DEVELOPMENTS IN LIKELIHOOD-BASED METHODS FOR STATE SPACE MODELS
Developments in Likelihood-Based Methods for State Space Models

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan de Vrije Universiteit Amsterdam, op gezag van de rector magnificus prof.dr. L.M. Bouter, in het openbaar te verdedigen ten overstaan van de promotiecommissie van de faculteit der Economische Wetenschappen en Bedrijfskunde op donderdag 18 november 2010 om 13.45 uur in de aula van de universiteit, De Boelelaan 1105

door
Borus Martinus Johannes Petrus Jungbacker
geboren te Velsen
promotor: prof.dr. S.J. Koopman
Acknowledgments

Writing a PhD thesis can be a very challenging undertaking. When I started my research project I tried not to think too much about it. When I did think about it I found it hard to believe I would be able to produce interesting research from scratch and in ‘only’ four years at that. Looking back at this past period I realise that, as so many things in life, doing a PhD is not something you do on your own. You need people to support you along the way. A supervisor to point out new angles to look at the problem you have been hopelessly stuck with for weeks and friends and family to cheer you up when the article you worked on for a year gets rejected by an editor. I am sure I would never have made it this far without the support of such wonderful people.

First I thank my supervisor Siem Jan Koopman. His deep insight in time series analysis and intimate knowledge of the state space methodology was invaluable for my research, but his moral support has been just as important. His optimism and encouragement played a major role in the successful completion of my thesis. I deeply admire him both for his scientific contributions as well as his energy and enthusiasm.

Besides my supervisor I had the fortune to work with a number of great colleagues. I would especially like to thank my roommates Brian, Irma, Bahar and Taoying. Their cheerfulness livened up some of the particularly bad days. I also thank the other PhD students often joining us for lunch, especially Michel, Bernd and Ting.

During the ten years I spent at the VU I have met many special people. In fear I will forget someone I will not try to name all of them. I do want to thank four people especially: Aukje, Tamar, Carolien and Robert. I have especially fond memories of my coffee sessions with Tamar sitting on the bench on the first floor looking out through the VU windows.

Finally I would like to thank the two most important people in my life. First my mother, Marian. I will be forever thankful for your support and love. I can safely say that I would never made it this far without you. I could not have wished for a better mother. Finally I would like to thank Shimrit, who has been with me for the last part of this journey. Doing a PhD might have been difficult, but you made sure it was never lonely. I very much enjoyed...
this time with you and look forward to many more years together.
Contents

1 Introduction 7

2 The Linear State Space Model 13
  2.1 Introduction ....................................................... 13
  2.2 Examples of Linear State Space Models ...................... 14
  2.3 The State Space Model in Matrix Form ....................... 16
  2.4 Minimum Mean Square Linear Estimation .................... 17
  2.5 Prediction ......................................................... 18
  2.6 State Smoothing .................................................... 20
  2.7 Filtering and Predicting the States .......................... 23
  2.8 Gaussian Likelihood .............................................. 23
  2.9 Univariate Treatment of Multivariate Time Series .......... 24
  2.10 Missing Observations ............................................. 25
  2.11 Unspecified Initial Conditions ................................. 26
    2.11.1 State Estimation ............................................ 27
    2.11.2 The Marginal Likelihood .................................... 30
    2.11.3 Regression Effects and the Marginal Likelihood ........ 30
  2.12 Diffuse Kalman filter ........................................... 31
    2.12.1 A First Transformation of the Model .................... 32
    2.12.2 Finding $\tilde{J}_1$ and $\tilde{J}_2$ .......................... 34
    2.12.3 Constructing a New State Space Model .................... 35
    2.12.4 Simplifying the Calculations ............................... 37
  2.A The Kalman Filter and the LDL decomposition ............... 39

3 Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models 43
  3.1 Introduction ....................................................... 43
  3.2 The Non-Gaussian State Space Model ........................... 44
  3.3 Simulation-Based Inference ....................................... 46
## CONTENTS

3.3.1 Importance Sampling ........................................ 46
3.3.2 Bayesian Analysis .......................................... 47
3.4 Computing the Posterior Mode ................................. 47
3.5 Sampling from the Proposal Density ....................... 50
3.6 Empirical Illustration ....................................... 53
3.A Proof of Theorem 3.2 ....................................... 56
3.C Derivation of Equation (3.24) ............................... 60
3.D Derivation of Equation (3.25) ............................... 61

4 Analysis of Dynamic Factor Models .......................... 63
4.1 Introduction .................................................. 63
4.2 Generalized Dynamic Factor Model with Covariates ...... 65
4.3 Parameter Estimation ....................................... 68
4.4 Estimation of Factors and Likelihood Evaluation ........ 70
  4.4.1 Transforming the Observation Equation ................ 70
  4.4.2 Estimation of factors .................................. 73
  4.4.3 Evaluation of the Gaussian Log-likelihood ............ 73
  4.4.4 Partial Concentration of Regression Coefficients .... 75
  4.4.5 Evaluation of the Marginal Gaussian Log-likelihood ... 76
  4.4.6 Estimation of Regression Coefficients ................. 77
  4.4.7 Computational Gains .................................. 79
4.5 Maximizing the Quasi-Likelihood Function ................. 79
  4.5.1 Calculating the Analytical Score ...................... 80
  4.5.2 The EM Algorithm ..................................... 81
4.6 An Empirical Illustration .................................. 82
  4.6.1 Parameter Estimation .................................. 83
  4.6.2 Factor Estimates and Factor Loadings ................. 84
  4.6.3 Diagnostic Checking .................................... 87
4.A Proof of Lemma 4.1 ....................................... 89
4.B Proof of Lemma 4.2 ....................................... 89
4.C Proof of Equation (4.30) .................................. 90
4.D Proof of Results of Section 4.4.6 ......................... 93
4.E The Score Function of Section 4.5.1 ....................... 94
# Dynamic Factor Models with Smooth Factor Loadings

5.1 Introduction .................................. 97
5.2 The Dynamic Factor Model .......................... 99
  5.2.1 Stationary Specification ......................... 102
  5.2.2 Nonstationary Specification ....................... 103
  5.2.3 Parameter Estimation and Signal Extraction .......... 105
5.3 Dynamic Factor Model with Smooth Factor Loadings ............. 106
  5.3.1 Model Specification .......................... 106
  5.3.2 Selecting Knots via a Wald Test Procedure .......... 107
  5.3.3 A General Version of the SDFM .................. 109
5.4 Dynamic Factor Models for the Term Structure of Interest Rates ..... 110
  5.4.1 Functional Signal Plus Noise Model ............... 111
  5.4.2 Nelson-Siegel Model ........................ 112
  5.4.3 Arbitrage-free Nelson-Siegel Model ............... 113
  5.4.4 Gaussian Affine Term Structure Model ............. 114
5.5 Data Description .................................. 115
5.6 Empirical Results ................................ 116
  5.6.1 General Model Assumptions ....................... 117
  5.6.2 Estimation Results for the DFM .................. 117
  5.6.3 Imposing the SDFM Restrictions ................... 119
  5.6.4 Estimation Results for Other Term Structure Models .... 120
  5.6.5 Model Fit ................................ 122
  5.6.6 Testing the Restrictions on the Factor Loadings .... 123
5.7 Conclusion ...................................... 124
5.A Tables and Figures ................................ 126

# Analysis of Dynamic Factor Models in the Presence of Missing Data

6.1 Introduction .................................. 141
6.2 The Dynamic Factor Model .......................... 142
  6.2.1 Model Specification .......................... 142
  6.2.2 Two State Space Representations ............... 143
  6.2.3 Missing Data ................................ 144
6.3 State Space Formulation in the Presence of Missing Data .......... 145
  6.3.1 Notation ................................ 145
  6.3.2 The Missing Data State Space Formulation ......... 146
  6.3.3 Discussion of the New Formulation ............... 147
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4</td>
<td>Signal Extraction and Likelihood Evaluation</td>
<td>148</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Estimation of States and Idiosyncratic Components</td>
<td>148</td>
</tr>
<tr>
<td>6.4.2</td>
<td>KFS with a Collapsed Observation Vector</td>
<td>149</td>
</tr>
<tr>
<td>6.4.3</td>
<td>A Collapsed KFS in Presence of Missing Data</td>
<td>150</td>
</tr>
<tr>
<td>6.4.4</td>
<td>Log-likelihood Evaluation</td>
<td>151</td>
</tr>
<tr>
<td>6.5</td>
<td>Maximum Likelihood Estimation</td>
<td>152</td>
</tr>
<tr>
<td>6.6</td>
<td>Computational Costs and Gains</td>
<td>153</td>
</tr>
<tr>
<td>6.7</td>
<td>Conclusions</td>
<td>153</td>
</tr>
</tbody>
</table>

**Samenvatting (Summary in Dutch)** | 163
Chapter 1

Introduction

In this thesis I discuss the state space approach to time series modelling. Suppose that we observe a multivariate time series and denote the observations by $y_1, \ldots, y_n$. In a state space model we assume that each vector $y_t$ depends in a prescribed way on the value of an unobserved stochastic process $\alpha_t$. This process $\alpha_t$ is known as the state process. In many applications $\alpha_t$ will have a clear economic interpretation. In a structural time series model, for instance, the states represent seasonal and trend components of the observed time series, see Harvey (1989, Chapter 2) for a discussion of this approach. In a stochastic volatility model, the unobserved state represents the time-varying risk of the asset under consideration and in a dynamic factor model we can often link the unobserved factors to the business cycle or other macro-economic developments. The state space model provides a convenient framework to construct intuitively appealing and economically sensible time series models.

In many empirical studies involving state space models the key objective is to predict future time series observations. However, because of the structure of the state space model, it is generally necessary to first obtain state estimates before we can generate predictions. These state estimates can also give us an idea of the nature of the time series dynamics or signal possible misspecification if we have a strong a-priori idea about the development of the state. This is especially true if the state process has an economic interpretation. Almost every non-trivial statistical model depends on a set of unknown model parameters. Estimation of these parameters is the first step in any empirical study that involves a state space model. These three issues, parameter estimation, state estimation and prediction are the main focus of this thesis.

The class of state space models is very broad. In practice, we therefore need to tailor the estimation methods to the characteristics of the state space model in question. Compu-
tationally efficient estimation methods are particularly well developed for the class of linear state space models. For these models we can obtain mean square optimal predictions using an algorithm known as the Kalman filter and mean square optimal state estimates can be found using the related Kalman smoother recursions. If the model has Gaussian innovations we can also use the Kalman filter to evaluate the likelihood. We obtain maximum likelihood estimators of the model parameters by simply maximizing the likelihood function using a numerical optimization algorithm. In this thesis I consider two classes of state space models for which this methodology is not directly applicable: dynamic factor models and non-Gaussian state space models.

Dynamic factor models belong to the class of linear state space models. This means that we could in theory use the Kalman filter and Kalman smoother recursions to estimate the underlying states and a maximum likelihood procedure to estimate the parameters. However, this approach is inapplicable for high-dimensional time series. Dynamic factor models for such time series panels typically contain a large number of parameters and numerical optimization of the likelihood function, obtained from the Kalman filter, is therefore infeasible. This is an important issue, since panels consisting of hundreds of time series are becoming increasingly common in applied econometric research. I discuss in this thesis new and computationally efficient methods for likelihood-based analysis of high-dimensional dynamic factor models. These new results enable us to routinely estimate dynamic factor models even for the high-dimensional data sets found in recent applied macroeconomic research. I illustrate the new methods in an empirical study of the time series properties of the term structure of interest rates for US treasuries.

For non-Gaussian state space models we generally do not have closed-form expressions for the likelihood and the mean square optimal predictions and state estimates. The most common solution to this problem is to use numerical methods based on simulation. In this thesis I extend such a method, originally proposed by Shephard and Pitt (1997) and Durbin and Koopman (1997), to a wider class of models. I prove the surprising result that all necessary computations can be performed by applying the Kalman filter methods to a state space model with ‘covariance’ matrices that are not positive semi-definite. Although such a model with ‘negative variances’ is clearly not well-defined, I show that the results obtained from the Kalman filter methods still have a clear interpretation and are of practical use.

I conclude this introduction with short descriptions of the content of each chapter.
Chapter 2: The Linear State Space Model

This chapter reviews all concepts related to the linear state space model which are needed for the developments in the remaining chapters. Most of the results in this chapter are well-known and can also be found in standard textbooks on linear state space models such as Anderson and Moore (1979), Harvey (1981), West and Harrison (1989) and Durbin and Koopman (2001). The new results in this chapter are Lemma 2.A and Corollary 2.1 that generalize similar results in Harvey (1981) and Durbin and Koopman (2001). Lemma 2.A was originally presented in Jungbacker and Koopman (2007). Although the results in this chapter are well-known, the discussion is somewhat non-standard in its focus on the interpretation of the Kalman filter and Kalman smoother methods in terms of an LDL decomposition of the variance matrix of the observations. The derivation of the diffuse Kalman filter is new.

Chapter 3: Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models

In this chapter I consider inference methods for a class of non-Gaussian state space models. The model specifies the states $\alpha_t$ as a linear Gaussian process while the observations $y_t$ are modelled by means of a family of conditional densities $p(y_t|\theta_t, \psi)$, where $\theta_t$ is a linear function of $\alpha_t$ and $\psi$ is a set of unknown model parameters. Analytical expressions for the likelihood and mean square optimal state estimators and predictors are generally unavailable for this type of models. The standard method to overcome this problem is the use of simulation-based methods. Importance sampling methods can be used for instance to approximate the likelihood and obtain state estimates for a fixed value of $\psi$.

Let $\theta$ denote the vector $(\theta_1', \ldots, \theta_n')'$ and let $p(\theta|y, \psi)$ denote the density of $\theta_1, \ldots, \theta_n$ conditional on $y_1, \ldots, y_n$. Importance sampling requires the choice of an importance density $f(\theta, y; \psi)$ that can be easily sampled and evaluated and that is in some sense close to $p(\theta|y, \psi)$. Shephard and Pitt (1997) and Durbin and Koopman (1997) proposed, in respectively Bayesian and frequentist settings, to choose a Gaussian density $f(\theta, y; \psi)$ with the same mode and curvature around the mode as $p(\theta|y, \psi)$. They also presented numerical methods that can be used to implement this approach for log-concave densities $p(y_t|\theta_t, \psi)$. In this chapter I show that these methods can be extended relatively easily such that they can be used if $p(y_t|\theta_t, \psi)$ is not log-concave. The key to this extension is a set of results that show how the output of the Kalman filter, Kalman smoother and simulation smoother recursions can be interpreted if we have ‘variance’ matrices that are not positive semi-definite. I demonstrate the applicability of this method by estimating a stochastic volatility with leverage model using maximum likelihood. This chapter is based on the results of Jungbacker...
and Koopman (2007). This approach to maximum likelihood estimation of non-linear non-
Gaussian state space models was also applied in Koopman, Jungbacker, and Hol (2005),

Chapter 4: Likelihood-based Analysis for Dynamic Factor Models

In the dynamic factor models considered in this chapter it is assumed that a large number of
time series depends on a relatively small number of unobserved stochastic processes known
as factors. The purpose of this chapter is to show how to obtain maximum likelihood
estimators of the model parameters, how to efficiently compute states estimates and how
to obtain predictions of future observations. Since the models considered can be written
in state space form we can use the Kalman filter to evaluate the likelihood if the model is
Gaussian. For non-Gaussian models this likelihood can be considered a quasi-likelihood and
the quasi-maximum likelihood estimators obtained by maximizing this function can be shown
to be consistent and asymptotically normal. Similarly we can use the Kalman smoother to
obtain state estimates. Such an approach, however, is infeasible in practice because of the
high dimensions involved. A typical dynamic factor model can include more than hundred
time series and more than thousand parameters, see also the empirical illustration in this
chapter. I present new results that can be used to efficiently compute the (quasi-)likelihood
and the score. This allows us to obtain (quasi-)maximum likelihood estimators for the model
parameters in a matter of minutes even for the large models that are common in recent
applied econometric research. The developments in this chapter are based on Jungbacker

Chapter 5: Dynamic Factor Models with Smooth Loadings

The main assumption of the dynamic factor model of chapter 4 is that a potentially large
number of time series depends linearly on a small set of unobserved stochastic processes.
The coefficients that determine how the unobserved factors influence the individual time
series are generally referred to as the factor loadings. In many applications it is reasonable
to assume that time series with common features will react in the same way to changes in
the underlying factor. In this chapter I discuss the example of yields of zero coupon bonds.
If we model a time series panel consisting of monthly yield curves as a dynamic factor model,
it is likely that yields corresponding to bonds with similar maturities will also have similar
factor loadings. Specifically, we expect that the factor loadings are a smooth function of the
time to maturity of the associated bonds. In this chapter I present a unified approach to
the modelling of such time series panels. I propose a new class of dynamic factor models
that imposes these smoothness restrictions on the factor loadings in an intuitive manner. Further, I present a statistical method to choose the optimal set of restrictions based on sets of Wald statistics. Maximum likelihood estimators of the model parameters can be obtained using the computationally efficient methods of chapter 4.

To illustrate the applicability of this new class of dynamic factor models I perform an extensive empirical study of the time series properties of the US treasuries term structure of interest rates. In this study I consider a number of popular term structure models that can be viewed as dynamic factor models with smoothness restrictions imposed on the factor loadings. I show that the restrictions imposed by these models are strongly rejected by standard likelihood ratio tests. I proceed by showing that we can construct a dynamic factor model with smooth factor loadings using the new methodology introduced in this chapter. This model is parsimonious and the restrictions are not rejected by the likelihood ratio test. The developments in this chapter are based on Jungbacker, Koopman, and van der Wel (2009b).

Chapter 6: Dynamic Factor Analysis in the Presence of Missing Data

In this chapter I extend the discussion of Chapter 4 to data sets where some of the observations are missing. I focus on dynamic factor models where the idiosyncratic component follows an vector autoregressive process. In Chapter 4 this type of model is written in state space form, which allows us to use efficient Kalman filter and smoother methods to estimate the factors and evaluate the likelihood. When some of the observations are missing, this state space formulation is however no longer valid. One possible solution proposed in the literature is to view all idiosyncratic components as unobserved states. This approach slows down the Kalman filter methods dramatically however. I show how to construct an alternative state space model that retains the low state dimension of the state space model of Chapter 4. Also, I demonstrate how the computational devices of Chapter 4 can be applied in this context. The results in this chapter show that likelihood-based analysis is a feasible option in empirical applications of the dynamic factor model, even if some observations are missing. This chapter is based on the results in Jungbacker, Koopman, and van der Wel (2009a).
Chapter 2

The Linear State Space Model

2.1 Introduction

In this chapter I introduce the linear state space model and discuss some important issues involved in the analysis of such models. The aim of this chapter is to discuss all concepts and results necessary for the reader to understand the developments of the following chapters. For more in-depth discussions of the linear state space model I refer to the monographs by Anderson and Moore (1979), Harvey (1981), West and Harrison (1989) and Durbin and Koopman (2001).

Let $y_1, \ldots, y_n$ denote observations from a $N$ dimensional time series. In a linear state space model, these observations $y_t$ depend linearly on a $p$ dimensional unobserved stochastic process $\alpha_1, \ldots, \alpha_n$

$$ y_t = c_t + Z_t \alpha_t + \varepsilon_t, \quad (2.1) $$

where $c_1, \ldots, c_n$ are $N$ dimensional vectors and $Z_1, \ldots, Z_n$ are $N \times p$ dimensional matrices. The stochastic variables $\varepsilon_1, \ldots, \varepsilon_n$ are uncorrelated, have zero mean and $\text{Var}(\varepsilon_t) = H_t$ for $t = 1, \ldots, n$. Throughout this chapter I will assume that $H_t$ is non-singular for all $t = 1, \ldots, n$. The key assumption of the linear state space model is that the system matrices $c_t, Z_t$ and $H_t$ are independent of $\alpha_t$ for $t = 1, \ldots, n$. These system matrices may however depend on a set of fixed parameters, see also the examples of Section 2.2. I will denote the vector containing all model parameters by $\psi$.

The state process $\alpha_t$ is assumed to be generated by a vector autoregressive (VAR) process

$$ \alpha_{t+1} = d_t + T_t \alpha_t + R_t \eta_t, \quad (2.2) $$

13
where $d_1, \ldots, d_{n-1}$ are $p$ dimensional vectors, $T_1 \ldots, T_{n-1}$ are $p \times p$ matrices, $R_1, \ldots, R_{n-1}$ are $p \times r$ dimensional matrices and $\eta_1, \ldots, \eta_{n-1}$ is a $r$ dimensional sequence of zero mean uncorrelated stochastic variables with $\text{Var}(\eta_t) = Q_t$ for non-singular matrices $Q_t$ and $t = 1, \ldots, n - 1$. Finally, the state disturbances $\eta_t$ are assumed to be uncorrelated with the observation disturbances $\varepsilon_t$. This formulation allows a broad range of possible dynamics for the time series $y_1, \ldots, y_n$. Note that even though the state process possesses the Markov property, this is generally not the case for the time series $y_1, \ldots, y_n$. In the state space literature equation (2.2) is often referred to as the transition equation while (2.1) is known as the observation equation.

The state space formulation is completed by specifying the first and second moment of $\alpha_1$. In a linear state space model the initial state is assumed to have finite second moment

$$\mathbb{E}(\alpha_1) = a_{1|0}, \quad \text{Var}(\alpha_1) = P_{1|0}.$$ 

Both $a_{1|0}$ and $P_{1|0}$ are fixed but may depend on $\psi$. Finally, we assume that the initial state $\alpha_1$ is uncorrelated with both $\varepsilon_1, \ldots, \varepsilon_n$ and $\eta_1, \ldots, \eta_{n-1}$. If the state process is stationary $a_{1|0}$ is generally set to the mean of the process and $P_{1|0}$ to the stationary variance.

In Section 2.2 I give two examples of state space formulations to demonstrate the flexibility of the state space approach. In the remainder of the chapter we will see how to estimate the state process and predict future values of $y_t$. For Gaussian models I will show to efficiently evaluate the likelihood function. Finally, I discuss the analysis of state space models where the initial state does not have a finite second moment and show how the methods developed there can be applied in the analysis of regression models.

### 2.2 Examples of Linear State Space Models

**Example 1: The Regression Model with Time-Varying Coefficients**

The classical regression model is given by

$$y_t = x_t \beta + \varepsilon_t,$$

for $t = 1, \ldots, n$, where $x_1, \ldots, x_n$ are $1 \times p$ vectors of exogenous variables, $\beta$ is a $p \times 1$ vector of regression coefficients and $\varepsilon_1, \varepsilon_2, \ldots$ is a white noise sequence with $\text{Var}(\varepsilon_t) = \sigma_\varepsilon^2$. In some situations we might believe that $y_t$ depends linearly on the regression coefficients $x_t$, but that the coefficient $\beta$ varies smoothly over time. Such considerations motivate the regression
2.2. EXAMPLES OF LINEAR STATE SPACE MODELS

model with time-varying coefficients, given by the observation equation

\[ y_t = x_t \beta_t + \varepsilon_t, \]

and state equation

\[ \beta_{t+1} = \beta_t + \eta_t, \]  \hspace{1cm} (2.3)

where \( \eta_1, \eta_2, \ldots \) is a \( p \) dimensional white noise sequence with \( \text{Var}(\eta_t) = \Sigma_\eta \). The initial state is given by \( \beta_1 = \beta \) for a fixed \( p \times 1 \) vector \( \beta \).

This model can be extended in a multitude of ways. For example, we can incorporate a linear restriction of the form \( s' \beta_t = c \), with a \( p \times 1 \) vector \( s \) and \( c \) a scalar, in the model by choosing \( \Sigma_\eta \) as follows

\[ \Sigma_\eta = \sigma^2 \left( I - \frac{1}{s's}ss' \right), \]  \hspace{1cm} (2.4)

where \( \sigma^2 \) is a positive scalar. It is easily seen that this choice of \( \Sigma_\eta \) implies that \( s' \eta_t = 0 \). If we restrict \( \beta_1 \) to satisfy \( s' \beta_1 = c \) it follows that \( s' \beta_t = c \) for \( t = 1, \ldots, n \). Expression (2.4) equals \( \text{Var}(\eta_t | s' \eta_t = 0) \) if \( \text{Var}(\eta_t) = \sigma^2 I \). This approach can be generalized to allow for more general variance structures for the restricted model.

**Example 2: The ARMA(1,1) Model**

A scalar time series \( y_t \) follows an autoregressive moving average (ARMA) process of order \( (p, q) \) if it is stationary and is generated by the recursive relation

\[ y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \]  \hspace{1cm} (2.5)

where \( \varepsilon_1, \varepsilon_2, \ldots, \) is a white noise sequence with \( \text{Var}(\varepsilon_t) = \sigma^2_\varepsilon \) and \( \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \) and \( \sigma^2_\varepsilon \) are fixed scalars. Every ARMA process can be written in state space form, see e.g. Harvey (1981). For example, the ARMA(1,1) model can be written as

\[ y_t = (1, 0)\alpha_t, \]

where \( \alpha_t = (y_t, \theta \varepsilon_t)' \) and

\[ \alpha_{t+1} = T \alpha_t + R \eta_t, \]
CHAPTER 2. THE LINEAR STATE SPACE MODEL

with \( \text{Var}(\eta_t) = \sigma^2 \) and

\[
T = \begin{pmatrix} \phi & 1 \\ 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 \\ \theta \end{pmatrix}.
\]

(2.6)

For the initial state we set \( a_{1|0} = 0 \) and

\[
P_{1|0} = \sigma^2 \frac{\phi \theta + \theta + 1}{1 - \phi^2}.
\]

This state space formulation can be useful, since it allows us to apply the wide range of numerical methods developed for state space models to the ARMA model. Furthermore, we can use the ARMA process as a building block in more general state space models. For example, we can extend the regression model of the previous section to include ARMA disturbances

\[
y_t = x_t \beta_t + u_t,
\]

where \( \beta_t \) is modelled by (2.3) and \( u_t \) is an ARMA\((p,q)\) process. This simple example should convince the reader of the ease with which new and intuitively appealing models can be built.

2.3 The State Space Model in Matrix Form

Let \( y \) and \( \alpha \) denote respectively the stacked vectors \( y = (y_1', \ldots, y_n')' \) and \( \alpha = (\alpha_1', \ldots, \alpha_n')' \). In this section I give expressions for the first two moments of these random variables. These expressions will be used extensively in the remainder.

If we denote \( Z = \text{diag}(Z_1, \ldots, Z_n) \) and \( c = (c_1', \ldots, c_n')' \) then \( \alpha \) and \( y \) are linked through the equation

\[
y = c + Z\alpha + \varepsilon,
\]

where \( \varepsilon = (\varepsilon_1', \ldots, \varepsilon_n')' \). Further, if we define the matrix \( T \) as follows

\[
T = \begin{pmatrix} I & 0 & 0 & \cdots & 0 \\ T_1 & I & 0 & \cdots & 0 \\ T_2T_1 & T_2 & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ T_{n-1}\cdots T_1 & T_{n-1}\cdots T_2 & \cdots & T_{n-1} & I \end{pmatrix},
\]
then \( \alpha \) is given by

\[
\alpha = T(d + R\eta),
\]

where

\[
d' = (d'_1, \ldots, d'_n), \quad \eta' = (\alpha'_1 - a'_{1|0}, \eta'_1, \ldots, \eta'_{n-1}), \quad R = \text{diag}(I, R_1, \ldots, R_{n-1}).
\]

It follows that

\[
\mathbb{E}(\alpha) = Td, \quad \Omega \overset{\text{def.}}{=} \text{Var}(\alpha) = TRQR'T', \quad \text{Cov}(\alpha, y) = \Omega Z'
\]

\[
\mathbb{E}(y) = c + ZTd, \quad \Sigma \overset{\text{def.}}{=} \text{Var}(y) = Z\Omega Z' + H,
\]

where \( Q = \text{diag}(P_{1|0}, Q_1, \ldots, Q_{n-1}) \) and \( H = \text{diag}(H_1, \ldots, H_n) \).

### 2.4 Minimum Mean Square Linear Estimation

Forecasting future observations and estimating the unobserved states are especially important in the analysis of state space models. We will see that for a linear state space model linear minimum mean square estimators of the states and future observations can be obtained efficiently.

Suppose \( X \) and \( Y \) are respectively \( N_X \) and \( N_Y \) dimensional stochastic variables. The minimum mean square linear estimator (MMSLE) of \( Y \) conditional on \( X \), \( P(Y|X) \), is defined as follows

\[
P(Y|X) = \arg\min_{\hat{y} \in \mathbb{S}} \mathbb{E}\|Y - \hat{y}\|^2,
\]

where \( \mathbb{S} \) is the space of all variables of the form \( b + BX \) for \( N_Y \) dimensional vectors \( b \) and \( N_Y \times N_X \) matrices \( B \). Note that, contrary to what its name suggests, the estimator \( P(Y|X) \) is actually an affine function of the variables \( X \).

The following well-known result gives a closed form expression for MMSLEs. For proofs and a more in-depth discussion see e.g. Ruud (2000).

**Lemma 2.1.** Let \( X \) and \( Y \) be two stochastic variables with finite second moment, then \( \hat{y} = P(Y|X) \) if and only if \( \hat{y} \in \mathbb{S} \) and

\[
\mathbb{E}[(Y - \hat{y})X'] = 0.
\] (2.7)
Furthermore, if \( X \) has a non-singular variance matrix then
\[
P(Y|X) = \mathbb{E}(Y) + \text{Cov}(Y, X) \text{Var}(X)^{-1}(X - \mathbb{E}(X)),
\]
and
\[
\mathbb{E} \left[ \{Y - P(Y|X)\} \{Y - P(Y|X)\}' \right] = \text{Var}(Y) - \text{Cov}(Y, X) \text{Var}(X)^{-1} \text{Cov}(X, Y).
\]

In the following sections I will focus on linear prediction of \( y_t \) as well as the calculation of \( a_t|s \) and \( Q_t|s \) given by
\[
a_t|s = P(\alpha_t|y_1,\ldots,y_s), \quad Q_t|s = \mathbb{E} \left[ (\alpha_t - a_t|s)(\alpha_t - a_t|s)' \right],
\]
for \( s, t = 1,\ldots,n \). Lemma 2.1 combined with the expressions of Section 2.3 allows us to compute these MMSLEs analytically. Note however that the matrices involved generally have very high dimensions and naive application of Lemma 2.1 will therefore be infeasible.

If the time series \( y_1,\ldots,y_n \) is Gaussian the MMSLE \( a_t|s \) coincides with the minimum mean square estimator (MMSE) \( \mathbb{E}(\alpha_t|y_1,\ldots,y_s) \). In general, however, analytical expressions for MMSEs are unavailable and evaluation of these estimators requires the use of simulation based methods, such as those described in Chapter 3.

### 2.5 Prediction

Recall from Section 2.3 that \( \Sigma = \text{Var}(y) \), where \( y = (y_1',\ldots,y_n')' \). In this section we will see that the problem of obtaining mean square optimal linear predictions of future observations in a state space model is intimately related to the form of the LDL decomposition of \( \Sigma \). This decomposition can be obtained very efficiently due to the special structure of \( \Sigma \).

Write
\[
\Sigma = LDL',
\]
with \( L \) and \( D \) of the form
\[
L = \begin{pmatrix}
I & 0 & \cdots & 0 \\
L_{21} & I & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
L_{n1} & L_{n2} & \cdots & I
\end{pmatrix}, \quad D = \text{diag}(F_1,\ldots,F_n), \quad (2.8)
\]
where \( L_{ij} \) is \( N \times N \) for \( i = 2, \ldots, n \) and \( j = 1, \ldots, n-1 \) and \( F_1, \ldots, F_n \) are \( N \times N \) symmetric matrices. Now define \( v \) as follows

\[
v = L^{-1}[y - \mathbb{E}(y)].
\]

It follows immediately that

\[
\mathbb{E}(v) = 0, \quad \text{Var}(v) = D. \tag{2.9}
\]

Let \( v = (v'_1, \ldots, v'_n)' \) where each \( v_i \) is \( N \) dimensional. Since \( Lv = y - \mathbb{E}(y) \), the triangular structure of \( L \) implies that for \( t = 1, \ldots, n \)

\[
v_t = y_t - \mathbb{E}(y_t) - \sum_{j=1}^{t-1} L_{tj}v_j. \tag{2.10}
\]

Recall that \( v_t \) has zero mean and is uncorrelated with \( v_1, \ldots, v_{t-1} \). Since (2.10) implies that every \( y_j \) is an affine function of \( v_1, \ldots, v_j \) it follows that for all \( i < t \) we have \( \mathbb{E}(v_ty'_j) = 0 \).

From equation (2.7) of Lemma 2.1 it follows that

\[
P(y_t|y_1, \ldots, y_{t-1}) = \mathbb{E}(y_t) + \sum_{j=1}^{t-1} L_{tj}v_j, \tag{2.11}
\]

for \( t = 2, \ldots, n \). The vectors \( v_1, \ldots, v_n \) are known as prediction errors or innovations.

In Appendix 2.A, I derive the following expressions

\[
L_{ij} = Z_iT_{i-1} \cdots T_{j+1}K_j, \quad F_k = Z_kP_kZ'_k + H_k, \tag{2.12}
\]

for \( i > j \) with \( i = 2, \ldots, n \), \( j = 1, \ldots, n-1 \) and where \( k = 1, \ldots, n \) with \( K_1, \ldots, K_{n-1} \) and \( P_1, \ldots, P_n \) obtained from the recursions

\[
K_t = T_tP_tZ'_tF_t^{-1}, \quad P_{t+1} = T_tP_t(T'_t - Z'_tK'_t) + R_tQ_tR'_t, \tag{2.13}
\]

with \( P_1 = P_{1|0} \). If we define \( a_t \) as follows

\[
a_t = a_{1|0} + \sum_{j=1}^{t-1} (T_{t-1} \cdots T_{j+1}) K_jv_j,
\]
and using the easily checked identity

$$E(y_t) = c_t + Z_t \sum_{j=1}^{t-1} (T_{t-1} \cdots T_{j+1}) d_t,$$

it follows after some minor algebra that the innovations can be computed using the following recursions

$$v_t = y_t - Z_t a_t - c_t, \quad a_{t+1} = d_t + T_t a_t + K_t v_t. \quad (2.14)$$

The set of equations (2.13) and (2.14) is known as the Kalman filter.

Combining (2.11) and (2.12) we find

$$P(y_t|y_{t-1}, \ldots, y_1) = E(y_t) + Z_t \sum_{j=1}^{t-1} (T_{t-1} \cdots T_{j+1}) K_j v_j$$

$$= c_t + Z_t a_t.$$

Finally, we note that

$$E[(y_t - c_t - Z_t a_t)(y_t - c_t - Z_t a_t)'] = F_t,$$

for \( t = 1, \ldots, n \).

### 2.6 State Smoothing

State smoothing consists of the calculation of the quantities \( a_{1|n}, \ldots, a_{n|n} \) and \( Q_{1|n}, \ldots, Q_{n|n} \) given by

$$a_{t|n} = P(\alpha_t|y_1, \ldots, y_n), \quad Q_{t|n} = E \left[ (\alpha_t - a_{t|n})(\alpha_t - a_{t|n})' \right]$$

for \( t = 1, \ldots, n \), where \( P(\alpha_t|y_1, \ldots, y_n) \) is the minimum mean square linear estimator of \( \alpha_t \) given \( y_1, \ldots, y_n \), see Section 2.4. Smoothed estimates of the disturbances \( \varepsilon_1, \ldots, \varepsilon_n \) and \( \eta_1, \ldots, \eta_{n-1} \) can also be useful for diagnostic purposes. In this section we will see how these smoothed estimates can be found very efficiently.

Recalling that \( \Sigma = \text{Var}(y) \) with \( y = (y_1', \ldots, y_n')' \) the stacked vector of observations I define

$$u = \Sigma^{-1} [y - E(y)].$$
It follows from Lemma 2.1 that
\[ P(\varepsilon|y) = Hu, \quad P(\eta|y) = QR'T'Z'u, \quad P(\alpha|y) =Td + \Omega Z'u, \] (2.15)
where \( \varepsilon, \eta, H, Q, T, Z, d, \) and \( \Omega \) were defined in Section 2.3. As in Section 2.5, we write \( \Sigma = LDL' \) and \( v = L^{-1}[y - E(y)] \). It follows that
\[ L'u = D^{-1}v. \]
Using (2.12) we find
\[ u_k = F^{-1}_k v_k - K'_k \sum_{j=k+1}^{n} (T'_{k+1} \cdots T'_{j-1}) Z'_j u_j = F^{-1}_k v_k - K'_k r_k, \]
for \( k = 1, \ldots, n-1 \) where \( r_k \) is given by the following backward recursion
\[ r_{k-1} = Z'_k u_k + T'_k r_k, \] (2.16)
with \( r_n = 0 \).

The vectors \( u_1, \ldots, u_n \) suffice to calculate \( P(\varepsilon|y) \), since
\[ P(\varepsilon_t|y) = H_t u_t, \]
for \( t = 1, \ldots, n \). Also, it is not hard to see that \( r = T'Z'u \) and therefore
\[ P(\eta_t|y) = Q_t R'_t r_t, \quad P(\alpha_t|y) = a_{t|0} + P_{t|0} r_0, \]
for \( t = 1, \ldots, n-1 \). Recall from Section 2.3 that \( \alpha = T(d + R\eta) \). It follows that \( P(\alpha|y) = T[d + RP(\eta|y)] \). We can therefore calculate the smoothed state estimates \( \alpha_{t|n} \) as follows
\[ a_{t+1|n} = d_t + T_t a_{t|n} + R_t Q_t R'_t r_t, \] (2.17)
for \( t = 1, \ldots, n-1 \).

Calculating the smoothed state estimates using expression (2.17) requires an additional forward recursion. At the cost of storing \( a_1, \ldots, a_n \) and \( P_1, \ldots, P_n \) we can calculate the smoothed estimates more efficiently. First note that we can simplify recursion (2.16) as
follows
\[ r_{k-1} = Z_k F_k^{-1} v_k + L_k' r_k, \]  
(2.18)

where \( L_k = T_k - K_k Z_k \). I will now show that
\[ a_{s|n} = a_s + P_s r_{s-1}, \]  
(2.19)

for \( s = 1, \ldots, n \). We already know that this expression is valid for \( s = 1 \). Suppose that the expression is valid for \( s = t - 1 \), expression (2.19) then follows from an induction argument and

\[
a_{t|n} = d_{t-1} + T_{t-1} a_{t-1|n} + R_{t-1} Q_{t-1} R_{t-1}' r_{t-1} \\
= a_t + (T_{t-1} P_{t-1} L_{t-1}' + R_{t-1} Q_{t-1} R_{t-1}') r_{t-1} \\
= a_t + P_t r_{t-1}.
\]

We can calculate mean squared errors for the smoothed estimates by calculating \( \Sigma^{-1} = (LDL')^{-1} \) and using Lemma 2.1. Some tedious algebra gives us

\[ Q_t|n = P_t - P_t N_t P_t, \quad N_t = Z_t F_t^{-1} Z_t + L_t' N_{t+1} L_t, \]  
(2.20)

for \( t = 1, \ldots, n \), with \( N_{n+1} = 0 \). Recursions (2.18), (2.19) and (2.20) are collectively known as the Kalman smoother. For the mean squared errors of the smoothed disturbances we have the following expressions

\[
\mathbb{E} 
\left[
\begin{array}{c}
\eta_t - P(\eta_t|y) \\
\eta_t - P(\eta_t|y)
\end{array}
\right]' = Q_t - Q_t R_t' N_t R_t Q_t,
\]

and

\[
\mathbb{E} 
\left[
\begin{array}{c}
\epsilon_t - P(\epsilon_t|y) \\
\epsilon_t - P(\epsilon_t|y)
\end{array}
\right]' = H_t - H_t (F_t^{-1} + K_t' N_t K_t) H_t,
\]
2.7 Filtering and Predicting the States

The state estimates \( a_{1|1}, \ldots, a_{n|n} \) are known as filtered estimates. From (2.19) and (2.20) it follows that

\[
a_{t|t} = a_t + P_t Z'_t F_t^{-1} v_t, \quad Q_{t|t} = P_t - P_t Z'_t F_t^{-1} Z_t P_t, \tag{2.21}
\]

for \( t = 1, \ldots, n \). From the state equation (2.2) we find the following relations between \( a_{t-1|t-1} \) and \( a_{t-1|t-1} \) respectively \( Q_{t-1|t-1} \) and \( Q_{t-1|t-1} \)

\[
a_{t-1|t-1} = d_{t-1} + T_{t-1} a_{t-1|t-1}, \quad Q_{t|t-1} = T_{t-1} Q_{t-1|t-1} T'_{t-1} + R_{t-1} Q_{t-1|t-1} R'_{t-1}. \tag{2.22}
\]

If we combine (2.21) and (2.22) we find

\[
a_{t|t-1} = d_{t-1} + T_{t-1} a_{t-1|t-1} + T_{t-1} P_{t-1} Z'_t F_t^{-1} v_{t-1}, \tag{2.23}
\]

and

\[
Q_{t|t-1} = T_t P_t (T'_t - Z'_t F_t^{-1} Z_t P_t T'_t) + R_{t-1} Q_{t-1|t-1} R'_{t-1}. \tag{2.24}
\]

If we compare (2.23) and (2.24) to the Kalman filter recursions (2.13) and (2.14) we find that \( a_{t|t-1} = a_t \) and \( Q_{t|t-1} = P_t \) for \( t = 1, \ldots, n \). Considering the structure of the Kalman filter this result should not come as a surprise.

2.8 Gaussian Likelihood

If the disturbances and the initial state are Gaussian, the linearity of the linear state space model implies that the observations \( y_1, \ldots, y_n \) are Gaussian as well. In this case the Kalman filter can be used to efficiently calculate the likelihood.

Let \( \ell(y) \) denote the log-likelihood function of a time series \( y_1, \ldots, y_n \) generated by a Gaussian linear state space model. Further, recall from Section 2.5 that the vector of innovations \( v \) is defined as \( v = L^{-1} [y - \mathbb{E}(y)] \), where \( y = (y_1, \ldots, y_n) \) and \( L \) is a lower triangular matrix with ones on the diagonal. If we denote the log-likelihood of the innovations by \( \ell(v) \) we have \( \ell(y) = \ell(v) \) since \( |L| = 1 \). Finally, it follows from \( \text{var}(v) = \text{diag}(F_1, \ldots, F_n) \) that

\[
\ell(v) = -\frac{nN}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \log |F_i| - \frac{1}{2} \sum_{i=1}^{n} v'_i F_i^{-1} v_i. \tag{2.25}
\]
We can therefore calculate the log-likelihood of \( y_1, \ldots, y_n \) by running the Kalman filter and using (2.25). Since \( v_i \) is the prediction error for observation \( y_i \), equation (2.25) is called the prediction-error decomposition. Expression (2.25) was first given by Schweppe (1965). Note that

\[
\log p(y_t|y_{t-1}, \ldots, y_1) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |F_t| - \frac{1}{2} v_t'F_t^{-1}v_t,
\]

where \( p(y_t|y_{t-1}, \ldots, y_1) \) is the likelihood of \( y_t \) conditional on \( y_1, \ldots, y_{t-1} \).

Equation (2.25) allows us to evaluate the log-likelihood of a linear Gaussian state space model. If the time series model is non-Gaussian we need to use computationally intensive numerical methods, such as the simulation-based algorithms of Chapter 3, to evaluate the log-likelihood.

2.9 Univariate Treatment of Multivariate Time Series

The Kalman filter recursions (2.13) require the inverses of the matrices \( F_1, \ldots, F_n \). Even for multi-dimensional time series with a moderate value of \( N \) this can be computationally intensive. It is however possible to eliminate the need for these matrix inversions by rewriting the model as a univariate state space model, which can result in considerable computational savings. This approach was proposed in Anderson and Moore (1979) and Fahrmeir and Tutz (1994) and further developed in Koopman and Durbin (2000).

Consider the linear state space model of equations (2.1) and (2.2). We can perform LDL decompositions of \( H_1, \ldots, H_n \) and write \( H_t = S_tV_tS_t' \) for \( t = 1, \ldots, n \), where \( S_t \) is a lower triangular matrix with ones on the diagonal and \( V_t \) is a diagonal matrix. Defining \( y_t^* = S_t^{-1}y_t \) for \( t = 1, \ldots, n \) we have

\[
y_{it}^* = c_{it}^* + Z_{it}^*\alpha_{i,t} + \varepsilon_{it}^*,
\]

where \( y_{it}^* \) is the \( i \)th element of \( y_t^* \), \( c_{it}^* \) is the \( i \)th element of \( S_t^{-1}c_t \), \( Z_{it}^* \) is the \( i \)th row of \( S_t^{-1}Z_t \) and \( \varepsilon_{it}^* \) is the \( i \)th element of \( S_t^{-1}\varepsilon_t \). If we define the stochastic process \( \alpha_{i,t} \) as follows

\[
\alpha_{i,t+1} = \begin{cases} 
T_i\alpha_{N,t} + \eta_t & i = 1, \\
\alpha_{i,t} & i = 2, \ldots, N,
\end{cases}
\]

with \( \alpha_{1,1} = \alpha_1 \), it follows that

\[
y_{it}^* = c_{it}^* + Z_{it}^*\alpha_{i,t} + \varepsilon_{it}^*.
\]
Note that the random variables $\varepsilon^*_it$ for $i = 1, \ldots, N$ and $t = 1, \ldots, n$ are uncorrelated. The model consisting of (2.27) and (2.28) is a univariate state space model, with

$$P(\alpha_{1,t}|y_{11}^*, \ldots, y_{N,s}^*) = P(\alpha_t|y_1, \ldots, y_s),$$

for $s, t = 1, \ldots, n$. Also, since the transformation matrices $S^{-1}_t$ are full rank and have determinant 1, we have $\ell(y^*) = \ell(y)$ where $\ell(y^*)$ and $\ell(y)$ denote the Gaussian log-likelihood functions of respectively $y^*$ and $y$. It follows that state estimation and evaluation of the likelihood for the original model can be done by applying the Kalman filter and smoother recursions to (2.27) and (2.28). Since this new model is univariate, the computational cost can be significantly smaller than for the original model. Note however that we do need to obtain LDL decompositions of the matrices $H_1, \ldots, H_n$. In practice the additional cost associated with the construction of these decompositions will often be small, for example because the model is time-invariant or because the variance matrices have a convenient structure.

