sum_{t=16}^{T} v_{l\,t}., \quad l = i, j.$$

The first five panels in figure 6 depict the univariate autocorrelation functions $\Gamma(k)_{ii}$ for $i = 1, 2, \ldots, 5$. The picture in the lower right corner is intended to give an overall measure of autocorrelation of the residuals. For each lag k, we plotted for each model the norm $||\Gamma(k)||$ of the autocorrelation matrix $\Gamma(k)$ against the lag k. We have defined this norm as[22]

$$||\Gamma(k)|| := \max_{i,j} |\Gamma(k)_{i,j}|. \tag{5.21}$$

That is, for each lag k, the figure contains the largest (in absolute value) element of the 25 elements of the autocorrelation matrix.

The first thing to be noted is that the five univariate autocorrelation functions do not differ much across models. At lag 1 the ACFs assume their maximum with $\Gamma(1)_{ii}$ amounting to a level between about 0.2 and 0.3. For higher lags, the ACFs fluctuate around zero. Using the popular bounds of $\pm 2/\sqrt{T}$ which corresponds to the interval $[-0.097, 0.097]$ here, it turns out that for $i = 1$ (residuals of three-month yields) eight of the estimated autocorrelations fall outside this interval. This is the case for all three models. For 60-month

[21] Note that we have dropped the first 15 observations in order to remove any dependence on the initialisation of the filters.

[22] Of course, there are several alternatives to define the norm of a matrix.

yields, only three of the autocorrelations fall outside the interval. Overall, the autocorrelations of residuals appear to be a little too high, but the observed patterns do not point towards strong misspecification.

Up to now we have said little about the factors that drive the term structure. Recall that with the filtering techniques at hand we are able to estimate the path of the unobservable factors. Figure 7 depicts the estimated paths of the first and second factor for our two-factor models. That is, we have drawn the first and second component of the filtered state vector $a_{t|t}$ (multiplied by 1200) against time. The first thing to note is that the results for the Gaussian model and the mixture model are similar. Second, comparing with figure 1 above, the path of the first factor seems to resemble the pattern of the evolution of the level of the yield curve.[23] In fact, the correlation between the filtered factor process and yields is high for each maturity. For both two-factor models, it reaches from 0.80 (correlation with the three-month yield) to 0.99 (correlation with the ten-year yield). Similar results are obtained for the three-factor model where the correlation is between 0.78 and 0.99. Against this background, the first factor may be referred to as a level factor.

This interpretation is supported if we look at the estimated factor loadings of the two-factor Gaussian model in figure 8.[24] The factor loading of the ith factor on the n-month yield is given by the ith component of the vector B_n/n. Note that for all models considered, the vector B_n/n only depends on the κ_i parameters. In our models with diagonal $\mathcal{K}$ matrices, the ith component of B_n is simply given by κ_i^n. The interpretation of an arbitrary point on one of the curves of factor loadings is as follows: if that factor is increased, ceteris paribus, by one unit, the yield with time to maturity n is increased by the amount given on the axis of ordinates. Here, an increase in the first factor shifts up yields of all maturities nearly proportionally. Hence, the name 'level factor' is justified. The second factor leads to a shift in the term structure that is strong at the short end of the yield curve and becomes weaker as time to maturity rises. Accordingly, the second factor may be referred to as a twisting factor.

For the three-factor model, the same type of picture is drawn. Figure 9 shows that the first two factors can be given the same interpretation as before for the two-factor models. The additional factor works mostly at the short end of the yield curve.

In their paper, Cassola and Luis (2003) try to match the term structure of volatility. For them, the term structure of observed volatility consists of the pairs $(n_i, Var^{emp}(y_t^{n_i}))$ where $Var^{emp}(y_t^{n_i})$ is the empirical variance of the n_i-month yield. This is compared to the theoretical term structure of volatility, i.e.

[23] Of course it is parallel shifted by some amount, since the factor process has mean zero by assumption.

[24] The results of the two-factor mixture model imply nearly the same picture, thus it is not shown here.