### 2.10 Missing Observations

One of the attractions of the Kalman filter and smoother recursions is that they can easily handle missing observations. Suppose that for a univariate time series $y_1, \ldots, y_n$ and a set of indices $\tau_1, \ldots, \tau_k$ the observations $y_{\tau_1}, \ldots, y_{\tau_k}$ are missing. The modified recursions can be easily derived by making small modifications to the developments in Section 2.5 and Section 2.6.

In the Kalman filter we only need to set $v_{\tau_1}, \ldots, v_{\tau_k}$ and $K_{\tau_1}, \ldots, K_{\tau_k}$ to zero. It follows that for $t = \tau_1, \ldots, \tau_k$ we have

$$a_{t+1} = d_t + T_t a_t, \quad P_{t+1} = T_t P_t T_t' + R_t Q_t R_t'.$$

The missing observations do not enter the Gaussian log-likelihood and we have

$$\ell(y) = -\frac{n-k}{2} \log 2\pi - \frac{1}{2} \sum_{t \notin \{\tau_1, \ldots, \tau_k\}} \left[ \log |F_t| + v_t' F_t^{-1} v_t \right].$$

The filtered estimates and the corresponding mean square errors can be obtained as before, with the exception that for $t = \tau_1, \ldots, \tau_k$

$$a_{t|t} = a_t, \quad Q_{t|t} = P_t.$$
For the Kalman smoother the modified recursions for \( t = \tau_1, \ldots, \tau_k \) are given by

\[
\begin{align*}
  r_{t-1} &= T_t' r_t, \\
  N_{t-1} &= T_t' N_t T_t.
\end{align*}
\]

The missing observations can be estimated from the output of the Kalman filter and smoother as follows

\[
P(y_t | y) = c_t + Z_t a_t | n, \\
\text{Var}(y_t | y) = Z_t Q_t | n Z_t' + H_t,
\]

for \( t = \tau_1, \ldots, \tau_k \).

If the time series is multivariate and for some value of \( t \) only a subset of the elements of \( y_t \) is missing, we can simply remove the relevant rows and columns in \( c_t, Z_t \) and \( H_t \) and use the classical Kalman filter and smoother recursions. Note that the resulting state space model has an observation equation of time-varying dimension.

### 2.11 Unspecified Initial Conditions

In the preceding section we assumed that the first and second moment of \( \alpha_1 \) were finite and known. In this section we consider the more general case where \( \alpha_1 \) is given by

\[
\alpha_1 = A \delta + R \eta_0,
\]

where \( A \) is a \( p \times q \) matrix, \( \delta \) is a \( q \times 1 \) vector with an unspecified distribution and possibly with an infinite variance, \( R \) is a \( p \times (p - q) \) matrix and \( \eta_0 \) is a stochastic variable with

\[
\mathbb{E}(\eta_0) = a, \quad \text{Var}(\eta_0) = I,
\]

for a fixed and known vector \( a \). A specification such as \( (2.30) \) is common for models with a (partially) non-stationary state. In this case \( \delta \) corresponds to the non-stationary part of the state. Since the distribution of \( y_1, \ldots, y_n \) is no longer well-defined we can not use the result of Section 2.4 to obtain MMSLEs. I will discuss in this section a method to obtain linear estimators of the states and predictions of future observations for this type of model. Furthermore, I will discuss the marginal likelihood which is often used as the basis of likelihood-based inference if the initial state is of the form \( (2.30) \).

Throughout this section I will illustrate the concepts involved in the context of the local level model.
2.11. UNSPECIFIED INITIAL CONDITIONS

Example: The Local Level Model

Probably the simplest non-trivial example of a non-stationary state space model is the local level model. In a local level model we assume \( y_1, \ldots, y_n \) are noisy observations of an unobserved underlying random walk \( \alpha_1, \ldots, \alpha_n \). Using the by now familiar notation, the observation and state equation for this model are given by

\[
y_t = \alpha_t + \varepsilon_t, \tag{2.31}
\]

for \( t = 1, \ldots, n \) where \( \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2 \) and

\[
\alpha_{t+1} = \alpha_t + \eta_t, \quad \alpha_1 = \delta, \tag{2.32}
\]

for \( t = 1, \ldots, n - 1 \) with \( \text{Var}(\eta_t) = \sigma_\eta^2 \) and \( \delta \) a random variable with unknown distribution. Note that the initial state is of the form (2.30) with \( A = 1 \) and \( R = 0 \). For more information on the analysis of the local level model and examples of applications I refer to Chapter 2 of Durbin and Koopman (2001).

2.11.1 State Estimation

In this section I discuss a class of linear estimators for the unobserved states \( \alpha_1, \ldots, \alpha_n \) for which the initial state is modelled as in (2.30). See Ansley and Kohn (1985) for more information. To simplify the notation I assume that the vectors \( c_1, \ldots, c_n \) and \( d_1, \ldots, d_{n-1} \), in the state space formulation of (2.1) and (2.2), are zero.

Suppose that the initial state of a state space model is of the form (2.30) and denote by \( \alpha_1^*, \ldots, \alpha_n^* \) the stochastic process generated by the same state equation as \( \alpha_t \), with the same innovations, but with \( \alpha_1^* = R\eta_0 \). The process \( \alpha_t^* \) can be seen as the part of \( \alpha_t \) that does not depend on the unspecified variable \( \delta \). We can write \( \alpha = (\alpha_1', \ldots, \alpha_n')' \) as follows

\[
\alpha = BA\delta + \alpha^*,
\]

where \( \alpha^* = (\alpha_1^*, \ldots, \alpha_n^*) \) and

\[
B = \begin{bmatrix}
I & T_1' & T_2T_1' & \cdots & \left( \prod_{i=1}^{n-1} T_i \right)'
\end{bmatrix}'. \tag{2.33}
\]
It follows that the stacked vector \( y = (y'_1, \ldots, y'_n)' \) is given by
\[
y = ZBA\delta + Z\alpha^* + \varepsilon, \tag{2.34}
\]
where \( Z = \text{diag}\{Z_1, \ldots, Z_n\} \) and \( \varepsilon = (\varepsilon'_1, \ldots, \varepsilon'_n)' \). In the remainder I will assume that \( ZBA \) has full column rank. If this is not the case, some linear combination of the elements of \( \delta \) can not be identified from the data.

Let \( J_1 \) and \( J_2 \) denote respectively \((Nn - q) \times Nn\) and \( q \times Nn \) dimensional matrices such that \((J'_1, J'_2)'\) is non-singular and
\[
J_1(ZBA) = 0, \quad J_2(ZBA) = I_q. \tag{2.35}
\]
We can now write \( \alpha_t \) as follows
\[
\alpha_t = G_tJ_2y + (\alpha_t - G_tJ_2y) = G_tJ_2y + (\alpha^*_t - G_tJ_2\omega), \tag{2.36}
\]
where \( G_t \) is the \( t \)th row of \( BA \) and \( \omega = Z\alpha^* + \varepsilon \). Equation (2.36) expresses \( \alpha_t \) as the sum of two terms, with only the first term \( G_tJ_2y \) depending on \( \delta \). The second term, \( \alpha^*_t - G_tJ_2\omega \) has a well defined distribution with finite second moment. Motivated by (2.36) we therefore consider the following linear state estimators
\[
\tilde{a}_{t|n} = G_tJ_2y + P(\alpha_t - G_tJ_2y|J_1y) = G_tJ_2y + P(\alpha^*_t - G_tJ_2\omega|J_1y), \tag{2.37}
\]
for \( t = 1, \ldots, n \). Predictions of future values of the observed time series \( y_{n+1}, y_{n+2}, \ldots \) denoted by \( \hat{y}_{n+1}, \hat{y}_{n+2}, \ldots \) can be obtained as follows
\[
\hat{y}_{n+h} = Z_{n+h}G_{n+h}J_2y + P(Z_{n+h}\alpha^*_{n+h} - Z_{n+h}G_{n+h}J_2\omega^{(n+h)}|J_1y), \tag{2.38}
\]
for \( h = 1, 2, \ldots \), where \( \omega^{(n+h)} = (\omega_1, \ldots, \omega_{n+h})' \).

If the MMSLEs of \( \alpha_t \) and \( y_{t+h} \) are well-defined they generally do not coincide with (2.37) and (2.38). For example for \( \alpha_t \) we have
\[
P(\alpha_t|y) = G_tJ_2y + P(\alpha^*_t - G_tJ_2\omega|J_2y, J_1y),
\]
it follows that \( P(\alpha_t|y) \) and \( \tilde{a}_{t|n} \) only coincide if \( J_2y \) is uncorrelated with \( \alpha^*_t - G_tJ_2\omega \). Note that in this case \( \hat{y}_{n+h} \) also coincides with \( P(y_{n+h}|y) \). This type of assumption is standard in the analysis of autoregressive integrated moving average processes, see the discussion in
Suppose now that $\delta$ is a variable with $\mathbb{E}(\delta) = 0$ and $\text{Var}(\delta) = \kappa I$ for a positive scalar $\kappa$. In this case $\tilde{a}_{t|n}$ does not coincide with $P(\alpha|y)$. However, it is not hard to show using the results of Section 2.4 that
\[
\lim_{\kappa \to \infty} P_\kappa(\alpha^*_t - G_t J_2 \omega | y) = P(\alpha^*_t - G_t J_2 \omega | J_1 y),
\]
(2.39)
where $P_\kappa(\alpha^*_t - G_t J_2 \omega | y)$ denotes the MMSLE of $\alpha^*_t - G_t J_2 \omega$ in the linear state space model with $\text{Var}(\delta) = \kappa I$. It follows that in models where $\text{Var}(\delta) = \kappa I$ for large $\kappa$, the estimator $\tilde{a}_{t|n}$ will be approximately equal to the MMSLE of $\alpha_t$. This gives an intuitive interpretation of the estimator $\tilde{a}_{t|n}$. We can easily obtain a similar result for $\hat{y}_{n+1}$, $\hat{y}_{n+2}$, ... Variables with $\text{Var}(\delta) = \kappa I$ for arbitrarily large $\kappa$ are often referred to as being diffuse.

The most natural interpretation of $\tilde{a}_{t|s}$ comes from its connection to generalized least squares (GLS) estimation. First note that (2.34) can be seen as a linear model with regression coefficient $\delta$ and variance matrix $\text{Var}(Z\alpha^* + \varepsilon)$. Let $\delta_{GLS}$ denote the GLS estimator of $\delta$ given $y_1, \ldots, y_n$. It can be shown that $\tilde{a}_{t|n}$ is equal to the MMSLE of $\alpha_t$ if $\delta$ is set to $\delta_{GLS}$ and $\text{Var}(\delta) = \text{Var}(\delta_{GLS})$, see e.g. de Jong (1991).

Example (continued): The Local Level Model

I will now show how the concepts of this section translate to the case of the local level model. First we note that
\[ ZBA = BA = (1, \ldots, 1)' . \]
The simple form of both $BA$ and $ZBA$ means that finding suitable matrices $J_1$ and $J_2$ is easy. Particularly convenient choices for $J_1$ and $J_2$ are given by
\[
J_1 = \begin{pmatrix} -i_{n-1} & I_{n-1} \end{pmatrix}, \quad J_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix},
\]
where $i_{n-1}$ is a $n-1 \times 1$ dimensional vector of the form $(1, \ldots, 1)'$. It is easily checked that $J_1 ZA = 0$ and $J_2 ZA = 1$. Recall that $G_t$ denotes the $t$th row of $BA$. If follows that $G_t$ is 1 and thus $G_t J_2 = J_2$. Finally, note that $\alpha^*_t$ was defined as $\alpha_t$ but with $\delta$ set to zero. For the local level model this means that $\alpha^*_1 = 0$. Expression (2.37) is now considerably simplified
\[
\tilde{a}_{t|n} = G_t J_2 y + P(\alpha_t - G_t J_2 y | J_1 y) = y_t + P(\alpha_t - y_1 | J_1 y),
\]
(2.40)
and thus
\[ \tilde{a}_t | n = y_1 + P(\alpha_t^* - \varepsilon_1 | J_1 y), \]
where I used the fact that \( \alpha_t = \delta + \alpha_t^* \) and \( y_1 = \delta + \varepsilon_1 \). Note that by subtracting \( y_1 \) from \( \alpha_t \) in (2.40) we remove the influence of \( \delta \). This illustrates the role of the term \( G_t J_2 y \) in equation (2.37).

### 2.11.2 The Marginal Likelihood

Suppose that the disturbances \( \varepsilon_1, \ldots, \varepsilon_n \) and \( \eta_1, \ldots, \eta_{n-1} \) are Gaussian. Since the distribution of \( \delta \) is not well-defined, we cannot calculate the likelihood of \( y_1, \ldots, y_n \). It is customary in these cases to use the marginal likelihood as the basis for inference. Choose \( J_1 \) as before but with the additional assumptions that \( |J'_1 J_1| = 1 \) and that \( J_1 \) is independent of any unknown parameters. The marginal likelihood is defined as the likelihood of \( w = J_1 y \). This likelihood is clearly well-defined since (2.35) implies that \( w \) does not depend on \( \delta \). The assumption on the determinant \( |J'_1 J_1| \) is inconsequential and is only made to uniquely determine the likelihood. The assumption that \( J_1 \) is independent of the parameters ensures that different specifications of the same model produce equivalent marginal likelihoods.

Suppose that \( \delta \sim N(0, \kappa I) \). Following the discussion of Section 2.11.1 we might suspect that for large \( \kappa \) the likelihood of \( y_1, \ldots, y_n \) is in some way related to the likelihood of \( J_1 y \). Ansley and Kohn (1985) show that the marginal likelihood is proportional to the diffuse likelihood \( \ell_d(y) \) defined by
\[
\ell_d(y) = \lim_{\kappa \to \infty} \left[ \frac{n}{2} \log \kappa + \ell_\kappa(y) \right],
\]
where \( \ell_\kappa(y) \) is the log-likelihood of \( y \) for a fixed value of \( \kappa \).

Finally, de Jong (1991) shows that we can also obtain the marginal likelihood by setting \( \delta \sim N[\delta_{GLS}, \text{Var}(\delta_{GLS})] \) and using the expressions of Section 2.8.

### 2.11.3 Regression Effects and the Marginal Likelihood

In Section 2.11.2 I discussed the use of the marginal likelihood for non-stationary state space models. The marginal likelihood can also be defined for stationary models with regression effects. In this case we rewrite the state space model, such that the regression coefficients are modelled as states with undefined initial conditions. The marginal likelihood can then be interpreted as the likelihood of the data with the influence of the regression effects removed.
2.12. DIFFUSE KALMAN FILTER

There is evidence that likelihood-based inference of variance parameters is more precise if such a likelihood concept is used, especially in small samples, see e.g. Tunnicliffe-Wilson (1989) and Shephard (1993).

Suppose that in equations (2.1) and (2.2) we have

\[ c_t = c_t^* + X_t^q \beta, \quad d_t = d_t^* + X_t^p \beta, \]

for \( t = 1, \ldots, n \), fixed vectors \( c_t^* \) and \( d_t^* \), matrices of exogenous variables \( X_t^q \) and \( X_t^p \) and a vector of regression coefficients \( \beta \). We can now write the state space model as

\[ y_t = c_t^* + (Z_t \ X_t^q) \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \varepsilon_t, \quad (2.41) \]

for \( t = 1, \ldots, n \) and

\[ \begin{pmatrix} \alpha_{t+1} \\ \beta_{t+1} \end{pmatrix} = \begin{pmatrix} T_t & X_t^p \\ 0 & I \end{pmatrix} \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \begin{pmatrix} \eta_t \\ 0 \end{pmatrix}, \quad (2.42) \]

where

\[ \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 0 \\ I \end{pmatrix} \beta + \begin{pmatrix} I \\ 0 \end{pmatrix} \eta_0, \quad (2.43) \]

with \( \mathbb{E}(\eta_0) = a_{1|0} \) and \( \text{Var}(\eta_0) = P_{1|0} \). If we consider \( \beta \) as a random variable with an unspecified distribution, it follows that the marginal likelihood can be obtained using the results from Section 2.12. Also, the linear estimators for the states \( \beta_t \) introduced in Section 2.12 coincide with the GLS estimator of \( \beta \) as was discussed in Section 2.11.

2.12 Diffuse Kalman filter

In this section I show how to obtain the linear estimators \( \tilde{a}_{ts} \) for \( s, t = 1, \ldots, n \) defined in Section 2.11.1. To simplify some of the notation I focus on state space models of the form (2.1) and (2.2) with the intercepts \( c_1, \ldots, c_n \) and \( d_1, \ldots, d_{n-1} \) and \( a \), the mean of the initial state, set to zero. Also, to avoid some technicalities, I restrict the discussion to univariate time series. The developments of Section 2.9 imply that this assumption is not restrictive.

Two different algorithms have been proposed in the literature. The diffuse Kalman filter of Ansley and Kohn (1985) and Koopman (1997) exploits the connection with diffuse variables by analytically evaluating the limit in (2.39). The augmented Kalman filter proposed
by de Jong (1991) evaluates the estimators $\tilde{a}_{t|n}$ by efficiently calculating the GLS estimator of $\delta$. The approach presented here is more direct. I show that for a smart choice of $J_1$, $J_1y$ is also generated by a state space model with state equal to $\alpha_t$. It follows that the estimates $\tilde{a}_{t|n}$ can be obtained by running the classical Kalman filter on this new model. It can be shown that the resulting algorithm is equivalent to the diffuse Kalman filter.

### 2.12.1 A First Transformation of the Model

We saw that for the local level model it is relatively easy to find suitable matrices $J_1$ and $J_2$. This is not true for the general linear state space model. In this section we see how to transform a linear state space model in such a way that it becomes straightforward to find suitable matrices $J_1$ and $J_2$.

Suppose that $\tilde{B}$ is a matrix of full column rank such that $\tilde{B}$ has the same column space as $BA$, then $Z\tilde{B}$ also has the same column space as $ZBA$. If a matrix $\tilde{J}_1$ satisfies $\tilde{J}_1(Z\tilde{B}) = 0$ we also have $\tilde{J}_1(ZBA) = 0$. Further, there is a non-singular matrix $C$ such that $\tilde{B} = BAC$ and therefore $Z\tilde{B} = ZBAC$. Suppose we choose $\tilde{J}_2$ such that $\tilde{J}_2Z\tilde{B} = I_q$ then a possible choice for $J_2$ would be $J_2 = C\tilde{J}_2$. In practice we are not interested in $J_2$ itself but only in $G_tJ_2$ for $t = 1, \ldots, n$. We have

$$G_tJ_2y = G_tC\tilde{J}_2y = \tilde{B}_t\tilde{J}_2y,$$

where $\tilde{B}_t$ is the $t$th row of $\tilde{B}$. We can thus calculate the estimators $\tilde{a}_{t|n}$ as follows

$$P(\alpha_t|y) = \tilde{B}_t\tilde{J}_2y + P(\alpha_t^* - \tilde{B}_t\tilde{J}_2\omega|\tilde{J}_1y).$$

It follows that the estimators $\tilde{a}_{t|n}$ remain unchanged if we replace $BA$ in (2.34) by $\tilde{B}$. Note also that the marginal likelihood is given by the likelihood of $\tilde{J}_1y$.

We will now construct a matrix $\tilde{B}$ for which it is particularly easy to find a suitable matrix $\tilde{J}_1$. First we obtain an LDL decomposition $LDL' = ZBA\bar{B}'\bar{B}'Z'$ using the Kalman filter recursions. The matrix $LD$ clearly has the same column space as $ZBA$. Further we will see that $LD$ can be written as $Z\tilde{B}$ for a matrix $\tilde{B}$ with the same column space as $BA$. In the next section we choose $\tilde{B}$ equal to this matrix $\tilde{B}$ and I show how to choose suitable $\tilde{J}_1$ and $\tilde{J}_2$.

Define a state space model as follows

$$\tilde{y}_t = Z_t\delta_t, \quad \delta_{t+1} = T_t\delta_t,$$
with $\mathbb{E}(\delta_1) = 0$ and $\text{Var}(\delta_1) = AA'$. The state space model is non-standard in that it has no noise terms, but it is well defined nonetheless. If $\tilde{y}$ denotes $\tilde{y} = (\tilde{y}_1^t, \ldots, \tilde{y}_n^t)'$ then $\text{Var}(\tilde{y}) = ZBA'A'B'Z'$. It follows from Corollary 2.1 in the appendix that $\text{Var}(\tilde{y}) = LDL'$ where

$$L = [L_1, \ldots, L_n], \quad D = \text{diag}\{F_{\infty,1}, \ldots, F_{\infty,n}\},$$

and

$$L_i = \begin{bmatrix} 0 & \cdots & 0 & M_{\infty,i}^t T_i' Z_i' \cdots & M_{\infty,i}^t \prod_{j=i}^{n-1} T_j' Z_i' \end{bmatrix},$$

for $i = 1, \ldots, n$ where the first $(i-1)$ rows of each $L_i$ consist of zeros and

$$P_{\infty,t+1} = T_t P_{\infty,t} L_{\infty,t}', \quad L_{\infty,t} = T_t - K_{\infty,t} Z_i', \quad F_{\infty,t} = Z_t P_{\infty,t} Z_i', \quad K_{\infty,t} = T_t M_{\infty,t}$$

where

$$M_{\infty,t} = \begin{cases} P_{\infty,t} Z_i' F_{\infty,t}^{-1} & F_{\infty,t} \neq 0, \\ 0 & \text{else.} \end{cases}$$

It follows that the matrix $LD$ is given by

$$LD = Z[\bar{B}_1, \ldots, \bar{B}_n],$$

where

$$\bar{B}_i = \begin{bmatrix} 0 & \cdots & 0 & Z_i P_{i,\infty} F_{\infty,i} M_{\infty,i}^t T_i' \cdots & F_{\infty,i} M_{\infty,i}^t \prod_{j=i}^{n-1} T_j' \end{bmatrix}'. $$

Since $\text{Var}(\tilde{y}) = LDL'$ has rank $q$, where $q$ is the dimension of $\delta$, it follows that there are precisely $q$ indices $s_1, \ldots, s_q$ such that $F_{\infty,s_i} \neq 0$ for $i = 1, \ldots, q$. Since $F_{\infty,t} = 0$ implies that $Z_i P_{\infty,i} = 0$, only $\bar{B}_{s_1}, \ldots, \bar{B}_{s_q}$ are non-zero. We therefore define

$$\bar{B} = \begin{bmatrix} \bar{B}_{s_1} & \cdots & \bar{B}_{s_q} \end{bmatrix}. \quad (2.44)$$

Note that the columns of $Z\bar{B}$ span the same space as the columns of $ZBA$. It can also be checked that the columns of $\bar{B}$ span the same space as those of $BA$, I omit further details.

**Example (continued): The Local Level Model**

For the local level model we find

$$P_{\infty,1} = 1, \quad L_{\infty,1} = 0, \quad F_{\infty,1} = 1, \quad K_{\infty,1} = 1. \quad (2.45)$$
Note that \(L_{\infty,t} = 0\) means that \(P_{\infty,t} = 0\) for \(t = 2, \ldots, n\). This can also be understood from the fact that \(\delta\) is one-dimensional and therefore only one \(F_{\infty,t}\) and one \(P_{\infty,t}\) can be non-zero. It follows that \(\bar{B}\) is given by
\[
\bar{B} = (1, \ldots, 1)' = BA.
\]
For other linear state space models \(\bar{B}\) and \(BA\) will generally not coincide. It is easily checked that this is only the case if there are \(q\) linearly independent rows in \(Z_1\), this implies that the whole vector \(\delta\) can be identified from \(y_1\).

### 2.12.2 Finding \(\tilde{J}_1\) and \(\tilde{J}_2\)

I choose \(\tilde{B}\) to be equal to \(\bar{B}\) constructed in the previous section. We will now see how to obtain \(\tilde{J}_1\) such that \(\tilde{J}_1Z\tilde{B} = 0\) and \(\tilde{J}_2\) such that \(\tilde{J}_2Z\tilde{B} = I_q\).

Note that \(Z\tilde{B}\) has a triangular form with the first non-zero entry of column \(i\) equal to \(F_{\infty,s_i}\). Let us define \(q\) matrices \(\tilde{J}(q)\) of dimension \(n \times n\) where the element \((i, j)\) of \(\tilde{J}(k)\) is given by
\[
\tilde{J}^{(ij)}(k) = \begin{cases} 
1 & i = j \\
-Z_i \left( \prod_{m=s_k}^{i-1} T_m \right) M_{\infty,s_k} & i > j = s_k \\
0 & \text{else,}
\end{cases}
\]
for \(k = 1, \ldots, q\). It is easily checked that
\[
J(k)\bar{B}_k = (0, \ldots, 0, F_{\infty,s_k}, 0, \ldots, 0)',
\]
where only the \(k\)th element is non-zero, and \(J(k)\bar{B}_j = B_j\) for \(j \neq k\). If we define \(\tilde{J} = \tilde{J}(q)\tilde{J}(q-1) \cdots \tilde{J}(1)\) then
\[
\tilde{J} \left( Z\tilde{B} \right) = \left( F_{\infty,s_1}E_{s_1} \cdots F_{\infty,s_q}E_{s_q} \right),
\]
where \(E_i\) denotes the \(i\)th column of a \(n \times n\) identity matrix. Further, if \(\tilde{J}^{(ij)}\) denotes element
(i, j) of the matrix \( \tilde{J}^{(ij)} \) we can choose \( \tilde{J}_1 \) and \( \tilde{J}_2 \) as follows

\[
\tilde{J}_1 = \begin{pmatrix}
\tilde{J}^{(11)} & \tilde{J}^{(12)} & \ldots & \tilde{J}^{(1n)} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{J}^{(s_k-1,1)} & \tilde{J}^{(s_k-1,2)} & \ldots & \tilde{J}^{(s_k-1,n)} \\
\tilde{J}^{(s_k+1,1)} & \tilde{J}^{(s_k+1,2)} & \ldots & \tilde{J}^{(s_k+1,n)} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{J}^{(n1)} & \tilde{J}^{(n2)} & \ldots & \tilde{J}^{(nn)}
\end{pmatrix}, \quad \tilde{J}_2 = \begin{pmatrix}
F_{\infty,s_1}^{-1} \tilde{J}^{(s_1,1)} & \ldots & F_{\infty,s_1}^{-1} \tilde{J}^{(s_1,n)} \\
\vdots & \ddots & \vdots \\
F_{\infty,s_q}^{-1} \tilde{J}^{(s_q,1)} & \ldots & F_{\infty,s_q}^{-1} \tilde{J}^{(s_q,n)}
\end{pmatrix}.
\]

Note that \( \tilde{J}_1 \) can be seen as \( \tilde{J} \) with rows \( s_1, \ldots, s_q \) removed.

This choice of \( \tilde{J}_1 \) and \( \tilde{J}_2 \) is not unique. Different choices lead to different algorithms for the computation of \( \tilde{a}_{1|n}, \ldots, \tilde{a}_{n|n} \). It is not clear if other choices of \( \tilde{J}_1 \) and \( \tilde{J}_2 \) can lead to more efficient algorithms than the diffuse Kalman filter.

**Example (continued): The Local Level Model**

Using the identities in (2.45) we find that \( \tilde{J}_1 \) and \( \tilde{J}_2 \) coincide with the matrices \( J_1 \) and \( J_2 \) that we proposed earlier for the local level model.

### 2.12.3 Constructing a New State Space Model

Recall that we are interested in obtaining the estimators \( \tilde{a}_{t|n} \) given by

\[
\tilde{a}_{t|n} = G_t J_2 y + P(\alpha_t - G_t J_2 y | J_1 y) \\
= \tilde{B}_t \tilde{J}_2 y + P(\alpha_t - \tilde{B}_t \tilde{J}_2 y | \tilde{J}_1 y)
\]

for \( t = 1, \ldots, n \), where \( \tilde{B}_t, \tilde{J}_1 \) and \( \tilde{J}_2 \) were constructed in the preceding sections. We will now see that the elements of \( \tilde{J}_1 y \) can be viewed as a time series generated by a state space model with states \( \alpha_t - \tilde{B}_t \tilde{J}_2 y \). Therefore we can obtain the MMSLEs \( P(\alpha_t - \tilde{B} \tilde{J}_2 y | \tilde{J}_1 y) \) by means of the traditional Kalman filter and smoother.

Let \( w_t = \tilde{J}^{(t)} y \), where \( \tilde{J}^{(t)} \) is the \( t \)th row of the matrix \( \tilde{J} \) constructed recursively in the previous section. From the definition of \( \tilde{J} \) it follows that \( w_t \) are related through the following recursions

\[
w_t = y_t - Z_t \sum_{k:s_k < t} \left( \prod_{m=s_k}^{t-1} T_m \right) M_{\infty,s_k} w_{s_k}, \tag{2.47}
\]
for \( t = 1, \ldots, n \). We also have
\[
\tilde{B}_t \tilde{J}_2 y = \sum_{k:s_k \leq t} \left( \prod_{m=s_k}^{t-1} T_m \right) M_{\infty,s_k} w_{s_k}, \tag{2.48}
\]
for \( t = 1, \ldots, n \), which follows from the fact that \( \tilde{J}_2 y = (F_{\infty,s_1}^{-1} w_{s_1}, \ldots, F_{\infty,s_q}^{-1} w_{s_q})' \) and the definition of \( \tilde{B} \) in (2.44). If we define \( \bar{\alpha}_t = \alpha_t - \tilde{B}_t \tilde{J}_2 y \) for \( t = 1, \ldots, n \), we find that for \( t \notin \{s_1, \ldots, s_q\} \)
\[
w_t = Z_t \bar{\alpha}_t + \epsilon_t. \tag{2.49}
\]
Equation (2.49) is the observation equation of our new state space model. Recall that our goal is construct a state space model for \( \tilde{J}_1 y \) and since \( w_{s_1}, \ldots, w_{s_q} \) are not included in this vector we can leave the observation equation undefined for these values of \( t \). In practice \( w_{s_1}, \ldots, w_{s_q} \) will be considered missing ‘observations’.

We will now see how \( \bar{\alpha}_t \) can be written as a linear process. For the initial state \( \bar{\alpha}_1 \) we find
\[
\bar{\alpha}_1 = \begin{cases} \alpha_1 - M_{\infty,1}(Z_1 \alpha_1 + \epsilon_1) & \text{if } s_1 = 1, \\ \alpha_1 & \text{else.} \end{cases}
\]
We transformed \( BA \) to \( \tilde{B} \). Therefore if \( s_1 = 1 \) we have \( \alpha_1 = P_{\infty,1} Z_1' \delta + \alpha_1^* \) and else \( \alpha_1 = \alpha_1^* \), where \( \text{Var}(\alpha_1^*) = RR' \). It follows that
\[
\text{Var}(\bar{\alpha}_1) = \begin{cases} (I - M_{\infty,1}Z_1)RR'(I - M_{\infty,1}Z_1)' + H_1 & \text{if } s_1 = 1, \\ RR' & \text{else.} \end{cases}
\]
It follows from (2.48) and the definition of \( \bar{\alpha}_t \) that
\[
\bar{\alpha}_{t+1} = T_t \bar{\alpha}_t - X_t + \eta_t, \tag{2.50}
\]
with
\[
X_t = \begin{cases} M_{\infty,t+1}[Z_{t+1}(T_t \bar{\alpha}_t + \eta_t) + \epsilon_{t+1}], & \text{if } t \in I_{s-1}, \\ 0, & \text{else}, \end{cases}
\]
where \( I_{s-1} = \{s_1 - 1, \ldots, s_q - 1\} \).

The linear process of (2.50) can thus also be written as
\[
\bar{\alpha}_{t+1} = T_t^* \bar{\alpha}_t + \eta_t^*, \tag{2.51}
\]
where $\mathbb{E}(\eta_t^*) = 0$ and $\text{Var}(\eta_t^*) = Q_t^*$ with

$$T_t^* = \begin{cases} (I - M_{\infty,t+1}Z_{t+1})T_t & t \in I_{s-1}, \\ T_t & \text{else} \end{cases}$$

and

$$Q_t^* = \begin{cases} Q_t + M_{\infty,t+1}[Z_{t+1}Q_tZ_{t+1}' + H_{t+1}] M'_{\infty,t+1} & t \in I_{s-1}, \\ Q_t & \text{else}. \end{cases}$$

**Example (continued): The Local Level Model**

For the local level model we find that $w_t = y_t - y_1$ for $t = 2, \ldots, n$. Also, since $\tilde{B}_t\tilde{J}_2y = y_1$ we have $\tilde{\alpha}_t = \alpha_t - y_1$ for $t = 1, \ldots, n$. Note further that for all $t$ we have $T_t^* = T_t$ and $Q_t^* = Q_t$, since $s_1 = 1$. The process definition of $\alpha_t$ and $\tilde{\alpha}_t$ only differs in the specification of the initial state. We have

$$\tilde{\alpha}_1 = -\varepsilon_1,$$

The new state space model is therefore given by

$$w_t = \tilde{\alpha}_t + \varepsilon_t, \quad \alpha_{t+1} = \tilde{\alpha}_t + \eta_t,$$

with $\mathbb{E}(\tilde{\alpha}_t) = 0$ and $\text{Var}(\tilde{\alpha}_t) = \sigma^2_{\varepsilon}$.

**2.12.4 Simplifying the Calculations**

It follows from the discussion of the preceding section that if we define the following state space model

$$w_t = Z_t\tilde{\alpha}_t + \varepsilon_t, \quad \alpha_{t+1} = T_t^*\tilde{\alpha}_t + \eta_t^*,$$

with $w_{s_1}, \ldots, w_{s_q}$ treated as missing, then

$$\tilde{\alpha}_{t|n} = \tilde{B}_t\tilde{J}_2y + P(\tilde{\alpha}_t|w_1, \ldots, w_n),$$

for $s, t = 1, \ldots, n$. However, this requires an extra recursion to evaluate $\tilde{B}_t\tilde{J}_2y, \ldots, \tilde{B}_n\tilde{J}_2y$. A more efficient approach is to incorporate the term $\tilde{B}_t\tilde{J}_2y$ directly into the state space model. I define $\tilde{\alpha}_t = \tilde{\alpha}_t + \tilde{B}_t\tilde{J}_2y$ for $t = 1, \ldots, n$, where $\tilde{B}_t\tilde{J}_2y$ is considered constant. It follows from
(2.47) and (2.48) and after some minor algebra that

\[ y_t = Z_t \tilde{\alpha}_t + \varepsilon_t, \tag{2.52} \]
\[ \tilde{\alpha}_{t+1} = \tilde{d}_t + T_t^* \tilde{\alpha}_t + \eta_t^*, \tag{2.53} \]

where \( y_{s_1}, \ldots, y_{s_q} \) are considered missing and

\[ \tilde{d}_t = \begin{cases} M_{\infty,t+1} y_{t+1} & t \in I_{s-1}, \\ 0 & \text{else}, \end{cases} \quad \tilde{\alpha}_1 = \begin{cases} \bar{\alpha}_1 + M_{\infty,1} y_1 & s_1 = 1, \\ \bar{\alpha}_1 & \text{else}, \end{cases} \]

were as before \( I_{s-1} = \{s_1 - 1, \ldots, s_q - 1\} \). It follows that if we run the Kalman filter and Kalman smoother on the state space model given by (2.52) and (2.53) that

\[ \tilde{a}_{t|s} = P(\tilde{\alpha}_t|w_1, \ldots, w_s), \]

for \( s, t = 1 \ldots, n \). Also note that the likelihood of model (2.52) and (2.53) coincides with the likelihood \( p(w_1, \ldots, w_n) \). It follows that the marginal likelihood can be computed using the Kalman filter for the new model.

It is a matter of simple algebra to check that the recursions obtained when running the Kalman filter and smoother recursions on model (2.52) and (2.53) are equivalent to the diffuse Kalman filter and smoother recursions of Ansley and Kohn (1985) and Koopman (1997).

**Example (continued): The Local Level Model**

For the local level model only the initial state changes. Since \( \tilde{\alpha}_t = \bar{\alpha}_t + y_1 \) we have \( \tilde{\alpha}_1 = y_1 - \varepsilon_1 \) and

\[ y_t = \tilde{\alpha}_t + \varepsilon_t, \quad \tilde{\alpha}_{t+1} = \tilde{\alpha}_t + \eta_t. \]

It follows that in a local level model we can obtain the estimators \( \tilde{a}_{1|n}, \ldots, \tilde{a}_{n|n} \) by using the usual Kalman smoother recursions but setting \( \mathbb{E}(\alpha_1) = y_1 \) and \( \text{Var}(\alpha_1) = \sigma_\varepsilon^2 \) and treating the first observation as missing.
2.A The Kalman Filter and the LDL decomposition

The following results relate the Kalman filter recursions to the LDL decomposition of the variance matrix $\Sigma$ defined in Section 2.3. Similar results are well-known in the literature, see e.g. Section 4.11 of Durbin and Koopman (2001). The result of Proposition 2.1 is however slightly more general since it does not require the matrices $H_1, \ldots, H_n$ to be positive semi-definite. This level of generality is needed for the developments in Chapter 3. For univariate models the result is extended to the case where the variance matrix $\Sigma$ is singular in a corrolary. This result is used in the discussion of the diffuse Kalman filter.

**Proposition 2.1.** Let $\Sigma = \Psi + H$ be a non-singular matrix, where $\Psi$ is defined in Section (2.3) where $H$ is given by $H = \text{diag}\{H_1, \ldots, H_n\}$ for a set of symmetric matrices $H_1, \ldots, H_n$. Assume that a decomposition for the symmetric matrix $\Sigma = LDL'$ exists, where $L$ and $D$ are of the form (2.8), then

$$L_{t,j} = Z_t (T_{t-1} \ldots T_{j+1}) K_j,$$

for $t = 2, \ldots, n$, $j = 1, \ldots, t - 1$ and $D = \text{diag}\{F_1, \ldots, F_n\}$ where $K_1, \ldots, K_{n-1}$ and $F_1, \ldots, F_n$, are obtained from the Kalman filter recursions (2.13) and (2.14).

**Proof.** Let $U = DL'$. The $(i, j)$ block of a matrix is labelled by subscript ‘$ij$’ for the range of $i, j = 1, \ldots, n$. The $m \times m$ matrix block $(i, j)$ of $\Sigma$ is given by

$$\Sigma_{ij} = \begin{cases} 
\sum_{k=1}^{i-1} L_{ik} U_{ki} + U_{ij}, & i = j, \\
\sum_{k=1}^j L_{ik} U_{kj}, & i > j, \\
\sum_{k=1}^{i-1} L_{ik} U_{kj} + U_{ij}, & i < j.
\end{cases}$$

(2.55)

From the expressions in Section 2.3, we have

$$\Omega_{ij} = \begin{cases} 
T_{i-1} \Omega_{i-1,i-1} T_{i-1}^\prime + R_i Q_i R_i^\prime, & i = j > 1, \\
T_{i-1} T_{i-2} \ldots T_{j} \Omega_{jj}, & i > j, \\
\Omega_{ii} (T_{j-1} T_{j-2} \ldots T_i)^\prime, & i < j,
\end{cases}$$

(2.56)

and $\Omega_{11} = P_{1|0}$. From the definition $\Sigma = \Psi + H = Z\Omega Z' + H$, we have further

$$\Sigma_{ij} = \begin{cases} 
Z_i \Omega_{ii} Z_i', & i = j, \\
Z_i \Omega_{ij} Z_j', & i \neq j.
\end{cases}$$

(2.57)

Equating (2.55) with (2.57) and considering (2.56), the following expressions for the block
element matrices of $U$ and $L$ are obtained:

$$U_{ij} = Z_i \Omega_{ii} Z_i' + H_i - \sum_{k=1}^{i-1} L_{ik} U_{ki}, \quad i = j,$$

$$L_{ij} = \left( Z_i T_{i-1} T_{i-2} \ldots T_j \Omega_{jj} Z_j' - \sum_{k=1}^{j-1} L_{ik} U_{kj} \right) U_{jj}^{-1}, \quad i > j, \quad (2.58)$$

$$U_{ij} = Z_i \Omega_{ii} (T_{j-1} T_{j-2} \ldots T_i)' Z_j' - \sum_{k=1}^{i-1} L_{ik} U_{kj}, \quad i < j,$$

which describe the typical triangular system for which solutions are obtained by forwards and backwards substitution algorithms; see Golub and Van Loan (1997, Chs 3, 4).

Given the block-diagonal structure of $Z$ and the definition $\Sigma = LU = Z\Omega Z' + H$, the lower block matrix $L_{ij}$ equals a matrix that is premultiplied by $Z_i$ for $i > j$, and the upper block matrix $U_{ij}$ equals a matrix that is postmultiplied by $Z_j$ for $i < j$, with $i, j = 1, \ldots, n$. Therefore, we obtain from the expression (2.58)

$$U_{ij} = Z_i Y_{ii} Z_i', \quad i = j,$$

$$L_{ij} = Z_i X_{ij} U_{jj}^{-1}, \quad i > j, \quad (2.59)$$

$$U_{ij} = Z_i Y_{ij} Z_j', \quad i < j,$$

where the matrices $X_{ij}$ and $Y_{ij}$ are defined by

$$Y_{ij} = \Omega_{ii} - \sum_{k=1}^{i-1} X_{ik} U_{kk}^{-1} Z_k Y_{ki}, \quad i = j,$$

$$X_{ij} = T_i T_{i-1} T_{i-2} \ldots T_j \Omega_{jj} Z_j' - \sum_{k=1}^{j-1} X_{ik} U_{kk}^{-1} U_{kj}, \quad i > j,$$

$$Y_{ij} = \Omega_{ii} (T_{j-1} T_{j-2} \ldots T_i)' - \sum_{k=1}^{i-1} X_{ik} U_{kk}^{-1} Z_k Y_{kj}, \quad i < j. \quad (2.60)$$

For $i = t + 1$ it follows that

$$X_{t+1,t} = T_t \Omega_{tt} Z_t' - \sum_{k=1}^{t-1} X_{t+1,k} U_{kk}^{-1} U_{kt}$$

$$= T_t (\Omega_{tt} - \sum_{k=1}^{t-1} X_{t+1,k} U_{kk}^{-1} Z_k Y_{kt}) Z_t'$$

$$= T_t Y_{tt} Z_t', \quad (2.61)$$

$$X_{t+1,j} = T_t T_{t-1} T_{t-2} \ldots T_j Y_{jj} Z_j'$$

$$L_{t+1,j} = Z_{t+1} T_t T_{t-1} \ldots T_j Y_{jj} Z_j' U_{jj}^{-1}$$

$$= Z_{t+1} T_t T_{t-1} \ldots T_{j+1} M_j,$$

where we define $M_j = T_j Y_{jj} Z_j' U_{jj}^{-1} = Y_{j,j+1} Z_j' U_{jj}^{-1}$ for $j = 1, \ldots, t + 1$. Note that $X_{t+1,t} = M_t U_{tt}$. The matrices $X_{ij}$ and $L_{ij}$ depend on state space matrices and on $Y_{ij}$ for $i > j$.

Given the earlier definitions and results, we continue the proof by deriving a recursion
for $Y_{ij}$ for $i = t + 1$ and $j = t + 2, \ldots, n$. It follows that

\[
Y_{t+1,j} = Y_{t+1,t+1} (T_{j-1} T_{j-2} \ldots T_{t+1})',
\]

\[
U'_{t+1,j} = Z_j T_{j-1} \ldots T_{t+2} M_{t+1} U_{t+1,t+1},
\]

\[
Y_{t+1,t+1} = \Omega_{t+1,t+1} - \sum_{k=1}^{t} X_{t+1,k} U_{kk}^{-1} Z_k Y_{k,t+1}
\]

\[
= T_l \left( \Omega_l - \sum_{k=1}^{t-1} X_{t,k} U_{kk}^{-1} Z_k Y_{k,l} \right) T_l' + R_l Q_l R_l' - X_{t+1,t} U_{tt}^{-1} Z_t Y_{t,t+1}
\]

\[
= T_l Y_l T_l' + R_l Q_l R_l' - M_l U_l M_l',
\]

(2.62)

for $j = t + 2, \ldots, n$. The expression (2.54) follows from (2.60) and (2.62) on noting that $P_l = Y_{ll}$, $K_l = M_l$ and $F_l = U_{ll}$, where $P_l$, $K_l$ and $F_l$ are obtained from the Kalman filter recursions of equation (2.13). Since $U = DL'$, we have $U'_{ij} = L_{ij} U_{ii}$, and thus

\[
D = \text{diag}\{U_{11}, \ldots, U_{nn}\} = \text{diag}\{F_1, \ldots, F_n\}.
\]

Corollary 2.1. Let $\Sigma = \Psi + H$ as in Proposition 2.1, but for a univariate state space model. Assume that $\Sigma$ is positive semi-definite, then $\Sigma = LDL'$, where $L$ is a lower triangular matrix with ones on the diagonal and $D$ is a diagonal matrix. The elements of $L$ are given by

\[
L_{t,j} = \left\{ \begin{array}{ll}
Z_t (T_{t-1} \ldots T_{j+1}) K_j & F_j > 0, \\
0 & F_j = 0,
\end{array} \right.
\]

(2.63)

for $t = 2, \ldots, n$, $j = 1, \ldots, t - 1$ and $D = \text{diag}\{F_1, \ldots, F_n\}$ where $K_1, \ldots, K_{n-1}$ and $F_1, \ldots, F_n$, are obtained from the Kalman filter recursions (2.13) and (2.14).

Proof. The existence of the LDL decomposition follows immediately from Lemma 1 in Ansley and Kohn (1983). The result is proved by repeating the proof of Proposition 2.1, but setting $U_{jj}^{-1} = 0$ if $U_{jj} = 0$ in the expressions of (2.58).
Chapter 3

Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models

3.1 Introduction

In this chapter we will consider nonlinear non-Gaussian state space models. As before, let $y_1, \ldots, y_n$ be a $N$ dimensional time series. The observation equation is now replaced by a family of observation densities

$$p(y_t|\theta_t; \psi), \quad t = 1, \ldots, n,$$

that define the distributions of the observations $y_t$ conditional on $m$ dimensional latent variables $\theta_t$. The vector $\psi$ consists of a set of unknown parameters. I assume that the latent signals $\theta_1, \ldots, \theta_n$ are linear functions of a linear Gaussian state process $\alpha_1, \ldots, \alpha_n$. The state process has the same specification as the state process of the linear Gaussian state space model of Chapter 2, see equation (2.2). Note that (3.1) allows the relationship between $\theta$ and $y$ to be nonlinear.

As discussed in Chapter 2, we can generally not calculate the likelihood for a non-Gaussian model analytically, nor can we obtain mean square optimal estimators of the states or of future observations. If the model is non-linear we can also no longer use the Kalman filter and smoother to obtain mean square optimal linear estimators. Simulation-based methods present a possible solution to these problems. In this chapter I will focus on the use of importance sampling to evaluate the likelihood and smoothed estimators of the states. The results presented carry over easily to Markov chain Monte Carlo methods that can be used for Bayesian analysis of the non-Gaussian state space model. Relevant previous references

Let \( p(\theta \mid y; \psi) \) denote the density of the signal \( \theta = (\theta_1', \ldots, \theta_n')' \) conditional on the observations \( y = (y_1', \ldots, y_n') \). Both importance sampling and the Markov chain Monte Carlo methods require a so-called proposal density that is close to the density \( p(\theta \mid y; \psi) \) and from which samples can be obtained efficiently. In this chapter I focus on a Gaussian proposal density with the same mode as \( p(\theta \mid y; \psi) \) and with the same curvature around the mode. This proposal density has been adopted in a Bayesian analysis by Shephard and Pitt (1997) and in a classical analysis by Durbin and Koopman (1997). The results in these papers were only valid for the state space model (3.1) with a log-concave observation density \( p(y \mid \theta; \psi) \). In this chapter we will see that with small modifications the algorithms can also be used for the general model.

3.2 The Non-Gaussian State Space Model

The \( m \times 1 \) signal vectors \( \theta_t \) are modeled as linear functions of \( p \times 1 \) state vectors \( \alpha_t \),

\[
\theta_t = c_t + Z_t \alpha_t, \tag{3.2}
\]

for \( t = 1, \ldots, n \) with fixed and known system vectors \( c_t \) and system matrices \( Z_t \), possibly depending on \( \psi \). The unobserved process \( \alpha_t \) is of the same form as in the linear state space model of Section 2.1

\[
\alpha_{t+1} = d_t + T_t \alpha_t + R_t \eta_t, \quad \eta_t \sim N(0, Q_t), \tag{3.3}
\]

for \( t = 1, \ldots, n \), where the system vectors \( d_t \) and the system matrices \( T_t \) and \( Q_t \) may depend on \( \psi \). The state vector \( \alpha_t \) and the disturbance vector \( \eta_t \) have dimension \( p \times 1 \). The system matrices have appropriate dimensions and the variance matrix \( Q_t \) is positive semidefinite. For the initial state \( \alpha_1 \), I assume that

\[
\alpha_1 \sim N(a_{1|0}, P_{1|0}).
\]

The disturbances \( \eta_t \) are serially independent and are independent of the initial state vector.
3.2. THE NON-GAUSSIAN STATE SPACE MODEL

It follows that $\alpha = (\alpha_1', \ldots, \alpha_n')'$ has a multivariate Gaussian distribution

$$\alpha \sim N(m, \Omega),$$

with

$$m = Td, \quad \Omega = TRQR'T',$$

where $T$, $d$, $R$ and $Q$ were defined in Section 2.3. Furthermore, if we denote $\theta = (\theta_1', \ldots, \theta_n')'$ then

$$\theta \sim N(\mu, \Psi), \quad (3.4)$$

with

$$\mu = c + Zm, \quad \Psi = Z\Omega Z',$$ \quad (3.5)

where $c$, $Z$ and $\Omega$ were defined in Section 2.3.

In this chapter I consider nonlinear non-Gaussian observation model for which the conditional independence assumption applies; that is

$$p(y|\theta; \psi) = \prod_{t=1}^{n} p(y_t|\theta_t; \psi), \quad t = 1, \ldots, n, \quad (3.6)$$

where $p(y|\theta; \psi)$ is the density of $y$ conditional on the signal $\theta$. Examples of densities $p(y_t|\theta_t; \psi)$ are all members of the exponential family densities and the stochastic volatility models; see Durbin and Koopman (2001, Ch. 10). System variables and densities may depend on a parameter vector $\psi$. Classical and Bayesian estimation of $\psi$ is discussed in Section 3.3.

Note that the linear state space models of Chapter 2 are a special case of the state space models considered in this chapter. Recall from Section 2.1 that in a linear Gaussian state space model $y_1, \ldots, y_n$ are given by

$$y_t = c_t + Z_t \alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t), \quad (3.7)$$

for $t = 1, \ldots, n$, where the $\varepsilon_t$ are mutually independent and independent of the state disturbances and initial state. The variance matrices $H_1, \ldots, H_n$ are fixed and known. System variables $c_t$, $Z_t$ and $H_t$ may partly depend on $\psi$. In matrix form, the model is given by

$$y = c + Z\alpha + \varepsilon, \quad \varepsilon \sim N(0, H), \quad (3.8)$$

where $H = \text{diag}\{H_1, \ldots, H_n\}$. 

3.3 Simulation-Based Inference

3.3.1 Importance Sampling

The purpose of importance sampling methods is to numerically evaluate integrals through simulation. Suppose we want to evaluate \( \bar{\kappa}(y; \psi) \) given by

\[
\bar{\kappa}(y; \psi) = \int \kappa(\theta, \psi) p(\theta | y; \psi) d\theta,
\]

where \( \kappa(\theta, \psi) \) can be any function of signal \( \theta \) and parameter vector \( \psi \) for which the integral exists. Let \( f(\theta; y, \psi) \) denote a density that is in some sense close to \( p(\theta | y; \psi) \). We have the following identity

\[
\bar{\kappa}(y; \psi) = \int \kappa(\theta, \psi) p(\theta | y; \psi) f(\theta; y, \psi) d\theta = \frac{1}{p(y; \psi)} \int \kappa(\theta, \psi) p(y, \theta; \psi) f(\theta, y, \psi) d\theta,
\]

where \( p(y; \psi) \) is the likelihood of \( y_1, \ldots, y_n \). We can write the last term as a quotient of two expectations with respect to the density \( f(\theta; y, \psi) \)

\[
E_f \left[ \frac{\kappa(\theta, \psi) p(y, \theta; \psi)}{f(\theta; y, \psi)} \right] = \int \kappa(\theta, \psi) p(y, \theta; \psi) f(\theta; y, \psi) d\theta,
\]

\[
E_f \left[ \frac{p(y, \theta; \psi)}{f(\theta; y, \psi)} \right] = p(y, \psi).
\]

If we sample \( M \) vectors \( \theta^1, \ldots, \theta^M \) from \( f(\theta; y, \psi) \) we can estimate \( \bar{\kappa}(y; \psi) \) by \( \hat{\kappa}(y; \psi) \) given by

\[
\hat{\kappa}(y; \psi) = \left\{ \sum_{i=1}^{M} q(\theta^i; \psi) \right\}^{-1} \sum_{i=1}^{M} \kappa(\theta^i, \psi) q(\theta^i; y; \psi),
\]

where

\[
q(\theta^i; y; \psi) = \frac{p(y, \theta^i; \psi)}{f(\theta^i; y, \psi)}.
\]

This estimator is subject to Monte Carlo error. Under certain regularity conditions \( \hat{\kappa} \) converges to \( \bar{\kappa} \) as \( M \to \infty \) with a rate of convergence depending on the precision of the proposal density (Geweke, 1989). Note that the likelihood \( p(y; \psi) \) is estimated as follows

\[
\hat{p}(y; \psi) = \frac{1}{M} \sum_{i=1}^{M} q(\theta^i; \psi).
\]

We can thus obtain (simulated) maximum likelihood estimators by maximizing \( \hat{p}(y; \psi) \) with respect to \( \psi \).

Smoothed estimators of the signals can be obtained by setting \( \kappa(\theta, \psi) = \theta \). Smoothed
estimator of the state vector can be obtained using the identity,

\[ E[\alpha_t | y; \psi] = \int E[\alpha_t | y, \theta; \psi] p(\theta | y; \psi) d\theta, \]

which can be evaluated by (3.10) with \( \kappa(\theta, \psi) \) set to \( E(\alpha_t | \theta; \psi) \). The conditional expectation \( E_\psi(\alpha_t | \theta) \) can be computed by applying the Kalman filter and smoother to the Gaussian state space model (3.8) with \( y = \theta \) and \( H = 0 \). The same argument can be adopted for the evaluation of \( \text{Var}_\psi(\alpha_t | y) \) for \( t = 1, \ldots, n \). We note that the smoothed state estimator is not evaluated via the simulation of the state \( \alpha \) but via the simulation of the signal \( \theta \). This is a direct application of the Rao-Blackwellisation argument; see Casella and Robert (1996, §3.7.3).

### 3.3.2 Bayesian Analysis

In a Bayesian analysis our prior believes on the value of \( \psi \) are summarized by a prior density \( p(\psi) \) and the focus is on the posterior density \( p(\psi | y) \). This posterior is generally not available in closed form for non-Gaussian state space models. This can be resolved by the use of Markov chain Monte Carlo methods as described in Shephard and Pitt (1997). In this case samples from the density \( p(\psi, \theta | y) \) are obtained by means of the Gibbs sampler. This amounts to successively sampling \( \theta \) conditional on \( y \) and \( \psi \), and then \( \psi \) conditional on \( y \) and \( \theta \). Within the Gibbs sampler, \( \theta \) is sampled via a Metropolis-Hastings step where the density \( f(\theta; y, \psi) \) can be used as the proposal density. Sampling \( \theta \) as one block gives a low acceptance rate for the Metropolis-Hastings step and introduces large correlations between successive draws. This can be solved by sampling the \( N + 1 \) individual blocks \((\theta_1, \ldots, \theta_{k_1-1}), (\theta_{k_1}, \ldots, \theta_{k_2-1}), \ldots, (\theta_{k_N}, \ldots, \theta_n)\) as part of an additional Gibbs sampler sweep. The algorithms presented in this chapter can be easily adapted for carrying out this Metropolis-Hastings step. The number of blocks is crucial for the performance of the algorithm. The correlation between samples is high when \( N \) is large while the Metropolis-Hastings acceptance rate is low when \( N \) is small. It is possible to let the size and location of the blocks vary between sweeps of the Gibbs sampler; see Shephard and Pitt (1997) for more details.

### 3.4 Computing the Posterior Mode

In this chapter the proposal density \( f(\theta; y; \psi) \) is chosen as the Gaussian density with the same mode as the target density \( p(\theta | y; \psi) \) as well as the same curvature around this mode. For nonlinear non-Gaussian state space models an analytical expression for the posterior mode
\( \hat{\theta} \) of \( p(\theta|y; \psi) \) is usually not available. I will show in this section how to obtain the mode by maximizing \( p(\theta|y; \psi) \) with respect to \( \theta \) using the Newton-Raphson method, see Nocedal and Wright (1999) for a treatment of numerical optimization methods. The dimension of \( \theta \) is \( mn \times 1 \) so that matrix dimensions are typically high, straightforward matrix manipulations become infeasible and efficient algorithms need to be considered.

For a given guess \( g \) of the optimum \( \hat{\theta} \), the Newton-Raphson method produces a new guess as follows

\[
g^+ = g - \left\{ \frac{\hat{p}(\theta|y; \psi)|_{\theta=g}}{\hat{p}(\theta|y; \psi)|_{\theta=g}} \right\}^{-1} \hat{p}(\theta|y; \psi)|_{\theta=g}, \tag{3.11}
\]

where

\[
\hat{p}(\cdot|\cdot) = \frac{\partial \log p(\cdot|\cdot)}{\partial \theta}, \quad \hat{p}(\cdot|\cdot) = \frac{\partial^2 \log p(\cdot|\cdot)}{\partial \theta \partial \theta^t}. \tag{3.12}
\]

Since,

\[
\log p(\theta|y; \psi) = \log p(y|\theta; \psi) + \log p_G(\theta; \psi) - \log p(y; \psi),
\]

were \( p_G(\theta; \psi) \) is the density of the Gaussian variable \( \theta \), we have

\[
\hat{p}(\theta|y) = \hat{p}(y|\theta) - \Psi^{-1}(\theta - \mu), \quad \hat{p}(\theta|y) = \hat{p}(y|\theta) - \Psi^{-1}. \tag{3.13}
\]

The conditional independence assumption in (3.6) implies that \( \hat{p}(y|\theta) \) is block-diagonal.

By substitution of (3.13) into (3.11), the Newton-Raphson updating step becomes

\[
g^+ = g - \left\{ \frac{\hat{p}(y|\theta)|_{\theta=g}}{\hat{p}(y|\theta)|_{\theta=g}} - \Psi^{-1}(g - \mu) \right\}^{-1} \left\{ \frac{\hat{p}(y|\theta)|_{\theta=g}}{\hat{p}(y|\theta)|_{\theta=g}} - \Psi^{-1}(g - \mu) \right\}
= (\Psi^{-1} + A^{-1})^{-1} (A^{-1} x + \Psi^{-1} \mu), \tag{3.14}
\]

where

\[
A = - \left\{ \frac{\hat{p}(y|\theta)|_{\theta=g}}{\hat{p}(y|\theta)|_{\theta=g}} \right\}^{-1}, \quad x = g + A \hat{p}(y|\theta)|_{\theta=g}.
\]

Now consider the linear Gaussian state space model of equation (3.7). Using Lemma 2.1 and \( \operatorname{Var}(\theta) = \operatorname{Cov}(\theta, y) = \Psi \), it follows that

\[
\mathbb{E}[\theta|y; \psi] = \mu + \Psi \Sigma^{-1} (y - \mu), \tag{3.15}
\]

where

\[
\Sigma = \operatorname{Var}_\psi(y) = \Psi + H,
\]

and the subscript \( \psi \) is added to emphasize the dependence on the value of the parameters \( \psi \). After some minor manipulations, it follows from equation (3.15) that the conditional mean
3.4. COMPUTING THE POSTERIOR MODE

is given by

\[ \hat{\theta} = (\Psi^{-1} + H^{-1})^{-1} (H^{-1}y + \Psi^{-1}\mu). \] (3.16)

Note the similarity between (3.14) and (3.16). If \( A \) is positive definite it follows that if we apply the Kalman filter and smoother to the Gaussian state space model given by

\[
\begin{align*}
\mathbf{x}_t &= c_t + Z_t \alpha_t + \epsilon_t, & \epsilon_t &\sim N(0, A_t), \\
\alpha_{t+1} &= d_t + T_t \alpha_t + R_t \eta_t, & \eta_t &\sim N(0, Q_t),
\end{align*}
\] (3.17)

then

\[ g^+ = \mathbb{E}_G [\theta | x; \psi], \]

where \( x = (x_1', \ldots, x_n')', \theta, = (\theta_1', \ldots, \theta_n') \) with \( \theta_i \) defined in equation 3.2, and the subscript \( G \) indicates that the expectation is taken with respect to the model given by equations (3.17) and (3.18).

This iterative approach of computing the posterior mode is adopted by Shephard and Pitt (1997), Durbin and Koopman (1997) and So (2003, §2). The method is clearly not valid when \( \hat{p}(y|\theta) \) is not negative definite since this will imply that the variance matrix \( H \) of the linear Gaussian model (3.8) is not positive semidefinite.

The following theorem states that the classical Kalman filter and smoother can be used to evaluate equation (3.14) even if the matrices \( A_t \) are not semi-positive definite.

**Theorem 3.1.** Let \( \mu \) and \( \Psi \) be of the form (3.5), \( y = (y_1', \ldots, y_n') \) for \( N \times 1 \) dimensional vectors \( y_t \) and for a set of nonsingular symmetric matrices \( H_1, \ldots, H_n \) let \( H = \text{diag}\{H_1, \ldots, H_n\} \) then

\[ \theta^* = (\Psi^{-1} + H^{-1})^{-1} (H^{-1}y + \Psi^{-1}\mu), \] (3.19)

with \( \theta^* = (\theta^*_1', \ldots, \theta^*_n') \) where

\[ \theta^*_t = c_t + Z_t a_t|n, \]

for \( t = 1, \ldots, n \) and \( a_t|n, \ldots, a_n|n \) are obtained from the Kalman smoother recursions (2.19) and (2.18).

To proof this theorem, first note that

\[
(\Psi^{-1} + H^{-1})^{-1} (H^{-1}y + \Psi^{-1}\mu) = \mu + \Psi \Sigma^{-1}(y - \mu) = c + Z [m + \Omega \Sigma^{-1}(y - \mu)],
\]

where...
where \( m \) and \( \Omega \) were defined in Section (3.2). In Section 2.6 we saw already that

\[
\hat{a} = m + \Omega \Sigma^{-1} (y - \mu),
\]

(3.20)

where \( \hat{a}' = (a'_{1:n}, \ldots, a'_{n:n}) \) if the matrices \( H_1, \ldots, H_n \) are positive definite. Expression (3.20) can be proved for the general case by following essentially the same steps as in Sections 2.5 and 2.6 and using the fact that Proposition 2.1 does not require \( H_1, \ldots, H_n \) to be positive definite.

The intuition behind Theorem 3.1 is that the Kalman filter and smoother construct an LDL decomposition of \( \Sigma \), as explained in Section 2.5. This remains true even if the ‘variance matrices’ \( H_1, \ldots, H_n \) are not semi-positive definite, see also Proposition 2.1. Note however, that \( \hat{a} \) can no longer be interpreted as the smoothed state of a state space model.

When implementing the Newton-Raphson optimization procedure it is useful to be able to evaluate the derivative \( p(\theta|y; \psi) \) with respect to \( \theta \), which we denote by \( \dot{p}(\theta|y; \psi) \). This derivative can be used to determine if the optimization algorithm has converged or to implement a line-search. Furthermore, the derivative is required when using a quasi-Newton algorithm, such as the BFGS algorithm or the conjugate gradient algorithm to find the mode, see Nocedal and Wright (1999). Fortunately, \( \dot{p}(\theta|y; \theta) \) can be evaluated analytically. Recall from equation 3.13 that

\[
\dot{p}(\theta|y) = \dot{p}(y|\theta) - \Psi^{-1}(\theta - \mu).
\]

We can usually find an analytical expression for \( \dot{p}(y|\theta) \) with relative ease. It follows from Theorem 3.1 that the term \( \Psi^{-1}(\theta - \mu) \) can be evaluated by applying the Kalman filter smoother to the linear Gaussian state space model (3.8) with \( y = \theta \) and \( H = 0 \), since \( \Sigma = \Psi \) when \( H = 0 \).

### 3.5 Sampling from the Proposal Density

The importance density \( f(\theta; y; \psi) \) is given by

\[
\log f(\theta; y; \psi) = -\frac{mn}{2} \log 2\pi - \frac{1}{2} \log |V| - \frac{1}{2}(\theta - \hat{\theta})'V^{-1}(\theta - \hat{\theta}),
\]

(3.21)

where \( \hat{\theta} \) is the mode of \( p(\theta|y, \psi) \) and

\[
V = -[\dot{p}(\theta|y; \psi)]_{\theta=\hat{\theta}}^{-1} = (A^{-1} + \Psi^{-1})^{-1},
\]
3.5. SAMPLING FROM THE PROPOSAL DENSITY

with

$$A = \{ \tilde{p}(y|\theta; \psi)|_{\theta = \hat{\theta}} \}^{-1}. $$

After some minor matrix manipulations, it follows that

$$V = (A^{-1} + \Psi^{-1})^{-1} = \Psi - \Psi \Sigma^{-1} \Psi = A - A \Sigma^{-1} A, \quad (3.22)$$

where $\Sigma = \Psi + A$.

Suppose $A$ is positive definite and consider the linear Gaussian state space model given by equations (3.17) and (3.18). It follows from the discussion of Section 3.4 and equations (3.22) and (3.15) that

$$E_G(\theta|x) = \hat{\theta}, \quad \text{Var}_G(\theta|x) = V,$$

where as before the subscript $G$ is used to emphasize that the variance is calculated under the model consisting of (3.17) and (3.18). If we denote the density of $\theta$ conditional on $x$ for the model consisting (3.17) and (3.18) by $p_G(\theta|y)$, then $p_G(\theta|y)$ and $f(\theta; y; \psi)$ coincide since both are Gaussian and have equal mean and variance.

The simulation smoothers of de Jong and Shephard (1995) and Durbin and Koopman (2002) can be used to generate draws from $p_G(\theta|y)$, and thus from $f(\theta; y; \psi)$, in a computationally efficient way. However, this approach requires a log-concave $p(y|\theta; \psi)$ so that $A$ is positive definite and the state space model with observation equation (3.17) is well-defined. This condition is not necessary for Theorem 3.2, the proof of which is in Appendix 3.A. Theorem 3.2 only requires that matrix $\tilde{p}(\theta|y)$ is invertible.

**Theorem 3.2.** Let $\mu$ and $\Psi$ be of the form (3.5), $y = (y_1', \ldots, y_n')'$ for $N \times 1$ dimensional vectors $y_t$ and for a set of nonsingular symmetric matrices $H_1, \ldots, H_n$ let $H = \text{diag}\{H_1, \ldots, H_n\}$. Denoting $V = \Psi - \Psi \Sigma^{-1} \Psi$, where $\Sigma = \Psi + H$, we have

$$(y - u) \sim N(\theta^*, V),$$

where $\theta^*$ is given in (3.19) and $u = (u_1', \ldots, u_n')'$ is obtained from

\[
\begin{align*}
   u_t &= H_t(w_t + F_t^{-1} v_t - K_t^{-1} \tilde{r}_t), \quad w_t \sim N(0, C_t), \\
   \tilde{r}_{t-1} &= Z_t' H_t^{-1} u_t - R_t' w_t + T_t' \tilde{r}_t, \quad C_t = H_t^{-1} - F_t^{-1} - K_t^{-1} \tilde{N}_t K_t, \quad (3.23) \\
   R_t &= C_t^{-1}(H_t^{-1} Z_t - K_t^{-1} \tilde{N}_t T_t), \quad \tilde{N}_{t-1} = R_t' C_t R_t - Z_t' H_t^{-1} Z_t + T_t' \tilde{N}_t T_t, 
\end{align*}
\]

for $t = n, n - 1, \ldots, 1$, with $\tilde{r}_n = 0$ and $\tilde{N}_n = 0$ and where $v_1, \ldots, v_n$, $F_1, \ldots, F_n$ and $K_1, \ldots, K_n$ are obtained from the Kalman filter recursions of Section 2.5.
Although the matrices $A$ and $\Sigma$ can be non-positive definite, the matrix $V$ is evaluated at the optimum $\hat{\theta}$ and is therefore positive definite by definition. I show in Appendix 3.B that the new simulation smoothing equations (3.23) are equivalent to the recursions of the simulation smoother of de Jong and Shephard (1995).

Suppose $\theta^1, \ldots, \theta^M$ are samples from the proposal density $f(\theta^i; y)$. To use these samples to estimate the likelihood we need to calculate the values $f(\theta^i; y; \psi)$. Suppose each sample $\theta^i$ is obtained from the simulation smoother (3.23) and $w^1_i, \ldots, w^n_i$ are the Gaussian samples used. In Appendix 3.C I prove the following expression for $\log f(\theta^i; y; \psi)$

$$\log f(\theta^i; y; \psi) = -\frac{nm}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} \log |A_t|^2 - \sum_{t=1}^{n} \log |C_t| - \frac{1}{2} \sum_{t=1}^{n} w^t_i' C^{-1}_t w^t_i. \quad (3.24)$$

Alternatively, we can use the following expression, which is proved in Appendix 3.D

$$\log f(\theta^i; y; \psi) = \log \tilde{p}(x, \theta^i; \psi) - \log \tilde{p}(x; \psi), \quad (3.25)$$

where $p_G(\theta^i)$ is the (Gaussian) density of $\theta$,

$$\log \tilde{p}(x, \theta^i; \psi) = -\frac{nm}{2} \log 2\pi - \frac{1}{4} \sum_{t=1}^{n} \log |A_t|^2 - \frac{1}{2} \sum_{t=1}^{n} (x_t - \theta^t_i)' A^{-1}_t (x_t - \theta^t_i), \quad (3.26)$$

and

$$\log \tilde{p}(x; \psi) = -\frac{nm}{2} \log 2\pi - \frac{1}{4} \sum_{t=1}^{n} \log |F_t|^2 - \frac{1}{2} \sum_{t=1}^{n} v^t_i' F^{-1}_t v_i,$$

where $F_1, \ldots, F_n$ and $v_1, \ldots, v_n$ are obtained by applying the Kalman filter to the state space model consisting of equations (3.17) and (3.18) with $x_t$ and $A_t$ evaluated in $\hat{\theta}$. This last expression is especially convenient when evaluating the importance weights $q(\theta^i; y; \psi)$ of equation (3.10), because

$$q(\theta^i; y; \psi) = \frac{p(y|\theta^i; \psi)}{\tilde{p}(x, \theta^i; \psi)} \tilde{p}(x; \psi).$$

If the matrices $A_1, \ldots, A_n$ are positive definite $\log \tilde{p}(x; \psi)$ is the likelihood of $x_1, \ldots, x_n$ assuming the true model is that consisting of (3.17) and (3.18) and $\tilde{p}(x, \theta^i; \psi)$ is the density of $x_1, \ldots, x_n$ conditional on $\theta_1, \ldots, \theta_n$.

Specific applications of importance sampling and Markov Chain Monte Carlo methods may require simulations from a density that is close to the density $p(\alpha, \theta|y; \psi)$. The proposal density $f(\theta; y; \psi)$ can be extended to a proposal density for $p(\alpha, \theta|y; \psi)$, denoted by
f(\alpha, \theta; y; \psi). Note that conditional on \theta the states are independent of y. It follows that

\[ p(\alpha, \theta|y; \psi) = p(\alpha|\theta; \psi)p(\theta|y; \psi) \]

where \(p(\alpha|\theta; \psi)\) is the density of \(\alpha\) conditional on \(\theta\). A natural choice for \(f(\alpha, \theta; y; \psi)\) is therefore

\[ p(\alpha|\theta; \psi) = f(\theta; y; \psi) \]

where \(p(\alpha|\theta; \psi)\) is the density of \(\alpha\) conditional on \(\theta\). A natural choice for \(f(\alpha, \theta; y; \psi)\) is therefore

\[ p(\alpha|\theta; \psi) f(\theta; y; \psi) \]

A two-step procedure: first draw \(\theta^i\) from \(f(\theta; y; \psi)\) using Theorem 3.2 and then sample \(\alpha^i\) from \(p(\alpha|\theta^i; \psi)\). To sample from \(p(\alpha|\theta^i; \psi)\) we first sample \(\tilde{\alpha}^i\) from \(p(\alpha; \psi)\). We can then obtain a sample \(\alpha^i\) from \(p(\alpha|\theta^i; \psi)\) as follows

\[ \alpha^i = \mathbb{E}(\alpha|\theta^i; \psi) + \tilde{\alpha}^i - \mathbb{E}(\alpha|\tilde{\theta}^i; \psi), \quad (3.27) \]

where \(\tilde{\theta}^i = c + Z\tilde{\alpha}^i\) and where \(\mathbb{E}(\alpha|\tilde{\theta}^i; \psi)\) and \(\mathbb{E}(\alpha|\theta^i; \psi)\) are evaluated by the Kalman filter smoother applied to the linear Gaussian state space model (3.8) with \(H = 0\) and respectively \(y = \tilde{\theta}^i\) and \(y = \theta^i\). This approach of simulating from conditional densities by mean adjustment is proposed by Journel (1974) in the geostatistics literature and by Durbin and Koopman (2002) for the linear Gaussian state space model.

### 3.6 Empirical Illustration

I will now illustrate the results presented in this chapter, in the context of maximum likelihood estimation of a stochastic volatility. Consider a univariate time series of daily asset log-returns \(y_t\). The log-variance process, \(h_t\), is modeled by a stationary autoregressive process

\[ h_{t+1} = \mu(1 - \phi) + \phi h_t + \sigma \zeta_t, \quad \zeta_t \sim N(0, 1), \quad (3.28) \]

for \(t = 1, \ldots, n - 1\), with \(h_1 \sim N\{0, \sigma^2(1 - \phi^2)^{-1}\}\) and with \(\zeta_1, \ldots, \zeta_{n-1}\) mutually independent and independent of \(h_1\). The observations \(y_t\) are modeled as

\[ y_t = \exp\left(\frac{1}{2}h_t\right) \varepsilon_t, \quad \varepsilon_t \sim N(0, 1), \]

for \(t = 1, \ldots, n\), where \(\varepsilon_1, \ldots, \varepsilon_n\) are mutually independent. Note that the stochastic volatility model has a nonlinear observation equation because of the multiplicative term \(\exp(\frac{1}{2}h_t)\varepsilon_t\). Finally, I assume that the state and observation disturbances are correlated as follows

\[ \text{Cov}(\varepsilon_t, \zeta_t) = \rho, \]
for \( t = 1, \ldots, n - 1 \). This correlation accounts for the leverage effect, the phenomenon that volatility will increase as the asset price decreases. The correlation \( \rho \) is therefore typically negative, see Black (1976), Nelson (1991) and Yu (2005), among others. The parameter vector \( \psi \) consists of \( \phi, \sigma, \zeta, \mu \) and \( \rho \). For more information on the stochastic volatility model and its extensions the reader is referred to the selected articles in Shephard (2005).

The most natural way to write the SV model with leverage effect in state space form is to set the signal vector \( \theta_t \) equal to \((h_t, \zeta_t)\) for \( t = 1, \ldots, n \). Note however that in this case the log-variances \( h_2, \ldots, h_n \) are linear functions of \( h_1 \) and the disturbances \( \eta_1, \ldots, \eta_{n-1} \). We can therefore not use a Newton-Raphson scheme to maximize \( p(\theta|y; \psi) \) with respect to \( \theta \). I therefore formulate the model in state space form as follows

\[
y_t = \exp\left(\frac{1}{2} h_t\right) \left\{ \varepsilon_t^* + \text{sign}(\rho)\zeta_{2t} \right\}, \quad \varepsilon_t^* \sim N(0, 1 - |\rho|),
\]

for \( t = 1, \ldots, n \) where

\[
h_{t+1} = \mu(1 - \phi) + \phi h_t + \sigma \zeta_t (\zeta_{1t} + \zeta_{2t}), \quad \zeta_{1t} \sim N(0, 1 - |\rho|),
\]

with \( h_1 \sim N\left\{0, \sigma^2(1 - \phi^2)^{-1}\right\} \) and \( \zeta_{2t} \sim N(0, |\rho|) \) for \( t = 1, \ldots, n \). The disturbances \( \varepsilon_t^*, \zeta_{1t} \) and \( \zeta_{2t} \) are assumed mutually and serially independent and independent of the initial state \( h_1 \). The signal vector is now given by \( \theta_t = (h_t, \zeta_{2t})' \). Note that the variance matrix \( \Psi \) is now nonsingular and we can find the optimum of \( p(\theta|y; \psi) \) as discussed in Section 3.4.

Here, I estimate the parameter vectors \( \psi \) for two time series of daily returns. These two series have originally been analyzed by Yu (2005). The first series contains 2022 daily returns of the Standard & Poor’s 500 index from January 1980 to December 1987 while the second one contains 2529 daily returns of the Center for Research in Security Prices index from January 1986 to December 1995. Yu (2005) estimated \( \psi \) as part of a Bayesian analysis using Markov chain Monte Carlo methods. His estimation results are reproduced in Table 1, which reports the mean of the posterior distribution \( p(\theta, \psi|y) \) and the associated 95% Bayes credible intervals. I complement the analysis of Yu by reporting the maximum likelihood estimates of \( \psi \) as well as the associated confidence intervals in Table 3.1. The maximum likelihood estimates are obtained by maximizing the likelihood \( \hat{p}(y; \psi) \) using a quasi-Newton method while the 95% confidence intervals are calculated from the information matrix. The maximum likelihood likelihood and the Markov chain Monte Carlo estimates with their confidence and credible intervals are very similar. This confirms that importance sampling methods are a viable option for the analysis of the a stochastic volatility with leverage model based on time series of this sample size.
3.6. EMPIRICAL ILLUSTRATION

Table 3.1: Estimation results

Estimates of $\psi = (\phi, \sigma_\zeta, \mu, \rho)'$ for the stochastic volatility with leverage model using the Standard and Poor’s 500 index series and Center for Research in Security Prices index series from Yu (2005). The maximum likelihood estimates are obtained by maximizing $\log \hat{p}(y, \psi)$ using a quasi-Newton method. The 95% confidence intervals, reported between brackets below the estimates, are based on the information matrix. The Markov chain Monte Carlo estimates and their 95% Bayes credible intervals are obtained from Yu (2005). The acronyms S&P500 and CRSP refer to the Standard and Poor’s 500 index series and the Center for Research in Security Prices index series, respectively. The simulated maximum likelihood and the Markov chain Monte Carlo estimation methods are referred to as SML and MCMC in Table 1, respectively.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>S&amp;P500</th>
<th>CRSP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SML</td>
<td>MCMC</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.9685</td>
<td>0.9720</td>
</tr>
<tr>
<td></td>
<td>(0.947,0.990)</td>
<td>(0.951,0.987)</td>
</tr>
<tr>
<td>$\sigma_\zeta$</td>
<td>0.1578</td>
<td>0.1495</td>
</tr>
<tr>
<td></td>
<td>(0.111,0.205)</td>
<td>(0.114,0.193)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$-0.0830$</td>
<td>$-0.0688$</td>
</tr>
<tr>
<td></td>
<td>($-0.310,0.144$)</td>
<td>($-0.308,0.201$)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$-0.3298$</td>
<td>$-0.3179$</td>
</tr>
<tr>
<td></td>
<td>($-0.498,-0.161$)</td>
<td>($-0.475,-0.143$)</td>
</tr>
</tbody>
</table>
3. A Proof of Theorem 3.2

To proof Theorem 3.2 we need the following two lemmas.

**Lemma 3.1.** Let $\Sigma = \Psi + H$, where $\Psi$ is defined in Section 2.3 and for a set of symmetric invertible matrices $H_1, \ldots, H_n$ let $H = \text{diag}\{H_1, \ldots, H_n\}$. Assume that a decomposition for the symmetric matrix $\Sigma = LDL'$ exists, where $L$ is a lower block-unit-triangular matrix and $D$ is a block diagonal matrix. Further let

$$L' H^{-1} L - D^{-1} = E' CE.$$  

where $E$ is a lower block-unit-triangular matrix and $C$ is a block-diagonal matrix. The block elements of $C$ and $E$ can then be evaluated as follows

$$C_i = H_i^{-1} - F_i^{-1} - K_i' \tilde{N}_i K_i, \quad E_{ij} = R_i T_{i-1} \ldots T_{j+1} K_j,$$  

(3.29)

$$R_i = C_i^{-1}(A_i^{-1} Z_i - K_i' N_i T_i), \quad \tilde{N}_i = R_i C_i R_i - Z_i' H_i^{-1} Z_i + T_i' \tilde{N}_i T_i,$$  

(3.30)

for $i, j = n, \ldots, 1$ and $j < i$ and where $N_n = 0$.

**Proof.** From Proposition 2.1 we have

$$L_{i,i-1} = Z_i K_{i-1}, \quad L_{ij} = Z_i T_{i-1} \ldots T_{j+1} K_j,$$  

for $i = 2, \ldots, n$ and $j = 1, \ldots, i - 2$. Given the block structures of the matrices $E$ and $C$, the $(k,l)$ block of $E' CE$ is given by

$$ (E' CE)_{kl} = \begin{cases} 
  C_k + \sum_{m=k+1}^n E_{mk}' C_m E_{mk}, & k = l, \\
  C_k E_{kl} + \sum_{m=k+1}^n E_{mk}' C_m E_{ml}, & k > l, \\
  (E' CE)'_{lk}, & k < l 
\end{cases}$$  

(3.31)

From Proposition 2.1 it follows that the block $(k,l)$ of matrix $L' H^{-1} L - D^{-1}$ is given by

$$ (L' H^{-1} L - D^{-1})_{kl} = \begin{cases} 
  H_k^{-1} - F_k^{-1} + K_k' N_k^{[a]} K_k, & k = l, \\
  (H_k^{-1} Z_k + K_k' N_k^{[a]} T_k) T_{k-1} \ldots T_{j+1} K_j, & k > l, \\
  (D' A^{-1} D - F^{-1})_{lk}, & k < l, 
\end{cases}$$  

(3.32)

where

$$N_k^{[a]} = \sum_{m=k+1}^n T_{m+1}' \ldots T_m' H_m^{-1} Z_m T_{m-1} \ldots T_{k+1}.$$
3.A. PROOF OF THEOREM 3.2

can be evaluated by the backwards recursion

\[ N_{k-1}^{[a]} = Z_i H_k^{-1} Z_k + T_k N_k^{[a]} T_k, \]

for \( k = n, \ldots, 1 \) and with \( N_n^{[a]} = 0 \). The proof now proceeds by an induction argument. It is easily checked that (3.29) is valid for \( i = n \). Suppose (3.29) and (3.30) hold for \( i = m+1, \ldots, n \) and \( j < i \). We need to verify (3.29) and (3.30) for \( i = m \) and \( j < m \). If we substitute the equation (3.29) for \( E_{ij} \) into (3.31) we have

\[ \sum_{k=i+1}^{n} E'_{ki} C_k E_{ki} = K'_{i} N_{i}^{[c]} K_{i}, \]

\[ \sum_{k=i+1}^{n} E'_{ki} C_k E_{kj} = K'_{i} N_{i}^{[c]} T_{i-1} \cdots T_{j+1} K_{j}, \]

(3.33)

for \( i = m, \ldots, n \) and \( j > i \), where

\[ N_{i}^{[c]} = \sum_{k=i+1}^{n} T_{i+1} \cdots T_{k-1} R_{k} C_k R_{k} T_{k-1} \cdots T_{i+1}, \]

can be evaluated by the backwards recursion

\[ N_{i-1}^{[c]} = R_{i} C_{i} R_{i} + T_{i} N_{i}^{[c]} T_{i}, \]

for \( i = n, \ldots, 1 \) and with \( N_{n}^{[c]} = 0 \). By substituting the first equation of (3.33) into the first row of (3.31) and by equating (3.31) and (3.32) for \( k = l \), first equation in (3.29) for \( i = m \) follows from the fact that

\[ \tilde{N}_i = N_i^{[c]} - N_i^{[a]}, \]

for \( i = n, \ldots, 1 \). Similarly, the second equation in 3.29, for \( i = m \) and \( j < m \), is established as follows

\[ E_{mj} = C_m^{-1} \left[ (H_m^{-1} Z_m + K'_m N_m^{[a]} T_m) T_{m-1} \cdots T_{j+1} K_j - \sum_{k=m+1}^{n} E'_{km} C_k E_{kj} \right] \]

\[ = C_m^{-1} \left[ H_m^{-1} Z_m - K'_m (N_m^{[c]} - N_m^{[a]}) T_m \right] T_{m-1} \cdots T_{j+1} K_j \]

\[ = R_m T_{m-1} \cdots T_{j+1} K_j. \]

\[ \square \]

**Lemma 3.2.** Let \( y, \Sigma, \Psi \) and \( H \) be defined as in Lemma 3.1. If \( u = (u'_1, \ldots, u'_n)' \) where
CHAPTER 3. NONLINEAR NON-GAUSSIAN STATE SPACE MODELS

$u_1, \ldots, u_n$ are obtained from the recursions of Theorem 3.2 then

$$u \sim N \left[ H\Sigma^{-1}(y - \mu), H - H\Sigma^{-1}H \right],$$

where $\mu$ is defined in equation (3.5).

**Proof.** Let $L$ and $D$ as in Proposition 2.1 and

$$E'CE = L'H^{-1}L - D^{-1},$$

where $E$ is a lower block-unit-triangular matrix and $C$ is a block-diagonal matrix. The proof proceeds by showing that

$$u = HL'\Sigma^{-1}(u^+ + D^{-1}v), \quad (3.34)$$

where $u^+ = E'w$ for $w = (w'_1, \ldots, w'_n)'$. It follows that

$$u^+ \sim N(0, L'H^{-1}L - D^{-1}).$$

since $w \sim N(0, C)$ and thus

$$\text{Var}(HL'\Sigma^{-1}u^+) = H - H\Sigma^{-1}H,$$

and

$$HL'\Sigma^{-1}D^{-1}v = HL'\Sigma^{-1}D^{-1}L^{-1}(y - \mu) = H\Sigma^{-1}(y - \mu),$$

from Proposition 2.1. If $u^+ = (u^+_1, \ldots, u^+_n)'$ then we have from Lemma 3.1

$$u^+_t = w_t + \sum_{i=t+1}^{n} E'_{it}w_i = w_t + \sum_{i=t+1}^{n} K'_{t}T'_{i+1} \cdots T'_{i-1}R'_{i}w_i,$$

It follows that $u^+_t$ can be evaluated by the backwards recursion

$$u^+_t = w_t + K'_{t}r^{'[c]}_{t}, \quad r^{'[c]}_{t-1} = R'_{t}w_t + T'_{t}r^{'[c]}_{t},$$

with $r^{'[c]}_{n} = 0$ and $t = n, \ldots, 1$. Define $u^x = L'\Sigma^{-1}(u^+_t + D^{-1}v)$. It follows, see Section 2.6,

$$u^x_t = F^{-1}_t v_t + u^+_t - K'_{t}r^{'[a]}_{t}, \quad r^{'[a]}_{t-1} = Z'_{t}u^x_t + T'_{t}r^{'[a]}_{t},$$
where \( u^x = (u^x_1, \ldots, u^x_n)' \) with \( r_n^{[a]} = 0 \) and thus
\[
\begin{align*}
u_t^x &= F_t^{-1}v_t + w_t - K'_t(r_t^{[a]} - r_t^{[c]}), &r_t^{[a]} - r_t^{[c]} &= Z_t'u_t^x - R'_tw_t + T'_t(r_t^{[a]} - r_t^{[c]}).
\end{align*}
\]

Defining \( \tilde{r}_t = r_t^{[a]} - r_t^{[c]} \) and noting that for \( t = 1, \ldots, n \) we have \( u_t = H_tu_t^x \) and expression (3.34) follows.

Having proved Lemma 3.2, the proof of Theorem 3.2 is now relatively easy.

**Proof of Theorem 3.2.** Clearly, \( y - u \) is a Gaussian variable, it therefore suffices to show that it has the correct mean and variance. We have
\[
\mathbb{E}(y - u) = y - H\Sigma^{-1}(y - \mu) = \mu + \Psi\Sigma^{-1}(y - \mu),
\]
and
\[
\text{Var}(y - u) = H - H\Sigma^{-1}H = \Psi - \Psi\Sigma^{-1}\Psi,
\]
both following from Lemma 3.2 and \( H = \Sigma - \Psi \). This concludes the proof of Theorem 3.2.

### 3.B Equivalence of Theorem 3.2 to Simulation Smoother of de Jong and Shephard (1995)

**Proposition 3.1.** The recursions for \( \tilde{r}_t \) and \( \tilde{N}_t \) of Theorem 3.2 are equivalent to the following recursions given in de Jong and Shephard (1995)
\[
\begin{align*}
\tilde{r}_{t-1} &= Z_tF_t^{-1}v_t - W_tC_t^{-1}u_t + L_t\tilde{r}_t, &N_{t-1} &= Z_tF_t^{-1}Z_t + W_tC_t^{-1}W_t + L_t\tilde{N}_tL_t, \\
W_t &= F_t^{-1}Z_t - K'_t\tilde{N}_tL_t, &L_t &= T_t - K_tZ_t,
\end{align*}
\]
for \( t = n, n - 1, \ldots, 1 \) and where \( F_t, v_t, C_t, u_t \) and \( K_t \) are the same as in Theorem 3.2 and where \( \tilde{r}_n = 0 \) and \( \tilde{N}_n = 0 \).

**Proof.** Define \( S_t = F_t^{-1} + K'_tN_tK_t \) for \( t = 1, \ldots, n \). From the definition of \( C_t \) it follows that
We have
\[ R_t = C_t^{-1}(H_t^{-1}Z_t - K_t'N_t T_t) \]
\[ = C_t^{-1}\{(C_t + S_t)Z_t - K_t'N_t T_t\} \]
\[ = C_t^{-1}\{(C_t + F_t^{-1})Z_t - K_t'N_t L_t\} \]
\[ = C_t^{-1}(C_t Z_t + W_t), \]
for \( t = 1, \ldots, n \). It follows that
\[ W_t = C_t (R_t - Z_t) \text{ and } (C_t + S_t)Z_t - C_t R_t - K_t'N_t T_t = 0. \]
The equivalence of the recursion for \( \tilde{N}_t \) is now shown by
\[ N_{t-1} = R_t'\tilde{C}_t R_t - Z_t' H_t^{-1} Z_t + T_t' N_t T_t \]
\[ = R_t'\tilde{C}_t R_t - Z_t' (C_t + S_t) Z_t + T_t' N_t T_t \]
\[ + Z_t' \{(C_t + S_t) Z_t - C_t R_t - K_t' N_t T_t\} + \{Z_t' (C_t + S_t) - R_t'\tilde{C}_t - T_t' N_t K_t\} Z_t \]
\[ = (R_t - Z_t)' C_t (R_t - Z_t) + Z_t' F_t^{-1} Z_t + L_t' N_t L_t \]
\[ = W_t' C_t^{-1} W_t + Z_t' F_t^{-1} Z_t + L_t' N_t L_t. \]
The equivalence of the recursion for \( r_t \) follows from
\[ r_{t-1} = Z_t' H_t^{-1} w_t - R_t' w_t + T_t' r_t \]
\[ = Z_t' F_t^{-1} w_t + (Z_t - R_t)' w_t - Z_t' K_t' r_t + T_t' r_t \]
\[ = Z_t' F_t^{-1} w_t - W_t' C_t^{-1} w_t + L_t' r_t. \]

\[ \square \]

3.C Derivation of Equation (3.24)

The logarithm of the importance function \( f(\theta^i; y) \) is given by
\[ \log f(\theta^i; y) = -\frac{mn}{2} \log 2\pi - \frac{1}{2} \log |V| - \frac{1}{2} (\theta^i - \hat{\theta})' V^{-1} (\theta^i - \hat{\theta}), \]
where \( V = (A^{-1} + \Psi^{-1})^{-1} \). Writing \( \Sigma = LDL' \) and \( L'A^{-1}L - D^{-1} = E'CE \) as in respectively Lemmas 2.1 and 3.1 we have that
\[ V = AL'^{-1}E'CEL^{-1}A, \]
where matrices $A$ and $D$ are block-diagonal and matrices $L$ and $E$ are lower block-unit-triangular matrices. From expression (3.34) in the proof of Lemma 3.2 we have

$$w^i = E'^{-1}A^{-1}L(y - \theta^i) - D^{-1}v,$$

where $w = (w^i_1, \ldots, w^i_n)'$. Since $E'^{-1}A^{-1}L$ is a lower triangular matrix with $A^{-1}_1, \ldots, A^{-1}_n$ on the diagonal it follows that

$$|E'^{-1}A^{-1}L| = \prod_{i=1}^n \frac{1}{|A_i|}.$$ 

The result follows from

$$f(\theta^i; y; \psi) = |J|p(w^i),$$

where $p(w^i)$ is the density of the Gaussian variables $w^i_1, \ldots, w^i_n$ and $J = |E'^{-1}H^{-1}L|$.