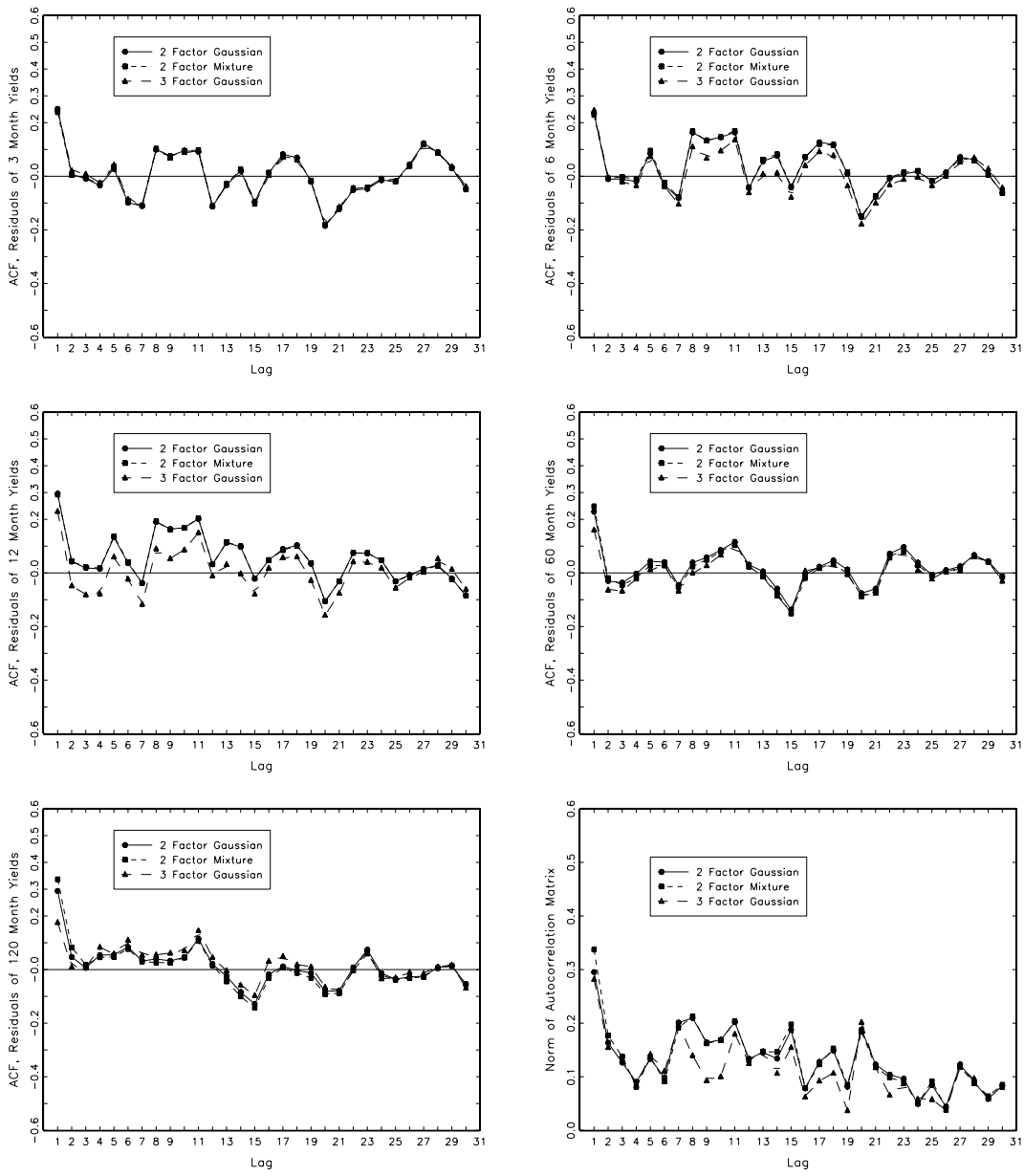

Figure 6: First five panels (from left to right and top to bottom): ACF of the residuals of 3-, 6-, 12-, 60-, and 120-month yields. Right panel in the last row: norm of the multivariate autocorrelation matrices plotted against lags.

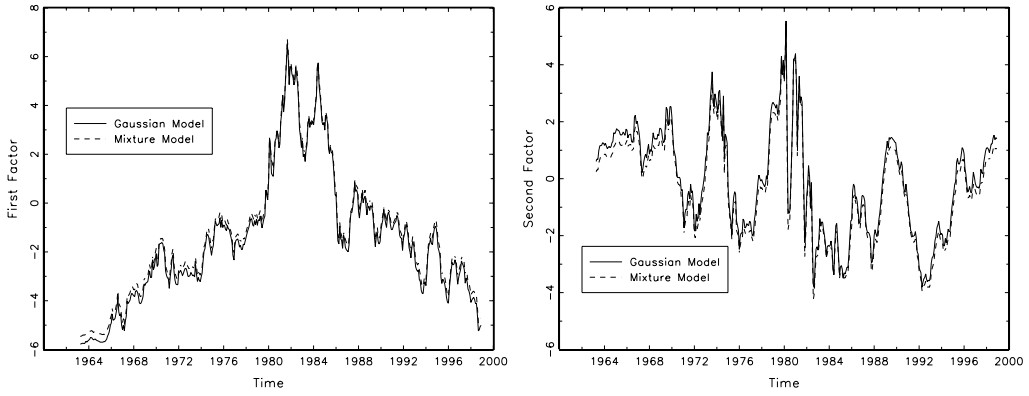

Figure 7: Filtered process of the first factor (left panel) and the second factor (right panel)

the variances implied by the model. Cassola and Luis come up with the results that the variance of yields implied by their two-factor model is unreasonable with respect to the observed variance. Therefore, they alter their estimation approach by including observed variances in the measurement equation. This leads to parameter estimates that are such that the observed volatility matches the theoretical volatility.

However, due to the fact that yields are highly autocorrelated, the empirical variance is a biased estimate of the true variance of a time series of yields. Thus, it may not be sensible to compare the estimated theoretical variance with the empirical one for small samples. Therefore, instead of trying to match the volatility curve of yield *levels*, we have a look at the volatilities of yields *in first differences*. As seen in table 3 in section 2, their autocorrelation is low.

Figure 10 shows standard deviations of first differences in yields that are part of our data set. The solid line connects the empirical standard deviations computed from the data. These are drawn together with the standard deviations that the estimated models imply for these yield changes.[25] The values in the picture are computed according to the formulas in the appendix, where the parameters are replaced by the maximum likelihood estimates. The figure shows that all models imply a volatility curve that is decreasing in time to maturity. For maturities of two, five and ten years, the three-factor model comes closer to the observed volatility, but it overestimates the volatility at the short end. The two-factor models, in contrast, underestimate the volatility curve for maturities that exceed one year. However, all of these comparisons have to be made with caution since even for differenced yields we do not know how well the empirical standard deviations estimate the true ones.

Our last comments on the estimation results focus on the difference between

[25] See Lemke (2005) for the respective formulas.