### 3.D Derivation of Equation (3.25)

Recall that

$$\log f(\theta^i; y) = -\frac{mn}{2} \log 2\pi - \frac{1}{2} \log |V| - \frac{1}{2} (\theta^i - \hat{\theta})'V^{-1}(\theta^i - \hat{\theta}),$$

where $V = (A^{-1} + \Psi^{-1})^{-1}$. With some straightforward, but tedious algebra we can show that

$$(\theta^i - \hat{\theta})'V^{-1}(\theta^i - \hat{\theta}) = (\theta^i - \mu)'\Sigma^{-1}(\theta^i - \mu) - (x - \mu)'\Sigma^{-1}(x - \mu) + (x - \hat{\theta})'A^{-1}(x - \hat{\theta}),$$

the result follows on noting that since $|V|$ is positive, we have $\log |V| = \frac{1}{2} \log |V|^2$ and

$$\frac{1}{2} \log |V|^2 = \frac{1}{2} \log |H - H\Sigma^{-1}H|^2 = \frac{1}{2} \log |H|^2 + \frac{1}{2} \log |I - \Sigma^{-1}H|^2$$

$$= \frac{1}{2} \log |H|^2 + \frac{1}{2} \log |\Sigma^{-1}\Psi|^2$$

$$= \frac{1}{2} \log |H|^2 - \frac{1}{2} \log |\Sigma|^2 + \log |\Psi|. $$
Writing $\Sigma = LDL'$ where $D$ is block-diagonal and $L$ is lower block-unit-triangular we have

$$(x - \mu)'\Sigma^{-1}(x - \mu) = (x - \mu)'L' \, D^{-1} \, L^{-1} \, (x - \mu) = v'D^{-1}v = \sum_{t=1}^{n} v_t'F_t^{-1}v_t,$$

and

$$|\Sigma| = |D| = \prod_{i=1}^{n} |F_i|,$$

both direct consequences of Proposition 2.1.
Chapter 4

Likelihood-based Analysis for Dynamic Factor Models

4.1 Introduction

In this chapter I consider the dynamic factor model given by

\[ y_{it} = \mu_i + x_{it}\beta + \sum_{j=0}^{q_A} X'_{ij}f_{t-j} + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, n, \]  

(4.1)

where \( y_{it} \) denotes the observed value for the \( i \)th time series at time \( t \), \( \mu_i \) is a fixed and unknown constant, \( x_{it} \) is a \( 1 \times K \) vector of covariates, \( \beta \) is a \( K \times 1 \) vector of regression coefficients, \( f_t \) is an \( r \times 1 \) vector of common factors, \( \lambda_{ij} \) is an \( r \times 1 \) vector of loadings associated with the common factors at lag \( j \) and \( u_{it} \) is the idiosyncratic component. The factors are modeled by linear dynamic processes and the idiosyncratic components by autoregressive processes with mutually correlated zero mean innovations. We particularly focus on the case where a high-dimensional panel of \( N \) time series depends on a relatively small number of \( r \) common dynamic factors.

We will see in Section 4.2 that the dynamic factor models presented here can be written in state space form. Minimum mean square linear estimators (MMSLEs) of the common factors \( f_t \) can therefore be obtained using the Kalman smoother recursions of Chapter 2. Also, when the idiosyncratic components \( u_{it} \) and the common factors \( f_t \) are assumed Gaussian we can evaluate the likelihood function efficiently by means of the Kalman filter. In case the innovations are non-Gaussian, the Gaussian likelihood can be regarded as a quasi-likelihood.

The Gaussian likelihood function can be numerically maximized to obtain maximum likelihood or quasi-maximum likelihood (QML) parameter estimates. This is the approach
taken by Engle and Watson (1981) for a Gaussian model with one common factor. Watson and Engle (1983) use the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) to find the optimum of the likelihood, see also Shumway and Stoffer (1982) and Quah and Sargent (1993). However, in many of the recent applications of the dynamic factor model, the high-dimensional panel of time series and the resulting large number of parameters make such an approach infeasible.

In this chapter we will see some new results that lead to computationally efficient methods for a likelihood-based analysis of high-dimensional dynamic factor models. Both signal extraction as well as likelihood evaluation will be covered. Finding the optimum of a likelihood function is not straightforward if there is a large number of parameters. New devices for an effective implementation of the optimization methods are also presented.

For models including regression effects it was argued in Section 2.11 that the marginal likelihood, or equivalently the diffuse likelihood, is to be preferred over the standard likelihood. However, if the number of regression coefficients is very large the state augmentation approach for calculating the marginal likelihood presented in Section 2.11 is no longer feasible. Efficient evaluation of the marginal likelihood for dynamic factor models is discussed in Section 4.4.5.

The key insight behind the results of this chapter is that the observed time series can be split into a low-dimensional vector series and a high-dimensional vector series. For the estimation of the factors and the evaluation of the likelihood function, we need to apply the computationally intensive Kalman filter methods only to the low-dimensional series while simple regression-style calculations suffice for the high-dimensional part. This results in large computational gains.

The likelihood-based approach has a number of advantages over alternative approaches such as the principal components method. Since the factors are explicitly modeled and the estimation method takes account of the model specification, the factors can represent aspects of economic theory. Hypothesis tests can be formulated and tested. The techniques of this paper allow real-time estimation of the underlying factors, estimation of past factors as well as prediction of factors and future observations. The Kalman filter further produces mean squared errors of the factor estimates without an extra computational effort. Unbalanced data-sets are also easily handled in this framework. Finally, in case the data generating process can be represented as a Gaussian dynamic factor model, the parameter estimators are asymptotically efficient. Moreover, Doz, Giannone, and Reichlin (2006) show, under mild conditions, that the factor estimates from the QML procedure are consistent for the true factors when $T \to \infty$ and $N \to \infty$, even if the model is misspecified. They also
present Monte Carlo evidence that the QML factor estimates are often more precise than the principal component estimates.

### 4.2 Generalized Dynamic Factor Model with Covariates

Suppose that the dynamic characteristics of a time series of observed $N \times 1$ vectors $y_1, \ldots, y_n$ can be described by the dynamic factor model (4.1). The vector form of model (4.1) is given by

$$y_t = \bar{\mu} + \bar{X}_t \beta + \Lambda(L) f_t + u_t, \quad t = 1, \ldots, T, \tag{4.2}$$

where $y_t = (y_{1t}, \ldots, y_{Nt})'$, $u_t = (u_{1t}, \ldots, u_{Nt})'$, $\bar{\mu} = (\mu_1, \ldots, \mu_N)'$, $\bar{X}_t = (x_{1t}', \ldots, x_{Nt}')'$ and matrix lag polynomial $\Lambda(L) = \Lambda_0 + \sum_{j=1}^{q_\Lambda} \Lambda_j L^j$ with $\Lambda_j = (\lambda_{1j}, \ldots, \lambda_{Nj})'$ for $j = 0, \ldots, q_{\Lambda}$, lag-operator $L$ and non-negative integer $q_{\Lambda}$. The vector $f_t$ of common factors is modeled by the vector autoregressive moving average (VARMA) process

$$\Phi(L) f_t = \Theta(L) \eta_t, \tag{4.3}$$

where $\eta_t$ is a vector of innovations and the matrix lag polynomials are $\Phi(L) = I - \sum_{j=1}^{q_\Phi} \Phi_j L^j$ and $\Theta(L) = I + \sum_{j=1}^{q_{\Theta}} \Theta_j L^j$ with $r \times r$ autoregressive coefficient matrices $\Phi_j$ for $j = 1, \ldots, q_\Phi$ and $r \times r$ moving average coefficient matrices $\Theta_j$ for $j = 1, \ldots, q_{\Theta}$. The idiosyncratic component vector $u_t$ in (4.2) is modeled as the vector autoregressive (VAR) process

$$\Psi(L) u_t = \varepsilon_t, \tag{4.4}$$

where $\varepsilon_t$ is a vector of innovations and the matrix lag polynomial is $\Psi(L) = I - \sum_{j=1}^{q_{\Psi}} \Psi_j L^j$ with $N \times N$ autoregressive coefficient matrix $\Psi_j$ for $j = 1, \ldots, q_{\Psi}$. Finally, we denote the set of all parameters in the model by $\psi$. The set of parameters in $\psi$ excluding $\bar{\mu}$ and $\beta$, are denoted by $\theta$, that is

$$\psi = (\bar{\mu}', \beta', \theta').$$

We adopt the following set of assumptions for model (4.2) – (4.4):

(i) The permissible parameter space $S_\psi$ is a compact sub-set of the Euclidean space. The true parameter $\psi_0$ is an interior point of $S_\psi$.

(ii) For all $\psi$ in $S_\psi$ and $|z| \leq 1$, we have $|\Phi(z)| \neq 0$ and $|\Psi(z)| \neq 0$. 

(iii) Denote by $\mathcal{F}_t$ the $\sigma$-algebra generated by $y_1, \ldots, y_t$, with $\mathcal{F}_0$ the trivial $\sigma$-algebra, then

$$
\mathbb{E}(\varepsilon_t | \mathcal{F}_{t-1}) = 0, \quad \mathbb{E}(\eta_t | \mathcal{F}_{t-1}) = 0, \quad \mathbb{E}(\varepsilon_t \varepsilon'_t | \mathcal{F}_{t-1}) = \Sigma_{\varepsilon}, \quad \mathbb{E}(\eta_t \eta'_t | \mathcal{F}_{t-1}) = \Sigma_{\eta},
$$

for $t = 1, \ldots, n$. We assume that $\Sigma_{\varepsilon}$ is a nonsingular matrix.

(iv) The vector sequences $\{\varepsilon_t\}$ and $\{\eta_t\}$ are uncorrelated and have finite fourth moments.

(v) The covariate sequence $\{\bar{X}_t\}$ is independent of the innovation sequences $\{\varepsilon_t\}$ and $\{\eta_t\}$.

(vi) Matrix $\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \bar{X}_n \bar{X}'_{n+j}$ exists and is finite for every non-negative integer $j$.

(vii) Let $\Gamma_y(h; \psi) = \text{Cov}_\psi(y_t, y_{t+h})$ for $\psi \in S_\psi$, then for any $\psi^* \in S_\psi$ such that $\psi^* \neq \psi_0$, $\Gamma_y(s; \psi) \neq \Gamma_y(s; \psi_0)$ for at least one value of $s \in \mathbb{Z}_+$.

(viii) Denote $\bar{\mu}_0$ and $\beta_0$ as the true values of $\bar{\mu}$ and $\beta$, respectively. The process $y_t - \bar{\mu}_0 - \bar{X}_t \beta_0$ can be written as a VAR process $\Pi(L; \psi_0)(y_t - \bar{\mu}_0 - \bar{X}_t \beta_0) = \tilde{u}_t$, where $\Pi(z; \psi_0) = I - \sum_{i=1}^{\infty} \Pi_i(\psi_0)z^i$, $\mathbb{E}(\tilde{u}_t | \mathcal{F}_{t-1}) = 0$ and the elements of $\Pi_1, \Pi_2, \ldots$ are absolutely summable.

Assumption (ii) implies that the dynamic factor model is stationary for all admissible parameter vectors. The assumption in (iii) of $\Sigma_{\varepsilon}$ nonsingular is not restrictive since any dynamic factor model with a singular matrix $\Sigma_{\varepsilon}$ can be rewritten to satisfy assumption (iii).

Assumption (vii) is an identifiability assumption. In practice, for this assumption to hold, we need to put restrictions on $\Lambda_0, \Lambda_1, \ldots, \Lambda_q, \Phi_1, \ldots, \Phi_q, \Theta_1, \ldots, \Theta_q$, and $\Sigma_{\zeta}$. Parameter restrictions are common in the literature on factor models, see e.g. Geweke and Zhou (1996) for further discussions. Examples of the general model specification (4.2) – (4.4) are given in Illustrations 1 and 2 below.

Denoting $F_t = (f'_t, f'_{t-1}, \ldots, f'_{t-s})'$, the dynamic factor model (4.2) with idiosyncratic component (4.4) can be expressed in static form as follows

$$
y_t = \mu + d_t + X_t \beta + \Lambda F_t + \varepsilon_t,
$$

for $t = q_\Psi + 1, \ldots, n$, where

$$
\mu = \Psi(I)\bar{\mu}, \quad d_t = \sum_{j=1}^{q_\Psi} \Psi_j y_{t-j}, \quad X_t = \Psi(L)\bar{X}_t, \quad \Lambda = \begin{pmatrix}
\Lambda_0 \\
\Lambda_1^+ - \Psi_1^+ \Lambda_0^+
\vdots
\Lambda_s^+ - \sum_{j=1}^{s} \Psi_j^+ \Lambda_{s-j}^+
\end{pmatrix},
$$
with \( s = q_\Lambda + q_\Psi \) and

\[
\Lambda^+_i = \begin{cases} 
\Lambda_i, & i \leq q_\Lambda, \\
0, & \text{else,}
\end{cases} \quad \Psi^+_i = \begin{cases} 
\Psi_i, & i \leq q_\Psi, \\
0, & \text{else,}
\end{cases}
\]

for \( i, k = 1, \ldots, s \). The number of static factors in \( F_t \) is given by \( m = r (s + 1) \).

Every VARMA process can be written as a state space model without measurement noise, see e.g. Harvey (1989) and Durbin and Koopman (2001). It follows that the static factors can be written as as \( F_t = G \alpha_t \) for a suitable matrix \( G \) and

\[
\alpha_t = T \alpha_{t-1} + R \eta_t, \quad (4.6)
\]

for a \( p \times 1 \) state vector \( \alpha_t \). The system matrices \( T \) and \( R \) are sparse matrices and contain the coefficient matrices in the polynomials \( \Phi(L) \) and \( \Theta(L) \). Matrices \( T \) and \( R \) can be constructed such that \( G \) consists of rows of the unity matrix and has full row rank. The dimension of the state \( \alpha_t \) is generally higher than the dimension of \( F_t \) when the latent VARMA process (4.3) with non-zero orders \( q_\Phi \) and \( q_\Theta \) is specified in state space form. Model (4.5) can be expressed in terms of the state vector \( \alpha_t \) via the observation equation

\[
y_t = \mu + d_t + X_t \beta + Z \alpha_t + \varepsilon_t, \quad (4.7)
\]

for \( t = q_\Psi + 1, \ldots, n \), with

\[
Z = \Lambda G. \quad (4.8)
\]

To handle the initial stretch of observations \( y_1, \ldots, y_{q_\Psi} \) explicitly, we need to consider the observation equation (4.7) with different system matrices \( Z \) and \( \Sigma_\varepsilon \) for \( t = 1, \ldots, q_\Psi \). An example is given in Illustration 1 below. In the remainder of this chapter I assume for convenience that all system matrices are time-invariant. However, all results hold for time-varying system matrices subject to some minor modifications.

**Illustration 1.** Consider the dynamic factor model

\[
y_t = \Lambda_0 f_t + u_t, \quad (4.9)
\]

for \( t = 1, \ldots, n \) with \( N \times r \) factor loading matrix \( \Lambda_0 \) and where the \( r \times 1 \) vector \( f_t \) follows a VAR(1) process, that is equation (4.3) with \( q_\Phi = 1 \) and \( q_\Theta = 0 \). Furthermore, the idiosyncratic components \( u_t \) are modeled as independent AR(1) processes, that is equation (4.4) with \( q_\Phi = 1 \) and both \( \Psi_1 \) and \( \Sigma_\varepsilon \) diagonal. To ensure that all parameters are identified \( \Lambda_0 \) is
assumed to be of the form \( \Lambda_0 = (\bar{\Lambda}_1', \bar{\Lambda}_2')' \) where \( \bar{\Lambda}_1 \) is an \( r \times r \) lower triangular matrix and \( \bar{\Lambda}_2 \) is an \((N - r) \times r\) full matrix. The diagonal elements of \( \bar{\Lambda}_1 \) are set to one. Additionally, the variance matrix \( \Sigma_\zeta \) is restricted to be diagonal.

The state vector \( \alpha_t \) is specified as \( \alpha_t = F_t = (f_t', f_{t-1}')' \) so that \( G = I_{2p} \) in (4.8). The matrices \( T \) and \( R \) in (4.6) are given by

\[
T = \begin{bmatrix} \Phi & 0 \\ I_r & 0 \end{bmatrix}, \quad R = \begin{bmatrix} I_r \\ 0 \end{bmatrix}.
\]

Further, we have \( \mathbb{E}(\alpha_1) = 0 \) and \( \text{Var}(\alpha_1) \) is set to the unconditional variance of the stationary vector series \((f_t', f_{t-1}')'\). The observation equation (4.7) for \( t = 2, \ldots, n \) has

\[
\mu = 0, \quad d_t = \Psi_1 y_{t-1}, \quad Z = \Lambda = (\Lambda_0, -\Psi_1 \Lambda_0),
\]

and \( \beta = 0 \). For \( t = 1 \) we have

\[
Z = (\Lambda_0, 0), \quad \text{Var}(\varepsilon) = (I_N - \Psi_1^2)^{-1} \Sigma_\varepsilon,
\]

and \( d_1 = 0 \), since \( \mathbb{E}(u_1) = 0 \) and \( \mathbb{E}(u_1 u_1') = (I_N - \Psi_1^2)^{-1} \Sigma_\varepsilon \).

**Illustration 2.** Suppose \( y_t \) is modeled by the observation equation (4.9), but with independent idiosyncratic components, that is \( q_\Psi = 0 \) and \( u_t = \varepsilon_t \) in (4.4). Furthermore, \( f_t \) is modeled by the VARMA\((1,1)\) process defined as (4.3) with \( r = q_\Phi = q_\Theta = 1 \). We have

\[
f_t = \Phi f_{t-1} + \zeta_t + \Theta \zeta_{t-1}, \quad \mathbb{E}(\eta_t | \mathcal{F}_{t-1}) = 0, \quad \mathbb{E}(\eta_t^2 | \mathcal{F}_{t-1}) = \Sigma_\eta,
\]

where \( \Phi, \Theta \) and \( \Sigma_\zeta \) are unknown matrices. Identifiability of parameters is guaranteed by restricting \( \Lambda_0 \) and \( \Sigma_\eta \) as in Illustration 1.

If case \( r = 1 \) then \( f_t \) is an ARMA\((1,1)\) as described in the example of Section 2.2. Setting \( \Phi = \phi, \Theta = \theta \) and \( \Sigma_\eta = \sigma_\eta^2 \) we can write \( \alpha_t = (f_t, \theta_\xi_t)' \) with \( T \) and \( R \) are given (2.6). It follows that \( Z = (\Lambda_0, 0) \).

### 4.3 Parameter Estimation

The state space model (4.6) and (4.7) can be written in the form

\[
y = d + X(\theta)\gamma + \xi, \quad (4.10)
\]
where $y = (y_1', \ldots, y_n')'$, $d = (d_1', \ldots, d_n')'$, $\gamma = (\mu', \beta')'$, $\tilde{X}(\theta)$ is a $Nn \times (K + N)$ matrix valued function of $\theta$, $E(\xi) = 0$ and $E(\xi'\xi') = \Sigma(\theta)$ for a matrix valued function $\Sigma(\theta)$. The Gaussian log-likelihood function is defined as follows

$$\ell(y; \psi) = c - \frac{1}{2} \log |\Sigma(\theta)| - \frac{1}{2} \{y - \tilde{X}(\theta)\gamma - d\}'\Sigma(\theta)^{-1}\{y - \tilde{X}(\theta)\gamma - d\}, \quad (4.11)$$

where $c$ is a constant independent of $\gamma$ and $\theta$. In case the disturbances $\varepsilon_t$ and $\zeta_t$ in the dynamic factor model (4.2) – (4.4) are Gaussian, equation (4.11) is the exact log-likelihood function. In other cases, the log-likelihood function is generally intractable. If the Gaussian assumption does not apply, the likelihood is designated as a quasi-likelihood. Quasi-maximum likelihood (QML) estimators of the parameters are obtained by maximizing (4.11) with respect to $\psi$. These QML estimators are strongly consistent as $n \to \infty$ under the assumptions of Section 4.2. Additionally, the QML estimators are asymptotically Gaussian, see Hannan, Dunsmuir, and Deistler (1980) for details and proofs.

Small sample sizes are not uncommon in macroeconomic applications of the dynamic factor model. For example, Quah and Sargent (1993) analyze a data-set where $n$ is as small as 42. Following the discussion of Section 2.11.3 inference based on the marginal likelihood is preferred in these cases. Since the marginal likelihood does not depend on $\gamma$, we need to estimate these parameters separately, for example by generalized least squares (GLS). The GLS estimator and its variance are given by

$$\hat{\gamma}(\theta) = \{\tilde{X}'\Sigma^{-1}\tilde{X}\}^{-1}\tilde{X}'\Sigma^{-1}y, \quad \text{Var}_{\theta}(\hat{\gamma}(\theta)) = \{\tilde{X}'\Sigma^{-1}\tilde{X}\}^{-1}, \quad (4.12)$$

where the dependence of $\tilde{X}$ and $\Sigma$ on $\theta$ is suppressed. Suppose $\hat{\theta}$ denotes the estimator of $\theta$ obtained by maximizing $\ell_d(y; \theta)$. We can then obtain an estimator of $\gamma$ by substituting $\hat{\theta}$ in (4.12). In case the QML estimators of $\gamma$ and $\theta$, $\bar{\gamma}$ and $\bar{\theta}$, respectively, are obtained by maximizing (4.11), we have $\hat{\gamma}(\bar{\theta}) = \bar{\gamma}$.

In principle we can evaluate the Gaussian likelihood (4.11) using the Kalman filter and the marginal likelihood and GLS estimators using the diffuse Kalman filter as explained in respectively Sections 2.8 and 2.11.3. However, if the dimension $N$ of $y_t$ is very large, the Kalman filter methods are computationally infeasible, even when the dimension $p$ of the state vector $\alpha_t$ is modest. The recursion can be made computationally more efficient by processing the elements of $y_t$ individually rather than the whole vector at once as explained in Section 2.9. This modification leads to substantial computational gains, but they are not sufficient for the dimensions common in recent applications of dynamic factor models.

In the next section we present computationally efficient algorithms for evaluating the
Gaussian log-likelihood function (4.11), the marginal Gaussian log-likelihood function and the GLS estimator (4.12) for any value of $\theta$.

### 4.4 Estimation of Factors and Likelihood Evaluation

In this section I present new results that allow for the computationally efficient evaluation of the likelihood functions and GLS estimator of Section 4.3. Furthermore, I show how we can efficiently obtain the state estimates and mean squared errors given by

$$a_{t|s} = P(\alpha_t|y_1, \ldots, y_s; \psi), \quad Q_{t|s} = \mathbb{E} \left[ (\alpha_t - a_{t|s})(\alpha_t - a_{t|s})' | y_1, \ldots, y_s; \psi \right],$$  \hspace{1cm} (4.13)

for $s, t = 1, \ldots, n$, where $P(\alpha_t|y_1, \ldots, y_s; \psi)$ denotes the minimum mean squared error linear estimator (MMSLE) of $\alpha_t$ based on $y_1, \ldots, y_s$ for given $\psi$. Note that generally $a_{t|s}$ and $Q_{t|s}$ will depend on the parameter vector $\psi$, this is however suppressed in the notation.

#### 4.4.1 Transforming the Observation Equation

Consider the state space model (4.6) and (4.7) for a given parameter vector $\psi$. Define $y_t^+ = Ay_t$, for $t = 1, \ldots, n$, for some non-singular matrix $A$. The MMSLEs of $\alpha_1, \ldots, \alpha_n$ in (4.13) are not affected if $y_1, \ldots, y_n$ is replaced with $y_1^+, \ldots, y_n^+$. Furthermore, the log-likelihood functions of $y_1, \ldots, y_n$ and $y_1^+, \ldots, y_n^+$ differ only by the Jacobian term $\log |A|^n$. We will show that for certain choices of $A$, factor estimates and likelihood functions can be computed more efficiently based on $y_1^+, \ldots, y_n^+$ rather than $y_1, \ldots, y_n$.

Suppose we partition $N \times N$ matrix $A$ and $N \times 1$ vector $y_t^* = A(y_t - \mu - d_t - X_t\beta)$ as

$$A = \begin{bmatrix} A^L \\ A^H \end{bmatrix}, \quad y_t^* = \begin{pmatrix} y_t^L \\ y_t^H \end{pmatrix},$$  \hspace{1cm} (4.14)

where

$$y_t^L = A^L(y_t - \mu - d_t - X_t\beta), \quad y_t^H = A^H(y_t - \mu - d_t - X_t\beta),$$

with $m \times N$ matrix $A^L$ and $(N-m) \times N$ matrix $A^H$. The observation vectors $y_t^L$ and $y_t^H$ have dimensions $m \times 1$ and $(N-m) \times 1$, respectively. We aim to choose $A$ such that $y_t^L$ and $y_t^H$ are uncorrelated and only $y_t^L$ depends on $\alpha_t$. More specifically, the model for $y_t^*$ will be of the form

$$y_t^L = A^LZ\alpha_t + e_t^L, \quad y_t^H = e_t^H,$$  \hspace{1cm} (4.15)
where
\[ E(e_t^L | \mathcal{F}_{t-1}) = 0, \quad E(e_t^H | \mathcal{F}_{t-1}) = 0, \]
and
\[ E(e_t^L e_t^{L'} | \mathcal{F}_{t-1}) = \Sigma_L, \quad E(e_t^H e_t^{H'} | \mathcal{F}_{t-1}) = \Sigma_H, \quad E(e_t^H e_t^{L'} | \mathcal{F}_{t-1}) = 0, \]
for \( t = 1, \ldots, n \), with \( \Sigma_L = A^L \Sigma_e A^L' \) and \( \Sigma_H = A^H \Sigma_e A^H' \). A suitable matrix \( A \) needs to fulfill the following conditions:

(i) \( A \) is full rank,

(ii) \( A^H \Sigma_e A^L' = 0 \),

(iii) Row\{\( A^H \}\} = Col\{\( Z \}\) \( \perp \),

where Col\{\( X \}\} and Row\{\( X \)\} denote the row and column spaces of a matrix \( X \), respectively, and the superscript \( \perp \) denotes the orthogonal complement. Condition (i) prevents any loss of information due to the transformation \( A_y \). Condition (ii) ensures that \( e_t^L \) and \( e_t^H \) in (4.15) are uncorrelated and condition (iii) implies that the second equation does not depend on \( \alpha_t \). Condition (iii) is stronger than strictly necessary. The transformed model will still be of the form (4.15) if condition (iii) is replaced with \( A^H Z = 0 \). In its current form however, condition (iii) ensures that the reduction in dimension is as large as possible, in the sense that the dimension of \( y_t^H \) cannot be enlarged without compromising the special form of (4.15).

Finally, we add the condition

(iv) \( |\Sigma_H| = 1 \).

Condition (iv) is not restrictive but it simplifies various calculations. For example, we can express the determinant of \( A \) in terms of \( A^L \) and \( \Sigma_e \) since

\[ |A|^2 = |\Sigma_e|^{-1} |A \Sigma_e A'| = |\Sigma_e|^{-1} |A^L \Sigma_e A^L'| |A^H \Sigma_e A^H'| = |\Sigma_e|^{-1} |\Sigma_L|. \tag{4.16} \]

The conditions (i)–(iii) imply a closed form for \( A^L \), which is given in the following lemma.

**Lemma 4.1.** Consider model (4.6) – (4.7). Suppose matrix \( A \) is of the form (4.14) and \( A^H \) satisfies (iii), then \( A \) satisfies (i)–(iii) if and only if

\[ A^L = \Lambda^\dagger \Sigma_e^{-1}, \tag{4.17} \]

where the columns of the \( N \times r_\Lambda \) matrix \( \Lambda^\dagger \) form a basis for the column space of \( \Lambda \).

**Remarks**
(a) The columns of $\Lambda^\dagger$ also form a basis of the column space of $Z$, which follows from the fact that $Z = \Lambda G$, for a full row rank matrix $G$. It is therefore easily verified that any matrix $A$ with $A^L$ given by (4.17) and $A^H$ satisfying (iii), fulfills conditions (i)–(iii). I prove the necessity part of Lemma 4.1 in Appendix 4.A.

(b) Since column rank deficiency of $\Lambda$ is rare in practice, we can generally choose

$$\Lambda^\dagger = \Lambda C,$$

(4.18)

for any $r_A \times r_A$ nonsingular matrix $C$. In case $\Lambda$ does not have full column rank, it is often straightforward to construct a suitable $\Lambda^\dagger$. An example of such a situation is Illustration 1 if $\Psi_1 = \varphi I_N$ with scalar $-1 < \varphi < 1$.

(c) A closed form expression for $A^H$ is generally not available. For $A^H$ to satisfy (iii), we need to choose $A^H$ such that its rows form a basis for the null space of $\Lambda^{\dagger\prime}$. Condition (iv) can then be satisfied by rescaling the rows. Finding a basis for the null space of a matrix requires computationally intensive numerical methods. Fortunately, we will see that matrix $A^H$ is not required for any of the computations.

(d) The results below are based on transformation (4.14) and model (4.15). Although our results are more general and are developed for different purposes, a similar transformation as (4.14) for a different class of factor models is considered by (Fiorentini, Sentana, and Shephard 2004, section 2.4.1).

Illustration 3. Consider the dynamic factor model $y_t = \Lambda f_t + \varepsilon_t$ of Illustration 2. Apply transformation (4.14) to $y_t$ where matrix $A^L$ is given by (4.17) and (4.18) with $C = (\Lambda'\Sigma^{-1}_\varepsilon \Lambda)^{-1}$. For this choice of $C$, vector $y^L_t$ is effectively the GLS estimator of $f_t$ in the “regression model” $y_t = \Lambda f_t + \varepsilon_t$, for each $t$. We have

$$y^L_t = (\Lambda'\Sigma^{-1}_\varepsilon \Lambda)^{-1} \Lambda'\Sigma^{-1}_\varepsilon y_t, \quad t = 1, \ldots, T.$$

In case $r = r_A = 1$, model (4.15) for the univariate time series $y^L_t$ is then given by

$$y^L_t = G\alpha_t + \varepsilon^L_t, \quad \mathbb{E}(\varepsilon^L_t\varepsilon^L_t'|F_{t-1}) = C,$$

for $t = 1, \ldots, n$ where vector $G = (1, 0)$. 
4.4.2 Estimation of factors

By considering a matrix $A$ that satisfies the conditions (i)–(iv) in Section 4.4.1, we are able to efficiently compute MMSLEs of the factors. Since matrix $A$ has full rank, we have $P(\alpha_t|y_1, \ldots, y_s; \psi) = P(\alpha_t|y_1^*, \ldots, y_s^*; \psi)$. Furthermore, from (4.15) it follows that $y_t^L$ and $y_t^H$ are uncorrelated and that $y_t^H$ does not depend on $\alpha_t$. Hence,

$$a_{ts} = P(\alpha_t|y_1^*, \ldots, y_s^*; \psi) = P(\alpha_t|y_1^L, \ldots, y_s^L; \psi),$$

for $s, t = 1, \ldots, T$. The MMSLEs of the states can therefore be obtained by applying the Kalman filter recursions of Chapter 2 to the low-dimensional model

$$y_t^L = A^L Z \alpha_t + e_t^L,$$

where

$$\mathbb{E}(e_t^L | \mathcal{F}_{t-1}) = 0, \quad \mathbb{E}(e_t^L e_t^{L'} | \mathcal{F}_{t-1}) = \Sigma_L,$$

for $t = 1, \ldots, n$. The high-dimensional matrix $A^H$ and vector $y_t^H$ are not required for the estimation of $\alpha_t$. In case of Illustration 3 with $r = 1$, the estimation of $\alpha_t$ is simply carried out by univariate Kalman filter recursions. The low-dimensional Kalman filter recursions also produces the correct mean squared error matrices $Q_{ts}$ in (4.13) for $s, t = 1, \ldots, n$. The Kalman filter and smoother recursions provide solutions for prediction ($s = t - 1$), filtering ($s = t$), smoothing ($s = n$) and forecasting ($t > s$) of observation and state vectors.

The procedures of this section can still be used if observed vectors $y_t$ do not all have the same dimension due to, for example, missing values. In this case, a different matrix $A$ must be constructed for $t = 1, \ldots, n$. This solution is also adopted in cases where the system matrices of the state space form (4.6) – (4.7) vary over time.

4.4.3 Evaluation of the Gaussian Log-likelihood

We saw in Section 2.8 that the Gaussian log-likelihood $\ell(y; \psi)$ defined in (4.11) can be evaluated via the prediction error decomposition,

$$\ell(y; \psi) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} \log |F_t| - \frac{1}{2} \sum_{t=1}^{n} v_t' F_t^{-1} v_t,$$

where the quantities $v_t$ and $F_t$ are obtained from the Kalman filter of Section 2.5 applied to the state space model (4.6) and (4.7). A computationally more efficient way to evaluate (4.20) is to choose a matrix $A$ that satisfies the conditions (i)–(iv) in Section 4.4.1, to transform $y_t$...
as in (4.14) and to consider model (4.15). We then have
\[
\ell(y; \psi) = \ell(y^L; \psi) + \ell(y^H; \psi) + n \log |A|,  \tag{4.21}
\]
where \(y^L = (y^L_1, \ldots, y^L_n)'\) and \(y^H = (y^H_1, \ldots, y^H_n)'\). The first term \(\ell(y^L; \psi)\) can be evaluated by the Kalman filter applied to the low-dimensional model (4.19). The second term is given by
\[
\ell(y^H; \psi) = -\frac{(N-m)n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} y_t^H \Sigma_H^{-1} y_t^H,  \tag{4.22}
\]
since \(|\Sigma_H| = 1\). Lemma 4.2 shows that the last term in equation (4.22) can be calculated without constructing \(A^H\). The proof is given in Appendix 4.B.

**Lemma 4.2.** For the state space model (4.6) – (4.7), transformation (4.14) and resulting model (4.15), with \(A^L\) given by (4.17), we have the identity
\[
y_t^H \Sigma_H^{-1} y_t^H = \epsilon_t^t \Sigma_e^{-1} \epsilon_t,  \tag{4.23}
\]
where \(\epsilon_t = \left[I_N - \Lambda^t (\Lambda^t \Sigma_e^{-1} \Lambda^t)^{-1} \Lambda^t \Sigma_e^{-1}\right] (y_t - d_t - \mu - X_t \beta)\).

Given the expression for \(|A|^2\) in (4.16), log-likelihood function (4.21) can be expressed as
\[
\ell(y; \psi) = c + \ell(y^L; \psi) - \frac{n}{2} \log \frac{|\Sigma_e|}{|\Sigma_L|} - \frac{1}{2} \sum_{t=1}^{n} \epsilon_t^t \Sigma_e^{-1} \epsilon_t,  \tag{4.24}
\]
where \(c\) is a constant independent of both \(y\) and \(\psi\). It follows that for the evaluation of the log-likelihood, computation of matrix \(A^H\) and vectors \(y_t^H\), for \(t = 1, \ldots, n\), is not required. Expression (4.24) is instrumental for a computationally feasible approach to the quasi-likelihood based analysis of the dynamic factor model.

**Remarks**

(a) The vectors \(\epsilon_t\) in Lemma 4.2 have an intuitive interpretation as the residuals of a GLS regression of \(y_t - X_t \beta - \mu - d_t\) on the columns of \(\Lambda^\dagger\) with variance matrix \(\Sigma_e\). Since the columns of \(\Lambda^\dagger\) also form a basis of the column space of \(Z\), this is equivalent to regressing \(y_t - X_t \beta - \mu - d_t\) on the columns of \(Z\).

(b) The concluding remarks of Section 4.4.2 concerning missing values and time-varying specifications of the dynamic factor model apply to the evaluation of the log-likelihood via (4.21) as well.
4.4. ESTIMATION OF FACTORS AND LIKELIHOOD EVALUATION

Illustration 4. In Illustration 3, the transformation (4.14) is based on the matrix $A^L$ defined in (4.17) and (4.18) with $C = (\Lambda^\dagger \Sigma_e^{-1} \Lambda^\dagger)^{-1}$. However, it can be more convenient to choose $C$ such that $C'C = (\Lambda^\dagger \Sigma_e^{-1} \Lambda^\dagger)^{-1}$ with $C$ upper-triangular. For this choice, the variance matrix $\Sigma_L$ in (4.19) is the identity matrix and the loading matrix in (4.19) is $A^L \Lambda^\dagger = C^{-1}$. We obtain the model

$$y_t^L = C^{-1}G\alpha_t + e_t^L, \quad E(e_t^L|\mathcal{F}_{t-1}) = 0, \quad E(e_t^L e_t^L'|\mathcal{F}_{t-1}) = I, \quad t = 1, \ldots, n.$$ 

We can now transform the model to a univariate state space model as explained in Section 2.9. Note that this does not require additional computations since $\Sigma_L = I$. Furthermore, the log-likelihood function (4.24) reduces to

$$\ell(y; \psi) = c + \ell(y^L; \psi) - \frac{T}{2} \log |\Sigma_e| - \frac{1}{2} \sum_{t=1}^T e_t' \Sigma_e^{-1} e_t.$$ 

The computations for $|\Sigma_e|$ and $\Sigma_e^{-1}$ can exploit special structures in $\Sigma_e$ such as the matrix being diagonal or having Toeplitz, spatial or block structures.

4.4.4 Partial Concentration of Regression Coefficients

Maximizing the Gaussian log-likelihood function $\ell(y; \psi)$ is computationally intensive, since the dimension of $\psi$ is generally very high. It is therefore attractive to concentrate the regression coefficients out of the likelihood and maximize the resulting profile likelihood function. In this section I show how the constant vector $\mu$ can be partially concentrated out of the likelihood with minimum effort.

Choose matrix $A$ such that conditions (i)–(iii) in Section 4.4.1 are satisfied and define for $t = 1, \ldots, n$

$$\tilde{y}_t^L = A^L (y_t - d_t - X_t\beta), \quad \tilde{y}_t^H = A^H (y_t - d_t - X_t\beta), \quad (4.25)$$

such that $y_t^L = \tilde{y}_t^L - \mu^L$ and $y_t^H = \tilde{y}_t^H - \mu^H$ where $\mu^L = A^L \mu$ and $\mu^H = A^H \mu$. In the likelihood function (4.21), $\ell(y^L; \psi)$ does not depend on $\mu^H$ while $\mu^H$ only appears in the second term of $\ell(y^H; \psi)$ which can be expressed as

$$-\frac{1}{2} \sum_{t=1}^n (\tilde{y}_t^H - \mu^H)' \Sigma_{H^{-1}}(\tilde{y}_t^H - \mu^H) = -\frac{1}{2} \sum_{t=1}^n (\tilde{e}_t - M_A \mu)' \Sigma_{e^{-1}}(\tilde{e}_t - M_A \mu), \quad (4.26)$$

where $M_A = I_N - \Lambda^\dagger (\Lambda^\dagger \Sigma_e^{-1} \Lambda^\dagger)^{-1} \Lambda^\dagger \Sigma_e^{-1}$ and $\tilde{e}_t = M_A (y_t - d_t - X_t\beta)$ such that $e_t = \tilde{e}_t - M_A \mu$. 


The equality in (4.26) is justified by Lemma 4.2. It follows from equation (4.42) in Appendix 4.B that $M_A\mu$ is a linear function of $\mu^H$. Concentrating out $\mu^H$ is therefore equivalent to concentrating out $M_A\mu$ from the likelihood function. The GLS estimator of $M_A\mu$, denoted by $\hat{\mu}_{\perp A}(\beta, \theta)$, is given by
\[
\hat{\mu}_{\perp A}(\beta, \theta) = \frac{1}{n} \sum_{t=1}^{n} \tilde{e}_t. \tag{4.27}
\]
The (partial) profile log-likelihood function is given by (4.24) where the last term is replaced by $-0.5 \sum_{t=1}^{n} \tilde{e}_t^{m'} \Sigma_{\epsilon}^{-1} \tilde{e}_t^{m}$ where $\tilde{e}_t^{m} = \tilde{e}_t - \hat{\mu}_{\perp A}(\beta, \theta)$ for $t = 1, \ldots, n$.

The QML estimator of $\mu$ can be obtained via the identity
\[
\mu = P_A^L \mu^L + M_A \mu, \quad \text{where} \quad P_A = \Lambda^\dagger (\Lambda^\dagger \Sigma_{\epsilon}^{-1} \Lambda^\dagger)^{-1}. \tag{4.28}
\]
The QML estimator of $\mu$ is then given by
\[
\hat{\mu} = \hat{\mu}_{\perp A}(\tilde{\beta}, \tilde{\theta}) + P_A \hat{\mu}^L, \tag{4.29}
\]
where $\tilde{\theta}, \tilde{\beta}$ and $\hat{\mu}^L$ are the QML estimators of $\theta$, $\beta$ and $\mu^L$, respectively, which are obtained by maximizing the profile Gaussian log-likelihood function with respect to $\theta$, $\beta$ and $\mu^L$.

4.4.5 Evaluation of the Marginal Gaussian Log-likelihood

In Section 2.11 we saw that the diffuse Kalman filter can be used to evaluate the marginal likelihood for state space models with regression effects. The diffuse Kalman filter also produces GLS estimators of the regression effects. The number of time series in a dynamic factor model can be high and direct application of the diffuse Kalman filter is often infeasible. We can use the earlier results to compute the marginal likelihood based on a much smaller dimension. As a result, inference based on the marginal likelihood becomes feasible for a high-dimensional dynamic factor model (4.6) – (4.7). Let $A$ denote a matrix satisfying conditions (i)–(iv) in Section 4.4.1. Pre-multiplying the observations by $A$, we obtain the model
\[
\bar{y}_t^L = \mu^L + X_t^L \beta + A^L Z \alpha_t + e_t^L, \\
\bar{y}_t^H = \mu^H + X_t^H \beta + e_t^H,
\]
where $\bar{y}_t^L = A^L (y_t - d_t)$, $\bar{y}_t^H = A^H (y_t - d_t)$, $X_t^L = A^L X_t$ and $X_t^H = A^H X_t$ with $\mu^L$ and $\mu^H$ defined below (4.25) and the disturbances $e_t^L$ and $e_t^H$ defined below (4.15). In the remainder of this section I show that the evaluation of the marginal likelihood can be carried out in
4.4. ESTIMATION OF FACTORS AND LIKELIHOOD EVALUATION

two steps: first processing the original time series $y_t$ and second applying the diffuse Kalman filter to the time series $\tilde{y}^L_t$.

Since $\tilde{y}^L = (\tilde{y}^L_1, \ldots, \tilde{y}^L_n)'$ and $\tilde{y}^H = (\tilde{y}^H_1, \ldots, \tilde{y}^H_n)'$ both depend on coefficient vector $\beta$, the marginal Gaussian likelihood function cannot be easily expressed in two independent parts. Denote the marginal log-likelihood for a given value of $\theta$ by $\ell_m(y; \theta)$. I show in Appendix 4.C that for any given parameter vector $\theta$

$$
\ell_m(y; \theta) = c + L_m(\tilde{y}^H; \theta) + L_m(\tilde{y}^L; \theta) - \frac{n - 1}{2} \log \frac{|\Sigma_e|}{|\Sigma_L|},
$$

(4.30)

where $c$ is a constant independent of $\theta$ and $L_m(\tilde{y}^H; \theta)$ and $L_m(\tilde{y}^L; \theta)$ are obtained by the following two-step algorithm.

**Step 1.** Define

$$
b = B^{-1} \sum_{t=1}^{n} \tilde{X}_t^m \Sigma_e^{-1} \tilde{e}_t^m, \quad B = \sum_{t=1}^{n} \tilde{X}_t^m \Sigma_e^{-1} \tilde{X}_t^m,
$$

(4.31)

where

$$
\tilde{e}_t^m = M_\Lambda(y_t - d_t - \bar{y}), \quad \tilde{X}_t^m = M_\Lambda(X_t - \bar{X}),
$$

(4.32)

for $t = 1, \ldots, n$, with $\bar{y} = n^{-1} \sum_{t=1}^{n} (y_t - d_t)$ and $\bar{X} = n^{-1} \sum_{t=1}^{n} X_t$. Then, compute

$$
L_m(\tilde{y}^H) = -\frac{1}{2} \log |B| - \frac{1}{2} \sum_{t=1}^{n} \tilde{e}_t^* \Sigma_e^{-1} \tilde{e}_t^*,
$$

(4.33)

where $\tilde{e}_t^* = \tilde{e}_t^m - \tilde{X}_t^m b$.

**Step 2.** Set $L_m(\tilde{y}^L; \theta)$ equal to the Gaussian diffuse log-likelihood for the model

$$
\tilde{y}^L_t = \mu^L + X_t^L \tilde{\beta} + A^L Z \alpha_t + e_t^L,
$$

(4.34)

where $\mu^L$ is treated as an unknown regression coefficient vector and $\tilde{\beta}$ is a random effect with mean $b$ and variance $B^{-1}$. The evaluation of $L_d(\tilde{y}^L; \theta)$ is carried out by the diffuse Kalman filter.

4.4.6 Estimation of Regression Coefficients

For the dynamic factor model (4.6) – (4.7) define $\hat{\mu}(\theta)$ and $\hat{\beta}(\theta)$ as the GLS estimators of $\mu$ and $\beta$ as functions of $\theta$. Note that $\hat{\gamma}(\theta) = [\hat{\mu}(\theta)', \hat{\beta}(\theta)'']'$ where the GLS estimator $\hat{\gamma}(\theta)$ is given by (4.12). The GLS estimators are based on the data-set $y_1, \ldots, y_n$. The two-step algorithm of the previous section is also instrumental for computing $\hat{\mu}(\theta)$ and $\hat{\beta}(\theta)$ in a
computationally efficient way. The application of the diffuse Kalman filter in the second step of the algorithm produces the MMSLEs of $\hat{\beta}$ and $\mu^L$, that is $P(\hat{\beta}|\bar{y}_1^L, \ldots, \bar{y}_n^L; \theta)$ and $P(\mu^L|\bar{y}_1^L, \ldots, \bar{y}_n^L; \theta)$, respectively, as well as their mean squared errors. In Appendix 4.D I prove that for given

$$\hat{\beta}(\theta) = P(\tilde{\beta}|\bar{y}_1^L, \ldots, \bar{y}_n^L; \theta), \quad \hat{\mu}^L(\theta) = P(\mu^L|\bar{y}_1^L, \ldots, \bar{y}_n^L; \theta),$$

(4.35)

where $\hat{\mu}^L(\theta)$ is defined as the GLS estimator of $\mu^L$ based on $y_1, \ldots, y_n$ as a function of $\theta$. The GLS estimator $\hat{\mu}(\theta)$ follows from (4.28) and is given by

$$\hat{\mu}(\theta) = P_\Lambda \hat{\mu}^L(\theta) + \hat{\mu}_{\bot \Lambda}(\theta),$$

(4.36)

where matrix $P_\Lambda$ is defined in (4.28) and $\hat{\mu}_{\bot \Lambda}(\theta)$ denotes $\hat{\mu}_{\bot \Lambda}(\beta, \theta)$, as given by (4.27), evaluated in $\beta = \hat{\beta}$. The variance matrix of $\hat{\mu}(\theta)$ is given by

$$\text{Var}_\theta(\hat{\mu}(\theta)) = \text{Var}_\theta\{\hat{\mu}_{\bot \Lambda}(\theta) + P_\Lambda \hat{\mu}^L(\theta)\}$$

$$= \text{Var}_\theta\{\hat{\mu}_{\bot \Lambda}(\theta)\} + P_\Lambda \text{Cov}_\theta\{\hat{\mu}_{\bot \Lambda}(\theta), \hat{\mu}^L(\theta)\}' + \text{Cov}_\theta\{\hat{\mu}_{\bot \Lambda}(\theta), \hat{\mu}^L(\theta)\} P_\Lambda'$$

$$+ P_\Lambda \text{Var}_\theta\{\hat{\mu}^L(\theta)\} P_\Lambda',$$

where the dependence of variances and covariances on $\theta$ is made explicit in the notation $\text{Var}_\theta(\cdot)$ and $\text{Cov}_\theta(\cdot, \cdot)$. To evaluate $\text{Var}_\theta\{\hat{\mu}(\theta)\}$, we require the expressions

$$\text{Var}_\theta\{\hat{\mu}_{\bot \Lambda}(\theta)\} = M_\Lambda \left[ \bar{X} \text{Var}_\theta\{\hat{\beta}(\theta)\} \bar{X}' + \frac{1}{n} \Sigma_e \right] M_\Lambda',$$

and

$$\text{Cov}_\theta\{\hat{\mu}_{\bot \Lambda}(\theta), \hat{\mu}^L(\theta)\} = -M_\Lambda \bar{X} \text{Cov}_\theta\{\hat{\beta}(\theta), \hat{\mu}^L(\theta)\},$$

where $M_\Lambda$ and $\bar{X}$ are defined below (4.26) and (4.32), respectively. The variance matrices $\text{Var}_\theta\{\hat{\beta}(\theta)\}$ and $\text{Var}_\theta\{\hat{\mu}^L(\theta)\}$ are equal to the mean squared error matrices of $\hat{\beta}(\theta)$ and $\hat{\mu}^L(\theta)$ in (4.35), respectively. These two variance matrices, together with the covariance matrix $\text{Cov}_\theta\{\hat{\beta}(\theta), \hat{\mu}^L(\theta)\}$, are evaluated by the diffuse Kalman filter from Step 2 of the algorithm in Section 4.4.5. Derivations and more details are given in Appendix 4.D. Estimators of $\mu$ and $\beta$ can be obtained by substituting the QML estimator of $\theta$, found by maximizing the Gaussian marginal likelihood, in (4.35) and (4.36).
4.4.7 Computational Gains

The main purpose of the results of the previous sections is to obtain computationally efficient inference procedures for the class of dynamic factor models discussed in Section 4.2. In this section I present evidence of the gains in computing times that are achieved by the algorithms of this chapter. The gains are relative to the standard application of the Kalman filter based on \( y_1, \ldots, y_n \).

The computational gains depend primarily on the panel dimension \( N \) and state vector dimension \( p \). To get some insight in the size of these gains, I calculate the Gaussian log-likelihood and marginal Gaussian log-likelihood functions for different values of \( N \) and \( p \). The calculations are performed using the Kalman filter and the methods described in Sections 4.4.3 and 4.4.5 for the basic factor model given by

\[
y_{it} = \mu_i + \lambda_i f_t + \varepsilon_{it},
\]

where \( f_t \) is modeled by the VARMA process (4.3) with \( q_\Phi = 1 \) and \( q_\Theta = 0 \) while the innovations \( \varepsilon_{it} \) are uncorrelated. For the different model representations in Section 4.2, we have \( \alpha_t = F_t = f_t \) and \( p = m = r \). In the first panel of Table 4.1 we see the ratios of CPU times needed for the evaluation of the two log-likelihood functions. The results are encouraging. If \( N = 250 \) and \( p = 5 \), the log-likelihood is evaluated 15 times faster using the methods of this chapter than when using the standard Kalman filter. Furthermore, the computational savings are substantial for moderate values of \( N \) and relatively small values of \( p \), say, 5 or 10. If \( p \) is relatively large, say, 25, the gains are less dramatic but still substantial by any means.

Even more computational gains can be achieved if we evaluate the marginal log-likelihood using the method of Section 4.4.5. The reported ratios in the second panel of Table 4.1 are so high because the Kalman filter based on \( y_1, \ldots, y_n \) requires an \( N \) dimensional augmentation for the constant vector \( \mu \). The Kalman filter used in Section 4.4.5 and based on the observation equation (4.34) requires a limited \( p \) dimensional augmentation for the constant vector \( \mu^L \).

4.5 Maximizing the Quasi-Likelihood Function

With the results of Sections 4.4.3 and 4.4.5 in hand we can evaluate the Gaussian log-likelihood and the marginal Gaussian log-likelihood functions efficiently even for high-dimensional dynamic factor models. Numerical optimization procedures, such as the quasi-Newton BFGS
CHAPTER 4. ANALYSIS OF DYNAMIC FACTOR MODELS

Table 4.1: Computational Gains

The two panels below present the gains in computing time when evaluating the Gaussian likelihood respectively the marginal likelihood functions of a basic dynamic factor model. The model considered is $y_t = \mu_i + \lambda_i f_t + \epsilon_t$, where $f_t$ is modeled as a VAR(1) process, $\epsilon_t \sim \text{IID}(0, \sigma^2)$, for some positive scalar $\sigma$ and $\mu_i$ is a scalar. The ratio $d_1/d_2$ is reported: $d_1$ is the CPU time for the standard Kalman filter respectively augmented Kalman filter and $d_2$ is CPU time for the algorithms of Sections 4.4.3 and 4.4.5. The ratios are reported for different panel dimensions $N$ and different state vector dimensions $p$.

<table>
<thead>
<tr>
<th>$N \backslash p$</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.0</td>
<td>1.3</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>10.4</td>
<td>2.3</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>50</td>
<td>5.7</td>
<td>4.7</td>
<td>3.1</td>
<td>1.5</td>
<td>–</td>
<td>50.6</td>
<td>40.0</td>
<td>18.0</td>
<td>3.4</td>
<td>–</td>
</tr>
<tr>
<td>100</td>
<td>6.7</td>
<td>7.5</td>
<td>5.6</td>
<td>2.5</td>
<td>1.5</td>
<td>55.0</td>
<td>62.0</td>
<td>47.2</td>
<td>13.5</td>
<td>3.2</td>
</tr>
<tr>
<td>250</td>
<td>8.7</td>
<td>14.8</td>
<td>12.4</td>
<td>5.5</td>
<td>3.0</td>
<td>79.0</td>
<td>82.2</td>
<td>82.9</td>
<td>63.6</td>
<td>22.6</td>
</tr>
<tr>
<td>500</td>
<td>12.5</td>
<td>15.9</td>
<td>21.2</td>
<td>10.2</td>
<td>5.4</td>
<td>107.5</td>
<td>108.9</td>
<td>109.5</td>
<td>108.7</td>
<td>69.7</td>
</tr>
</tbody>
</table>

The algorithm described in Nocedal and Wright (1999), can be adopted to maximize the log-likelihood function with respect to $\psi$ or $\theta$. These methods require evaluation of the score vector however. Since the number of parameters can be as high as 2,000, evaluating the score vector numerically is infeasible, even if the results of Section 4.4.3 are used. Fortunately, it can be shown that the exact score vector can be obtained by a single Kalman smoother or diffuse Kalman smoother applied to the low-dimensional model (4.19) or (4.34). Alternatively, the EM algorithm can be used to obtain the QML estimates. In Section 4.5.2 we will see that each EM step requires only a single application of the (diffuse) Kalman smoother.

4.5.1 Calculating the Analytical Score

Koopman and Shephard (1992) develop analytical expressions for the score function of the parameters in a state space model. They adopt the results in Louis (1982) and Ruud (1991) and in particular the identity

$$
\frac{\partial \ell(y; \psi)}{\partial \psi} \bigg|_{\psi = \psi^*} = \frac{Q(\psi^*|\psi)}{\partial \psi} \bigg|_{\psi = \psi^*},
$$

(4.37)

where $Q(\psi^*|\psi)$ is the expected complete Gaussian log-likelihood function, given by

$$
Q(\psi^*|\psi) = \mathbb{E} \left[ \log p(y, \alpha; \psi) | y; \psi^* \right],
$$

...
and \( p(y, \alpha; \psi) \) is the joint density of \( y \) and \( \alpha_1, \ldots, \alpha_n \). For the state space model (4.6) – (4.7), with Gaussian innovations \( \varepsilon_t \) and \( \eta_t \), \( Q(\psi^*|\psi) \) is given by

\[
Q(\psi^*|\psi) = c - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{tr} Q_\varepsilon - \frac{n-1}{2} \log |\Sigma_\eta| - \frac{1}{2} \text{tr} Q_\eta \\
- \frac{1}{2} \log |P| - \frac{1}{2} \text{tr} [P^{-1} (a_{1|n} - a) (a_{1|n} - a)' + Q_{1|n}],
\]

where \( a = \mathbb{E}(\alpha_1) \), \( P = \mathbb{E}[(\alpha_1 - a)(\alpha_1 - a)'] \) and \( c \) is a constant independent of \( \psi \) and \( \varepsilon_t = \mathbb{E}(\varepsilon_t|y) \), \( \text{Var}(\varepsilon_t|y) \), \( \hat{\eta}_t = \mathbb{E}(\eta_t|y) \) and \( \text{Var}(\eta_t|y) \) can be expressed in terms of \( a_{jn} \) and \( Q_{jn} \) for \( n = 1, \ldots, T \), which can be evaluated using the Kalman smoother recursion discussed in Section 2.6. Since the estimation of factors can be based on the low-dimensional model (4.19) while matrix \( A^H \) and time series \( y_t^H \) are not needed, these computations can be performed efficiently. Expressions for the derivatives of (4.38) with respect to the system matrices, evaluated at \( \psi = \psi^* \), are given in Appendix 4.E. The score vector with respect to \( \psi \) is then obtained using the chain rule. The score vector of the marginal Gaussian log-
likelihood function \( \ell_m(y; \theta) \) with respect to \( \theta \) and evaluated at \( \theta = \theta^* \) can be obtained in the same way with the difference that \( Q(\theta^*|\theta) \) is obtained using the diffuse Kalman filter as described in Section 4.4.5, see Durbin and Koopman (2001, section 7.3).

### 4.5.2 The EM Algorithm

The well-known EM algorithm, introduced by Dempster, Laird, and Rubin (1977), is an iterative algorithm that repeatedly performs two types of calculations: (E)xpectation and (M)aximization. For a given value of \( \psi = \psi^* \), the E and M steps are given by

- **E step:** determine the expected complete log-likelihood function \( Q(\psi^*|\psi) \) in (4.38).
- **M step:** maximize \( Q(\psi^*|\psi) \) with respect to \( \psi \).

The M step produces a vector \( \psi^+ \) with the property \( \ell(y; \psi^+) \geq \ell(y; \psi^*) \). If the EM steps are continuously repeated, convergence to a (local) optimum of \( \ell(y; \psi) \) is guaranteed, see Wu (1983) for a more detailed discussion. Shumway and Stoffer (1982) and Watson and Engle (1983) have proposed the use of the EM algorithm in the context of state space models. A feasible EM algorithm for high-dimensional dynamic factor models is obtained by applying the methods of Sections 4.4.1 and 4.4.2 in the E step. The details of the EM algorithm are
CHAPTER 4. ANALYSIS OF DYNAMIC FACTOR MODELS

Table 4.2: List of sectors

This table lists the 15 sectors in the data-set that we consider in Section 4.6. For each sector, the code, a short description and the number of series in the sector are given. More detailed descriptions of the 132 time series can be found in Appendix A of Stock and Watson (2005).

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Number of Time Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Real Output and Income</td>
<td>17</td>
</tr>
<tr>
<td>B</td>
<td>Employment and Hours</td>
<td>30</td>
</tr>
<tr>
<td>C</td>
<td>Real Retail</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>Manufacturing and Trade Sales</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>Consumption</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>Housing Starts and Sales</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>Real Inventories</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>Orders</td>
<td>7</td>
</tr>
<tr>
<td>I</td>
<td>Stock Prices</td>
<td>4</td>
</tr>
<tr>
<td>J</td>
<td>Exchange Rates</td>
<td>5</td>
</tr>
<tr>
<td>K</td>
<td>Interest Rates and Spreads</td>
<td>17</td>
</tr>
<tr>
<td>L</td>
<td>Money and Credit Quantity Aggregates</td>
<td>11</td>
</tr>
<tr>
<td>M</td>
<td>Price Indexes</td>
<td>21</td>
</tr>
<tr>
<td>N</td>
<td>Average Hourly Earnings</td>
<td>3</td>
</tr>
<tr>
<td>O</td>
<td>Miscellanea</td>
<td>1</td>
</tr>
</tbody>
</table>

specific to the particular specification of the dynamic factor model. The EM algorithm can also be adopted to find a (local) optimum of the marginal Gaussian log-likelihood function \( \ell_d(y; \theta) \). In this case we use the two-step algorithm of Section 4.4.5 to evaluate \( Q(\psi^*|\psi) \) in the E step.

4.6 An Empirical Illustration

In this section I consider the dynamic factor model given by

\[
y_t = \bar{\mu} + \Lambda f_t + u_t, \tag{4.40}
\]
\[
f_t = \Phi_1 f_{t-1} + \eta_t, \tag{4.41}
\]

for \( t = 1, \ldots, n \) where \( y_t \) is the \( N \times 1 \) observation vector and \( f_t \) is the \( r \times 1 \) vector of factors. The intercept vector \( \bar{\mu} \) is fixed and unknown. The factor loading matrix \( \Lambda \) is unknown and its \( r \) top rows form a lower-triangular matrix with the diagonal elements restricted to be one. Further I assume that the disturbances \( \eta_t \) are IID with diagonal variance matrix \( \Sigma_\eta \).
4.6. AN EMPIRICAL ILLUSTRATION

The idiosyncratic term is modeled as

\[ u_t = \Psi_1 u_{t-1} + \varepsilon_t, \]

where \( \Psi_1 \) is assumed to be diagonal and the disturbances \( \eta_t \) are assumed IID with a diagonal variance matrix \( \Sigma_\eta \).

To demonstrate the applicability of the results of this chapter I estimate two special cases of the model given by equations (4.40) and (4.41) on the data-set of Stock and Watson (2005). Model I has \( r = 7 \) and \( \Psi_1 = 0 \) (such that \( u_t = \varepsilon_t \) ) and Model II has \( r = 4 \) and a non-zero diagonal matrix \( \Psi_1 \). Model I contains over a thousand parameters. With the methods of this chapter quasi-maximum likelihood estimators of these parameters can be obtained in minutes.

Model I is motivated by Stock and Watson (2005) who adopt the procedure of Bai and Ng (2002) to conclude that seven static factors are present in this data-set. Model II is motivated by an analysis of Bai and Ng (2007) based on the same data-set and where they advocate 4 dynamic factors which may span over 7 static factors.

Since in Model II the disturbances follow autoregressive processes of order 1, the 4 dynamic factors translate to a total of 8 static factors. It follows that the number of static factors in models I and II are comparable (7 and 8, respectively) and it is interesting to compare the empirical findings for the two model specifications. The dimensions of the different model specifications and the dimensions of sub-vectors \( \bar{\mu}, \bar{\beta} \) and \( \theta \) of the parameter vector \( \psi \) are reported in Table 4.3.

From the data-set I constructed a balanced panel of \( N = 132 \) monthly US macroeconomic time series from 1960:1 through 2003:12 (44 years, \( T = 528 \)). The data is transformed and differenced to obtain a stationary set of time series; the details of each series and its transformation are given in Appendix A of Stock and Watson (2005). The 132 series are categorized into 15 sectors as presented in Table 4.2. Each sector is indexed by a code (A...O). Table 4.2 also reports the number of time series in each sector. For all series, observations larger than 6 times the standard deviation of the series, \( \sigma \), (in absolute value) are set to \( \pm 6\sigma \). In total, 46 (out of 69,696) observations are Winsorized in this way (0.066%). Finally, each time series is scaled such that its sample variance equals one.

4.6.1 Parameter Estimation

Tables 4.4 and 4.5 present the QML estimates of the VAR coefficients in \( \Phi_1 \) together with the corresponding eigenvalues for Models I and II. The factors in the models are organized
CHAPTER 4. ANALYSIS OF DYNAMIC FACTOR MODELS

Table 4.3: Dynamic factor model specifications

The table reports dimensions for two dynamic factor model specifications that we consider in the empirical analyses as well the corresponding parameter vectors $\psi$ and its components. The observed series, $y_t$, are modeled as $y_t = \bar{\mu} + \Lambda f_t + u_t$ where $f_t$ is a VAR(1) process, see Section 4.2 for more details. The state space formulation is discussed in Section 4.2. In model I, the innovation vector $u_t$ is an IID process with mean zero and diagonal variance matrix $\Sigma_\varepsilon$. In model II, the vector $u_t$ is modeled as a VAR(1) process with a diagonal autoregressive coefficient matrix $\Psi_1$ and a diagonal variance matrix $\Sigma_\varepsilon$. The dimension of $\theta$ is the total of all parameter dimensions excluding $\mu$ and $\beta$ since $\psi = (\bar{\mu}', \beta', \theta')'$. 

<table>
<thead>
<tr>
<th>Model formulation §4.2</th>
<th>State Space</th>
<th>Parameter vector $\psi$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$ $q_r$ $q_\Phi$ $q_\Theta$ $q_\Psi$</td>
<td>$s$ $m$ $p$ $\bar{\mu}$ $\beta$ $\Lambda$ $\Phi$ $\Theta$ $\Psi$ $\Sigma_\varepsilon$ $\theta$</td>
</tr>
<tr>
<td>I</td>
<td>7 0 1 0 0</td>
<td>0 7 7 132 0 903 49 0 0 132 1084</td>
</tr>
<tr>
<td>II</td>
<td>4 0 1 0 1</td>
<td>1 8 8 132 0 522 16 0 132 132 802</td>
</tr>
</tbody>
</table>

In descending order of the eigenvalues of $\Phi_1$. The parameters were estimated by numerically maximizing the Gaussian log-likelihood function $\ell(y; \psi)$ with respect to $\psi$ using the results of Sections 4.4 and 4.5. First the EM algorithm was used to find a point in the neighbourhood of the optimum. Afterwards, BFGS algorithm was used to maximize $\ell(y; \psi)$ with respect to $\psi$ by starting from the final iteration of the EM algorithm. The numerical maximization routine makes use of the analytical score calculations of Appendix 4.E. On a standard computer with 3 GB memory and a 2.2 GHz two-core processor, this took less than 15 minutes for Model I and less than 10 minutes for Model II.

In Tables 4.4 and 4.5 we see that the factors are estimated as stationary but highly persistent processes given the largest eigenvalue of 0.95. For both models, we find the presence of persistent cyclical behaviour in the factors since one conjugate pair of complex eigenvalues is obtained where the real part is equal to 0.94. The other eigenvalues range from large to small. As in any VAR analysis, it is hard to comment on individual coefficients in $\Phi_1$.

4.6.2 Factor Estimates and Factor Loadings

The latent factors can be estimated by applying the low-dimensional Kalman smoother to the transformed time-series as explained in Section 4.4.2. To facilitate a clear interpretation of the factors it is advisable to rotate the factors for example by means of the varimax method. The varimax method tries to construct a rotation such that the resulting factors are as distinct from each other as possible, see e.g. Lawley and Maxwell (1971) for details. To facilitate comparisons, I show only Factors 1, 2, 3 and 5 of Model I and all factors of Model II in Figure 4.1. The first two estimated factors of the two models are similar although
Table 4.4: QML estimates of VAR coefficients for Model I

This table presents the QML estimates of the coefficients in the $7 \times 7$ matrix $\Phi_1$ for Model I. The eigenvalues of the estimates of $\Phi_1$ are reported in descending order. For complex eigenvalues present both the real and imaginary (img) components are presented.

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>real</th>
<th>img</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>-0.15</td>
<td>0.18</td>
<td>-0.031</td>
<td>-0.14</td>
<td>0.062</td>
<td>-0.031</td>
<td>0.95</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-0.36</td>
<td>0.84</td>
<td>-0.017</td>
<td>0.03</td>
<td>0.099</td>
<td>0.028</td>
<td>0.031</td>
<td>0.94</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>0.065</td>
<td>0.074</td>
<td>0.9</td>
<td>0.048</td>
<td>0.19</td>
<td>0.0069</td>
<td>0.034</td>
<td>0.94</td>
<td>-0.08</td>
</tr>
<tr>
<td>4</td>
<td>0.068</td>
<td>0.051</td>
<td>0.034</td>
<td>0.92</td>
<td>0.045</td>
<td>-0.031</td>
<td>-0.017</td>
<td>0.91</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>-0.075</td>
<td>0.025</td>
<td>0.014</td>
<td>-0.073</td>
<td>0.25</td>
<td>-0.1</td>
<td>0.043</td>
<td>0.28</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.003</td>
<td>-0.022</td>
<td>-0.029</td>
<td>0.036</td>
<td>0.003</td>
<td>-0.33</td>
<td>-0.012</td>
<td>0.042</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>-0.024</td>
<td>-0.027</td>
<td>-0.038</td>
<td>-0.0002</td>
<td>-0.049</td>
<td>-0.028</td>
<td>0.97</td>
<td>-0.33</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.5: QML estimates of VAR coefficients for Model II

This table shows the QML estimates of the coefficients in the $4 \times 4$ matrix $\Phi_1$ for Model II. The eigenvalues of the estimates of $\Phi_1$ are reported in descending order. For complex eigenvalues present both the real and imaginary (img) components are presented.

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>real</th>
<th>img</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.29</td>
<td>-0.23</td>
<td>-0.12</td>
<td>-0.1</td>
<td>0.94</td>
<td>0.094</td>
</tr>
<tr>
<td>2</td>
<td>-0.38</td>
<td>0.44</td>
<td>0.031</td>
<td>0.13</td>
<td>0.94</td>
<td>-0.094</td>
</tr>
<tr>
<td>3</td>
<td>0.086</td>
<td>-0.43</td>
<td>0.96</td>
<td>0.17</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-0.64</td>
<td>0.33</td>
<td>-0.019</td>
<td>0.56</td>
<td>0.046</td>
<td>0</td>
</tr>
</tbody>
</table>
the amount of noise in the factors is somewhat different. The first set of factors seems to be associated with the business cycle and is therefore displayed with the NBER business cycle reference dates of peaks and troughs. The NBER dates do not coincide perfectly with the peaks and troughs of the first factor but close enough to justify referring to it as the “business cycle” factor. The Factors 3 of both models appear to pick up the first and second oil crisis periods in the mid 1970’s and the early 1980’s. The turbulence of the interest rates in the early 1980’s are present in the estimated Factor 5 of Model I and the estimated Factor 4 of Model II.

**Figure 4.1: Estimated factors**
These figures show a selection of estimated factors extracted from the observed series by applying the Kalman smoother as described in Section 4.4.6. In the left panel the estimated Factors 1, 2, 3 and 5 for Model I are displayed and in the right panel all four estimated factors for Model II are presented.

The actual estimates of $\Lambda$ are not easy to interpret and therefore Stock and Watson (2002) propose to focus on the $R^2$ goodness-of-fit statistics which are obtained by regressing the univariate time series $y_{it}$ (for each $i = 1, \ldots, N$) on a constant and a particular principal
component estimate (or diffusion index). The series of $N$ regressions can be repeated for each principal component and the resulting $N$ dimensional series of $R^2$ statistics can be displayed as an index plot for each principal component. I present the $N$ series of $R^2$ statistics for the seven factors, estimated using the Kalman smoother, of Model I in the left panel of Figure 4.2. The clustering of high $R^2$ statistics within one or more sectors is clearly visible. The first factor is highly correlated with the real variables in sector (A) real output & income and weakly correlated with the variables in the sectors (B) employment & hours and (H) orders. The second factor is mostly associated with the sectors (G) real inventories and (H) orders but also correlated with variables in the sectors (B) employment & hours and (F) housing starts & sales. The two individual indices for production and unemployment in sectors (A) and (B) are particularly highly correlated with Factor 2. The third and fifth factors are clearly connected with interest rates and spreads, respectively, from sector (K). The fourth factor does not contribute much to the analysis. Factors 6 and 7 can be interpreted as the price index and the housing market index, respectively. The $R^2$ statistics for the four factors in Model II are presented in the right panel of Figure 4.2. The third and fourth factors in Model II are strongly connected with interest rates and spreads, respectively, from sector (K). They are similar to Factors 3 and 5 of Model I. However, the first two factors of Model II are not very distinctive and can be regarded as a mix of the first two factors of Model I. In other words, they are associated with the “real” sectors (A) real output & income, (B) employment & hours, (F) housing starts & sales, and (H) orders.

#### 4.6.3 Diagnostic Checking

An appealing feature of a model-based analysis is that model misspecification tests and diagnostics concerning normality, heteroskedasticity and serial correlation can be considered as an effective tool for model selection. In the practice of time series analysis, diagnostic test statistics are applied to standardized one-step ahead prediction errors. If the model is correctly specified these errors should be IID. It is hard to argue that a dynamic factor model is the appropriate specification for a joint analysis of 132 time series. However, the model misspecification diagnostics may indicate how far we are from a reasonable specification.

The Kalman filter allows us to compute the prediction errors for all 132 series in a few seconds. More specifically, I computed the generalized least squares residuals as advocated by Harvey (1989, section 5.4) to allow for the intercept vector $\bar{\mu}$ in both model specifications I and II. The residuals are standardized. To illustrate the effectiveness of residual diagnostics in the context of dynamic factor analysis, I computed for each residual series the serial correlation portmanteau $\chi^2$ test of Ljung and Box (1978). The Box-Ljung $Q(q)$ statistic
These figures show $R^2$ statistics obtained by regressing the 132 time series on each of the estimated factors from Model I and Model II.

The displayed Box-Ljung values are truncated at 100. It is evident that for many series the null hypothesis of no serial correlation in the residuals is rejected. The current dynamic factor models are therefore not fully satisfactory for this panel of macroeconomic time series. We can conclude however that Model II is more successful in capturing the collective dynamics in the data-set compared to Model I.
4.A Proof of Lemma 4.1

From Conditions (i), (ii) and (iii) in Section 4.4.1 and using the fact that $\Sigma_e A^L$ has full column rank, we obtain

$$\text{Col}\{\Sigma_e A^L\} = \text{Row}\{A^H\}^\perp = \text{Col}\{Z\}.$$ 

Define $\Lambda^\dagger = \Sigma_e A^L$ then $A^L = \Lambda^\dagger \Sigma_e^{-1}$ and $\text{Col}\{\Lambda^\dagger\} = \text{Col}\{Z\}$. Since $Z = \Lambda G$, with $G$ of full row rank, we have $\text{Col}\{\Lambda^\dagger\} = \text{Col}\{\Lambda\}$. This proves the necessity part of Lemma 4.1.

4.B Proof of Lemma 4.2

We have

$$y_t^H \Sigma^{-1}_H y_t^H = (y_t - d_t - X_t \beta - \mu)' A^H (A^H \Sigma_e A^H)^{-1} A^H (y_t - d_t - X_t \beta - \mu)$$

$$= (y_t - d_t - X_t \beta - \mu)' J^H \Sigma^{-1}_e (y_t - d_t - X_t \beta - \mu),$$
where $J^H \overset{\text{def.}}{=} \ A^H(\Sigma^H A^H)^{-1} A^H \Sigma^e$ is the projection matrix for a GLS with covariate matrix $A^H$ and variance matrix $\Sigma_{e}^{-1}$. Similarly, define

$$J^L \overset{\text{def.}}{=} \ A^L(\Sigma^L A^L)^{-1} A^L \Sigma^e,$$

as the GLS projection matrix for covariate matrix $A^L$ and variance matrix $\Sigma_{e}^{-1}$. Since the transformation matrix $A = (A^L, A^H)'$ is full rank and $A^L \Sigma^e A^H' = 0$, we must have

$$J^H = I - J^L.$$

The definition of $A^L$ implies that $J^H = I - \Sigma_{e}^{-1} \Lambda^\dagger (\Lambda^\dagger \Sigma_{e}^{-1} \Lambda^\dagger )^{-1} \Lambda^\dagger$ and

$$J^H = \Sigma_e^e A^H(\Sigma^H A^H)^{-1} A^H = I - \Lambda^\dagger (\Lambda^\dagger \Sigma_{e}^{-1} \Lambda^\dagger )^{-1} \Lambda^\dagger \Sigma_{e}^{-1} \overset{\text{def.}}{=} M_{\Lambda}. \quad (4.42)$$

The proof of (4.23) is completed by the identity $J^H \Sigma_{e}^{-1} = J^H \Sigma_{e}^{-1} J^H$ and the definition $e_t \overset{\text{def.}}{=} M_{\Lambda}(y_t - d_t - X_t \beta - \mu)$ as the GLS residual for data vector $y_t - d_t - X_t \beta - \mu$, covariate $\Lambda^\dagger$ and variance matrix $\Sigma_{e}$.

### 4.C Proof of Equation (4.30)

The following two lemmas are required for the main proof.

**Lemma 4.3.** Consider the regression model (4.10). The marginal Gaussian loglikelihood $\ell_m(y; \theta)$ is given by

$$\ell_m(y; \theta) = c - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \log |\tilde{X}' \Sigma^{-1} \tilde{X}| - \frac{1}{2} \epsilon^T \Sigma^{-1} \epsilon, \quad (4.43)$$

where $c$ is a constant independent of $\theta$ and $\epsilon^x$ is the residual vector from a GLS regression on $y$ with covariate matrix $\tilde{X}$ and variance matrix $\Sigma$.

**Proof.** See e.g. Harville (1974). □

**Lemma 4.4.** Consider the linear regression model

$$y_1 = X_1 \beta + \epsilon_1, \quad \epsilon_1 \sim N(0, \Omega_1), \quad (4.44)$$
$$y_2 = X_2 \beta + \epsilon_2, \quad \epsilon_2 \sim N(0, \Omega_2), \quad (4.45)$$
where \( \varepsilon_1 \) and \( \varepsilon_2 \) are independent, the GLS estimator \( \hat{\beta}_{GLS} \) of \( \beta \) is given by

\[
\hat{\beta}_{GLS} = \hat{\beta}_{GLS}^2 + V X_1 F^{-1}(y_1 - X_1 \hat{\beta}_{GLS}^2), \quad \text{Var}(\hat{\beta}_{GLS}) = V - V X_1 F^{-1} X_1 V,
\]

where \( F = X_1 V X_1' + \Omega_1 \), \( \hat{\beta}_{GLS}^2 \) is the GLS estimator of \( \beta \) based on \( y_2 \) only and \( V \) is the associated variance

\[
\hat{\beta}_{GLS}^2 = (X_2 \Omega_2^{-1} X_2)^{-1} X_2 \Omega_2^{-1} y_2, \quad V = (X_2 \Omega_2^{-1} X_2)^{-1}.
\]

Furthermore,

\[
(y_1 - X_1 \hat{\beta}_{GLS})' \Omega_1^{-1} (y_1 - X_1 \hat{\beta}_{GLS}) + (y_2 - X_2 \hat{\beta}_{GLS})' \Omega_2^{-1} (y_2 - X_2 \hat{\beta}_{GLS}) = (y_2 - X_2 \hat{\beta}_{GLS})' \Omega_2^{-1} (y_2 - X_2 \hat{\beta}_{GLS}) + (y_1 - X_1 \hat{\beta}_{GLS})' F^{-1} (y_1 - X_1 \hat{\beta}_{GLS}). \tag{4.46}
\]

**Proof.** The results follow from regression theory. \( \square \)

**Proof of equation (4.30).** It can be verified from (4.43) that

\[
\ell_m(y; \theta) = \ell_m(y^L, \hat{y}^H; \theta) + (T - 1) \log |A| = \ell_m(y^L, \hat{y}^H; \theta) - \frac{T - 1}{2} \log \frac{|\Sigma_e|}{|\Sigma_L|},
\]

Denote by \( \hat{\beta}, \hat{\mu}^L \) and \( \hat{\mu}^H \) the GLS estimators of respectively \( \beta, \mu^L \) and \( \mu^H \) based on \( y \) then

\[
\ell_m(\hat{y}^L, \hat{y}^H; \theta) = c - \frac{1}{2} \log |\hat{X}_A \Sigma_A^{-1} \hat{X}_A| - \frac{1}{2} \log |\Sigma_A| - \frac{1}{2} \text{RSS}(\hat{\beta}, \hat{\mu}^L, \hat{\mu}^H), \tag{4.47}
\]

where \( \Sigma_A = (I_T \otimes A) \Sigma (I_T \otimes A)' \),

\[
X_A = \begin{pmatrix} E_L & 0 & X_L \\ 0 & E_H & X_H \end{pmatrix},
\]

with \( X^L = (X_1^L, \ldots, X_T^L)' \), \( X^H = (X_1^H, \ldots, X_T^H)' \), \( E_L = i_n \otimes I_m, E_H = i_n \otimes I_{N-m} \), with \( i_n = (1, \ldots, 1)' \) and suppressing the dependence on \( \theta \), further

\[
\text{RSS}(\beta, \mu^L, \mu^H) = \hat{e}^L' \Sigma_{yL}^{-1} \hat{e}^L + \hat{e}^H' \Sigma_{yH}^{-1} \hat{e}^H, \tag{4.48}
\]

with \( \hat{e}^L = (\hat{y}^L - X^L \beta - E_1 \mu^L) \) and \( \hat{e}^H = (\hat{y}^H - X^H \beta - E_2 \mu^H) \), and finally

\[
\Sigma_{yL} = \text{Var}(\hat{y}^L), \quad \Sigma_{yH} = \text{Var}(\hat{y}^H).
\]

Denote \( M_{\perp E_H} = I - E_H (E_H' \Sigma_{yH}^{-1} E_H)^{-1} E_H' \Sigma_{yH}^{-1} \), \( \bar{y}_{\perp E_H} = M_{\perp E_H} \bar{y}^H \) and \( X^H_{\perp E_H} = M_{\perp E_H} X^H \), we
have

$$RSS(\beta, \mu^L, \hat{\mu}^H) = \hat{e}^L L \Sigma_{gL}^{-1} \hat{e}^L + (\hat{y}_L^H - X^H_{\perp E_H} \hat{\beta})' \Sigma_{\hat{y}^H}^{-1}(\hat{y}^L_{\perp} - X_{\perp E_H}^H \beta). \quad (4.49)$$

Using the result of equation (4.42) we have

$$X^H_{\perp E_H} \Sigma_{gH}^{-1} X^H_{\perp E_H} = \sum_t (X_t - \bar{X})' A^H (A^H \Sigma A^H)'^{-1} A^H (X_t - \bar{X})$$

$$= \sum_t (X_t - \bar{X})' M^H_\beta \Sigma^{-1} M_\beta (X_t - \bar{X})$$

$$= \sum_t \tilde{X}_t \Sigma^{-1} \tilde{X}_t = B, \quad (4.50)$$

and similarly

$$X^H_{\perp E_H} \Sigma_{gH}^{-1} \hat{y}^H = \sum_t \tilde{X}_t \Sigma^{-1} \tilde{y}_t = Bb.$$ 

From (4.46) we have

$$RSS(\hat{\beta}, \mu^L, \hat{\mu}^H) = (\hat{y}^L - E_L \mu^L - X^L b)' F^{-1} (\hat{y}^L - E_L \mu^L - X^L b) +$$

$$(\hat{y}^H_{\perp E_H} - X^H_{\perp E_H} b)' \Sigma_{gH}^{-1} (\hat{y}^H_{\perp E_H} - X^H_{\perp E_H} b), \quad (4.51)$$

where $F = X^L B^{-1} X^L' + \Sigma_{gL}$. Minimizing (4.51) with respect to $\mu^L$, we find

$$\hat{\mu}^L = (E^L_F)^{-1} E^L_F F^{-1} (\hat{y}^L - X^L b). \quad (4.52)$$

Note that $\hat{\mu}^L$ is identical to the GLS estimator of $\mu^L$ from model (4.34). It follows from the definitions in section 4.4.5 and expression (4.43) that

$$C_1 + C_2 - \frac{1}{2} RSS(\hat{\beta}, \hat{\mu}^L, \hat{\mu}^H) = L_m(\hat{y}^L; \theta) + L_m(\hat{y}^H; \theta), \quad (4.53)$$

where $C_1$ is independent of $\theta$ and

$$C_2 = -\frac{1}{2} \log |B| - \frac{1}{2} \log |X^L B^{-1} X^L' + \Sigma_{gL}| - \frac{1}{2} \log |E^L_F (X^L B^{-1} X^L' + \Sigma_{gL})^{-1} E^L_F|.$$ 

Using a well known determinant identity we have

$$|X^L B^{-1} X^L' + \Sigma_{gL}| = |B + X^L \Sigma_{gL} X^L||B^{-1}||\Sigma_{gL}|, \quad (4.54)$$
and with the Woodbury matrix identity and results on determinants of block matrices, inverses of a block matrices and (4.42) we have

\[
|E'_{L}(X^{L}B^{-1}X^{L'} + \Sigma_{g^{L}})^{-1}E_{L}| = |E'_{L}\Sigma_{g^{L}}^{-1}E_{L} - E'_{L}\Sigma_{g^{L}}^{-1}X^{L}(B + X^{L'}\Sigma_{g^{L}}^{-1}X^{L})^{-1}X^{L'}\Sigma_{g^{L}}^{-1}E_{L}|
\]

\[
= \frac{1}{|D|} \begin{vmatrix}
E'_{L}\Sigma_{g^{L}}^{-1}E_{L} & 0 & E'_{L}\Sigma_{g^{L}}^{-1}X^{L} \\
0 & T\Sigma_{H}^{-1} & \sum_{t}\Sigma_{H}^{-1}X_{t}^{H} \\
X^{L'}\Sigma_{g^{L}}^{-1}E_{L} & \sum_{t}X_{t}^{H}\Sigma_{H}^{-1} & X^{L'}\Sigma_{g^{L}}^{-1}X^{L} + \sum_{t}X_{t}M_{t}^{L}_{A}\Sigma_{r}^{-1}M_{A}X_{t}
\end{vmatrix}
= \frac{|\hat{X}_{A}^{'}\Sigma_{A}^{-1}\hat{X}_{A}|}{|D|},
\]

(4.55)

where

\[
|D| = \begin{vmatrix}
T\Sigma_{H}^{-1} & \sum_{t}\Sigma_{H}^{-1}X_{t}^{H} \\
\sum_{t}X_{t}^{H}\Sigma_{H}^{-1} & X^{L'}\Sigma_{g^{L}}^{-1}X^{L} + \sum_{t}X_{t}M_{t}^{L}_{A}\Sigma_{r}^{-1}M_{A}X_{t}
\end{vmatrix} = |B + X^{L'}\Sigma_{g^{L}}X^{L}|T^{N-m}.
\]

Finally, we have

\[
|\Sigma_{g^{L}}| = |\Sigma_{g^{L}}||\Sigma_{g^{H}}| = |\Sigma_{A}|,
\]

(4.57)

since \(|\Sigma_{g^{H}}| = 1\) and \(\Sigma_{A}\) is block diagonal. Combining (4.47) and (4.53) with (4.54), (4.56) and (4.57) we obtain

\[
C_{3} + L_{m}(\bar{y}^{L}; \theta) + L_{m}(\bar{y}^{H}; \theta) = \ell_{m}(\bar{y}^{L}, \bar{y}^{H}; \theta),
\]

for a constant \(C_{3}\) independent of \(\theta\), which concludes the proof.

\[\square\]

4.D Proof of Results of Section 4.4.6

We saw in Appendix 4.C that the MMSLE of \(\mu^{L}\) from model (4.34) equals the GLS estimator of \(\mu^{L}\), see equation (4.52). Then,

\[
P(\hat{\beta}|\bar{y}^{L}) = E(\hat{\beta}) + \text{Cov}(\hat{\beta}, \bar{y}^{L})\text{Var}(\bar{y}^{L})^{-1}[y^{L} - E_{L}\mu_{GLS} - X^{L}E(\hat{\beta})]
= b + B^{-1}X_{L}^{'}(X^{L}B^{-1}X^{L} + \Sigma_{g^{L}})^{-1}(y^{L} - E_{L}\mu_{GLS} - X^{L}b)
\]
where $E_L$ is defined below equation (4.48) and the variances and covariances are evaluated assuming (4.34) is the true model. Consider the function $RSS(\beta, \mu^L, \hat{\mu}^H)$ defined in (4.49). The GLS estimators of $\hat{\beta}$ and $\hat{\mu}^L$ can be obtained by minimizing $RSS(\beta, \mu^L, \hat{\mu}^H)$ with respect to $\beta$ and $\mu^L$. Substituting $\hat{\mu}^L$ in $RSS(\beta, \mu^L, \hat{\mu}^H)$ and minimizing with respect to $\beta$, equation (4.35) follows from Lemma 4.4.

Finally, we need to prove that the diffuse Kalman filter produces the correct variances and covariances for the GLS estimators $\hat{\beta}$ and $\hat{\mu}^L$. From (4.49) it follows that the GLS estimator of $\gamma_L = (\mu^L, \beta')$ is given by $C^{-1}c$ where

$$C = \begin{pmatrix} E_L^{\prime} \Sigma^{-1}_L E_L & E_L^{\prime} \Sigma^{-1}_L X^L \\ X^L \Sigma^{-1}_L X^L & X^L \Sigma^{-1}_L X^L + B \end{pmatrix}, \quad c = \begin{pmatrix} E_L^{\prime} \Sigma^{-1}_L \bar{y}^L \\ X^L \Sigma^{-1}_L \bar{y}^L + Bb. \end{pmatrix},$$

Since the correct variance of the estimator is given by $C^{-1}$ we need to prove that

$$\mathbb{E}[(\hat{\gamma}_L - \gamma_L)(\hat{\gamma}_L - \gamma_L)'] = C^{-1}, \quad (4.58)$$

where $\hat{\gamma}_L = [P(\mu^L|\bar{y}^L)', P(\delta|\bar{y}^L)']'$ and the expectation is computed under the assumption that (4.34) is the true model. Since we already showed that $\hat{\gamma}_L = C^{-1}c$, we have

$$\mathbb{E}((\hat{\gamma}_L - \gamma_L)(\hat{\gamma}_L - \gamma_L)') = C^{-1}\mathbb{E}[(c - C\gamma_L)(c - C\gamma_L)'] C^{-1}.$$  

It can now be shown that $\mathbb{E}[(c - C\gamma_L)(c - C\gamma_L)'] = C$, which proves (4.58).

4.E The Score Function of Section 4.5.1

In this section I give the derivatives of (4.38) with respect to the system matrices $Z$, $T$, $\Sigma_\varepsilon$ and $\Sigma_\eta$ for the case where $\mu = 0$, $\beta = 0$ and $R = I$. For cases where $\mu \neq 0$, $\beta \neq 0$ and/or $R \neq I$, similar expressions can be obtained for the derivatives but the expressions are more lengthy and more intricate, see Koopman and Shephard (1992) for a detailed discussion.