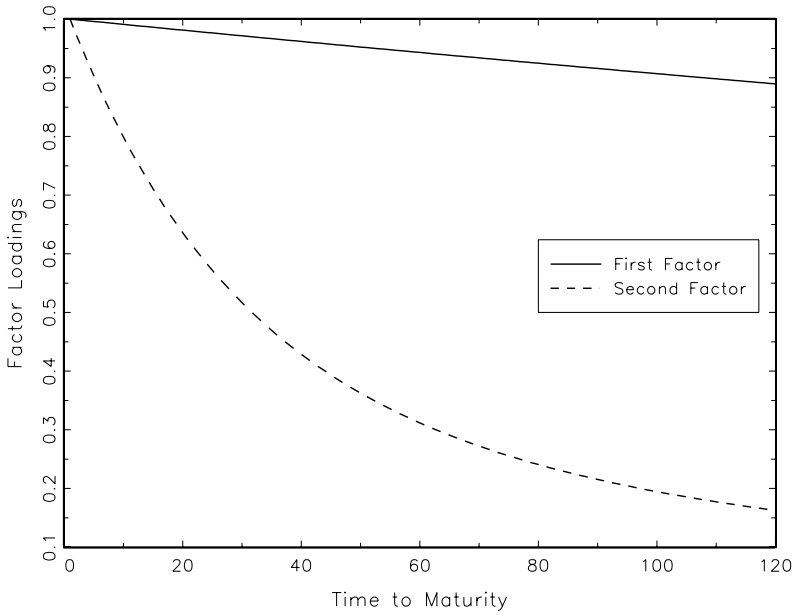

Figure 8: Factor loadings for the two-factor model

the Gaussian models and the mixture model. As already pointed out, table 3 shows that yield changes exhibit considerable excess kurtosis. Multifactor term structure models with Gaussian innovations, however, imply zero excess kurtosis for yields in levels and yields in first differences. Lemke (2005) provides the formula for the kurtosis of differenced yields implied by multifactor models with mixture innovations. Based on the maximum likelihood estimates of the parameters, the kurtosis has been computed for the maturities in the data set. These measures of kurtosis are graphed together with their empirical counterparts in figure 11. The important point to note is that the model in fact implies that the kurtosis is different from zero and that it decreases with maturity. Gaussian models imply a kurtosis which is identically zero for all maturities. One-factor models with mixture innovations, as discussed by Backus et al. (1998) are capable of generating excess kurtosis, but the latter is constant for all maturities. Thus, concerning the matching of fourth moments, our simple two-factor model can be regarded as a step into the right direction.

Up to now we have discussed to what extent our three models are able to capture the behavior of first, second and fourth moments of yields in levels or first differences. Now, we want to look at the distribution at the whole. This will be done exemplarily for the three-month yield, representing the short end of the yield curve, and the five-year yield, representing longer maturities. The analysis is done for first differences again.

The solid line in figure 12 depicts a kernel estimate for the distribution

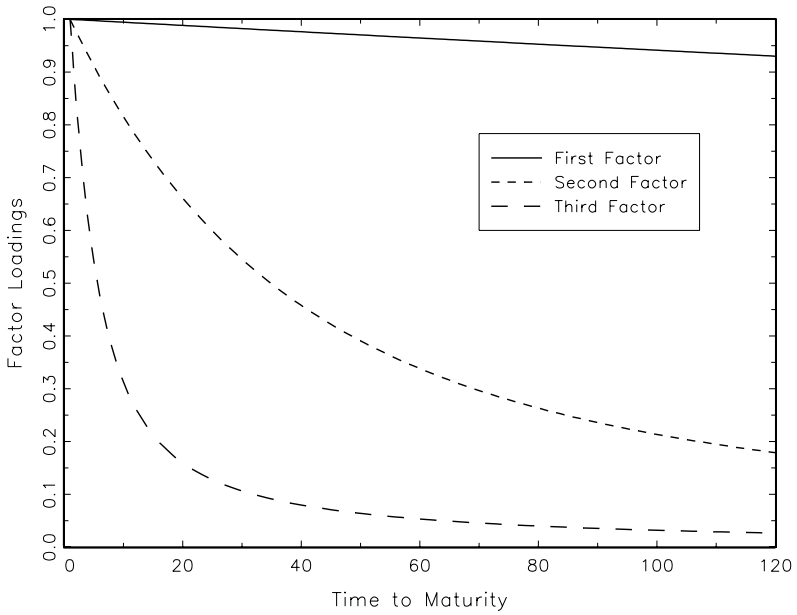

Figure 9: Factor loadings for the three-factor model

of Δy_t^3. It is based on our 444 observations and uses a Gaussian kernel with bandwidth $b = 1.364\hat{\sigma}N$, where N is the number of observations and $\hat{\sigma}$ is their standard deviation.[26] The other lines are the density functions implied by the estimated models. The following describes how they are computed.

For the mixture model, it is not so simple to derive the unconditional density of Δy_t^3. We therefore construct the density implied by the model using a Monte Carlo simulation. Based on the maximum likelihood estimates of the parameters, we generate 10,000 observations of Δy_t^3 from the two-factor mixture model.[27] Based on them, a kernel estimate of the density is constructed and drawn into figure 12. In order to work under the same conditions for all models, the densities for the Gaussian models have been generated by analoguous simulations.

The figure suggests that the two-factor mixture model captures the shape of the density best, followed by the two-factor Gaussian model. The density implied by the three factor model does not appear to capture the distribution well.

We also use QQ-plots for comparing the distributions implied by the models with that given from the data. The QQ-plots in figure 13 (three of them drawn into one picture) are based on the probabilities 0.01, 0.02, ..., 0.99. For each of

[26] This is the default bandwidth suggested by Gauss' TSM package.

[27] The observations are generated without superimposing a measurement error.

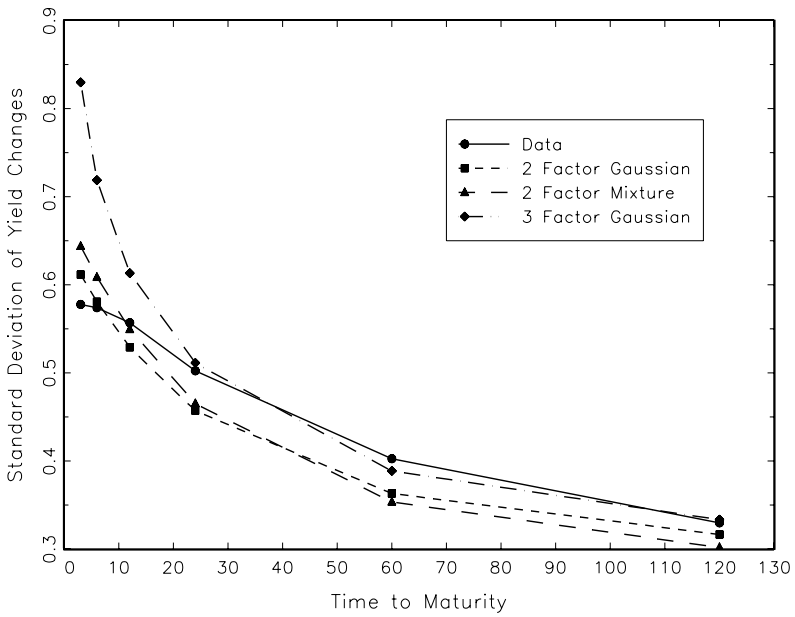

Figure 10: Standard deviation of yield changes

these probabilities, the corresponding quantile implied by the models is plotted against the empirical quantile of the data. If the points corresponding to a model were lying on the 45 degree line, this model would share the same quantiles with the data. Deviations from that line can be interpreted as a measure of distance between the two distributions. Like the density plot above, the QQ-plots suggest that the distribution implied by the two-factor mixture model comes closer to the distribution of the data than that implied by the two-factor Gaussian model. Again, the three-factor model performs worst.