The derivatives are given by

$$\frac{\partial \ell(y)}{\partial Z} = \Sigma_{\varepsilon}^{-1} \left( \sum_{t=1}^n \{y_t - d_t\} a_{t|n}' - ZS_{1:n}^0 \right), \quad \frac{\partial \ell(y)}{\partial \Sigma_\varepsilon} = Q_{\varepsilon}\Sigma_{\varepsilon}^{-1} - \frac{1}{2} \text{diag}(Q_{\varepsilon}^*\Sigma_{\varepsilon}^{-1}),$$

$$\frac{\partial \ell(y)}{\partial T} = \Sigma_{\eta}^{-1}(S_{2:n}^{(1)} - TS_{1:n-1}^{(0)}), \quad \frac{\partial \ell(y)}{\partial \Sigma_\eta} = Q_{\eta}\Sigma_{\eta}^{-1} - \frac{1}{2} \text{diag}(Q_{\eta}^*\Sigma_{\eta}^{-1}),$$
4.E. THE SCORE FUNCTION OF SECTION 4.5.1

where \( Q^*_\varepsilon = Q_\varepsilon - n \), \( Q^*_\eta = Q_\eta - n - 1 \), with \( Q_\varepsilon \) and \( Q_\eta \) defined in (4.39) and

\[
S^{(0)}_{j:k} = \sum_{t=j}^{k} a_t \alpha_t' + Q_{t|n}, \quad S^{(1)}_{j:k} = \sum_{t=j}^{k} a_t \alpha_{t-1}' + Q_{t,t-1|n},
\]

for \( j, k = 1, \ldots, n \) (\( j \leq k \)) and where

\[
Q_{t-1|n} = Cov(\alpha_{t-1}, \alpha_t|y),
\]

An expression for this is presented by de Jong and MacKinnon (1988) and given by

\[
Q_{t,t-1|n} = (Q_{t|t-1} N_{t-1} - I) L'_{t-1} Q_{t-1|t-2}, \quad t = 2, \ldots, n,
\]

where \( Q_{t|t-1}, N_{t-1} \) and \( L_{t-1} \) can be obtained from the Kalman filter and smoother recursions for \( t = 1, \ldots, n \).
Chapter 5

Dynamic Factor Models with Smooth Factor Loadings

5.1 Introduction

Dynamic factor models explain a large panel of time series in terms of a small number of unobserved common factors. The relationship between the time series and the common factors is generally assumed to be linear. In this case it is customary to refer to the weights of the individual factors as the factor loadings. In many applications where dynamic factor models are considered, the individual series in the panel have a natural ordering in terms of one or more variables or indicators. For example, if we consider a time series panel of bond yields we can order the yields according to the time to maturity of the bond. For this type of data set, it is often natural to assume that the factor loadings for a specific factor are a relatively smooth function of the variable that is used to order the time series. I present a new class of dynamic factor models that can be used to model this type of time series panels. I propose a dynamic factor model that imposes a set of smoothness constraints on the factor loadings. These smoothness constraints are imposed by assuming that the factor loadings are given by the value of a cubic spline evaluated at the associated ordering variables. The flexibility of the cubic spline is determined by the choice of a set of knots. I present a statistical procedure based on a sequence of Wald tests that allows us to systematically search for the optimal set of knots. The parameters of the dynamic factor model with smooth factor loadings can be estimated routinely using maximum likelihood procedures.

Although the dynamic factor model with smooth factor loadings can be applied to a wide variety of different problems, I focus in this chapter on the modelling of the term structure
of interest rates. Many models for the yield curve can be viewed as dynamic factor models with a set of restrictions imposed on the factor loadings. Almost always, these restrictions imply smoothness of the factor loadings when viewed as a function of time to maturity. In an empirical study I will investigate whether the restrictions imposed by a number of popular models for the term structure are supported by the data. As part of this study I compare the performance of these models to the performance of the unrestricted dynamic factor model and the new dynamic factor model with smooth factor loadings. For all the term structure models considered, with the exception of the new dynamic factor model, I have found that the restrictions imposed are rejected by a standard likelihood ratio test. Using the new methodology I construct a parsimonious dynamic factor model with smooth loadings that is not rejected. The fit of this model is close to the fit of the general model and the residual diagnostics are satisfactory.

Early contributions to the literature on dynamic factor models can be found in Sargent and Sims (1977), Geweke (1977), Engle and Watson (1981), Watson and Engle (1983) Connor and Korajczyk (1993) and Gregory, Head, and Raynauld (1997). Most of these papers consider time series panels with limited panel dimension. The increasing availability of high-dimensional data sets has intensified the quest for computationally efficient estimation methods. The strand of literature headed by Forni, Hallin, Lippi, and Reichlin (2000), Stock and Watson (2002) and Bai (2003) led to a renewed interest in dynamic factor analysis. These new methods are typically applied to large panels of time series. Exact maximum likelihood methods such as those proposed in Watson and Engle (1983) have traditionally been dismissed as being too computationally intensive for such high-dimensional panels. Jungbacker and Koopman (2008) however present new results that allow the application of exact maximum likelihood methods to large panels, see also the discussion in Chapter 4. Examples of recent papers employing likelihood-based methods for the analysis of dynamic factor models are Doz, Giannone, and Reichlin (2006) and Reis and Watson (2007).

A prominent class of time series models for term structures is based on the results of Nelson and Siegel (1987). It can be regarded as a special case of the dynamic factor model. Nelson and Siegel (1987) show in a seminal paper that a typical yield curve can be closely approximated by a curve obtained from a weighted sum of three smooth functions. The form of these three functions depends on a single parameter. Diebold and Li (2006) use the Nelson-Siegel set up to develop a two-step procedure with the purpose of forecasting future term structure yields. They show that forecasts obtained from this procedure are competitive with forecasts obtained from other standard prediction methods. Diebold, Rudebusch, and Aruoba (2006) integrate this two-step approach into a single state space model by viewing
the Nelson-Siegel weights as an unobserved vector autoregressive process. Generalizations of this state space approach are considered by De Pooter (2007) and Koopman, Mallee, and Van der Wel (2009). The former considers more coefficients for the yield curve while the latter allows the parameter governing the shape of the three Nelson-Siegel functions to vary over time and includes conditional heteroskedasticity in the innovations. A different approach is proposed by Bowsher and Meeks (2008). In their model the term structure is represented by a cubic spline that is observed with measurement noise. The parameters controlling the shape of the spline are modelled as a cointegrated vector autoregressive process. This way of modelling smooth functions that vary stochastically over time was introduced in Harvey and Koopman (1993).

A large number of papers have been concerned with the construction of models for yield curve dynamics that incorporate the restriction that the market is free of arbitrage opportunities, see e.g. Brigo and Mercurio (2006) for an extensive overview. Just as the time series models discussed earlier, these models are generally specified in terms of a small number of underlying factors. However, many of these models imply a nonlinear relation between the unobserved factors and the yields. An exception is the class of affine term structure models presented in Duffie and Kan (1996). The Gaussian specifications contained in this class of models can be shown to be special cases of the dynamic factor model. Another interesting arbitrage-free model of the term structure of interest rates is the arbitrage-free version of the Nelson-Siegel model of Diebold, Rudebusch, and Aruoba (2006), proposed by Christensen, Diebold, and Rudebusch (2007).

The structure of the chapter is as follows. The general dynamic factor model is presented and discussed in section 5.2. The new methodology to construct dynamic factor models with smooth factor loadings is developed in section 5.3. Section 5.4 discusses a selection of well-known term structure models that can be regarded as restricted versions of the general dynamic factor model. In section 5.5 I give a description of the data set and give the results of a preliminary data analysis. Section 5.6 presents the results of the empirical study. Section 5.7 concludes and provides suggestions for future research.

5.2 The Dynamic Factor Model

I consider a monthly time series panel of treasury yields for a set of \( N \) different maturities \( \tau_1, \ldots, \tau_N \). The yield at time \( t \) of the treasury with maturity \( \tau_i \) is denoted by \( y_t(\tau_i) \) for
$t = 1, \ldots, n$. The $N \times 1$ vector of all yields at time $t$ is given by

$$y_t = \begin{bmatrix} y_t(\tau_1) \\ \vdots \\ y_t(\tau_N) \end{bmatrix}, \quad t = 1, \ldots, n.$$  

I denote the vector of all observations by $y = (y'_1, \ldots, y'_n)'$.

The general dynamic factor model is given by

$$y_t = \mu_y + \Lambda f_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H), \quad t = 1, \ldots, n,$$  

(5.1)

where $\mu_y$ is an $N \times 1$ vector of constants, $\Lambda$ is the $N \times r$ factor loading matrix, $f_t$ is an $r$-dimensional stochastic process, $\varepsilon_t$ is the $N \times 1$ disturbance vector and $H$ is an $N \times N$ variance matrix. Throughout this chapter I restrict the variance matrix of the observation disturbances $H$ to be diagonal. This means that covariance between the yields of different maturities is explained solely by the common latent factor $f_t$. These latent factors are given by

$$f_t = U\alpha_t,$$  

(5.2)

where the $r \times p$ matrix $U$ contains appropriate weights that link $f_t$ to a $p$-dimensional unobserved state vector. This state vector $\alpha_t$ is modelled by the dynamic stochastic process

$$\alpha_{t+1} = \mu_\alpha + T\alpha_t + R\eta_t, \quad \eta_t \sim N(0, Q), \quad t = 1, \ldots, n,$$  

(5.3)

where $\mu_\alpha$ is the $p \times 1$ vector of constants, $T$ is the $p \times p$ transition matrix, $R$ is the $p \times q$ selection matrix (typically consisting of ones and zeros), $\eta_t$ is the $q \times 1$ disturbance vector and $Q$ is an $q \times q$ variance matrix. For the initial state vector I assume

$$\alpha_1 \sim N(a_1, P_1),$$  

(5.4)

with $p \times 1$ mean vector $a_1$ and $p \times p$ variance matrix $P_1$. Generally, we set the mean of the initial state $a_1$ to zero and choose the initial variance matrix $P$ to be a function of the system matrices. The Gaussian disturbances $\varepsilon_t$ and $\eta_t$ are serially and mutually uncorrelated for $t,s = 1, \ldots, n$ and are assumed independent of $\alpha_1$. Although dimensions $N, p, q$ and $r$ can be chosen freely we can assume without loss of generality that $r \leq p$ and $p \geq q$. Also, since the motivation of the dynamic factor model is to explain a multivariate time series using a small number of common components, we will generally have $N >> r$. The vectors $\mu_y$ and
5.2. THE DYNAMIC FACTOR MODEL

\[ \mu, \alpha \text{ and the matrices } \Lambda, H, U, T \text{ and } Q \text{ are referred to as system matrices. This general dynamic factor model can be regarded as a special case of the linear Gaussian state space model, see Harvey (1989) and Durbin and Koopman (2001) and the discussion in Chapter 2.} \]

The dynamic specification for \( f_t \) is general. All vector autoregressive moving average processes can be formulated as (5.2) and (5.3) which is known as the companion form; see, for example, Box, Jenkins, and Reinsel (1994). The family of time series processes that can be formulated as (5.2) and (5.3) is however much wider and includes a broad range of nonstationary time series processes. In this chapter I focus on models where \( f_t \) is either a vector autoregression or a cointegrated vector autoregression. I discuss the form that \( U, T \) and \( R \) take for these two specifications in sections 5.2.1 and 5.2.2.

The elements of the system matrices \( \mu_y, \Lambda, H, \mu_\alpha, T \text{ and } Q \) will generally contain parameters that need to be estimated from the data. To ensure identification we need to impose two sets of restrictions on respectively the parameters in the means of the yields, determined by \( \mu_y \text{ and } \mu_\alpha \), and the parameters in \( \Lambda, T \text{ and } Q \) that govern the covariance structure.

First, we cannot estimate both vectors \( \mu_y \text{ and } \mu_\alpha \) without restrictions. Diebold, Rudebusch, and Aruoba (2006) and Bowsher and Meeks (2008), among others, assume that \( \mu_y \) is zero and proceed by estimating \( \mu_\alpha \) only. Additional restrictions need to be imposed on \( \mu_\alpha \) in case the dynamic process of \( f_t \) is nonstationary, see Bowsher and Meeks (2008). In this chapter I leave \( \mu_y \) unrestricted and set \( \mu_\alpha \) to zero. I choose this more general model because the main concern is inference on the loading matrix \( \Lambda \) and therefore I prefer to avoid additional restrictions on the remaining parameters.

Second, restrictions on \( \Lambda \) are needed because only its column space can be identified uniquely. Several restrictions on \( \Lambda \) can be considered. I choose to set a selection of \( r \) rows of \( \Lambda \) equal to the rows of the \( r \times r \) identity matrix. In case \( r = 3 \) and \( N > r \), we may have

\[
\Lambda = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\lambda_{4,1} & \lambda_{4,2} & \lambda_{4,3} \\
\vdots & \vdots & \vdots \\
\lambda_{N,1} & \lambda_{N,2} & \lambda_{N,3}
\end{pmatrix}.
\]

(5.5)

In this example, we can interpret the elements of \( f_t \) as being a set of hypothetical mean-adjusted ‘true’ yields for the maturities \( \tau_1, \ldots, \tau_r \) which are observed at time \( t \) subject to
measurement noise in $\varepsilon_t$. We do not necessarily have to restrict the first $r$ rows. We can choose to impose the restrictions on each set of $r$ rows of $\Lambda$ to obtain a dynamic factor model that is observationally equivalent to the model with $\Lambda$ of the form (5.5). Since the rows of $\Lambda$ correspond to fixed maturities I prefer to distribute the rows of the identity matrix evenly over the full range of rows. This allows us to interpret the factors as short-term, medium-term and long-term components. If $f_t$ is a vector autoregression or a cointegrated vector autoregression, as I assume throughout this chapter, this choice of restrictions for $\Lambda$ allows us to leave the parameters in $T$ and $Q$ unrestricted.

5.2.1 Stationary Specification

The stationary dynamic factor model for time series of yields is defined by (5.1) where the $r \times 1$ vector $f_t$ is modelled by the vector autoregressive process

$$f_{t+1} = \sum_{j=0}^{k-1} \Gamma_j f_{t-j} + \zeta_t, \quad \zeta_t \sim NID(0, Q_\zeta),$$

(5.6)

with $r \times r$ coefficient matrices $\Gamma_j$ for $j = 0, \ldots, k - 1$ and variance matrix $Q_\zeta$. The dynamic process (5.6) is commonly known as a VAR($k$) model. I will refer to a dynamic factor model with VAR($k$) factors as a DFM-VAR($k$) model. I denote by $\Gamma(z)$ the characteristic polynomial of the VAR($k$) process given by $\Gamma(z) = I - \sum_{j=0}^{k-1} \Gamma_j z^j$. The stationarity of $f_t$ is ensured by imposing the restriction that $|\Gamma(z)| = 0$ has all its solutions outside the unit circle. The process $f_t$ is straightforwardly written in the form (5.2) – (5.3). In case $k = 1$, we have $\alpha_t = f_t$, $U = R = I_r$, $T = \Gamma_0$ and $Q = Q_\zeta$ where $I_m$ is the $m \times m$ identity matrix. In case $k > 1$, we have

$$\alpha_t = \begin{pmatrix} f_t \\ \vdots \\ f_{t-k+1} \end{pmatrix}, \quad U = R' = \begin{pmatrix} I_r & 0 & \cdots & 0 \end{pmatrix}, \quad T = \begin{pmatrix} \Gamma_0: \Gamma_{k-2} \\ \Gamma_{k-1} \\ I_{r(k-1)} \\ 0 \end{pmatrix},$$

(5.7)

and $Q = Q_\zeta$ where $\Gamma_{i,j} = (\Gamma_i \cdots \Gamma_j)$ for $i, j = 0, \ldots, k - 1$ and $i < j$. I choose the variance $P_1$ of the initial state $\alpha_1$ to be equal to the variance of the invariant distribution of $\alpha_t$. This implies that $P_1$ is the solution to the equation $P_1 = TP_1T' + Q$. The mean of the initial state is set to zero.
5.2.2 Nonstationary Specification

For nonstationary dynamic factor models for time series of yields I assume that the factors $f_t$ are generated by a cointegrated vector autoregressive process. In this case the factors are given by the error-correction specification of the VAR($k$) process

$$\Delta f_{t+1} = \beta' f_t + \sum_{j=1}^{k-1} \Gamma_j \Delta f_{t-j} + \zeta_t, \quad \zeta_t \sim N(0, Q_\zeta),$$

(5.8)

where $\Delta$ is the difference operator ($\Delta f_{t+1} = f_{t+1} - f_t$) and $r \times s$ matrices $\beta$ and $\gamma$ have full column rank. The remaining matrices are defined as in (5.6). The matrices $\beta$ and $\gamma$ are usually subject to a set of identifying and normalizing restrictions. Let $\Gamma(z)$ denote the characteristic polynomial associated with the process (5.8). To ensure that the factors are integrated of order one and $\gamma' f_t$ is stationary we need to impose the additional restrictions that all solutions to the equality $|\Gamma(z)| = 0$ are outside or on the unit circle and that

$$\det [\beta'_\perp \Gamma(1) \gamma'_\perp] \neq 0,$$

where $\beta_\perp$ and $\gamma_\perp$ are $r \times (r-s)$ matrices with their column spaces spanning the null spaces of $\beta'$ and $\gamma'$, respectively. A more detailed discussion of error-correction models is given by Johansen (1995). I will refer to (5.8) as the CVAR($k$) model and to the dynamic factor model with CVAR($k$) factors as the DFM-CVAR($k$) model.

All elements of $f_t$ are nonstationary processes when $s < r$. From a practical perspective it is advantageous to have a specification that decomposes the $r$ factors in $f_t$ into $s$ stationary and $r-s$ nonstationary components. For this purpose I propose an alternative but observationally equivalent specification for $f_t$ via factor rotation. The alternative specification changes the interpretation of the factors but does not alter the dynamic properties of the model. The factors of the new model are given by

$$\tilde{f}_t = \begin{pmatrix} \tilde{f}_t^N \\ \tilde{f}_t^S \end{pmatrix} = \begin{bmatrix} \beta_\perp \\ \gamma' \end{bmatrix} f_t,$$

(5.9)

where $(r-s) \times 1$ vector $\tilde{f}_t^N$ is the nonstationary component and $s \times 1$ vector $\tilde{f}_t^S$ is the stationary component. The process $\tilde{f}_t$ can be represented by the CVAR($k$) model

$$\Delta \tilde{f}_{t+1} = A \tilde{f}_t + \sum_{j=1}^{k-1} \tilde{\Gamma}_j \Delta \tilde{f}_{t-j} + \tilde{\zeta}_t, \quad \tilde{\zeta}_t \sim N(0, \tilde{Q}_\zeta),$$

(5.10)
where the \( r \times r \) matrix \( A \) is given by
\[
A = \begin{pmatrix} 0 & 0 \\ 0 & \Gamma_0 \end{pmatrix},
\]
and \( \Gamma_0, \ldots, \Gamma_{k-1} \) and \( \bar{Q}_\zeta \) are functions of \( \beta, \gamma, \Gamma_1, \ldots, \Gamma_{k-1} \) and \( Q_\zeta \). To ensure that the model remains observationally equivalent we also need to construct a new loading matrix \( \bar{\Lambda} \) by rotating the original matrix \( \Lambda \) into
\[
\bar{\Lambda} = \begin{bmatrix} \bar{\Lambda}^N & \bar{\Lambda}^S \end{bmatrix},
\]
where \( N \times (r - s) \) matrix \( \Lambda^N \) and \( N \times s \) matrix \( \Lambda^S \) are both of the form (5.5). Note that the rotation transfers parameters from the transition equation to the factor loading matrix.

The observation equation is given by
\[
y_t = \mu_y + \bar{\Lambda}^N \bar{f}^N_t + \bar{\Lambda}^S \bar{f}^S_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H),
\]
for \( t = 1, \ldots, n \). I use this specification to estimate the nonstationary models, but to facilitate comparison with the stationary models I generally present the results for the model with factors given by (5.8) and loadings given by (5.5). Note that the maximum likelihood estimators for this second model can be easily obtained from the estimators for the model given by (5.13). The factors \( \bar{f}_t \) can be written in the form (5.3) by choosing the state vector as follows \( \alpha_t = (\bar{f}_t' \Delta \bar{f}_t' \cdots \Delta \bar{f}_{t-k+1}')' \) and, for \( k > 1 \), taking the system matrices \( Q = \bar{Q}_\zeta, U = (I_r \ 0 \cdots \ 0) \) and
\[
T = \begin{pmatrix} I_r + A & \Gamma_{1:k-2} \ \Gamma_{k-1} \\ A & \Gamma_{1:k-2} \ \Gamma_{k-1} \\ 0 & I_r(k-1) \ 0 \end{pmatrix}, \quad R = \begin{pmatrix} I_r & I_r & 0 \cdots & 0 \end{pmatrix}',
\]
where \( \bar{\Gamma}_{i,j} = (\bar{\Gamma}_i \cdots \bar{\Gamma}_j) \) for \( i, j = 0, \ldots, k-1 \) and \( i < j \). The representation for \( k = 1 \) follows immediately.

If \( \alpha_t \) is nonstationary we cannot specify \( \alpha_1 \) as in section 5.2.1. Rosenberg (1973) advocates to consider the nonstationary part of the initial state as an additional set of parameters which can be estimated by maximum likelihood methods. If we choose specification (5.10) for the factors, only the first \( r - s \) elements of \( \alpha_t \) are nonstationary. Therefore, if we adopt the approach of Rosenberg (1973) we can set the first \( r - s \) rows and columns of the variance of \( \alpha_1, P_1 \), to zero. The remaining rows and columns of \( P_1 \) are set equal to the variance of
the invariant distribution of the stationary elements of $\alpha_t$. Further, since $\mu_y$ is unrestricted, I also set the means of the elements of $\alpha_t$ corresponding to nonstationary components to zero. I use this approach in the empirical section of this chapter. Alternatively, we can choose a reference or diffuse prior for the initial state of the nonstationary components, see the discussion in Durbin and Koopman (2001, Chapter 5). In this case we need to restrict the first $r - s$ elements of $\mu_y$ to be zero.

5.2.3 Parameter Estimation and Signal Extraction

The dynamic factor model consisting of (5.1), (5.2) and (5.3) can be regarded as a special case of the linear state space model. Given the set of system matrices, we can use the Kalman filter and related methods to evaluate minimum mean square linear estimators (MMSLE) of the state vector at time $t$ given the observation sets $\{y_1, \ldots, y_{t-1}\}$ (prediction), $\{y_1, \ldots, y_t\}$ (filtering) and $\{y_1, \ldots, y_n\}$ (smoothing). A detailed treatment of state space methods is given by Durbin and Koopman (2001).

For a given set of system matrices the Kalman filter can also be used to evaluate the loglikelihood function via the prediction error decomposition. The maximum likelihood estimators of the model parameters can then be obtained by numerical optimization. To generate the results in this chapter I used the BFGS algorithm to perform the optimization, see for example Nocedal and Wright (1999). An alternative approach would be to use the EM algorithm as developed for state space models by Watson and Engle (1983).

Efficient versions of the Kalman filter have been developed for multivariate models, see for example, Koopman and Durbin (2000). Furthermore, we can achieve considerable computational savings using the methodology of Jungbacker and Koopman (2008). Their method first maps the set of observations $y_t$ into a set of vectors which have the same dimensions as the latent factors $f_t$ in (5.2). We can then apply the Kalman filter to a typically much lower dimensional set of ‘observations’. I refer the reader to Chapter 4 for an in-depth discussion of these results. I have implemented this approach for all models discussed in sections 5.3 and 5.4. These efficient Kalman filter methods are also used to evaluate the closed form expressions for the score function given in Koopman and Shephard (1992) and Jungbacker and Koopman (2008). Despite the large number of parameters involved, this combination of efficient Kalman filter methods and analytical score allows us to estimate all the models considered in this chapter in a matter of seconds.
5.3 Dynamic Factor Model with Smooth Factor Loadings

The observation equation (5.1) of the dynamic factor model for the yields $y_t(\tau_i)$ can be written as

$$y_t(\tau_i) = \mu_y,i + \sum_{j=1}^{r} \lambda_{ij} f_{jt} + \varepsilon_{it}, \quad t = 1, \ldots, T, \quad i = 1, \ldots, N,$$

(5.14)

where $\mu_y,i$ is the intercept, $\lambda_{ij}$ is the factor loading of maturity $i$ and factor $j$, $f_{jt}$ is the $j$th element of $f_t$ and $\varepsilon_{it}$ is the $i$th element of $\varepsilon_t$. I propose to specify the model in terms of a set of functions $g_1(\cdot), \ldots, g_r(\cdot)$ defined on the interval $[\tau_1, \tau_n]$ and then define the factor loadings as follows

$$\lambda_{ij} = g_j(\tau_i), \quad i = 1, \ldots, N, \quad j = 1, \ldots, r.$$  

(5.15)

Since the yield curves tend to be relatively smooth functions of time to maturity and the disturbances $\varepsilon_{it}$ are mutually uncorrelated, it is reasonable to assume that the loading functions $g_1(\cdot), \ldots, g_r(\cdot)$ are also smooth functions of time to maturity $\tau_i$. In this section I develop a dynamic factor model that is directly specified in terms of a set of functions $g_1(\cdot), \ldots, g_r(\cdot)$. The model provides means to let the factor loadings be smooth over time to maturity in a straightforward and intuitive manner. I further show how to test for the validity of smoothness restrictions using a series of Wald tests. The resulting model will be referred to as the smooth dynamic factor model (SDFM).

5.3.1 Model Specification

The main assumption of the smooth dynamic factor model is that the loading functions are specified as cubic splines. Specifically, I assume that there is a set of $r$ cubic splines $g_1(\cdot), \ldots, g_r(\cdot)$ defined on $[\tau_1, \tau_n]$ such that $\lambda_{ij} = g_j(\tau_i)$ for $i = 1, \ldots, N$ and $j = 1, \ldots, r$. Such cubic splines can capture a wide variety of different shapes. It is therefore reasonable to assume that even if the loading functions of the data generating process are not truly cubic splines, they can still be very closely approximated by functions of this form.

A cubic spline is specified by selecting a set of knots and choosing the function values of the spline at each of these knots. The cubic spline is uniquely defined as the function that is (i) equal to a third-order polynomial between the knots and (ii) twice continuously differentiable at the knots; see, for example, Monahan (2001). It is therefore the location of the knots that determines how the factor loadings behave for varying maturities. In case
5.3. DYNAMIC FACTOR MODEL WITH SMOOTH FACTOR LOADINGS

a small number of knots for a column of Λ is chosen, the corresponding loadings lie on the same cubic polynomial for a considerable number of adjacent maturities. In case we choose the set of knots equal to the set of maturities, no restrictions are imposed on the factor loadings \( \lambda_{ij} \) and the model reduces to the general dynamic factor model of section 5.2. We can choose a different set of knots for each of the functions \( g_1(\cdot), \ldots, g_r(\cdot) \). To ensure a well-defined spline on the interval \([\tau_1, \tau_N]\), the first and last maturities \( \tau_1 \) and \( \tau_N \) are taken as knots for all functions \( g_1(\cdot), \ldots, g_r(\cdot) \).

In practice, it is convenient to formulate the splines that determine the factor loadings as linear functions of a set of parameters which correspond to the unknown values of the splines at their respective knots. Denote the \( k \)th knot for the \( j \)th column of \( \Lambda \) by \( s^j_k \) and suppose that the knots for each column are ordered by time to maturity, that is

\[
\tau_1 = s^j_1 < \cdots < s^j_{K_j} = \tau_N, \quad j = 1, \ldots, r,
\]

where \( K_j \) is the number of knots for the \( j \)th column of \( \Lambda \). Following Poirier (1976), we can specify the loading function \( g_j(\tau_i) \) as a linear function, that is

\[
g_j(\tau_i) = w_{ij} \delta_j, \quad \delta_j = \begin{pmatrix} g_j(s^j_1) \\ \vdots \\ g_j(s^j_{K_j}) \end{pmatrix}, \quad j = 1, \ldots, r, \tag{5.16}
\]

where \( w_{ij} \) is a \( 1 \times K_j \) vector that only depends on the location of the knots \( s^j_1, \ldots, s^j_{K_j} \) and \( \delta_j \) is treated as a \( K_j \times 1 \) unknown parameter vector that needs to be estimated. The resulting factor loading matrix \( \Lambda \) of the smooth dynamic factor model is given by

\[
\Lambda = \begin{bmatrix} W_1 \delta_1 & \cdots & W_r \delta_r \end{bmatrix}, \quad W_j = \begin{pmatrix} w_{ij} \\ \vdots \\ w_{Nj} \end{pmatrix}, \tag{5.17}
\]

for \( j = 1, \ldots, r \). Although the specification (5.17) of \( \Lambda \) is more parsimonious, we still need to impose restrictions on \( \Lambda \) such as in (5.5) to ensure that the model is identified.

5.3.2 Selecting Knots via a Wald Test Procedure

In this section I develop a statistic to test if a subset of knots is significantly contributing to model fit. I use the test statistic to systematically search for a suitable set of restrictions for the loading matrix \( \Lambda \) in the smooth dynamic factor model.
Suppose we have \( r \) sets of knots \( S_j = \{s_{j1}, \ldots, s_{jK_j}\} \) for \( j = 1, \ldots, r \). I denote the class of all splines associated with the set \( S_j \) by \( G_j \). We assume that the sets \( S_1, \ldots, S_r \) are sufficiently rich to capture the form of \( \Lambda \) of the data generating process. More formally, if \( g_j(\cdot) \) denotes the function that generates the \( j \)th column of \( \Lambda \) in the true data generating process, then \( g_j(\cdot) \in G_j \) for \( j = 1, \ldots, r \). My aim is to test whether a subset of knots can be removed from a given set \( S_j \).

Consider a new set of \( K_j^* \) knots denoted by \( S_j^* \) such that \( S_j^* \) is a subset of \( S_j \). Further assume that the set \( S_j^* \) is strictly smaller than \( S_j \) and therefore \( K_j^* < K_j \). Denote the family of splines determined by the knots in \( S_j^* \) by \( G_j^* \). It follows that \( G_j^* \subset G_j \).

For our purpose, the null-hypothesis \( H_0 \) and the alternate hypothesis \( H_1 \) are given by

\[
H_0 : g_j(\cdot) \in G_j^*, \quad H_1 : g_j(\cdot) \notin G_j^*.
\] (5.18)

The null-hypothesis is specifically for the \( j \)th spline (or the \( j \)th column of \( \Lambda \)) but it can be extended to more general settings and to all \( r \) splines jointly. Each spline function in \( G_j \) is uniquely determined by the value of \( \delta_j \) which represents the values of \( g_j(\cdot) \) at the knots in \( S_j \). It can therefore be shown that testing the hypotheses of (5.18) is equivalent to testing linear restrictions on \( \delta_j \).

Denote the \( j \)th column of \( \Lambda \) by \( g_j(\tau) = [g_j(\tau_1), \ldots, g_j(\tau_N)]' \). Then the null-hypothesis can be written as

\[
g_j(\tau) = W_j^* \delta_j^*,
\] (5.19)

where \( W_j^* \) is the spline weight matrix defined in (5.17) for set of knots \( S_j^* \) and \( \delta_j^* \) contains the values of the spline at the knots in \( S_j^* \). Since we assumed that \( g_j(\cdot) \) is an element of \( G_j \) we can also write \( g_j \) in terms of \( W_j \), the weight matrix associated with \( S_j \),

\[
g_j(\tau) = (W_j)_{\delta_j} \delta_j^*, \quad \delta_j^* = \begin{pmatrix} \delta_j^* \\ \delta_{j\setminus*} \end{pmatrix},
\] (5.20)

where matrix \( W_{j\setminus*} \) consist of columns of the spline weight matrix \( W_j \) that correspond to knots that are in \( S_j \) but not in \( S_j^* \), matrix \( W_j_{\delta_j} \) consists of the (remaining) columns in \( W_j \) corresponding to knots in \( S_j^* \) only and \( \delta_{j\setminus*} \) is a vector containing the value of \( g_j(\cdot) \) at the knots in \( S_j \) that are not in \( S_j^* \). Since a spline is uniquely determined by its value at the knots, the two expressions in (5.19) and (5.20) are equivalent if and only if

\[
\delta_{j\setminus*} = B_j \delta_j^*,
\]
where matrix $B_j$ consists of rows of $W_{j\lambda}$ corresponding to knots at maturities that are in $S_j$ but not in $S_j^*$. The hypotheses in (5.18) reduce to the linear hypotheses

$$H_0 : R_j\delta_j^\dagger = 0, \quad H_1 : R_j\delta_j^\dagger \neq 0, \quad R_j = (B_j - I).$$

Testing linear restrictions of the form (5.21) is standard in the context of maximum likelihood estimation; see, for example, Engle (1984). For our purposes, a Wald test is particularly convenient. Denote by $\hat{\delta}_j^\dagger$ the maximum likelihood estimator of $\delta_j^\dagger$ and by $\hat{V}_j$ a consistent estimator of the asymptotic variance of $\sqrt{nN}(\hat{\delta}_j^\dagger - \delta_j^\dagger)$. Under the null-hypothesis we have

$$n \cdot N \cdot \hat{\delta}_j^\dagger' R_j' \hat{V}_j^{-1} R_j \hat{\delta}_j^\dagger \overset{\text{d}}{\sim} \chi^2(K_j - K_j^*),$$

where $K_j - K_j^*$ is the number of restrictions imposed under the null-hypothesis. In practice a suitable estimator $\hat{V}_j$ can be constructed from the Hessian matrix of the log-likelihood function at the maximum likelihood estimator for $\delta_j^\dagger$.

The most important special case of (5.22) is the situation where $K_j - K_j^* = 1$, meaning that $S_j$ and $S_j^*$ differ by a single knot. I propose to use this test statistic to select the number of knots and their location by means of an iterative general-to-specific approach. At each step I calculate for all the knots in each column a Wald test with the null-hypothesis that the knot is not needed to form the true vector of factor loadings. I then remove the knot that has the smallest non-significant statistic among all the knots used to construct the loading matrix. The procedure is repeated until all selected knots have a statistically significant statistic. I start this iterative testing process with the unrestricted dynamic factor model.

### 5.3.3 A General Version of the SDFM

In this section I focussed on the application of the smooth dynamic factor model to yield curve data. However, this framework has a much wider applicability. We can use this model for any multivariate time series where we observe a smooth function that varies stochastically over time. Panels of implied volatilities, calculated from call and put contracts on a stock or index with different strikes, are examples of such data sets. These volatility smiles vary over time but tend to be smooth functions of time to maturity. With a slightly more general version of the model discussed in section 5.3.1 we can also handle a whole different class of problems.

Suppose we have an $N$ dimensional time series $z_1, \ldots, z_n$, where $z_t = (z_{1t}, \ldots, z_{Nt})$ for $t = 1, \ldots, n$ and we model this time series using a dynamic factor model with $r$ underlying
latent factors. Even if there is no smooth functional relationship apparent between the elements of \((z_{1t}, \ldots, z_{Nt})\), we might still be able to model the time series very effectively using a SDFM. Suppose that \(z_t\) is a very large time series panel containing house prices and let \(f_{1t}\) be the factor representing the business cycle. It is likely that houses that are alike have very similar factor loadings for \(f_{1t}\). We can model this by assuming that

\[
h_{it} = g_1(p_{it}), \quad i = 1, \ldots, N, \quad t = 1, \ldots, n,
\]

where \(h_{it}\) is the factor loading for house \(i\) and factor 1 at time \(t\), \(p_{it}\) is a regression variable that indicates the type of house and \(g_1(\cdot)\) is a smooth function defined for all values of \(p_{it}\). The variable \(p_{it}\) might for example contain the last price at which house \(i\) was sold. Just as before we can impose the smoothness restriction on the factor loadings function by assuming that \(g_1(\cdot)\) is a cubic spline with a limited set of knots. The general form of this type of SDFM is as follows

\[
z_{it} = \mu_{z,i} + \sum_{j=1}^{r} g_j(x_{ijt}) f_{jt} + \zeta_{it},
\]

for \(t = 1, \ldots, n\) and \(i = 1, \ldots, N\) where \(\mu_{z,i}\) is the intercept, \(x_{ijt}\) are regression variables for \(i = 1, \ldots, N, j = 1, \ldots, r\) and \(t = 1, \ldots, n\), \(g_1(\cdot), \ldots, g_r(\cdot)\) are cubic splines and \(\zeta_t = (\zeta_{1t}, \ldots, \zeta_{Nt})'\) are serially uncorrelated Gaussian random variables for \(t = 1, \ldots, n\). Note that this reduces to the model presented in section (5.3.1) if we set \(x_{ijt} = \tau_i\) for \(j = 1, \ldots, r\) and \(t = 1, \ldots, n\). This type of model can be especially useful for very large datasets, since it allows us to greatly reduce the number of parameters in the loading matrix without having to impose the potentially unrealistic assumption that large sets of loadings are equal.

### 5.4 Dynamic Factor Models for the Term Structure of Interest Rates

In this section I review a number of alternative models for the term structure of interest rates that have appeared in the literature. These models can all be regarded as special cases of the general formulation (5.1) – (5.3) with different restrictions imposed on the loading matrix \(\Lambda\). For some models, restrictions on the dynamics of the factors and the mean vector \(\mu_y\) are also required. I consider both the stationary specification for \(f_t\) as in (5.6) as well as the nonstationary specification for \(f_t\) as in (5.8).
5.4. DYNAMIC FACTOR MODELS FOR THE TERM STRUCTURE OF INTEREST RATES

5.4.1 Functional Signal Plus Noise Model

The functional signal plus noise (FSN) model is recently proposed by Bowsher and Meeks (2008) as a promising way to model the term structure. Their model is also based on cubic splines, just as the model of section 5.3, but it is used in a different and less flexible way. Consider $S_f$ as a set of $r$ knots and let $W_f$ denote the $N \times r$ spline weight matrix of Poirier (1976). The spline function is then defined by $g_f(\tau) = W_f \delta_f$ where vector $\delta_f$ contains the values of the spline function at the knot positions in $S_f$ and is treated as a parameter vector. Instead of using the spline function $g(\cdot)$ to smooth the loadings in each column of $\Lambda$, as proposed in the previous section, the spline can also be used to smooth the yield curve directly. In this case, the loading matrix $\Lambda$ is set equal to the weight matrix $W$ and parameter vector $\delta_f$ is replaced by $f_t$. As a result Bowsher and Meeks (2008) obtain a time-varying cubic spline function for the yield curve. The FSN model is then given by

$$y_t = \mu_y + W f_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, H),$$

(5.23)

where $\mu_y$ is the vector of intercepts, $f_t$ is the $r$-dimensional factor process and $H$ is assumed diagonal. The observed yield curve $y_t$ is now a noisy observation of an unobserved “true” term structure which is modelled by a stochastically time-varying cubic spline function. Finally I note that by construction the weight matrix $W$ has the same form as $\Lambda$ in (5.5). The rows of $W$ that correspond to knots are equal to rows of the identity matrix.

Bowsher and Meeks (2008) consider the CVAR($k$) specifications for the unobserved factor $f_t$ with additional restrictions imposed on the cointegration vectors. In this chapter I consider both stationary as well as nonstationary specifications for $f_t$. In case of the nonstationary CVAR($k$) specification, we assume that $f_t$ is of the form (5.10). The decomposition of $f_t$ into stationary and nonstationary components is achieved as in (5.9) where $f_t$ is transformed to $\bar{f}_t$ which consists of a nonstationary part $\bar{f}_t^N$ and a stationary part $\bar{f}_t^S$. The loading matrix for $\bar{f}_t$ is then given by $\bar{\Lambda} = WL$ where $L$ is the $r \times r$ matrix that transforms the factors $\bar{f}_t$ to the process $f_t$. This matrix $L$ contains parameters that need to be estimated and is of the same form as $\bar{\Lambda}$ in (5.12). This decomposition of $f_t$ is useful for interpretation purposes and for the exact handling of the initial state in the implementation of the Kalman filter and related methods.

When certain restrictions are imposed on the smooth dynamic factor model of section 5.3, it reduces to the FSN model. The key restriction is that all sets of knots $S_j$ for the columns of $\Lambda$ are set equal and that the number of knots is equal to the number of factors. The restriction that the number of knots equals the number of factors in $f_t$ is strong in
practice. For example, Bowsher and Meeks (2008) find that 6 or 7 knots are required to adequately fit the shapes of the term structure typically observed in financial markets. The FSN model therefore requires a vector \( f_t \) with at least 6 factors. This number contrasts sharply with empirical studies of, for example, Litterman and Scheinkman (1991) who argue that 3 factors are sufficient to describe the dynamics of the term structure. The SDFM of section 5.3 has the advantage that the number of factors and the number of knots can be chosen separately and the different sets of knots can be selected more flexibly. Furthermore, a general statistical methodology is provided for the selection of the knots. As a result, we can obtain a better fit using a relatively small number of factors.

5.4.2 Nelson-Siegel Model

In an important contribution Nelson and Siegel (1987) have shown that the term structure can surprisingly well be fitted by a linear combination of three smooth functions. The Nelson-Siegel yield curve, denoted by \( g_{ns}(\tau) \), is given by

\[
g_{ns}(\tau) = \xi_1 + \lambda^S(\tau) \cdot \xi_2 + \lambda^C(\tau) \cdot \xi_3, \tag{5.24}
\]

where

\[
\lambda^S(\tau) = \frac{1 - e^{-\lambda \tau}}{\lambda \tau}, \quad \lambda^C(\tau) = \frac{1 - e^{-\lambda \tau}}{\lambda \tau} - e^{-\lambda \tau}, \tag{5.25}
\]

and where \( \lambda, \xi_1, \xi_2 \) and \( \xi_3 \) are treated as parameters. The yield curve depends on these parameters which can be estimated by a least squares method based on the nonlinear regression model

\[
y_t(\tau_i) = g_{ns}(\tau_i) + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, n,
\]

where \( u_{it} \) is noise with zero mean and possibly different variances for different time to maturities \( \tau_i \). One of the attractions of the Nelson-Siegel curve is that the \( \xi \) parameters have a clear interpretation. The parameter \( \xi_1 \) clearly controls the level of the yield curve. The parameter \( \xi_2 \) can be associated with the slope of the yield curve since its loading \( \lambda^S(\tau) \) is high for a short maturity \( \tau \) and low for a long maturity. The loadings \( \lambda^C(\tau) \) for different time to maturities \( \tau \) form a U-shaped function and therefore \( \xi_3 \) can be interpreted as the curvature parameter of the yield. The decomposition of the yield curve into level, slope and curvature factors has also been highlighted by Litterman and Scheinkman (1991).

The Nelson-Siegel yield curve can also be incorporated in a dynamic factor model by
treating the \( \xi \) parameters as factors. We obtain

\[
y_t = \mu_y + \Lambda_{ns} f_t + \epsilon_t, \quad \epsilon_t \sim NID(0, H),
\]

where \( f_t \) is a \( 3 \times 1 \) vector \( (r = 3) \) and \( H \) is a diagonal variance matrix. The loading matrix \( \Lambda_{ns} \) consists of the three columns \( (1, \ldots, 1)' \), \( \lambda^S(\tau) \) and \( \lambda^C(\tau) \) respectively. Note the similarity between the Nelson-Siegel dynamic factor model and the smooth dynamic factor model (5.14). The slope and curvature loadings in the Nelson-Siegel model both depend on a single parameter \( \lambda \) and this framework is therefore somewhat more restrictive than the SDFM.

The dynamic factor representation of the Nelson-Siegel model is proposed by Diebold, Rudebusch, and Aruoba (2006). Their specification is slightly different as they set \( \mu_y \) in (5.26) to zero and include an intercept in the specification of the factors \( f_t \). Furthermore, they specify a stationary vector autoregressive model similar to (5.6) for the 3-dimensional factor \( f_t \). I will also consider a nonstationary Nelson-Siegel model with \( f_t \) of the form (5.10) and with \( f_t \) transformed to \( \bar{f}_t \) as in (5.9). The loading matrix for \( \bar{f}_t \) is then given by \( \bar{\Lambda} = \Lambda_{ns} L \) where \( L \) is the matrix that transforms \( \bar{f}_t \) to \( f_t \), see also the discussion of the previous section.

### 5.4.3 Arbitrage-free Nelson-Siegel Model

Absence of arbitrage opportunities imposes strict restrictions on the stochastic properties of the yield curve; see, for example, the discussion in Cox, Ingersoll, and Ross (1985). The dynamic factor models discussed so far do not satisfy such restrictions. This is unsatisfactory if we believe that such arbitrage possibilities do not exist in the real world. In this case imposing the no-arbitrage restrictions on the model might improve its performance. This was the motivation for Christensen, Diebold, and Rudebusch (2007) to develop an arbitrage-free version of their Nelson-Siegel dynamic factor model discussed in the previous section.

If the arbitrage-free Nelson-Siegel model is the true underlying data generating process then each \( y_t(\tau_i) \) is given by

\[
y_t(\tau_i) = \mu_{y,i} + f_{1t} + \lambda^S(\tau_i)f_{2t} + \lambda^C(\tau_i)f_{3t},
\]

where \( \mu_{y,i} \) is a correction term that is a deterministic function of the parameters determining the dynamics of the factors, see Christensen, Diebold, and Rudebusch (2007, p. 18) for details. The absence of measurement noise in (5.27) implies that the corrected yields can be exactly fitted using only \( f_{1t}, f_{2t} \) and \( f_{3t} \). Since observed yields never satisfy this restriction in practice it is customary to include measurement errors \( \epsilon_t \) in the model. Christensen,
Diebold, and Rudebusch (2007) model the factors $f_{1t}$, $f_{2t}$ and $f_{3t}$ in continuous-time as a multivariate Gaussian process. For evenly spaced observations in discrete time this process can be written as a stationary VAR(1)

$$f_{t+1} = \mu^*_f + \Gamma^*_0 (f_t - \mu^*_f) + \zeta_t, \quad \zeta_t \sim N(0, Q^*_\zeta),$$

where $\mu^*_f$ is a $3 \times 1$ mean vector, $\Gamma^*_0$ is the $3 \times 3$ autoregressive coefficient matrix and $Q^*_\zeta$ is the $3 \times 3$ variance matrix. For estimation purposes, it is in practice necessary to formulate the VAR(1) process in terms of the parameters of the original continuous-time process as these parameters appear in $\mu_{yi}$. I refer the reader to Christensen, Diebold, and Rudebusch (2007) for the functional relationship between these parameters and the VAR(1) matrices. I consider the most general form of the model proposed by Christensen, Diebold, and Rudebusch (2007). This model imposes no restrictions on the intercept $\mu^*_y$, the transition matrix $\Gamma^*_0$ and the variance matrix $Q^*_\zeta$. Note that this model can be seen as a restricted version of the standard Nelson-Siegel model. Specifically, the arbitrage-free Nelson-Siegel model imposes $N - 3$ restrictions on the intercept $\mu_y$ and restricts the factors to be generated by a VAR(1).