A similar ranking can be inferred by looking at five-year yields. For changes of the five-year yield, figure 14 contains the three densities implied by the models as well as the density estimated from the data. The QQ-plots in figure 15 suggest that for the five-year yield, the advantage of the two-factor mixture model over the other two models shows up quite clearly.

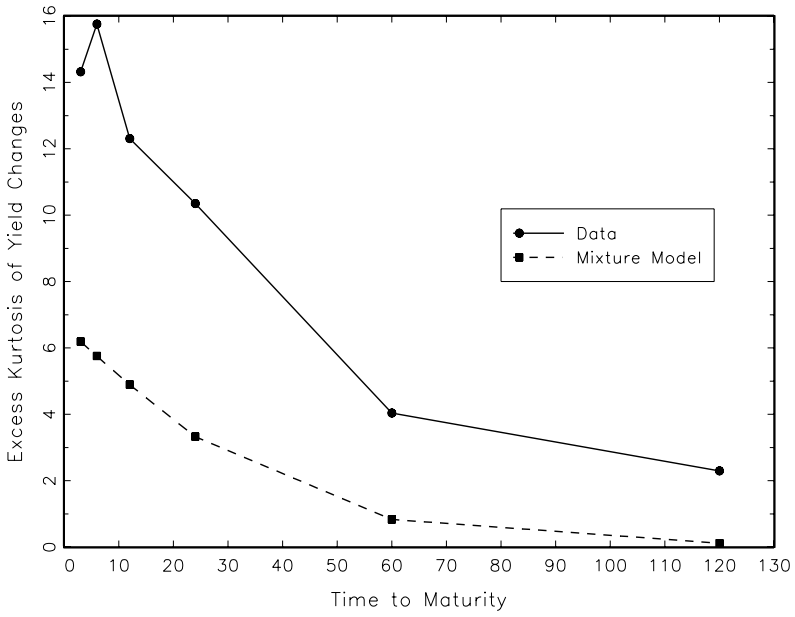

Figure 11: Excess kurtosis of yield changes

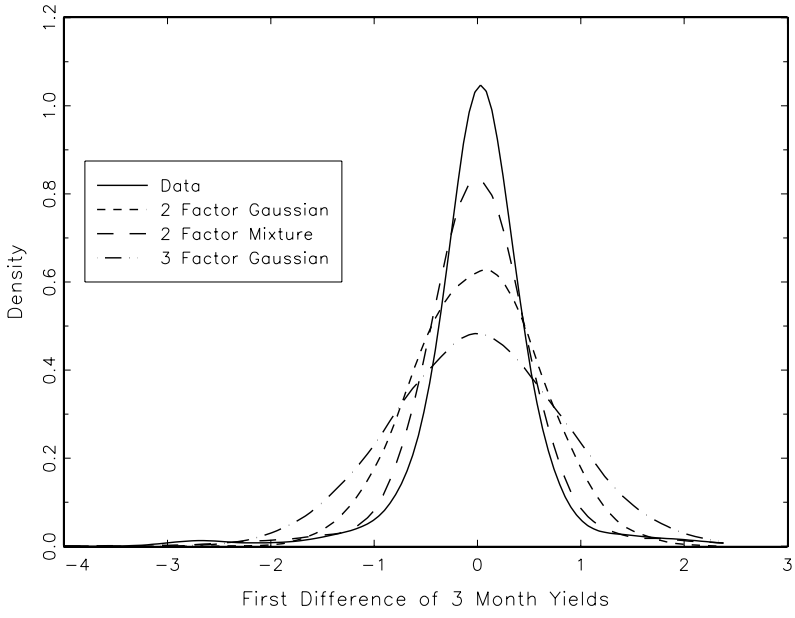

Figure 12: Density of monthly changes in three-month yield

	Two Factors, Gaussian	Two Factors, Mixture	Three Factors, Gaussian
κ_1	0.998 (2.36e-4)	0.998 (2.38e-4)	0.999 (1.48e-4)
v_1	0.000285 (8.27e-6)	0.000263 (1.14e-5)	0.000281 (1.13e-5)
κ_2	0.949 (1.20e-3)	0.950 (1.94e-3)	0.954 (1.34e-3)
v_2	0.000439 (9.54e-6)		0.000510 (2.13e-5)
$\lambda_2 v_2$	-0.142 (0.0560)		-0.0465 (0.0442)
δ	0.00669 (6.19e-4)	0.0121 (2.53e-3)	0.0117 (9.77e-4)
v_{22}		0.000230 (2.10e-5)	
$\lambda_2 v_{22}$		-0.049 (0.0158)	
c_{22}		26.10 (8.130)	
ω		0.138 (0.0386)	
κ_3			0.687 (0.0150)
v_3			0.000506 (2.20e-5)
$\lambda_3 v_3$			-0.340 (0.0518)
h^2	0.0346 (6.96e-4)	0.0345 (1.28e-3)	0.00855 (3.95e-4)
$\ln \mathcal{L}(\hat{\psi})$	-479.90	-404.07	157.36
AIC	973.79	826.13	-294.72

Table 5: Estimation results. Estimated standard errors are given in parentheses.

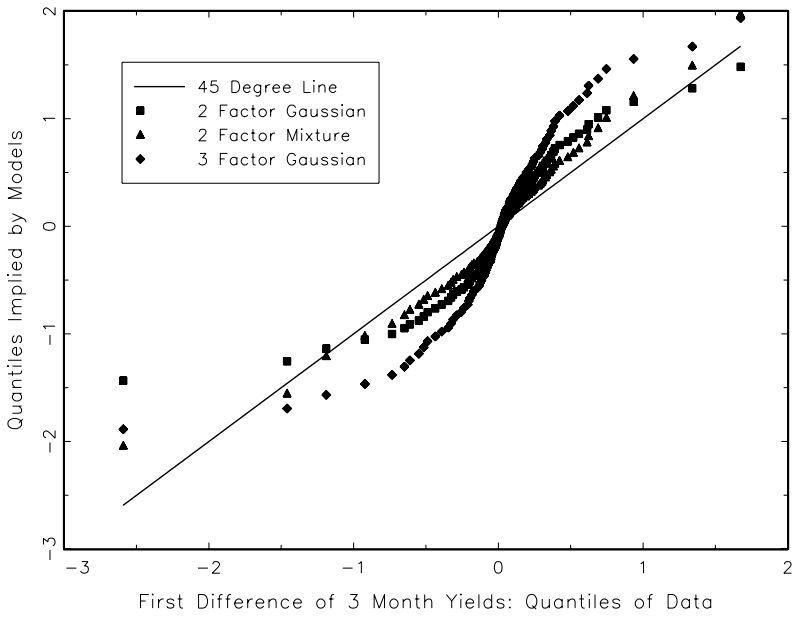

Figure 13: QQ-plots for monthly changes in three-month yield.