### 5.4.4 Gaussian Affine Term Structure Model

Let $r_t$ denote the short rate. The short rate can be thought of as the yield of a zero-coupon bond with infinitesimally short time to maturity. For models in the class of affine term structure (AfTS) models, Duffie and Kan (1996) assume that the short rate $r_t$ is an affine function of an unobserved $r \times 1$ dimensional stochastic process $f_t$

$$r_t = g_1 + g_2^t f_t,$$

where $g_1$ is a scalar parameter and $g_2$ is a $r \times 1$ vector of parameters. Using a no-arbitrage argument, they proceed to show that if the factors belong to a class of diffusions with affine volatility structure and the market price of risk for each factor is proportional to its volatility, the yields are given by

$$y_t(\tau) = F_1(\tau) + F_2(\tau)r_t,$$

where the functions $F_1(\tau)$ and $F_2(\tau)$ can be obtained from a set of ordinary differential equations, depending on the parameters governing the factor dynamics.

The class of affine term structure models includes a broad range of Gaussian and non-Gaussian specifications. In chapter I focus on the Gaussian case. For the Gaussian specifications it is possible to obtain closed form expressions for $F_1(\tau)$ and $F_2(\tau)$; see equations
5.5 DATA DESCRIPTION

(3.9) and (3.10) in Duffie and Kan (1996) or equations (9) and (10) in De Jong (2000). In discrete time we can write the factors as a VAR(1) process, after imposing suitable identifying restrictions, see De Jong (2000). This VAR process is of zero mean and has a diagonal transition matrix. Note that this implies that \( g_1 \) is the only free parameter in the intercept. Just as for the AFNS model it is unlikely that the observed term structure of interest rates can be fitted exactly by the relation (5.28). In practice I therefore include a vector of independent Gaussian measurement errors. These measurement errors are allowed to have a different variance for each maturity. The resulting factor model is clearly a restricted version of the DFM model of section 5.2. For more details on the formulation of the Gaussian AFTS model in dynamic factor form I refer the reader to De Jong (2000).

5.5 Data Description

The empirical study of the next section is based on the same data set considered in Diebold and Li (2006) who constructed a monthly data set of zero yields from the CRSP unsmoothed Fama and Bliss (1987) forward rates. I refer to Diebold and Li (2006) for a detailed discussion of the method that is used for the creation of this data set. I follow Diebold and Li (2006) in considering a subset of the data. The resulting data set consists of 17 maturities over the period from January 1985 up to December 2000. The maturities I analyze are 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108 and 120 months. This dataset has also been considered by Diebold, Rudebusch, and Aruoba (2006), Christensen, Diebold, and Rudebusch (2007) and Bowsher and Meeks (2008), though sometimes for different sample periods and number of maturities.

In Panel A of Figure 5.1 I present a three-dimensional plot of the data set. The data plot suggests the presence of an underlying factor structure. Although yields vary wildly over time for each of the maturities there is a strong common pattern in the way in which the 17 series develop over time. For most months, the yield curve is an upward sloping function of time to maturity. The overall level of the yield curve is mostly downward trending in the sample period. These findings are supported by the time series plots in Panel B of Figure 5.1. In these plots I also observe that volatility tends to be lower for the yields of bonds with a longer time to maturity.

Table 5.2 provides summary statistics for the dataset. For each each of the 17 time series I report mean, standard deviation, minimum, maximum and a selection of autocorrelation and partial-autocorrelation coefficients. The summary statistics confirm that the yield curve tends to be upward sloping and that volatility is lower for rates on the long end of the yield
In addition, there is a very high persistence in the yields: the first order autocorrelation for all maturities is above 0.95 for each maturity. Even the twelfth autocorrelation coefficient can be as high as 0.57. The partial-autocorrelation function suggests that autoregressive processes of limited lag order will fit the data well since only the first coefficient is significant for most maturities (to preserve space I do not display all coefficients). In Panel B of the Table 5.2 I present the sample correlations between yields of a selected number of maturities. The correlations are all well above 0.5, in accordance with the strong common pattern in the movements of the different yields that we have observed in Figure 5.1.

From the data plot in Figure 5.1, we observe some large breaks in the time series, specifically for the months of May 1985 and October 1987. In both cases, the breaks are apparent in the yield for all maturities and we even observe a drop of more than 1.25% for one of the yields. Also, in both cases the breaks in the yields have lasted for at least a few months.

Finally, Table 5.2 also provides some information that is relevant for the question whether the yield series are stationary. For this purpose I report Augmented Dickey-Fuller tests for each of the series, see Dickey and Fuller (1979). At a 5% significance level we reject the null hypothesis of a unit root for only 3 out of the 17 time series. We reject the null hypothesis for none of the time series if the significance level is lowered to 1%. These findings suggest that a nonstationary dynamic factor model might be a better representation of the yield curve data than a stationary model.

### 5.6 Empirical Results

In this section I investigate whether the restrictions imposed by the models presented in sections 5.2 – 5.4 are supported by the data presented in section 5.5. The results of the empirical study are presented as follows. In section 5.6.1 I review the general assumptions that are applicable to all models. In section 5.6.2 I discuss the estimation results for the general DFM. In section 5.6.3 I report the estimation results for the SDFM that is based on a suitable set of smoothing restrictions for the factor loadings as obtained from the Wald test procedure. Section 5.6.4 discusses the estimation results for the NS and FSN models as well as the arbitrage-free AFNS and AfTS models. In section 5.6.5 I assess the in-sample fit of the different models by investigating the properties of the residuals. Finally, in section 5.6.6 I test the validity of the different restrictions by performing a set of likelihood ratio tests. In the remainder of this section I will refer to the models by their acronyms which are listed in Table 5.1. The dynamic specification for the factors in (5.6) is referred to as \( \text{VAR}(k) \) while the nonstationary specification in (5.10) is referred to as \( \text{CVAR}(k) \). For ease
of reference I present in Table 5.1 the most important details of all the models discussed in this empirical section.

5.6.1 General Model Assumptions

I start by discussing some general assumptions. To facilitate a fair comparison with the Nelson-Siegel model, I restrict all models to include a total of three latent factors, that is $r = 3$. I justify this assumption by referring to a growing number of studies that find three factors adequate for explaining most of the variation in the cross-section of yields, see e.g. Litterman and Scheinkman (1991), Bliss (1997) and Diebold and Li (2006). However, some other studies have recommended more factors, see the discussion in De Pooter (2007).

For the DFM, SDFM, NS and FSN models the choice of the factor dynamics is arbitrary. To keep the discussion general, I consider $\text{VAR}(k)$ as well as $\text{CVAR}(k)$ factors for these models. Further, I assume for the CVAR specification that there are two cointegrating vectors for the factors. This means that there is only one random walk present in the cross-section of yields. This is the same assumption as made by Bowsher and Meeks (2008) and is consistent with the findings in Hall, Anderson, and Granger (1992). I will make no assumption on the lag order of the CVAR and VAR processes. Instead, I determine the optimal lag order by minimizing the Akaike Information Criterium (AIC). In this empirical study I find that the dynamic properties of the factors do not depend on the functional form of the factor loadings.

To account for the large shocks in the yield curve data for the months of May 1985 and October 1987, I include two sets of dummies in each of the models. Since the shocks were persistent and influenced the entire yield curve I included the dummies as intercepts in the unobserved factors. This adds a total of six parameters to each of the model specifications.

5.6.2 Estimation Results for the DFM

In this section I discuss results obtained from the maximum likelihood estimation of the general dynamic factor model. In Table 5.3 I give the values of the maximized log-likelihood functions for the $\text{VAR}(k)$ and $\text{CVAR}(k)$ factor specifications together with the corresponding Akaike Information Criterium (AIC) values for $k = 1, 2, 3, 4$. The maxima of the log-likelihood functions for the models with $k = 2$ are considerably higher than the corresponding values for $k = 1$. The improvements from additional lags ($k = 3, 4$) are however much smaller. The $\text{VAR}(2)$ and $\text{CVAR}(2)$ factor specifications give the smallest AICs values. I will therefore restrict the discussion to these two specifications in the remainder of the section.
Since we are mainly interested in the restrictions imposed on the factor loadings, it is of interest to investigate how the estimated factor loadings change with different choices of $k$. In Panel A of Figure 5.2 I plot the estimated factor loadings for the DFM model with CVAR($k$) factor specifications, for $k = 1, \ldots, 4$, as functions of time to maturity. The loading matrix $\Lambda$ is restricted to be of the form (5.5), with the rows of the identity matrix placed at the rows associated with the maturities of 3, 30 and 120 months. The estimated loadings are nearly identical for different values of $k$. I found similar results for the stationary DFM models. I therefore conclude that the increase of the log-likelihood due to adding extra lags in the VAR and CVAR models for the latent factors is mainly due to a better fit of the time series dynamics. The fit of the yield curve is not affected by different lag orders.

To clarify the effect of an increasing $k$ on the dynamics, I also present the factor loadings for the dynamic factor model with factor specification (5.10). Here the two stationary factors are separated from the random walk component. The resulting loadings are presented in Panel B of Figure 5.2. We now see clear differences between the factor loadings when the lag order $k$ changes. The loadings for the first (nonstationary) component shifts downward as $k$ increases. This implies that yield variation that is explained in the CVAR(1) model by a nonstationary factor is captured by a highly persistent but stationary component in the CVAR($k$) model with $k > 1$. I interpret such shifts as additional evidence that a CVAR(1) model cannot adequately capture the observed yield curve dynamics.

The autoregressive coefficient matrices $\Gamma_j$ for VAR($k$) and CVAR($k$) processes for $j = 0, \ldots, k - 1$ are generally difficult to interpret especially when $k > 1$. I therefore choose to report eigenvalues of the estimated transition matrix $T$ of the state space representation 5.3. In Table 5.4 I present the eigenvalues of the matrix $T$ for the DFM-VAR(2) and DFM-CVAR(2). For both models we see two eigenvalues close to one, or equal to one in the nonstationary case, with no imaginary part and two sets of eigenvalues that do have an imaginary component. We can therefore view the factors as a weighted sum of two highly persistent autoregressive (AR) processes (or one AR process and a random walk) and two cyclical components. The presence of two highly persistent factors in the estimated model is in line with the preliminary findings in section 5.5. Since the highest eigenvalue for the VAR(2) process, 0.992, is very close to one, this process is in practice almost a CVAR process. This explains why the remaining eigenvalues of both models are close to each other as well. Throughout this empirical section we will see that all the VAR specifications are very close to being nonstationary. In practice this means that estimation results for stationary and nonstationary models tend to be very similar.
5.6.3 Imposing the SDFM Restrictions

In the previous section I concluded that the VAR(2) and CVAR(2) factor specifications best represent the yield curve data. Next I apply the methodology of section 5.3.2 to find a suitable set of smoothness restrictions for the factor loadings of the DFM model with VAR and CVAR specifications.

To ensure that the SDFM specification is identified we need to impose restrictions on the knots and the associated parameters. I choose to restrict the loading matrix to be of the form (5.5) where the rows of the identity matrix are placed in the rows corresponding to the 3, 30, and 120 months of maturities. Since our interpolating cubic spline framework requires a knot at the first and last maturity (3 and 120 months), this implies that the knot at 30 months cannot be removed in the course of the selection procedure. Of course, the procedure can be repeated when the knot at 30 months is moved to another time to maturity. After some experimentation, I have concluded that the main results are not sensitive to moving this knot to maturities in the neighborhood of 30 months.

In Table 5.5 I present the Wald test-statistics for each knot in the unrestricted model with a CVAR(2) specification for the factors. I only give results for the CVAR(2) specification as the statistics for the VAR(2) model are almost identical for reasons given at the end of section 5.6.2. At the start of the procedure 12 out of 42 loading coefficients (or knots) are significant at the 5% significance level. This suggests that the number of parameters can be reduced enormously without affecting the fit. However, the test statistics are highly correlated and removing one knot will generally change the statistics of the neighbouring knots considerably. I then proceed by sequentially removing the knot with the lowest Wald-statistic and re-estimating the model after each step. The procedure is terminated when all statistics are significant at the 5% significance level.

In Table 5.6 I present for both the SDFM-VAR(2) and SDFM-CVAR(2) models the Wald-statistics for the knots that have remained after the final step. The final knot selections for the stationary and nonstationary models are different. However, the distribution of the knots over the interval \([\tau_1, \tau_N]\) is similar for both models. To let a cubic spline fit a certain shape, the distribution of the knots is generally more important than the exact location of the knots. Furthermore, I find that the procedure is especially successful in fitting the first column of factor loadings. The original set of 14 loading parameters is reduced to four remaining knot parameters. In total I reduced the number of parameters in the loading matrix \(\Lambda\) by 20 for the nonstationary and 21 for the stationary model, a reduction of, say, 50 percent. In Figure 5.3 I show the spline estimates for the factor loadings of the final smooth dynamic factor models. For both SDFMs the factor loadings are smooth and close to the estimated
loadings for the general DFM model. We have seen that we can achieve almost identical loadings using a much smaller set of parameters. This confirms that the true factor loadings are subject to smoothness restrictions. The results for the CVAR and VAR specifications are almost identical.

The construction of the two SDFM specifications in this section required the estimation of respectively 20 and 21 different dynamic factor models each containing around 100 parameters. This may appear computationally intensive from the outset. However, the computationally efficient methods, discussed in section 5.2.3, make such a procedure computationally feasible even for larger models.

5.6.4 Estimation Results for Other Term Structure Models

In this section I present estimation results for the term structure models discussed in section 5.4. In Table 5.3 I report log-likelihood and corresponding AIC values for the NS and FSN models with both VAR($k$) and CVAR($k$) specifications for $k = 1, \ldots, 4$. To generate these results we first need to specify a set of knots to construct the loading matrix $W$ for the FSN model. The location of the knots are selected using the same methodology as adopted in Bowsher and Meeks (2008). For each possible set of three knots, I fitted a spline through all observed yield curves. I then chose the knot configuration that produced the lowest average mean square error. In this setting we only need to choose one knot, since the other two knots are fixed at the first and last maturities. The results of Table 5.3 are consistent with the results for the general dynamic factor model reported in section 5.6.2. For both models, the AIC criteria favour both the VAR and CVAR specifications with $k = 2$. This is also in accordance with the results of Bowsher and Meeks (2008) who, working with a similar data set for a larger number of maturities, find that the FSN model with CVAR(2) factors performs best in their class of models.

In Table 5.7 I report the maxima of the log-likelihood functions and their corresponding AIC values for the AFNS and AfTS models. These log-likelihood values can only be compared with their corresponding values in Table 5.3 and Table 5.3 for the stationary VAR specifications of the factors since AFNS and AfTS models are defined as strictly stationary models. The maximized log-likelihood values for the arbitrage-free models are considerably smaller when compared to the dynamic factor model, with and without smooth factor loadings, and compared to the Nelson-Siegel model. The difference with the FSN model is however less pronounced, especially for the AfTS.

In Table 5.4 I present the eigenvalues of the estimated transition matrices for the NS and FSN models. The nonstationary and stationary specifications for the factor produce similar
5.6. EMPIRICAL RESULTS

results. This finding is consistent with the results of the DFM and SDFM models. The
eigenvalues for the DFM, NS and FSN specifications are almost identical. Table 5.7 presents
the eigenvalues for the two arbitrage-free models. In both cases the estimated parameters
imply a high level of persistence in the factor dynamics.

Next I investigate whether the choice of the lag order in the autoregressive factor dy-
namics influences the estimates of the factor loadings. This matter only applies to the
Nelson-Siegel model. The factor loadings of the FSN model only depend on the selection of
the knots and are therefore by definition independent of the dynamics of the factors. The
two arbitrage-free models are only defined for a VAR(1) process and therefore we do not
need to consider this issue for these models. For the Nelson-Siegel model, the factor loadings
depend on a single parameter $\lambda$. In Table 5.8 I report the maximum likelihood estimates of
this parameter $\lambda$ for different lag orders in the dynamic process of $f_t$. I find that the esti-
mates of $\lambda$ are almost identical for stationary and nonstationary factor specifications. Also
the estimates of $\lambda$ vary little for different choices of the lag order. I conclude that maximum
likelihood estimation of the factor loadings (here all functions of $\lambda$) is not influenced by the
dynamic specification of the factors.

In Figure 5.4 I present the estimated factor loadings for the NS and FSN models with
CVAR(2) factors. I rotated the loadings such that the loading matrix $\Lambda$ is of the form (5.5).
The rotation facilitates easier comparison with the estimated loadings of the general DFM.
It is clear that the rotated Nelson-Siegel loadings are similar to the smoothed versions of
the DFM loadings. This finding is surprising given that the loadings in the Nelson-Siegel
model depend on a single parameter $\lambda$ while the dynamic factor model relies on 42 factor
loadings. For the FSN model, the factor loadings have generally the same shape and form
of the loadings obtained from the DFM and NS models. However, the individual factor
loadings are quite different.

In Figure 5.5 I display the factor loadings for the arbitrage-free models. As the loadings of
the AFNS are similar to the Nelson-Siegel model I only show the loadings of the AfTS model.
These loadings have been rotated in the same way so they have the same form as the DFM-
VAR(1) loadings. As in the case of the Nelson-Siegel model, the factor loadings in the AfTS
are close to the unrestricted estimates. It is revealing to see how close the AfTS loadings are
compared to the DFM loadings while the eigenvalues of its transition matrix (its VAR(1)
coefficient matrix) presented in Table 5.7 are quite different from those of the DFM, SDFM,
NS and FSN models (based on VAR(2) and CVAR(2) coefficient matrices) reported in Table
5.4. Also, the factor loadings are estimated simultaneously with the parameters that govern
the factor dynamics. These findings suggest that the penalty on an incorrect yield curve
specification (determined by the loadings) is much larger than the penalty on an incorrect
dynamic specification of the factors. During the search for the optimum of the likelihood
function, the optimization algorithm can therefore almost ignore the time series dimension.
The parameters are then chosen such that the fit of the cross-section is as good as possible.
The same conclusion can be drawn from Tables 5.3 and 5.3. The difference between the
maximized log-likelihood values of the DFM-CVAR(1) and DFM-CVAR(2) models is much
smaller than the difference between the values for the DFM-CVAR(2) and NS-CVAR(2)
models while the factor loadings are very similar in all specifications, see Figure 5.4.

5.6.5 Model Fit

In case all considered dynamic factor models are good approximations of the data generating
process, we expect that the residuals are not serially correlated. To verify this, I computed
Ljung-Box statistics for all models considered and for all 17 standardized residual series. The
null hypothesis of the Ljung-Box tests is that residuals are a white noise sequence. If the null
is rejected, dynamic variation in the residuals remains to be explained by a linear process.
For a selection of the models I present the results of this procedure in Table 5.9. We see that
the Nelson-Siegel model and to a lesser extent the DFM model are less successful in fitting
the dynamics in the interest rates associated with times to maturity of 9 and 12 months. For
these yields the FSN model seems to outperform the two other models. The FSN appears
to explain the variation in the yields quite well for all maturities with the exception of the
shortest 3 month maturity. It is not surprising that the Ljung-Box test statistics for the
SDFM model are very similar to those for the general dynamic factor model. The arbitrage-
free models perform a lot worse than the DFM, NS and FSN models, especially for the
maturities from 6 to 24 months. For both the AFNS and AfTS models we reject the null
hypothesis of the Ljung-Box test for 6 out of 17 residual series, at the 5% significance level.
This lack of fit is surprising for the AFNS model since it is very similar to the standard
Nelson-Siegel model. This can be partly explained from the restrictions of a VAR(1) process
for the factors of the AFNS model. For comparison, I also present the Ljung-Box statistics
for the standard Nelson-Siegel with VAR(1) factors in Table 5.9. In this case, the model
only differs from the AFNS model in its restrictions imposed on the intercept. The NS-
VAR(1) model also performs significantly worse in capturing the dynamics compared to the
NS-CVAR(2) model. I stress however that it is not caused by the stationarity restriction of
a VAR process. Ljung-Box statistics for the models with CVAR(2) factors and the VAR(2)
specifications are very similar. In case of the AfTS model, the poor Ljung-Box statistics
can be explained from the VAR(1) restriction and the estimates of the parameters in the
transition matrix, see also the discussion at the end of section 5.6.4.

5.6.6 Testing the Restrictions on the Factor Loadings

In section 5.4 I have argued that all existing models considered in this chapter can be viewed as dynamic factor models with smoothness restrictions imposed on the parameters. Since all models are nested in the general dynamic factor model, we can test the validity of these restrictions by means of a likelihood ratio test. For each model I test the null hypothesis that the restrictions are correct versus the alternative hypothesis that the true model is a general DFM model with the same dynamic specification for the factors. In case of the arbitrage-free models we indirectly test the restrictions on the intercepts. In Table 5.10 I present the established likelihood ratio tests. For the NS and FSN models, I focus on the VAR(2) and CVAR(2) specifications which we have selected by minimizing the AIC. Similar results are obtained for the other dynamic specifications. For the SDFM models I consider the final specifications presented in section 5.6.3.

The likelihood ratio tests suggest that we should strongly reject the restrictions implied by NS and FSN and those implied by the arbitrage-free specifications. For the FSN model, the differences between the estimated FSN and DFM loadings may have been indicative. However, the results reported in section 5.6.4 and 5.6.5 have shown that the loadings for the standard Nelson-Siegel model are very close to those estimated for the general DFM model. It is therefore somewhat surprising that the Nelson-Siegel restrictions are so strongly rejected. To find a possible explanation, I take a close look at the factor loading estimates for the DFM-CVAR(2) model as presented in Figure 5.6 together with their 95% confidence intervals. It is clear that the factor loadings of the DFM are estimated very precisely. The confidence intervals are very narrow, especially for the third column of the factor loading matrix \( \Lambda \). It implies that small perturbations in the maximum likelihood estimates will cause large changes in the log-likelihood value. This may explain why the Nelson-Siegel model is rejected by the likelihood ratio test although it is seemingly similar to the DFM model.

The strong rejection of these model specifications might suggest that we can not impose smoothness restrictions on the factor loadings in dynamic factor models for the term structure of interest rates. However, for both the stationary as well as the nonstationary specifications, we cannot reject the restrictions imposed by the SDFM at any reasonable significance level. This finding confirms that we can impose a certain amount of smoothness on the factor loadings in a dynamic factor model for the yield curve. We do however require more flexibility in specifying the factor loadings than provided by the NS, FSN, AFSN and AfTS models. I have shown that the SDFM is sufficiently flexible for this purpose. A limited number of
knots is required for each column of the factor loading matrix \( \Lambda \). The exception may be the third column of \( \Lambda \). It is possibly explained by the very narrow confidence interval of the third factor loadings presented in Figure 5.6. Only a slight misspecification for the third column of \( \Lambda \) severely penalizes the log-likelihood value.

5.7 Conclusion

In this chapter I have introduced a new methodology to construct dynamic factor models with smooth factor loadings. The proposed class of models is broad and flexible. I also present a statistical method to find a suitable model specification. I expect that this modelling framework can be applied successfully to high-dimensional panels of time series in many different contexts.

I illustrated the new methodology by investigating the performance of dynamic factor models for a well-known data set containing US treasury zero coupon yields. I found very low standard errors for the estimated loadings in the general dynamic factor model. This means that the time series panel of observed yield curves allows us to very precisely estimate the factor loadings. As a result, the loadings of dynamic factor models for the term structure need to be very close to these unrestricted loadings to adequately fit the data. For some well-known term structure models, the imposed restrictions are too strong to allow the estimated factor loadings to closely match the factor loadings for the general dynamic factor model.

I find that this is not the case for some well-known term structure models. All the term structure models considered in this chapter are rejected on the basis of likelihood ratio tests. I also show however, using the new methodology, that it is possible to construct a parsimonious dynamic factor model with smooth factor loadings that is not rejected. The number of parameters in the loading matrix of the new dynamic factor model is almost 50% smaller than the number of parameters in the loading matrix of the general dynamic factor model. Despite the large reduction in the number of parameters I find that the fit of the new model is qualitatively the same as for the most general dynamic factor model. I also draw the following additional conclusions from the empirical study

(i) The dynamic factor model with 3 factors appears to describe the US yield curve for 17 maturities adequately.

(ii) Estimation results for nonstationary and stationary dynamic factor model are very similar and the estimated model parameters for the stationary specifications imply a very high level of persistence. The similarity between VAR and CVAR specifications suggests that a CVAR specification is likely the appropriate specification.
(iii) For all models considered in this chapter, the Akaike information criterion suggests CVAR and VAR factors with two lags. I also find that the residuals for the VAR(1) and CVAR(1) specifications show signs of remaining autocorrelation. I therefore conclude that the VAR(1) specification, popular in many recent papers considering dynamic factor models for the the term structure of interest rates, may not present the best possible description of the dynamics of the yield curve.

(iv) The Ljung-Box serial correlation test statistics indicate that most time series models considered in this chapter explain the variation in the term structure of interest rates reasonably well. A notable exception are the models that impose no-arbitrage. The residuals for these models show signs of remaining autocorrelation.
5.A Tables and Figures

Table 5.1: Model Summary

This table gives a summary of all the models considered in this paper. We give the acronym used to refer to the model in the text and the section in which the model specification is first discussed. An asterisk in one of the columns with headings $\mu_y$, $\Lambda$ and $T$ means that respectively the intercept, loading matrix and transition matrix, as defined in section (5.2), is restricted. An asterisk in the column with heading $VAR(1)$ means that not all CVAR($k$) and VAR($k$) are allowed but only a VAR(1) specification.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acronym</th>
<th>Section</th>
<th>Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Factor Model</td>
<td>DFM</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>Smooth Dynamic Factor Model</td>
<td>SDFM</td>
<td>5.3.1</td>
<td>$\ast$</td>
</tr>
<tr>
<td>Functional Signal plus Noise Model</td>
<td>FSN</td>
<td>5.4.1</td>
<td>$\ast$</td>
</tr>
<tr>
<td>Nelson-Siegel Model</td>
<td>NS</td>
<td>5.4.2</td>
<td>$\ast$</td>
</tr>
<tr>
<td>Arbitrage-Free Nelson-Siegel Model</td>
<td>AFNS</td>
<td>5.4.3</td>
<td>$\ast$ $\ast$ $\ast$</td>
</tr>
<tr>
<td>Gaussian Affine Term Structure Model</td>
<td>AFTS</td>
<td>5.4.4</td>
<td>$\ast$ $\ast$ $\ast$ $\ast$</td>
</tr>
</tbody>
</table>
Table 5.2: Summary Statistics

The table reports summary statistics for U.S. Treasury yields over the period 1985-2000. We examine monthly data, constructed using the unsmoothed Fama-Bliss method. Maturity is measured in months. In Panel A we show for each maturity mean, standard deviation (Sd), minimum, maximum and two (1 month and 12 month) autocorrelation (Acf, \( \hat{\rho}(1) \) and \( \hat{\rho}(12) \) respectively) and partial-autocorrelation (Pacf, \( \hat{\alpha}(1) \) and \( \hat{\alpha}(12) \)) coefficients. In addition we show the test-statistic and p-value from the Augmented Dickey-Fuller (ADF) unit-root tests. In Panel B we show the correlation matrix for some selected maturities.

### Panel A: Summary Statistics

<table>
<thead>
<tr>
<th>Maturity</th>
<th>Mean</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
<th>( \hat{\rho}(1) )</th>
<th>( \hat{\rho}(12) )</th>
<th>( \hat{\alpha}(1) )</th>
<th>( \hat{\alpha}(12) )</th>
<th>ADF</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5.63</td>
<td>1.48</td>
<td>2.73</td>
<td>9.13</td>
<td>0.98</td>
<td>0.57</td>
<td>0.98</td>
<td>0.01</td>
<td>-2.06</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>5.78</td>
<td>1.48</td>
<td>2.89</td>
<td>9.32</td>
<td>0.98</td>
<td>0.55</td>
<td>0.98</td>
<td>0.00</td>
<td>-2.46</td>
<td>0.13</td>
</tr>
<tr>
<td>9</td>
<td>5.91</td>
<td>1.49</td>
<td>2.98</td>
<td>9.34</td>
<td>0.97</td>
<td>0.54</td>
<td>0.97</td>
<td>0.01</td>
<td>-2.54</td>
<td>0.11</td>
</tr>
<tr>
<td>12</td>
<td>6.07</td>
<td>1.50</td>
<td>3.11</td>
<td>9.68</td>
<td>0.97</td>
<td>0.54</td>
<td>0.97</td>
<td>-0.02</td>
<td>-2.01</td>
<td>0.28</td>
</tr>
<tr>
<td>15</td>
<td>6.23</td>
<td>1.50</td>
<td>3.29</td>
<td>9.99</td>
<td>0.97</td>
<td>0.53</td>
<td>0.97</td>
<td>-0.01</td>
<td>-2.75</td>
<td>0.07</td>
</tr>
<tr>
<td>18</td>
<td>6.31</td>
<td>1.49</td>
<td>3.48</td>
<td>10.19</td>
<td>0.97</td>
<td>0.51</td>
<td>0.97</td>
<td>0.00</td>
<td>-2.85</td>
<td>0.05</td>
</tr>
<tr>
<td>21</td>
<td>6.37</td>
<td>1.48</td>
<td>3.64</td>
<td>10.27</td>
<td>0.96</td>
<td>0.50</td>
<td>0.96</td>
<td>0.00</td>
<td>-2.82</td>
<td>0.06</td>
</tr>
<tr>
<td>24</td>
<td>6.40</td>
<td>1.46</td>
<td>3.78</td>
<td>10.41</td>
<td>0.96</td>
<td>0.48</td>
<td>0.96</td>
<td>-0.01</td>
<td>-2.95</td>
<td>0.04</td>
</tr>
<tr>
<td>30</td>
<td>6.55</td>
<td>1.46</td>
<td>4.04</td>
<td>10.75</td>
<td>0.96</td>
<td>0.48</td>
<td>0.96</td>
<td>0.03</td>
<td>-3.03</td>
<td>0.03</td>
</tr>
<tr>
<td>36</td>
<td>6.64</td>
<td>1.44</td>
<td>4.20</td>
<td>10.79</td>
<td>0.96</td>
<td>0.47</td>
<td>0.96</td>
<td>-0.00</td>
<td>-3.04</td>
<td>0.03</td>
</tr>
<tr>
<td>48</td>
<td>6.84</td>
<td>1.44</td>
<td>4.31</td>
<td>11.27</td>
<td>0.95</td>
<td>0.46</td>
<td>0.95</td>
<td>0.00</td>
<td>-2.44</td>
<td>0.13</td>
</tr>
<tr>
<td>60</td>
<td>6.93</td>
<td>1.43</td>
<td>4.35</td>
<td>11.31</td>
<td>0.95</td>
<td>0.46</td>
<td>0.95</td>
<td>-0.00</td>
<td>-2.37</td>
<td>0.15</td>
</tr>
<tr>
<td>72</td>
<td>7.08</td>
<td>1.45</td>
<td>4.38</td>
<td>11.65</td>
<td>0.95</td>
<td>0.45</td>
<td>0.95</td>
<td>0.01</td>
<td>-2.36</td>
<td>0.16</td>
</tr>
<tr>
<td>84</td>
<td>7.14</td>
<td>1.42</td>
<td>4.35</td>
<td>11.84</td>
<td>0.95</td>
<td>0.45</td>
<td>0.95</td>
<td>0.01</td>
<td>-2.53</td>
<td>0.11</td>
</tr>
<tr>
<td>96</td>
<td>7.23</td>
<td>1.41</td>
<td>4.43</td>
<td>11.51</td>
<td>0.95</td>
<td>0.47</td>
<td>0.95</td>
<td>0.03</td>
<td>-2.20</td>
<td>0.21</td>
</tr>
<tr>
<td>108</td>
<td>7.27</td>
<td>1.42</td>
<td>4.43</td>
<td>11.66</td>
<td>0.95</td>
<td>0.48</td>
<td>0.95</td>
<td>-0.00</td>
<td>-2.27</td>
<td>0.18</td>
</tr>
<tr>
<td>120</td>
<td>7.25</td>
<td>1.43</td>
<td>4.44</td>
<td>11.66</td>
<td>0.95</td>
<td>0.47</td>
<td>0.95</td>
<td>0.02</td>
<td>-2.23</td>
<td>0.20</td>
</tr>
</tbody>
</table>

### Panel B: Correlation Matrix for Selected Maturities

<table>
<thead>
<tr>
<th>Maturity</th>
<th>3</th>
<th>12</th>
<th>36</th>
<th>60</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.00</td>
<td>0.97</td>
<td>0.88</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>0.95</td>
<td>0.88</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>1.00</td>
<td>0.98</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.3: Likelihoods and AICs for DFM

This table presents maximum likelihood estimation results for the dynamic factor model (DFM), Nelson-Siegel model (NS) and functional signal plus noise model (FSN). We present results for both models with VAR factors and CVAR factors. The models were estimated on the dataset discussed in section 5.5. We show the value of the loglikelihood evaluated at the maximum likelihood estimates, denoted by $\ell(\psi)$, and the value of the Akaike Information Criterion (AIC).

<table>
<thead>
<tr>
<th>Panel A: DFM</th>
<th>VAR</th>
<th>CVAR</th>
<th>Panel B: NS</th>
<th>VAR</th>
<th>CVAR</th>
<th>Panel C: FSN</th>
<th>VAR</th>
<th>CVAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model $\ell(\hat{\psi})$</td>
<td>AIC</td>
<td>Model $\ell(\hat{\psi})$</td>
<td>AIC</td>
<td>Model $\ell(\hat{\psi})$</td>
<td>AIC</td>
<td>Model $\ell(\hat{\psi})$</td>
<td>AIC</td>
</tr>
<tr>
<td>VAR(1)</td>
<td>3894.5</td>
<td>-7595</td>
<td>CVAR(1)</td>
<td>3899.0</td>
<td>-7606</td>
<td>VAR(1)</td>
<td>3446.9</td>
<td>-6784</td>
</tr>
<tr>
<td>VAR(2)</td>
<td>3918.5</td>
<td>-7625</td>
<td>CVAR(2)</td>
<td>3923.7</td>
<td>-7637</td>
<td>VAR(2)</td>
<td>3479.0</td>
<td>-6830</td>
</tr>
<tr>
<td>VAR(3)</td>
<td>3922.6</td>
<td>-7615</td>
<td>CVAR(3)</td>
<td>3927.7</td>
<td>-7627</td>
<td>VAR(3)</td>
<td>3483.4</td>
<td>-6821</td>
</tr>
<tr>
<td>VAR(4)</td>
<td>3932.2</td>
<td>-7616</td>
<td>CVAR(4)</td>
<td>3937.3</td>
<td>-7628</td>
<td>VAR(4)</td>
<td>3494.8</td>
<td>-6826</td>
</tr>
</tbody>
</table>

**Notes:**
- The table presents maximum likelihood estimation results for the dynamic factor model (DFM), Nelson-Siegel model (NS) and functional signal plus noise model (FSN).
- Results are shown for both models with VAR factors and CVAR factors.
- The models were estimated on the dataset discussed in section 5.5.
- The loglikelihood evaluated at the maximum likelihood estimates is denoted by $\ell(\hat{\psi})$.
- The Akaike Information Criterion (AIC) is calculated for each model.

**References:**

**Further Reading:**
Table 5.4: Eigenvalues Estimated Transition Matrices

In these two tables we present the eigenvalues of the estimated transition matrices for the DFM, SDFM, NS and FSN models. In Panel A we show results for the stationary VAR(2) specifications and in Panel B for the models with nonstationary CVAR(2) factors. The columns with heading ‘real’ contain the real part of the eigenvalues and the columns with heading ‘img.’ contain the imaginary parts. Eigenvalues are sorted by their norm in ascending order.

### Panel A: Stationary models

<table>
<thead>
<tr>
<th></th>
<th>DFM</th>
<th>SDFM</th>
<th>NS</th>
<th>FSN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>real</td>
<td>real</td>
<td>real</td>
<td>real</td>
</tr>
<tr>
<td></td>
<td>img.</td>
<td>img.</td>
<td>img.</td>
<td>img.</td>
</tr>
<tr>
<td>1</td>
<td>0.154</td>
<td>0.164</td>
<td>0.156</td>
<td>0.216</td>
</tr>
<tr>
<td></td>
<td>0.163</td>
<td>0.159</td>
<td>0.166</td>
<td>0.143</td>
</tr>
<tr>
<td>2</td>
<td>0.154</td>
<td>0.164</td>
<td>0.156</td>
<td>0.216</td>
</tr>
<tr>
<td></td>
<td>-0.163</td>
<td>-0.159</td>
<td>-0.166</td>
<td>-0.143</td>
</tr>
<tr>
<td>3</td>
<td>0.597</td>
<td>0.607</td>
<td>0.593</td>
<td>0.642</td>
</tr>
<tr>
<td></td>
<td>0.058</td>
<td>0.134</td>
<td>0.056</td>
<td>0.259</td>
</tr>
<tr>
<td>4</td>
<td>0.597</td>
<td>0.607</td>
<td>0.593</td>
<td>0.642</td>
</tr>
<tr>
<td></td>
<td>-0.058</td>
<td>-0.134</td>
<td>-0.056</td>
<td>-0.259</td>
</tr>
<tr>
<td>5</td>
<td>0.963</td>
<td>0.965</td>
<td>0.964</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Panel B: Nonstationary models

<table>
<thead>
<tr>
<th></th>
<th>DFM</th>
<th>SDFM</th>
<th>NS</th>
<th>FSN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>real</td>
<td>real</td>
<td>real</td>
<td>real</td>
</tr>
<tr>
<td></td>
<td>img.</td>
<td>img.</td>
<td>img.</td>
<td>img.</td>
</tr>
<tr>
<td>1</td>
<td>0.149</td>
<td>0.155</td>
<td>0.151</td>
<td>0.206</td>
</tr>
<tr>
<td></td>
<td>0.162</td>
<td>0.162</td>
<td>0.165</td>
<td>0.143</td>
</tr>
<tr>
<td>2</td>
<td>0.149</td>
<td>0.155</td>
<td>0.151</td>
<td>0.206</td>
</tr>
<tr>
<td></td>
<td>-0.162</td>
<td>-0.162</td>
<td>-0.165</td>
<td>-0.143</td>
</tr>
<tr>
<td>3</td>
<td>0.597</td>
<td>0.601</td>
<td>0.594</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td>0.092</td>
<td>0.123</td>
<td>0.099</td>
<td>0.258</td>
</tr>
<tr>
<td>4</td>
<td>0.597</td>
<td>0.601</td>
<td>0.594</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td>-0.092</td>
<td>-0.123</td>
<td>-0.099</td>
<td>-0.258</td>
</tr>
<tr>
<td>5</td>
<td>0.972</td>
<td>0.973</td>
<td>0.972</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 5.5: Wald-Statistics of Knots in Unrestricted SDFM specification

This table shows Wald-statistics for the knots in the unrestricted SDFM-CVAR(2) model. The symbol – indicates that for this knot no Wald-statistic was calculated. This is the case for the restricted knots corresponding to 3 months, 30 months and 120 months maturity. We use * respectively ** to indicate that a statistic is significant at the 5% and 1% significance level.

<table>
<thead>
<tr>
<th>Maturity</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>2.65</td>
<td>4.22*</td>
<td>6.08*</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>2.40</td>
<td>5.59*</td>
</tr>
<tr>
<td>12</td>
<td>0.23</td>
<td>1.35</td>
<td>4.28*</td>
</tr>
<tr>
<td>15</td>
<td>0.04</td>
<td>0.33</td>
<td>1.51</td>
</tr>
<tr>
<td>18</td>
<td>0.01</td>
<td>0.02</td>
<td>0.28</td>
</tr>
<tr>
<td>21</td>
<td>0.95</td>
<td>0.74</td>
<td>1.52</td>
</tr>
<tr>
<td>24</td>
<td>3.50</td>
<td>2.37</td>
<td>3.98*</td>
</tr>
<tr>
<td>30</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>36</td>
<td>1.14</td>
<td>1.50</td>
<td>6.68**</td>
</tr>
<tr>
<td>48</td>
<td>0.44</td>
<td>2.88</td>
<td>13.47***</td>
</tr>
<tr>
<td>60</td>
<td>1.20</td>
<td>5.00*</td>
<td>18.04**</td>
</tr>
<tr>
<td>72</td>
<td>2.59</td>
<td>5.76*</td>
<td>15.69**</td>
</tr>
<tr>
<td>84</td>
<td>2.60</td>
<td>4.59*</td>
<td>8.82**</td>
</tr>
<tr>
<td>96</td>
<td>0.77</td>
<td>1.68</td>
<td>1.79</td>
</tr>
<tr>
<td>108</td>
<td>0.01</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>120</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 5.6: Wald-Statistics of Knots in Final SDFM Specifications

This table shows Wald-statistics for the knots in the final SDFM-VAR(2) and SDFM-CVAR(2) models obtained using the iterative procedure discussed in section 5.3.2. The symbol $-$ indicates that for this knot no Wald-statistic was calculated. This is the case for knots that have been removed and for the restricted knots corresponding to 3 months, 30 months and 120 months maturity. We add a superscript $^*$ if the statistic is significant at the 5% significance level and $^{**}$ if the statistic is significant at the 1% level.

<table>
<thead>
<tr>
<th>Maturity</th>
<th>SDFM-VAR(2)</th>
<th>SDFM-CVAR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Factor 1</td>
<td>Factor 2</td>
</tr>
<tr>
<td>3</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>6</td>
<td>55.36$^{**}$</td>
<td>$-$</td>
</tr>
<tr>
<td>9</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>12</td>
<td>$-$</td>
<td>53.72$^{**}$</td>
</tr>
<tr>
<td>15</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>18</td>
<td>$-$</td>
<td>15.86$^{**}$</td>
</tr>
<tr>
<td>21</td>
<td>14.41$^{**}$</td>
<td>$-$</td>
</tr>
<tr>
<td>24</td>
<td>16.70$^{**}$</td>
<td>4.14$^{*}$</td>
</tr>
<tr>
<td>30</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>36</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>48</td>
<td>$-$</td>
<td>19.98$^{**}$</td>
</tr>
<tr>
<td>60</td>
<td>$-$</td>
<td>19.00$^{**}$</td>
</tr>
<tr>
<td>72</td>
<td>$-$</td>
<td>15.67$^{**}$</td>
</tr>
<tr>
<td>84</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>96</td>
<td>4.49$^{**}$</td>
<td>$-$</td>
</tr>
<tr>
<td>108</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>120</td>
<td>$-$</td>
<td>$-$</td>
</tr>
</tbody>
</table>
### Table 5.7: Arbitrage-Free Term Structure Models: Maximum Likelihood Results

This table presents the maximum likelihood estimation results for the two arbitrage-free term structure models: the AFNS and AfTS models. We show the maximum of the loglikelihood function in the column with heading $\ell(\hat{\psi})$. Further we give the number of parameters in the model $n_\psi$ and the Akaike information criterium (AIC). Finally we present the eigenvalues of the estimated transition matrix $T$. The eigenvalues are presented in descending order.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\ell(\hat{\psi})$</th>
<th>$n_\psi$</th>
<th>AIC</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFNS</td>
<td>3253.3</td>
<td>42</td>
<td>-6423</td>
<td>0.986</td>
<td>0.952</td>
<td>0.884</td>
</tr>
<tr>
<td>AfTS</td>
<td>3429.4</td>
<td>36</td>
<td>-6786.9</td>
<td>0.9997</td>
<td>0.998</td>
<td>0.984</td>
</tr>
</tbody>
</table>

### Table 5.8: Maximum Likelihood Estimates of Nelson-Siegel Parameter $\lambda$

This table presents maximum likelihood estimates of the Nelson-Siegel parameter $\lambda$, defined in (5.24), for models where the factors are given by VAR($p$) and CVAR($p$) processes and for varying values of $p$.

<table>
<thead>
<tr>
<th>Nelson-Siegel Factor Model Parameter $\lambda$</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR($p$)</td>
<td>0.07303</td>
<td>0.07211</td>
<td>0.07216</td>
<td>0.07193</td>
</tr>
<tr>
<td>CVAR($p$)</td>
<td>0.07302</td>
<td>0.07210</td>
<td>0.07213</td>
<td>0.07191</td>
</tr>
</tbody>
</table>
Table 5.9: Ljung-Box Statistics

This table shows Ljung-Box statistics calculated for the scaled residuals of some of the models discussed in this paper. Separate statistics are calculated for each maturity. We chose a number of 12 lags to calculate the test-statistics. The superscript $^*$ is used to indicate rejection of the null hypothesis at the 10% significance level and $^{**}$ is used for rejection at the 5% significance level. The headings ‘CVAR(2) factors’ and ‘VAR(1) factors’ indicate the specifications chosen for the factors.