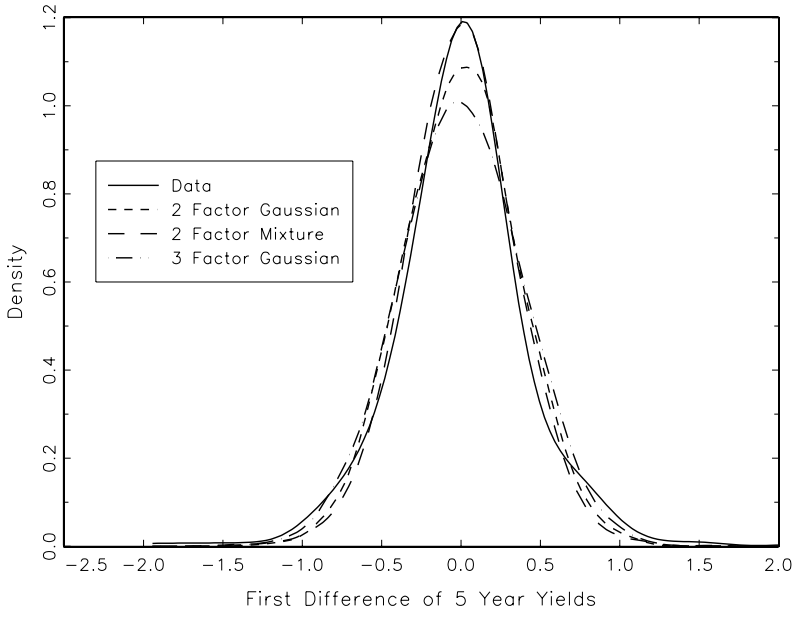

Figure 14: Density of monthly changes in five-year yield

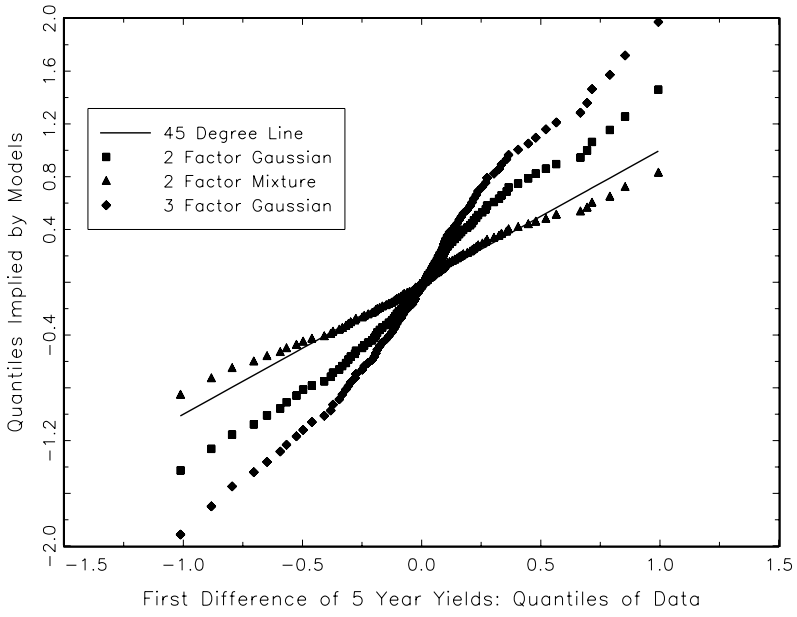

Figure 15: QQ-plots for monthly changes in five-year yield.

6 Conclusion

As a generalization of the one-factor model by Backus et al. (1998), we have introduced a d-factor model, for which the distribution of factor innovations is a Gaussian mixture with B components. This model allows for a flexible modeling of the distribution of yields in levels and first differences, while yields are affine functions of the factors.

For estimation, it has been shown how the theoretical model can be translated into the statistical state space form. The corresponding state space model has a transition equation for which the innovation is distributed as a Gaussian mixture. The exact filter associated with this type of state space model involves Gaussian mixtures with an exponentially growing number of components. In order to make estimation of the model numerically feasible, an approximation of the exact filter, the AMF(k) algorithm, has been introduced.

We have estimated two Gaussian models and one model involving a Gaussian mixture distribution. For the Gaussian models, maximum likelihood estimation based on the Kalman filter has been conducted. For the mixture model, we have employed the AMF(1) algorithm.

Parameter estimates are reasonable in size and have the correct signs. The autocorrelations of residuals do not point towards severe misspecification. The three-factor model is selected by the AIC. However, with respect to higher moments of the data, the two-factor mixture model appears to have an edge over the two Gaussian models.

For future research, an integration of the mixture specification into the three-factor model is imaginable. Furthermore, a more elaborate specification of the mixture distribution could be tried. More reliable standard errors for the parameter estimates may be obtained by using the bootstrap. Moreover, it would be instructive to use a different sample period and another country for the estimation. With regard to the focus of this paper, it would be particularly interesting to see how the estimated distributions of factor innovations change when the sample changes. Finally, it may be a worthwhile attempt to combine the structure of the affine models by Duffie and Kan (1996), in which volatility is level-dependent, with that of our model.

A The Exact Filter for the State Space Model with Mixture Innovations

Consider a state space model for which the transition equation is given by

$$\alpha_t = T\alpha_{t-1} + c + \eta_t, \tag{A.1}$$

where for the innovation vector η_t

$$\eta_t \sim i.i.d. \sum_{b=1}^{B} \omega_b \, N(\mu_b, \, Q_b), \quad \sum_{b=1}^{B} \omega_b = 1, \quad \sum_{b=1}^{B} \omega_b \mu_b = 0. \tag{A.2}$$

That is, the density of η_t is given by[28]

$$p(\eta_t) = \sum_{b=1}^{B} \omega_b \phi(\eta_t; \mu_b, Q_b). \tag{A.3}$$

For the variance-covariance matrix of η_t we have

$$Var(\eta_t) = \sum_{b=1}^{B} \omega_b \left(Q_b + \mu_b \, \mu_b'\right) =: Q. \tag{A.4}$$

The measurement equation is

$$y_t = M\alpha_t + d + \epsilon_t, \tag{A.5}$$

and the measurement error is normally distributed,

$$\epsilon_t \sim i.i.d. \, N(0, H). \tag{A.6}$$

The measurement error ϵ_t and the state innovation η_s are independent for all times s and t.

The weights ω_b as well as the system matrices and vectors T, c, M, d, H, μ_b, and Q_b are all assumed to be time-invariant.

The initial state is assumed to be normally distributed,

$$\alpha_0 \sim N(\bar{a}_0, \bar{P}_0), \tag{A.7}$$

and both, η_t and ϵ_t are independent from the initial state for all t.

We first assume that the system matrices are known and present the exact solution to the filtering problem and the one-step-prediction problem. Let $a_{t|t-1}$, $\hat{y}_{t|t-1}$ and $a_{t|t}$ denote the conditional expectations corresponding to the conditional densities $p(\alpha_t|\mathcal{Y}_{t-1})$, $p(Y_t|\mathcal{Y}_{t-1})$ and $p(\alpha_t|\mathcal{Y}_t)$, and denote by $\Sigma_{t|t-1}$,

[28] $\phi(x; \mu, Q)$ denotes the density function of $N(\mu, Q)$ evaluated at x.

F_t and $\Sigma_{t|t}$ the corresponding variance-covariance matrices. It turns out that for the mixture model, the filtering and prediction densities can be generated in an iterative fashion. They are all mixtures of normals, with the number of components increasing exponentially with time. The relationships between filtering and prediction densities are given by the following theorems.[29]

Theorem A.1 (Prediction density for the mixture model). *Let the filtering density at time $t-1$, $t = 1, 2, \ldots, T$, be given by a Gaussian mixture with l_{t-1} components,*

$$p(\alpha_{t-1}|\mathcal{Y}_{t-1}) = \sum_{i=1}^{l_{t-1}} \omega_{i,\,t-1|t-1} \cdot \phi(\alpha_{t-1};\, a_{i,\,t-1|t-1},\, \Sigma_{i,\,t-1|t-1}).$$

Then the one-step-prediction density for the state is

$$\begin{aligned} & p(\alpha_t|\mathcal{Y}_{t-1}) \\ = & \sum_{b=1}^{B} \sum_{i=1}^{l_{t-1}} \omega_{bi,\,t|t-1}\, \phi(\alpha_t;\, a_{bi,\,t|t-1}, \Sigma_{bi,\,t|t-1}) \end{aligned} \tag{A.8}$$

with

$$\omega_{bi,\,t|t-1} = \omega_b\, \omega_{i,\,t-1|t-1}, \tag{A.9}$$

$$a_{bi,\,t|t-1} = T a_{i,\,t-1|t-1} + c + \mu_b, \tag{A.10}$$

$$\Sigma_{bi,\,t|t-1} = T \Sigma_{i,\,t-1|t-1} T' + Q_b. \tag{A.11}$$

After reindexing and setting $l_t = B \cdot l_{t-1}$ the prediction density can be written as

$$p(\alpha_t|\mathcal{Y}_{t-1}) = \sum_{i=1}^{l_t} \omega_{i,\,t|t-1} \phi(\alpha_t;\, a_{i,\,t|t-1}, \Sigma_{i,\,t|t-1}). \tag{A.12}$$

The one-step-prediction density for the observation vector is

$$p(y_t|\mathcal{Y}_{t-1}) = \sum_{i=1}^{l_t} \omega_{i,\,t|t-1}\, \phi(y_t;\, \hat{y}_{i,\,t|t-1}, F_{i,\,t}) \tag{A.13}$$

with

$$\hat{y}_{i,\,t|t-1} = M a_{i,\,t|t-1} + d, \tag{A.14}$$

$$F_{i,\,t} = M \Sigma_{i,\,t|t-1} M' + H. \tag{A.15}$$

Theorem A.2 (Filtering density for the mixture model). *Let the prediction densities $p(\alpha_t|\mathcal{Y}_{t-1})$ and $p(y_t|\mathcal{Y}_{t-1})$ at time t, $t = 1, 2, \ldots, T$, be given by the Gaussian mixtures (A.12) and (A.13). Then the filtering density is*

$$p(\alpha_t|\mathcal{Y}_t) = \sum_{i=1}^{l_t} \omega_{i\,t|t}\, \phi(\alpha_t;\, a_{i,\,t|t}, \Sigma_{i,\,t|t}) \tag{A.16}$$

[29] The earliest derivation of these relations for the case of scalar measurement and transition equation may be attributed to Sorenson and Alspach (1971).

with

$$
\begin{aligned}
a_{i,\,t|t} &= a_{i,\,t|t-1} + K_{i,\,t}(y_t - \hat{y}_{i,\,t|t-1}), && \text{(A.17)}\\
\Sigma_{i,\,t|t} &= \Sigma_{i,\,t|t-1} - K_{i,\,t} M \Sigma_{i,\,t|t-1}, && \text{(A.18)}\\
K_{i,\,t} &= \Sigma_{i,\,t|t-1} M' F_{i,\,t}^{-1}, && \text{(A.19)}\\
\omega_{i,\,t|t} &= \frac{\omega_{i,\,t|t-1}\,\phi(y_t;\,\hat{y}_{i,\,t|t-1}, F_{i,\,t})}{\sum_{i=1}^{l_t} \omega_{i,\,t|t-1}\,\phi(y_t;\,\hat{y}_{i,\,t|t-1}, F_{i,\,t})}. && \text{(A.20)}
\end{aligned}
$$

For a proof, see Lemke (2005).