<table>
<thead>
<tr>
<th>Maturity</th>
<th>CVAR(2) factors</th>
<th>VAR(1) factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DFM</td>
<td>NS</td>
</tr>
<tr>
<td>3</td>
<td>5.8</td>
<td>6.2</td>
</tr>
<tr>
<td>6</td>
<td>7.1</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>19.2$^*$</td>
<td>19.3$^*$</td>
</tr>
<tr>
<td>12</td>
<td>22.5$^{**}$</td>
<td>29.2$^{**}$</td>
</tr>
<tr>
<td>15</td>
<td>15.9</td>
<td>17.8</td>
</tr>
<tr>
<td>18</td>
<td>12.8</td>
<td>13.0</td>
</tr>
<tr>
<td>21</td>
<td>12.2</td>
<td>12.0</td>
</tr>
<tr>
<td>24</td>
<td>10.2</td>
<td>11.2</td>
</tr>
<tr>
<td>30</td>
<td>9.3</td>
<td>9.4</td>
</tr>
<tr>
<td>36</td>
<td>8.7</td>
<td>9.1</td>
</tr>
<tr>
<td>48</td>
<td>6.2</td>
<td>6.0</td>
</tr>
<tr>
<td>60</td>
<td>5.9</td>
<td>5.7</td>
</tr>
<tr>
<td>72</td>
<td>5.7</td>
<td>5.5</td>
</tr>
<tr>
<td>84</td>
<td>8.4</td>
<td>9.4</td>
</tr>
<tr>
<td>96</td>
<td>7.7</td>
<td>7.6</td>
</tr>
<tr>
<td>108</td>
<td>9.2</td>
<td>8.6</td>
</tr>
<tr>
<td>120</td>
<td>10.1</td>
<td>9.7</td>
</tr>
</tbody>
</table>
This table presents the likelihood-ratio statistics for the null-hypothesis that the restrictions of the considered model are valid. The column $k$ contains the number of restrictions imposed by the model. In panel A we show the VAR(2) versions of the NS, FSN and SDFM models. Panel B gives the CVAR(2) versions of the NS, FSN and SDFM models. Finally, we give the likelihood ratio statistics for the arbitrage-free models in Panel C.

### Panel A: Stationary Models

<table>
<thead>
<tr>
<th>Model</th>
<th>$LR$</th>
<th>$k$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>220.2</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>FSN</td>
<td>879.0</td>
<td>42</td>
<td>0.000</td>
</tr>
<tr>
<td>SDFM</td>
<td>23.4</td>
<td>21</td>
<td>0.32</td>
</tr>
</tbody>
</table>

### Panel B: Nonstationary Models

<table>
<thead>
<tr>
<th>Model</th>
<th>$LR$</th>
<th>$k$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>220.4</td>
<td>41</td>
<td>0.000</td>
</tr>
<tr>
<td>FSN</td>
<td>879.8</td>
<td>42</td>
<td>0.000</td>
</tr>
<tr>
<td>SDFM</td>
<td>20.2</td>
<td>20</td>
<td>0.45</td>
</tr>
</tbody>
</table>

### Panel C: Arbitrage-Free Models

<table>
<thead>
<tr>
<th>Model</th>
<th>$LR$</th>
<th>$k$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFNS</td>
<td>1282.4</td>
<td>64</td>
<td>0.000</td>
</tr>
<tr>
<td>AFTS</td>
<td>930.2</td>
<td>76</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 5.1: Yield Curves from January 1985 up to December 2000

In this figure we show the U.S. Treasury yields over the period 1985-2000. We examine monthly data, constructed using the unsmoothed Fama-Bliss method. The maturities we show are 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108 and 120 months. Panel A presents a 3-dimensional plot, Panel B provides time-series plots for selected maturities.

(A) 3-Dimensional Term Structure Plot

(B) Time-Series for Selected Maturities
Figure 5.2: Maximum Likelihood Estimates Loadings for DFM

This figure shows the estimated factor loadings for the DFM-CVAR($k$) model, for $k = 1, \ldots, 4$ as a function of time to maturity. In panel A we see the results for the model where the factors are modelled by (5.8). The loadings are now restricted to be of the form (5.5) with the rows of the identity matrix at the maturities 3 months, 30 months and 120 months. Panel B shows the estimated loadings for the same model but with $f_t$ modelled as (5.10). In this case the first column of the loading matrix corresponds to the nonstationary factor and is scaled such that the first element is one. The sub-matrix consisting of the second and third columns is now of the form (5.5).
Figure 5.3: Estimated Factor Loadings for SDFM Model
This figure shows the estimated factor loadings for the SDFM-VAR(2) and SDFM-CVAR(2) models, obtained using the procedure of section 5.3.2, as functions of time to maturity. For ease of comparison we also show the maximum likelihood estimates of the loadings in the DFM model. The loadings are restricted to be of the form (5.5) with the rows of the identity matrix at the 3 months, 30 months and 120 months maturities.
This figure shows the estimated factor loadings as functions of time to maturity for the DFM-CVAR(2), NS-CVAR(2) and FSN-CVAR(2) models. The factor loadings are rotated such that loading matrix is of the form (5.5), where the rows of the identity matrix are at the maturities of 3 months, 30 months and 120 months. In panel A we show the loadings for the DFM and NS model and in panel B the loadings for the FSN model.
Figure 5.5: Estimated Factor Loadings Gaussian Affine Term Structure Model

This figure shows the estimated factor loadings for the Gaussian affine term structure model (AfTS) and the DFM-VAR(1) model as functions of time to maturity. The factor loadings for the AfTS are rotated such that the loading matrix is of the form (5.5), where the rows of the identity matrix are at the maturities of 3 months, 30 months and 120 months.
Figure 5.6: Estimated Factor Loadings DFM with Confidence Bounds
This figure shows the estimated factor loadings for the DFM-CVAR(2) model as functions of time to maturity with 95% confidence intervals.
Chapter 6

Analysis of Dynamic Factor Models in the Presence of Missing Data

6.1 Introduction

In this chapter we will see how to perform factor extraction and likelihood evaluation for the model of Chapter 4 in the presence of missing data. Recall that the time series $y_1, \ldots, y_n$ is generated by

$$y_t = \Lambda f_t + u_t, \quad t = 1, \ldots, n,$$  \hspace{1cm} (6.1)

where $y_t$ is $N \times 1$ dimensional, $f_t$ is an unobserved $q \times 1$ vector of common factors and $u_t$ is the $N \times 1$ vector of idiosyncratic terms. We are primarily interested in cases where $N$ is significantly larger than $q$. The factors are assumed to follow a Gaussian dynamic linear process and the idiosyncratic components in $u_t$ are modelled as autoregressive (AR) processes. The results are also applicable to more general models of this form. I will discuss these issues in detail.

In many applications, the dimension of $y_t$ is large and the model depends on a large number of parameters. The task of signal extraction and parameter estimation is therefore challenging in various respects. In Chapter 4 we saw that writing the model in state space form allows us to obtain minimum mean square estimates of the factors together with the corresponding mean square errors by means of the Kalman filter and smoother recursions. These methods can be implemented in a computationally efficient way. However, this state space representation is no longer valid in the presence of missing data. I address this problem by developing a low-dimensional linear state space model with time-varying state dimensions.
It is equivalent to the dynamic factor model (6.1) and is designed to allow for missing entries in the dataset.

Reis and Watson (2007) consider the dynamic factor model (6.1) and estimate the parameters by maximum likelihood using the approach of Watson and Engle (1983). This approach is not applicable when missing data is present. Banbura and Modugno (2008) propose a solution that overcomes the problem but is computationally demanding. Furthermore, their method requires modifications that can lead to numerical problems. In this chapter I present a computationally efficient method that leads to exact maximum likelihood parameter estimates. All methods provide minimum mean square estimates and corresponding mean square errors of the factors and the idiosyncratic components.

The remainder of the chapter is organized as follows. The dynamic factor model and its state space representations are discussed in detail in Section 6.2. In Section 6.3 I develop a new state space representation that remains valid when missing data is present. This state space representation allows the computationally efficient application of the Kalman filter and smoother recursions. Signal extraction and likelihood evaluation are explored in Section 6.4. Parameter estimation by maximum likelihood methods is discussed in Section 6.5. In Section 6.6 I discuss the computational cost of the method proposed. Section 6.7 concludes.

6.2 The Dynamic Factor Model

6.2.1 Model Specification

The dynamic factor model given in (6.1) links the observation $y_t$ to a set of unobserved factors $f_t$ for $t = 1, \ldots, n$. I take $f_1, \ldots, f_n$ to be linear combinations of an unobserved $p \times 1$ dimensional vector autoregressive process $\alpha_t$. Specifically, there is a $q \times p$ selection matrix $S$ such that

$$
 f_t = S\alpha_t. 
$$

(6.2)

The state process $\alpha_t$ is generated by the following transition equation

$$
 \alpha_{t+1} = T\alpha_t + \eta_t, \quad \eta_t \sim N(0, \Sigma_\eta),
$$

(6.3)

for $t = 1, \ldots, n - 1$ where the initial state vector $\alpha_1$ is specified as $\alpha_1 \sim N(0, P_{1|0})$ and where the $p \times p$ transition matrix $T$ and the $p \times p$ variance matrix $\Sigma_\eta$ are assumed fixed (non-stochastic). The matrices $S$, $T$ and $\Sigma_\eta$ may depend on a fixed and unknown vector of
model parameters $\psi$. In case $\alpha_t$ is a time-invariant stationary process, $P_{t|0}$ is chosen equal to the stationary variance of $\alpha_t$. It follows that the dynamic factor model (6.1) can be expressed in terms of the state vector

$$y_t = Z\alpha_t + u_t,$$

where $Z = \Lambda S$. The factor loading matrix $\Lambda$ is treated as fixed and may depend on the vector of model parameters $\psi$. The idiosyncratic component $u_t$ is modelled as a vector autoregressive process with $r$ lags

$$u_{t+1} = \phi_1 u_t + \cdots + \phi_r u_{t-r+1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma_\varepsilon),$$

where $\phi_1, \ldots, \phi_r$ and $\Sigma_\varepsilon$ are $N \times N$ matrices and the initial vector $u_1$ is specified as $u_1 \sim N(0, \Sigma_u)$. In general $\phi_1, \ldots, \phi_r$ will be chosen such that $u_t$ is a stationary process and $\Sigma_u$ will be set to the stationary variance of $u_t$.

In the remainder of the paper, we consider the dynamic factor model as specified above. However, the results below apply to more general settings. These generalizations are discussed in some detail in section 6.3.3.

### 6.2.2 Two State Space Representations

The dynamic factor model specification (6.4), (6.3) and (6.5) is close to the state space model introduced in Section 2.1. In this section I present two ways in which the dynamic factor model considered here can be written in state space form. We can therefore calculate the likelihood and obtain estimators of the factors $f_t$ by applying the Kalman filter and smoother recursions to these state space models. The two formulations are given as A and B below for the special case of $r = 1$. The higher-order case of $r > 1$ follows straightforwardly but is notationally more cumbersome.

#### A

A basic approach is to express the dynamic factor model in terms of $(1 - \phi_1 L)y_t$ where $L$ is the lag-operator. When the polynomial function $1 - \phi_1 L$ is applied to both sides of (6.4), we obtain

$$y_t = \phi_1 y_{t-1} + Z\alpha_t - \phi_1 Z\alpha_{t-1} + \varepsilon_t$$

$$= c_t + (Z, -\phi_1 Z) \begin{pmatrix} \alpha_t \\ \alpha_{t-1} \end{pmatrix} + \varepsilon_t,$$

(6.6)
with \( y_t = Z \alpha_1 + u_1 \) and where \( c_t = \phi_1 y_{t-1} \) for \( t = 2, \ldots, n \). The transition equation for the augmented state vector is given by

\[
\begin{pmatrix}
\alpha_{t+1} \\
\alpha_t
\end{pmatrix} =
\begin{bmatrix}
T & 0 \\
I & 0
\end{bmatrix}
\begin{pmatrix}
\alpha_t \\
\alpha_{t-1}
\end{pmatrix} +
\begin{pmatrix}
\eta_t \\
0
\end{pmatrix},
\]

for \( t = 1, \ldots, n - 1 \). The introduction of \( c_t \) in the observation equation does not cause further complications; it can be handled by the Kalman filter since \( c_t \) is known at time \( t \). Note that approach A was also used in Section 4.2.

**B. An alternative formulation is obtained by augmenting the state vector with \( u_t \) and is given by**

\[
y_t = (Z, I) \begin{pmatrix} \alpha_t \\ u_t \end{pmatrix}, \quad \begin{pmatrix} \alpha_{t+1} \\ u_{t+1} \end{pmatrix} =
\begin{bmatrix}
T & 0 \\
0 & \phi_1
\end{bmatrix}
\begin{pmatrix}
\alpha_t \\
u_t
\end{pmatrix} +
\begin{pmatrix}
\eta_t \\
\epsilon_t
\end{pmatrix},
\]

for \( t = 1, \ldots, n \). The initial condition for the state vector process is straightforwardly determined. The observation disturbance vector has disappeared from this formulation. This does not cause complications in the application of the Kalman filter.

Both formulations will lead to the same results when initialization issues are properly accounted for. Watson and Engle (1983) and, more recently, Reis and Watson (2007) have adopted formulation A while Banbura and Modugno (2008) have adopted formulation B.

### 6.2.3 Missing Data

In this chapter we consider the application of the Kalman filter and smoothing methods to the dynamic factor model in the presence of missing observations. The model formulation B is valid when \( y_t \) contains missing entries while formulation A is not valid since \( c_t \) cannot be determined when \( y_{t-1} \) is partly missing. An exact treatment of filtering and smoothing is therefore not possible when we adopt formulation A. The replacement of \( c_t \) by \( \hat{c}_t = \phi_1 E(y_t | y_1, \ldots, y_{t-1}) \) in formulation A may lead to a practical solution but it clearly does not lead to an exact solution. This assessment has led Banbura and Modugno (2008) to adopt formulation B for their dynamic factor analysis. This solution is however computationally inefficient since the dimension of the state vector \( \alpha_t \) can become very large when \( N \) increases. In the empirical study of Banbura and Modugno (2008), the observation dimension is close to \( N = 100 \) such that their state vector dimension is larger than 100. A high dimensional state vector slows down the Kalman filter enormously and may even lead to numerical inaccuracies.
Therefore formulation A is preferable, since the increase of the state dimension is moderate. The main contribution of this paper is a re-formulation of A that allows application of the Kalman filter and smoothing methods in the presence of missing data. Furthermore, I show that the developments of Chapter 4 can be used to speed-up the computations.

6.3 State Space Formulation in the Presence of Missing Data

In this section I will present a new way to write the model of Section 6.2 as a Gaussian state space model. For ease of notation we will focus on the special case of \( r = 1 \) but with a diagonal coefficient matrix \( \phi_1 \). Section 6.3.3 discusses the consequences of more general model specifications.

6.3.1 Notation

Consider some \( N \times 1 \) vector \( v_t \). The vector \( v_t(o_s) \) contains all elements of \( v_t \) that correspond to observed entries in the \( N \times 1 \) data vector \( y_s \) for \( t, s = 1, \ldots, n \). In a similar way, \( v_t(m_s) \) contains all elements of \( v_t \) that correspond to missing entries in \( y_s \). In case all entries in \( y_s \) are observed, \( v_t(m_s) \) is an empty vector. The vector \( v_t(o_s, m_{s'}) \) contains all elements of \( v_t \) that correspond only to observed entries in \( y_s \) and missing entries in \( y_{s'} \) for \( t, s, s' = 1, \ldots, n \). Using this notation we can split the vector \( v_t \) into four mutually exclusive sub-vectors \( v_t(o_s, o_{s'}) \), \( v_t(o_s, m_{s'}) \), \( v_t(m_s, o_{s'}) \) and \( v_t(m_s, m_{s'}) \). In case we have no missing data, vectors \( v_t(m_s) \) and \( v_t(m_s, m_{s'}) \) are empty while \( v_t = v_t(o_s) = v_t(o_s, o_{s'}) \). We further note that

\[
\{v_t\} = \{v_t(o_s), v_t(m_s)\} = \{v_t(o_s, o_{s'}), v_t(o_s, m_{s'}), v_t(m_s, o_{s'}), v_t(m_s, m_{s'})\}.
\]

To illustrate the notation, consider \( N = 5 \) and

\[
y_t = (1, m, 2, m, 3)', \quad y_{t-1} = (m, m, m, 4, 5)', \quad v_t = (6, 7, 8, 9, 10)',
\]

where \( m \) denotes a missing entry. It follows that

\[
v_t(o_t) = (6, 8, 10)', \quad v_t(m_t) = (7, 9)', \\
v_t(o_t, o_{t-1}) = 10, \quad v_t(o_t, m_{t-1}) = (6, 8)', \quad v_t(m_t, o_{t-1}) = 9, \quad v_t(m_t, m_{t-1}) = 7.
\]

This notation applies to matrices in a similar way. Consider the \( N \times k \) matrix \( V \). Matrix \( V(o_t; \cdot) \) contains selected rows of \( V \) that correspond to the observed entries in \( y_t \) while all
columns are retained. In case of a $k \times N$ matrix $V$, the selection $V(\cdot; o_t)$ applies to columns. In case of a $N \times N$ matrix, the selection $V(o_t; m_t)$ applies to both rows and columns.

### 6.3.2 The Missing Data State Space Formulation

We develop a state space formulation for the observation vector

$$y_t^o = \begin{pmatrix} y_t(o_t, o_{t-1}) \\ y_t(o_{t+1}, m_{t-1}) \end{pmatrix},$$

for $t = 1, \ldots, n$. The new state $\dot{\alpha}_t$ is given by

$$\dot{\alpha}_t = \begin{bmatrix} \alpha_t', \alpha_{t-1}', u_t(o_t, m_{t-1})', u_t(m_t, m_{t-1})', u_t(m_t, o_{t-1})' \end{bmatrix}'.

The state vector is augmented both with $\alpha_{t-1}$ and a selection of $u_t$. The new formulation below can therefore be interpreted as a mix of formulations A and B in section 6.2.2.

The observation equation that links the observation vector $y_t^o$ and the state vector $\dot{\alpha}_t$ is obtained straightforwardly as

$$y_t^o = c_t^o + \left[ \begin{array}{cc} Z(o_t, o_{t-1}; \cdot) & -\phi_t^o Z(o_t, o_{t-1}; \cdot) \\ Z(o_t, m_{t-1}; \cdot) & 0 \end{array} \right] \dot{\alpha}_t + \begin{pmatrix} \varepsilon_t(o_t, o_{t-1}) \\ 0 \end{pmatrix},$$

where $c_t^o = \{ \phi_t^o y_{t-1}(o_t, o_{t-1}) \}'$, $0'$ and $\phi_t^o = \phi_t(o_t, o_{t-1}; o_t, o_{t-1})$. Matrix $\phi_t^o$ is diagonal consisting of (a subset of), possibly reshuffled, diagonal elements of $\phi_t$. The specification for $y_t(o_t, o_{t-1})$ relies on formulation A while for $y_t(o_t, m_{t-1})$ it relies on formulation B. The major difference between the new formulation and B is that we only include those entries of $u_t$ in the state vector that correspond to missing entries in $y_t$ and/or $y_{t-1}$. For those entries of $y_t$ where both $y_t$ and $y_{t-1}$ are observed, we can compute the corresponding entries in $c_t^o$ and rely on formulation A.

The transition equation for the state process $\dot{\alpha}_t$ is obtained as follows. The updates for $\alpha_t$ and $\alpha_{t-1}$ are given as in (6.7) for formulation A. Next we develop equations for $u_{t+1}(o_{t+1}, m_t)$ and $u_{t+1}(m_{t+1}, m_t)$ which are effectively the selection $u_{t+1}(m_t)$ (re-ordered). The transition from $u_t(m_t)$ to $u_{t+1}(m_t)$ is the autoregressive update (6.5) with $r = 1$ in this specific case. We have

$$u_{t+1}(m_t) = \phi_t(m_t; m_t) u_t(m_t) + \varepsilon_t(m_t), \quad u_t(m_t) = \begin{pmatrix} u_t(m_t, m_{t-1}) \\ u_t(m_t, o_{t-1}) \end{pmatrix},$$
for $t = 1, \ldots, n - 1$. To place $u_{t+1}(m_t)$ into $\alpha_{t+1}$, we need to re-order it into

$$
\begin{pmatrix}
 u_{t+1}(o_{t+1}, m_t) \\
 u_{t+1}(m_{t+1}, m_t)
\end{pmatrix} = J_t u_{t+1}(m_t),
$$

where $J_t$ is implicitly defined as a selection matrix of ones and zeroes. The bottom part of $\alpha_{t+1}$ is $u_{t+1}(m_{t+1}, o_t)$ and corresponds to observed entries in $y_t$. Therefore, we have

$$
u_{t+1}(m_{t+1}, o_t) = \phi_t^* u_t(m_{t+1}, o_t) + \varepsilon_t(m_{t+1}, o_t)
= \phi_t^* [y_t(m_{t+1}, o_t) - Z_t^* \alpha_t] + \varepsilon_t(m_{t+1}, o_t),$$

where $\phi_t^* = \phi_t(m_{t+1}, o_t; m_{t+1}, o_t)$ and $Z_t^* = Z(m_{t+1}, o_t; \cdot)$. The transition equation for $\dot{\alpha}_t$ is therefore

$$
\dot{\alpha}_{t+1} = d_t + \begin{bmatrix}
 T & 0 & 0 & 0 \\
 I & 0 & 0 & 0 \\
 0 & 0 & 0 & J_t \phi_1(m_t; m_t) \\
 -\phi_t^* Z_t^* & 0 & 0 & 0
\end{bmatrix}
\begin{pmatrix}
 \alpha_t \\
 \alpha_{t-1} \\
 u_t(o_t, m_{t-1}) \\
 u_t(m_t)
\end{pmatrix} + \begin{pmatrix}
 \eta_t \\
 0 \\
 J_t \varepsilon_t(m_t) \\
 \varepsilon_t(m_{t+1}, o_t)
\end{pmatrix},
$$

(6.10)

where $d_t = [0, 0, 0, \{ \phi_t^* y_t(m_{t+1}, o_t) \}]'$, for $t = 1, \ldots, n - 1$.

### 6.3.3 Discussion of the New Formulation

The equations (6.9) and (6.10) define the state space model for the observed values while the missing observations are accounted for by including the relevant $u_t$’s in the state vector. In case we have no missing data, the vectors $u_t(o_t, m_{t-1})$ and $u_t(m_t)$ are empty and we return to formulation A. Entries of $u_t$ only appear in the state vector when they correspond to missing entries in $y_t$ or in $y_{t-1}$. In this way we keep the dimension of the state to a minimum while at all times we are able to produce optimal estimates using Kalman filter and smoothing.

In most cases the dimension of $\dot{\alpha}_t$ will be smaller than the dimension of $(\alpha_t', u_t')'$, the state vector in model formulation B. In case $\dot{\alpha}_t$ has a larger dimension than $(\alpha_t', u_t')'$, due to a large number of missings in $y_{t-1}$ or $y_t$, it is possible to reduce the dimension of $\dot{\alpha}_t$ by dropping $\alpha_{t-1}$ (partially and temporarily) from the state vector $\dot{\alpha}_t$. Since the resulting computational gains will be relatively small, I will not pursue this further.

The new formulation does imply time-varying system matrices in the observation and transition equations. In fact, the dimension of the state vector also varies over time. For-
fortunately, the Kalman filter can treat varying dimensions for the state vector. The implementation of such a Kalman filter requires attention but it comes with the benefit of a dynamic factor analysis that is computationally feasible when missing data is present. Some additional details are given in the next section.

The results presented in this paper apply to more general settings. For example, the dynamic specification of \( f_t \) may also depend on non-stationary time series processes such as a random walk. The Kalman filter and smoothing need to deal with the initialization problem but existing solutions can be applied straightforwardly. Lagged factors and explanatory variables can be included in the observation equation of the dynamic factor model, see the discussion in Jungbacker and Koopman (2008). The case of \( r > 1 \) is trivial but requires more notation in the expositions of sections 6.3 and 6.4. A particular concern is the case of a non-diagonal autoregressive coefficient matrix \( \phi_1 \) since it mixes the lag-dependence of idiosyncratic components associated with missing entries with those associated with observed entries. We therefore need to modify the system matrices in (6.10) accordingly. This exercise is straightforward but the notation is somewhat cumbersome.

### 6.4 Signal Extraction and Likelihood Evaluation

In this section I discuss computationally efficient approaches to signal extraction and likelihood evaluation. These methods are also relevant for parameter estimation as discussed in section 6.5.

#### 6.4.1 Estimation of States and Idiosyncratic Components

Given the state space formulation of the dynamic factor model, we can adopt the Kalman filter and associated smoothing methods (KFS) to obtain

\[
\hat{a}_{ts} = \mathbb{E}(\hat{\alpha}_t|Y_s), \quad \hat{Q}_{ts} = \text{Var}(\hat{\alpha}_t|Y_s),
\]

for \( t, s = 1, \ldots, n \) where \( Y_s = (y^o_1, \ldots, y^o_s) \), see the discussion of these methods in Chapter 2. Prediction refers to \( s = t - 1 \), concurrent filtering to \( s = t \) and smoothing to \( s = n \). The Kalman filter can also be used to evaluate the log-likelihood function using the prediction error decomposition result of Schweppe (1965), see Section 2.8.

In terms of the dynamic factor model (6.4), (6.3) and (6.5), KFS produces estimates (as well as the mean square errors) of \( \alpha_t \) and of those entries of \( u_t \) that are associated with missing entries in \( y_t \) and \( y_{t-1} \), that is \( u^m_t = [u_t(o_t, m_{t-1}'), u_t(m_t)']' \). We can also obtain estimates
and corresponding mean square errors of \( u_t^o = u_t(o_t, o_{t-1}) \) using the identity \( u_t = y_t - Z\alpha_t \) in (6.4). Let \( a_{t|s} = \mathbb{E}(\alpha_t|Y_s) \) and \( Q_{t|s} = \text{Var}(\alpha_t|Y_s) \) for \( t, s = 1, \ldots, n \). Obviously, \( a_{t|s} \) and \( Q_{t|s} \) are the upper (block) parts of \( \tilde{a}_{t|s} \) and \( \tilde{Q}_{t|s} \), respectively. It follows that

\[
\mathbb{E}(u_t^o|Y_s) = y_t^o - Z_t^o a_{t|s}, \quad \text{Var}(u_t^o|Y_s) = Z_t^o Q_{t|s} Z_t^{o'};
\]

\[
\text{Cov}(u_t^o, \alpha_t|Y_s) = -Z_t^o Q_{t|s}, \quad \text{Cov}(u_t^o, u_t^m|Y_s) = -Z_t^o \text{Cov}(\alpha_t, u_t^m|Y_s),
\]

where \( Z_t^o = Z(o_t, o_{t-1}; \cdot) \) and \( \text{Cov}(u_t^m, \alpha_t|Y_s) \) is part of \( \tilde{Q}_{t|s} \) for \( t, s = 1, \ldots, n \).

### 6.4.2 KFS with a Collapsed Observation Vector

The computational effort for the KFS depends on the dimensions of both the state and observation vectors. Consider the dynamic factor model (6.1) with \( q \times 1 \) vector \( f_t = S\alpha_t \) and state space representation (6.4) and (6.3) but with \( u_t \) replaced by \( \varepsilon_t \sim N(0, \Sigma_\varepsilon) \), that is

\[
y_t = Z\alpha_t + \varepsilon_t, \quad \alpha_{t+1} = T\alpha_t + \eta_t, \quad \text{(6.11)}
\]

for \( t = 1, \ldots, n \) with \( N \times 1 \) observation vector \( y_t \) and \( p \times 1 \) state vector \( \alpha_t \). In most practical applications of the dynamic factor model, the dimension of \( y_t \) is significantly larger than the dimension of \( \alpha_t \). In Chapter 4 we saw that in such circumstances, when \( N > q \), the computational efficiency of KFS can significantly be improved by a simple computational device. Recall that \( Z = \Lambda S \) and define the \( N \times N \) and \( q \times N \) matrices

\[
A = \begin{bmatrix} A^L \\ A^H \end{bmatrix}, \quad \Lambda = C^{-1} \Lambda \Sigma_\varepsilon^{-1},
\]

respectively, where \( C \) can be any invertible matrix and \( A^H \) is chosen such that matrix \( A \) is full rank and \( A^L \Sigma_\varepsilon A^H = 0 \). It follows that \( A^H Z = 0 \). We assume that \( \Lambda \) has full column rank. In most cases of practical interest this assumption will be valid. If matrix \( \Lambda \) does not have full rank, it can be replaced with any full rank matrix that spans the column space of \( \Lambda \), see the discussion in Chapter 4. Matrix \( A^H \) exists by construction but it does not need to be evaluated for our purposes. By choosing \( C \) such that \( CC' = Z' \Sigma_\varepsilon^{-1} Z \), we have

\[
Ay_t = \begin{pmatrix} A^L y_t \\ A^H y_t \end{pmatrix} = \begin{pmatrix} C' S \\ 0 \end{pmatrix} \alpha_t + \begin{pmatrix} A^L \varepsilon_t \\ A^H \varepsilon_t \end{pmatrix}, \quad \begin{pmatrix} A^L \varepsilon_t \\ A^H \varepsilon_t \end{pmatrix} \sim N \left( 0, \begin{bmatrix} I & 0 \\ 0 & A^H \Sigma_\varepsilon A^H \end{bmatrix} \right),
\]

for \( t = 1, \ldots, n \). The equation for \( \alpha_{t+1} \) is unaffected by the transformation. It follows that the part \( A^H y_t \) does not depend on \( \alpha_t \), it is not correlated with \( A^L y_t \) and therefore
does not need to be considered for the estimation of $\alpha_t$. Therefore, the KFS only need to be applied to the collapsed observation (low-dimensional) vector $A^L y_t$ for signal extraction. Since $\text{Var}(A^L \varepsilon_t) = I$, we can adopt the KFS devices discussed in Section 2.9 to further accelerate the computations.

The collapse can lead to high computational savings. To illustrate the reductions that we can achieve in practice, consider model (6.11) with $N = 100$ and $p = 10$. In this case, the observation vector relevant for the application of the KFS collapses from dimension $N = 100$ to dimension $p = 10$. Jungbacker and Koopman (2008) also demonstrate that likelihood evaluation can rely on the Kalman filter applied to $A^L y_t$, see section 6.4.4.

### 6.4.3 A Collapsed KFS in Presence of Missing Data

The computational device of Section 4.4 can be modified in the context of the state space formulation developed in section 6.3.2, in case missing data is present. Consider the observation equation (6.9). Since this formulation relies on time-varying system matrices, we require the collapsed transformations to vary over time as well.

We carry out a partial collapse of $y_o t$ and only consider the transformation of $y_t(o_t, o_{t-1})$ with dimension $N_{oo}^t$. For this purpose, we define

$$A^L_t = C_t^{-1} Z^+_t V_t^{-1}, \quad Z^+_t = [\Lambda(o_t, o_{t-1}; \cdot), -\phi^o_t \Lambda(o_t, o_{t-1}; \cdot)], \quad V_t = \Sigma_{\varepsilon}(o_t, o_{t-1}; o_t, o_{t-1}),$$

where $\phi^o_t = \phi_1(o_t, o_{t-1}; o_t, o_{t-1})$ and $C_t$ is chosen such that

$$C_t C_t' = Z^+_t V_t^{-1} Z^+_t,$$

for $t = 1, \ldots, n$. Again, we should make sure that $Z^+_t$ has full column rank. If this is not the case it is generally easy to find a new matrix with full column rank that spans the same column space. The transformation $A^L_t$ is applied to $y_t(o_t, o_{t-1})$ only and does not need to consider the elements of $\dot{\alpha}_t$ associated with $u_t$ since they do not affect $y_t(o_t, o_{t-1})$. We can extend the transformation towards $y_t(o_t, m_{t-1})$ but this will not lead to further reductions $y_t(o_t, m_{t-1})$.

Define matrix

$$A_t = \begin{bmatrix} A^L_t \\ A^H_t \end{bmatrix},$$

where $A^H_t$ is chosen such that $A^L_t V_t A^H_t' = 0$ and $A_t$ is a full rank matrix. The state space
model for the transformed observation vector $A_t y_t(o_t, o_{t-1})$ is given by
\[
\begin{pmatrix}
    A_t^L y_t(o_t, o_{t-1}) \\
    A_t^H y_t(o_t, o_{t-1}) 
\end{pmatrix}
= \begin{pmatrix}
    A_t^L c_t \\
    A_t^H c_t 
\end{pmatrix} + \begin{bmatrix}
    C_t' S & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 
\end{bmatrix} \hat{\alpha}_t + A_t \varepsilon_t(o_t, o_{t-1}),
\] (6.12)
where $q \times p$ matrix $S$ is defined in (6.2) and \( \text{Var}[A_t \varepsilon_t(o_t, o_{t-1})] \) is a block-diagonal variance matrix with the upper-block given by \( \text{Var}[A_t^L \varepsilon_t(o_t, o_{t-1})] = I \). It follows that we can remove $A_t^H y_t(o_t, o_{t-1})$ for the application of the KFS and for the same reasons as discussed in section 6.4.2. In particular, for the application of KFS we can replace (6.9) by the two observation equations
\[
\begin{pmatrix}
    A_t^L y_t(o_t, o_{t-1}) \\
    y_t(o_t, m_{t-1}) 
\end{pmatrix}
= \begin{pmatrix}
    A_t^L c_t' \\
    0 
\end{pmatrix} + \begin{bmatrix}
    C_t' S & 0 & 0 & 0 \\
    Z_t^{om} & I & 0 & 0 
\end{bmatrix} \hat{\alpha}_t + \begin{pmatrix}
    A_t^L \varepsilon_t(o_t, o_{t-1}) \\
    0 
\end{pmatrix},
\] (6.13)
where $Z_t^{om} = \{ Z(o_t, m_{t-1}; \cdot), 0 \}$. In most cases, the observation vector dimension of the collapsed model will be much lower than the dimension of $y_t^o$. However, when $y_t$ does contain many missing observations, it may become the case that the dimension of $y_t(o_t, o_{t-1})$ is lower than $2p$. In this case no computational gain can be achieved by transforming the model. The state space model (6.13) and (6.10) can be handled by the KFS devices discussed in Section 2.9.

In case both observation vectors $y_t$ and $y_{t-1}$ contain no missing entries, we can apply the time-invariant transformation as developed in section 6.4.2 and based on the state space formulation A of Section 6.2.2. We only require the modifications for collapsing the observation vector presented in this section when missing entries in the observation vectors $y_t$ or $y_{t-1}$ are present.

### 6.4.4 Log-likelihood Evaluation

For a set of observations $y_1, \ldots, y_n$, we define the log-likelihood function by
\[
\ell(y) = \log p(y_1^o, \ldots, y_n^o; \psi), \quad y = \{ y_t^o \}_{t=1}^n,
\] (6.14)
where $p(\cdot)$ is the Gaussian density function, $y$ is the set of observed data, and $\psi$ is the vector of parameters introduced in section 6.2. The prediction error decomposition result of Schweppe (1965) implies that
\[
\log p(y_1^o, \ldots, y_n^o; \psi) = \log p(y_1^o; \psi) + \sum_{t=2}^n \log p(y_t^o|Y_{t-1}; \psi)
\]
where $p(y_t^o|Y_{t-1}; \psi)$ can be evaluated by the Kalman filter.

In Chapter 4 we argued that the likelihood function $\ell(y)$ can be obtained by applying
the Kalman filter to the collapsed data vector only. In this case, we can limit the application of the Kalman filter to the observation equation (6.13). The log-likelihood function is then evaluated by
\[ \ell(y) = \text{constant} + \ell\left(y^L, y^{om}\right) + \ell\left(y^H\right), \]
where
\[ y^L = \{A^L_t y_t(o_t, o_{t-1})\}_{t=1}^n, \quad y^{om} = \{y_t(o_t, m_t-1)\}_{t=1}^n, \quad y^H = \{A^H_t y_t(o_t, o_{t-1})\}_{t=1}^n, \]
and the constant does not depend on \( \psi \) nor on the observations. The log-likelihood function \( \ell\left(y^L, y^{om}\right) \) is obtained from the Kalman filter applied to the state space model (6.13) and (6.10). The log-likelihood function \( \ell\left(y^H\right) \) can be evaluated by
\[ \ell\left(y^H\right) = \text{constant} - \frac{1}{2} \sum_{t=1}^n \log |V_t| - \frac{1}{2} \sum_{i=1}^n e_t^t V_t^{-1} e_t, \]
where \( e_t \) is given by
\[ e_t = \left( I - V_t A^L_t A^L_t' \right) [y_t(o_t, o_{t-1}) - \phi_t^o y_{t-1}(o_t, o_{t-1})], \]
for \( t = 1, \ldots, n \), see Lemma 4.2 in Chapter 4.

6.5 Maximum Likelihood Estimation

In Section 4.5 I discussed methods for likelihood maximization in a dynamic factor model. Both the Expectation-Maximization (EM) algorithm as well as direct maximization can be used in the presence of missing data. The main difference is that we use the new state space formulation. Also, the expressions for the gradient become slightly more difficult. Recall, that for the EM algorithm the maximization in the M-step can not be done analytically. In the presence of missing data we can use a modified version of the iterative scheme proposed in Watson and Engle (1983) for the M-step. Alternatively, we can perform the M-step via a quasi-Newton scheme. Since the gradient of \( Q(\psi|\psi^{(n)}) \) is available analytically, the necessary computations can be done computationally efficient.
6.6 Computational Costs and Gains

In this section I explore the computational gains that can be achieved using our new state space specification of Section 6.3.2 when applying the Kalman filter and associated smoothing algorithm (KFS). I compare the computational cost of the Kalman smoother applied to the new model to the cost for state space formulations A and B of Section 6.2.2. Below I will refer to the new model as formulation C. When no observations are missing the computational cost of the KFS will be the same for formulations A and C, since in this case the two specifications are equivalent. If some observations are missing, formulation A is not valid while formulation C is. The cost of the additional computations is modest when the number of missing entries is small. When \( y_t \) and \( y_{t-1} \) have a total of \( m \) unique missing entries (the entries that are both missing in \( y_t \) and \( y_{t-1} \) are counted once), the state vector \( \dot{\alpha}_t \) needs to be increased by \( m \) (temporarily). This will slow down the KFS computations but it will lead to exact results while formulation A cannot deal with missing entries. The increase in computing time depends on the number of missings in the data-set. In Table 6.1(a) I provide some indications of the computational costs for the dynamic factor model (6.1) with two dynamic factors \( (q = 2) \) which are modelled as stationary vector autoregressive processes. The comparisons are carried out for three different observation vector dimensions \( N = 10, 50, 100 \). The results reveal, for example, that for \( N = 50 \) and for 1% missing observations (missing entries are randomly chosen in the sample), the computations take 1.5 times longer than those for formulation A (instead of, say, 20 seconds, it takes 30 seconds). When the number of missings increases to 10%, the computations take 2.6 times longer.

The formulation B also provides exact results when data is missing and this is the approach adopted by Banbura and Modugno (2008). However, I have argued in this chapter that formulation C is computationally more efficient. In Table 6.1(b) I compare the computing times for formulations B and C. The gains of the new formulation compared to B are quite considerable. For the same model as described above with \( N = 50 \), the KFS for formulation C is almost 88 times faster when we have 1% missings while it is 44 times faster when 10% of the data is missing. These gains are considerable and they are even higher and more dramatic when \( N \) increases to higher values.

6.7 Conclusions

Likelihood-based analysis of dynamic factor models has seen renewed interest in the economics and finance literature recently. High-dimensional dynamic models with multiple factors contain many parameters that need to be estimated. For maximum likelihood estimation
CHAPTER 6. DYNAMIC FACTOR MODELS AND MISSING DATA

Table 6.1: Computational costs and gains

Table (a) presents ratios of computing times for the formulation of section 6.3.2 with missing data divided by those for the formulation without missing data (this is formulation A of Section 6.2.2). For example, the value 2 indicates that the computational demands are as twice as high. The table (b) presents ratios of computing times for the formulation B of section 6.2.2 divided by those for the formulation of section 6.3.2 with missing data. For example, the value 2 indicates that the new device is twice as fast. The ratios are presented for different dimensions $N$ of the observation vector $y_t$ and for different percentages of missing data.

(a) Costs relative to A

<table>
<thead>
<tr>
<th>$N$</th>
<th>percentage missing</th>
<th>1%</th>
<th>10%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.4</td>
<td>1.8</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.5</td>
<td>2.6</td>
<td>8.9</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.2</td>
<td>3.9</td>
<td>24.8</td>
<td></td>
</tr>
</tbody>
</table>

(b) Gains relative to B

<table>
<thead>
<tr>
<th>$N$</th>
<th>percentage missing</th>
<th>1%</th>
<th>10%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.1</td>
<td>1.5</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>87.9</td>
<td>43.7</td>
<td>11.3</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>625.5</td>
<td>197.5</td>
<td>25.8</td>
<td></td>
</tr>
</tbody>
</table>

to be feasible computationally efficient methods are of key importance. Various problems arise when these methods are applied to samples with missing observations. A standard solution requires the idiosyncratic component to be included in the state vector. This will lead to a high-dimensional state vector for a model with a high-dimensional observation vector and will slow down the Kalman filter and smoother algorithms considerably. To circumvent this problem, I propose a new state space formulation that allows for missing values and can exploit existing devices for computational efficiency. Only the idiosyncratic components associated with missing entries for the concurrent and previous time periods are accommodated in the state vector, all other ones are removed from the state vector. As a result, the dimension of the state vector is kept to a minimum. In the new formulation the dimension of the state vector varies over time and therefore the implementation of the Kalman filter and smoothing methods requires attention. However, we can obtain high computational savings even when the number of missing entries is moderate, see Table 6.1.
Bibliography


BIBLIOGRAPHY


Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for non-stationary


S&P 100 Stock Index using Historical, Realised and Implied Volatility Measurements.

Koopman, S. J., M. Mallee, and M. Van der Wel (2009). Analyzing the term structure of
interest rates using the dynamic Nelson-Siegel model with time-varying parameters. *J.

Koopman, S. J. and N. Shephard (1992). Exact score for time series models in state space

Lawley, D. N. and A. E. Maxwell (1971). *Factor analysis as a statistical method.* London:
Butterworths. 2nd edition.

of Fixed Income* 1(1), 54–61.

*Biometrika* 66, 67–72.

Statistical Society B* 44, 98–103.

Press.

60-4, 473–89.

*Econometrica* 59, 347–70.


Poirier, D. J. (1976). *The Econometrics of Structural Change: with Special Emphasis on
Spline Functions.* Amsterdam: North-Holland.

Cep discussion papers, Centre for Economic Performance, LSE.
Working paper.


series by stochastically convergent parameter regression. *Annals of Economic and So-
cial Measurement* 2, 399–428.

metrics* 49, 305–41.

Press.

Sargent, T. J. and C. A. Sims (1977). Business cycle modeling without pretending to have
too much a priori economic theory. In C. A. S. et al. (Ed.), *New methods in business
cycle research*. Minneapolis: Federal Reserve Bank of Minneapolis.

actions on Information Theory* 11, 61–70.


Press.


Shumway, R. H. and D. S. Stoffer (1982). An approach to time series smoothing and


*J. Business and Economic Statist.* 20, 147–162.

Discussion paper.

dynamic factor, MIMIC and varying coefficient regression. *J. Econometrics* 23, 385–
400.

Springer-Verlag.

Statistics* 11, 95–103.

Ontwikkelingen in Methoden voor State Space Modellen op Basis van de Likelihood

State space modellen zijn een klasse van tijdreeks modellen die pogen een geobserveerd fenomeen te verklaren aan de hand van een set ongeobserveerde stochastische processen. Vaak heeft zo een ongeobserveerd proces een economische interpretatie. Zo kan men de seizoenscomponent in een macro-economische tijdreeks expliciet modelleren als ongeobserveerde component.

De praktische toepasbaarheid van een state space model hangt voornamelijk af van de vraag of we onbekende grootheden zoals de model parameters en het ongeobserveerde state proces kunnen schatten. Dit probleem is de focus van dit proefschrift. Grof gezegd beschouwen we twee klassen modellen: lineaire en niet lineaire modellen. Voor lineaire modellen is het schattingsprobleem grotendeels op te lossen door gebruik van het beroemde Kalman filter. Voor deze modellen is de voornaamste uitdaging om state space modellen praktisch toepasbaar te maken voor hele grote datasets die omvangrijke modellen vereisen. Voor niet lineaire modellen is het Kalman filter niet toepasbaar en is zelfs het schatten van simpele modellen een uitdaging. In dit proefschrift presenteer ik nieuwe resultaten voor beide klassen modellen en illustreer ik de praktische toepasbaarheid aan de hand van empirische voorbeelden.

Chapter 2: The Linear State Space Model

Chapter 3: Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models

In dit hoofdstuk beschouw ik de klasse state space modellen waar de het state proces normaal verdeelde innovaties heeft maar waar de observaties niet lineair van de state afhangen. Het belangrijkste resultaat van dit hoofdstuk is een uitbreiding van de algoritmes in Shephard and Pitt (1997) and Durbin and Koopman (1997). Ik laat onder andere zien hoe men een Kalman filter kan interpreteren voor een state space model met negatieve ‘varianties’.

Chapter 4: Likelihood-based Analysis for Dynamic Factor Models

In dit hoofdstuk behandel ik dynamische factor modellen. Deze modellen vallen in de categorie van lineaire state space modellen. In principe is het Kalman filter dan ook toepasbaar op deze modellen. De grote hoeveelheid data en model parameters maken dit echter onpraktisch. Ik presenteer een aantal nieuwe resultaten die het mogelijk maken om het Kalman filter toe te passen op grote state space modellen met honderden tijdreeksen. Als bewijs van de praktische toepasbaarheid schat ik een dynamisch factor model met meer dan duizend parameters.

Chapter 5: Dynamic Factor Models with Smooth Loadings

Een van de aannames van een lineair state space model is dat een potentieel grote verzameling tijdreeksen een lineaire functie is van een klein aantal onderliggende processen. In veel toepassingen is het redelijk aan te nemen dat de系数enten in deze lineaire relaties niet te sterk varieren tussen de verschillende tijdreeksen. In dit hoofdstuk beschouw ik het voorbeeld van een model voor yields van obligaties met verschillende looptijden. De coefficienten voor verschillende looptijden zullen in het algemeen dicht bij elkaar liggen. In dit hoofdstuk presenteer ik een klasse state space modellen die deze verbanden tussen de coefficienten door middel van restricties expliciet oplegt. Ik laat ook zien hoe de juiste set van restricties kan worden gevonden door een systematische procedure op basis van Wald-tests.

Chapter 6: Dynamic Factor Analysis in the Presence of Missing Data

Dit hoofdstuk is een vervolg op hoofdstuk 4. Ik presenteer nieuwe resultaten die het mogelijk maken om de resultaten van hoofdstuk 4 toe te passen op tijdreeksen waar observaties missen.