A remark is in order that theorems A.1 and A.2 are in fact applicable to time $t = 1$. For the initial filtering density used in theorem A.1 we have $p(\alpha_0|\mathcal{Y}_0) = p(\alpha_0|1_N) = p(\alpha_0)$. Thus, technically speaking, the filtering density is the density of the initial state, that has been specified in (A.7) as a normal. It can be written as a mixture with one component, $l_0 = 1$, thus

$$
p(\alpha_0|\mathcal{Y}_0) = \sum_{i=1}^{l_0} 1\,\phi(\alpha_0;\,\bar{a}_0, \bar{P}_0).
$$

Hence, theorem A.1 can be applied to this density yielding $p(\alpha_1|\mathcal{Y}_0)$ and $p(y_1|\mathcal{Y}_0)$ as mixtures with B components. To these in turn, theorem A.2 can be applied yielding $p(\alpha_1|\mathcal{Y}_1)$.

With the conditional densities at hand, point estimators can be readily computed as the corresponding conditional expectations,

$$
\begin{aligned}
E(\alpha_t|\mathcal{Y}_{t-1}) &= \sum_{i=1}^{l_t} \omega_{i,\,t|t-1}\,a_{i,\,t|t-1} =: a_{t|t-1}, && \text{(A.21)}\\
E(y_t|\mathcal{Y}_{t-1}) &= \sum_{i=1}^{l_t} \omega_{i,\,t|t-1}\,\hat{y}_{i,\,t|t-1} =: \hat{y}_{t|t-1}, && \text{(A.22)}\\
E(\alpha_t|\mathcal{Y}_t) &= \sum_{i=1}^{l_t} \omega_{i,\,t|t}\,a_{i,\,t|t} =: a_{t|t}. && \text{(A.23)}
\end{aligned}
$$

The corresponding conditional variance-covariance matrices are given by

$$
\begin{aligned}
Var(\alpha_t|\mathcal{Y}_{t-1}) &= \sum_{i=1}^{l_t} \omega_{i,\,t|t-1}\,\left(\Sigma_{i,\,t|t-1} + (a_{i,\,t|t-1} - a_{t|t-1})(a_{i,\,t|t-1} - a_{t|t-1})'\right)\\
&=: \Sigma_{t|t-1} && \text{(A.24)}\\
Var(y_t|\mathcal{Y}_{t-1}) &= \sum_{i=1}^{l_t} \omega_{i,\,t|t-1}\,\left(F_{i,\,t} + (\hat{y}_{i,\,t|t-1} - \hat{y}_{t|t-1})(\hat{y}_{i,\,t|t-1} - \hat{y}_{t|t-1})'\right)\\
&=: F_{t|t-1} && \text{(A.25)}\\
Var(\alpha_t|\mathcal{Y}_t) &= \sum_{i=1}^{l_t} \omega_{i,\,t|t}\,\left(\Sigma_{i,\,t|t} + (a_{i,\,t|t} - a_{t|t})(a_{i,\,t|t} - a_{t|t})'\right)\\
&=: \Sigma_{t|t}. && \text{(A.26)}
\end{aligned}
$$

The latter results follow from the general properties of Gaussian mixtures. Note that the expectation is just the weighted average of the expectations of the normal densities that constitute the mixture, whereas the variance has an additional term taking the variation of the means into account.

The steps of the exact filter for the mixture model can be summarized as follows:

Given observations $\{y_1, \ldots, y_T\}$, and an initial density $\alpha_0 \sim N(\bar{a}_0, \bar{P}_0)$, the algorithm computes

- the sequences of conditional densities,

$$\begin{aligned} &p(\alpha_t|\mathcal{Y}_{t-1}), \quad t = 1, \ldots, T, \\ &p(y|\mathcal{Y}_{t-1}), \quad t = 1, \ldots, T, \\ &p(\alpha_t|\mathcal{Y}_t), \quad t = 1, \ldots, T, \end{aligned}$$

  each characterized by the corresponding components (weights, means, variances),

$$\begin{aligned} &\omega_{i,\,t|t-1}, a_{i,\,t|t-1}, \Sigma_{i,\,t|t-1}, \quad i = 1, \ldots, l_t \quad t = 1, \ldots, T, \\ &\omega_{i,\,t|t-1}, \hat{y}_{i,\,t|t-1}, F_{i,\,t}, \quad i = 1, \ldots, l_t \quad t = 1, \ldots, T, \\ &\omega_{i,\,t|t}, a_{i,\,t|t}, \Sigma_{i,\,t|t}, \quad i = 1, \ldots, l_t \quad t = 1, \ldots, T, \end{aligned}$$

- and the sequences of point estimates (conditional means) and corresponding variance covariance matrices

$$\begin{aligned} &a_{t|t-1}, \Sigma_{t|t-1}, \quad t = 1, \ldots, T, \\ &\hat{y}_{t|t-1}, F_t, \quad t = 1, \ldots, T, \\ &a_{t|t}, \Sigma_{t|t}, \quad t = 1, \ldots, T. \end{aligned}$$

These are computed according to the following scheme:

Algorithm A.1 (The exact filter).

- ***Step 1, Initialization***

  Set

$$a_{1,\,0|0} = \bar{a}_0, \quad \Sigma_{1,\,0|0} = \bar{P}_0, \quad \omega_{1,\,0|0} = 1, \quad l_0 = 1.$$

  Set $t = 1$.

- ***Step 2, Prediction step from*** $t-1$ ***to*** t

  Set $l_t = B^t$.

  Compute $\omega_{i,\,t|t-1}$, $a_{i,\,t|t-1}$, $\Sigma_{i,\,t|t-1}$, $\hat{y}_{i,\,t|t-1}$, *and* $F_{i,\,t}$ *for* $i = 1, \ldots, l_t$, *according to theorem A.1.*

  Use these quantities to compute $a_{t|t-1}$, $\hat{y}_{t|t-1}$, $\Sigma_{t|t-1}$, , *and* F_t *according to (A.21), (A.22), (A.24) and (A.25) respectively.*

- ***Step 3, Updating step at*** t

  Compute $\omega_{i,\,t|t}$, $a_{i,\,t|t}$, *and* $\Sigma_{i,\,t|t}$, *for* $i = 1, \ldots, l_t$, *according to theorem A.2.*

  Use theses quantities to compute $a_{t|t}$ *and* $\Sigma_{t|t}$, *according to (A.23) and (A.26), respectively.*

- ***Step 4***

  If $t < T$, *set* $t := t + 1$, *and go to Step 2;*

  else, STOP.

If the moments of the initial conditions are not known, one can proceed as in the case of a simple normal. If the state process is stationary, the filter can be initialized using the unconditional mean and variance-covariance matrix. The condition for stationarity of the state process is the same as in the Gaussian case: all eigenvalues of the transition matrix T have to have modulus less than one.

B The Approximate Filter AMF(k)

The following gives the algorithm for the approximate mixture filter of order k (AMF(k)).

Algorithm B.1 (The approximate filter AMF(k)).

- ***Step 1***

  Apply the exact filter to the sequence $\{y_1, \ldots, y_k\}$ *with initial condition* $\alpha_0 \sim N(\bar{a}_0, \bar{P}_0)$.

  Obtain the exact filtering densities

$$p(\alpha_t|\mathcal{Y}_t), \quad p(\alpha_t|\mathcal{Y}_{t-1}), \quad p(y_t|\mathcal{Y}_{t-1}), \quad t = 1, \ldots, k,$$

  with corresponding moments

$$a_{t|t},\ \Sigma_{t|t}, \quad a_{t|t-1},\ \Sigma_{t|t-1}, \quad \hat{y}_{t|t-1},\ F_t.$$

- ***Step 2***

  For $t = 1, \ldots, k$ *set:*

$$\tilde{p}(\alpha_t|\mathcal{Y}_t) = p(\alpha_t|\mathcal{Y}_t), \quad \tilde{a}_{t|t} = a_{t|t}, \quad \tilde{\Sigma}_{t|t} = \Sigma_{t|t}$$
$$\tilde{p}(\alpha_t|\mathcal{Y}_{t-1}) = p(\alpha_t|\mathcal{Y}_{t-1}), \quad \tilde{a}_{t|t-1} = a_{t|t-1}, \quad \tilde{\Sigma}_{t|t-1} = \Sigma_{t|t-1}$$
$$\tilde{p}(y_t|\mathcal{Y}_{t-1}) = p(y_t|\mathcal{Y}_{t-1}), \quad \tilde{\hat{y}}_{t|t-1} = \hat{y}_{t|t-1}, \quad \tilde{F}_t = F_t.$$

  Set $t = k + 1$.

- ***Step 3***

  Apply the exact filter to the sequence $\{y_{t-k+1}, \ldots, y_t\}$ *with initial condition* $\alpha_{t-k} \sim N(\tilde{a}_{t-k|t-k}, \tilde{\Sigma}_{t-k|t-k})$.

  Store the final filtering and prediction densities as $\tilde{p}(\alpha_t|\mathcal{Y}_t)$, $\tilde{p}(\alpha_t|\mathcal{Y}_{t-1})$, *and* $\tilde{p}(y_t|\mathcal{Y}_{t-1})$. *That is, store the corresponding components* $\tilde{\omega}_{i,\,t|t}$, $\tilde{a}_{i,\,t|t}$,$\tilde{\Sigma}_{i,\,t|t}$, $\tilde{\omega}_{i,\,t|t-1}$, $\tilde{a}_{i,\,t|t-1}$,$\tilde{\Sigma}_{i,\,t|t-1}$, $\tilde{\hat{y}}_{i,\,t|t-1}$, $\tilde{F}_{i,\,t}$, $\quad i = 1, \ldots, B^k$.

  Compute the corresponding means and variances $\tilde{a}_{t|t}$, $\tilde{\Sigma}_{t|t}$, $\tilde{a}_{t|t-1}$, $\tilde{\Sigma}_{t|t-1}$, $\tilde{\hat{y}}_{t|t-1}$, *and* $\tilde{F}_t$.

- ***Step 4***

  If $t < T$, *set* $t := t + 1$, *and go to Step 3;*

  else, STOP.

References

Babbs, S. H. and Nowman, K. B. (1998). An Application of Generalized Vasicek Term Structure Models to the UK Gilt-edged Market: a Kalman Filter Analysis. *Applied Financial Economics*, 8:637–644.

— (1999). Kalman Filtering of Generalized Vasicek Term Structure Models. *Journal of Financial and Quantitative Analysis*, 34:115–130.

Backus, D., Foresi, S., and Telmer, C. (1998). Discrete-Time Models of Bond Pricing. Working Paper 6736, NBER.

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

Bliss, R. R. (1997). Testing Term Structure Estimation Methods. *Advances in Futures and Options Research*, 9:197–231.

Brockwell, P. J. and Davis, R. A. (1996). *Time Series : Theory and Methods*. Oxford University Press, New York et al., 2. edition.

Campbell, J., Lo, A., and MacKinlay, A. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Cassola, N. and Luis, J. B. (2003). A Two-Factor Model of the German Term Structure of Interest Rates. *Applied Financial Economics*, 13:783–806.

Cochrane, J. (2001). *Asset Pricing*. Princeton University Press, Princeton et al.

de Jong, F. (2000). Time Series and Cross-Section Information in Affine Term-Structure Models. *Journal of Business and Economic Statistics*, 18:300–314.

Duan, J. and Simonato, J. (1999). Estimating and Testing Exponential-Affine Term Structure Models by Kalman Filter. *Review of Quantitative Finance and Accounting*, 13:111–135.

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

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

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford et al.

Geyer, A. L. J. and Pichler, S. (1999). A State-space Approach to Estimate and Test Multifactor Cox-Ingersoll-Ross Models of the Term Structure. *Journal of Financial Research*, 22:107–130.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Irle, A. (1998). *Finanzmathematik - Die Bewertung von Derivaten*. Teubner, Stuttgart.

Lemke, W. (2005). *Term Structure Modeling and Estimation in a State Space Framework*. Ph.D. thesis, University of Bielefeld.

McCulloch, J. H. and Kwon, H. C. (1993). U.S. Term Structure Data, 1947-1991. Working Paper 93-6, Ohio State University.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York et al.

Piazzesi, M. (2005). Affine Term Structure Models. In Y. Ait-Sahalia and L. P. Hansen, editors, *forthcoming: Handbook of Financial Econometrics*.

Schwaar, C. (1999). *Kalman-Filter basierte ML-Schätzung affiner, zeithomogener Faktormodelle der Zinsstruktur am bundesdeutschen Rentenmarkt*. Europäische Hochschulschriften. Peter Lang, Frankfurt et al.

Sorenson, H. W. and Alspach, D. L. (1971). Recursive Bayesian Estimation Using Gaussian Sums. *Automatica*, 7:465–479.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester et al.