Bayesian Forecasting of Stock Returns using Simultaneous Graphical Dynamic Linear Models

by

Nelson Kyakutwika

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematics in the Faculty of Science at Stellenbosch University

Supervisor: Dr. Bruce Bartlett

Co-supervisor: Prof. Ronnie Becker

December 2022

The financial assistance of Stellenbosch University towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to Stellenbosch University.

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2022 ....

Abstract

Bayesian Forecasting of Stock Returns using Simultaneous Graphical Dynamic Linear Models

Nelson Kyakutwika

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc

December 2022

Cross-series dependencies are crucial in obtaining accurate forecasts when forecasting a multivariate time series. Simultaneous Graphical Dynamic Linear Models (SGDLMs) are Bayesian models that elegantly capture cross-series dependencies. This study aims to forecast returns of a 40-dimensional time series of stock data using SGDLMs. The SGDLM approach involves constructing a customised dynamic linear model (DLM) for each univariate time series. Every day, the DLMs are recoupled using importance sampling and decoupled using mean-field variational Bayes. We summarise the standard theory on DLMs to set the foundation for studying SGDLMs. We discuss the structure of SGDLMs in detail and give detailed explanations of the proofs of the formulae involved. Our analyses are run on a CPU-based computer; an illustration of the intensity of the computations is given. We give an insight into the efficacy of the recoupling/decoupling techniques. Our results suggest that SGDLMs forecast the stock data accurately and respond to market gyrations nicely.

Uittreksel

Bayesian Voorspelling van Aandeelopbrengste deur gebruik te maak van Gelyktydige Grafiese Dinamiese Lineêre Modelle

("Bayesian Forecasting of Stock Returns using Simultaneous Graphical Dynamic Linear Models")

Nelson Kyakutwika

Departement Wiskuudige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc

Desember 2022

Kruisreeksafhanklikhede is van kardinale belang om akkurate voorspellings te verkry wanneer 'n meervariant tydreeks voorspel word. Gelyktydige grafiese dinamiese lineêre modelle (SGDLMs) is Bayesiaanse modelle wat kruisreeksafhanklikhede elegant vaslê. Hierdie studie het ten doel om opbrengste van 'n 40-dimensionele tydreeks van voorraaddata met behulp van SGDLMs te voorspel. Die SGDLM-benadering behels die konstruksie van 'n pasgemaakte dinamiese lineêre model (DLM) vir elke eenvariant tydreeks. Elke dag word die DLM's herkoppel met behulp van belangrikheidsteekproefneming en ontkoppel met behulp van gemiddelde-veld variasie Bayes. Ons som die standaardteorie oor DLM's op om die grondslag te lê vir die bestudering van SGDLM'e. Ons bespreek die struktuur van SGDLM'e in detail en gee gedetailleerde verduidelikings van die bewyse van die betrokke formules. Ons ontledings word op 'n SVE-gebaseerde rekenaar uitgevoer; 'n illustrasie van die intensiteit van die berekeninge word gegee. Ons gee 'n insig in die doeltreffendheid van die herkoppeling/ontkoppelingstegnieke. Ons resultate dui daarop dat SGDLM's die voorraaddata akkuraat voorspel en mooi reageer op markwisselings.

Acknowledgements

Firstly, I am profusely grateful to my supervisor, Dr. Bruce Bartlett, for the many things he has done for me throughout my studies. I am grateful for the time you invested in applying for my funding and correcting me whenever I was going astray with my work. I thank you for your outstanding mentorship. Similarly, I am immensely thankful to my co-supervisor, Prof. Ronnie Becker, first, for accepting to supervise me for a second Master's degree, for proposing the research topic, and for your advice throughout my studies. My supervisors, without the lengthy meetings which we used to have and your patience with me, I would not have managed to write this thesis. I thank you for the momentous impact that you have had on my life.

I thank Prof. Mike West of Duke University for clarifying several concepts about the SGDLM; I cannot forget that you always responded immediately whenever I contacted you. I am also appreciative to Dr. Lesley Wessels for translating the abstract to Afrikaans. I say thank you to Mary Ndimuwaki and Tifu Waiswa for their help during my studies. I thank the entire staff of the mathematics division of Stellenbosch University for being a highly receptive and supportive team.

Furthermore, my flatmates and my office mates, I am very grateful for your academic and social support.

In addition, I thank Stellenbosch University for providing me with a partial scholarship through the PSP and the Department of Mathematical Sciences for providing me with the top-up.

Last but not least, I thank my mother and my siblings for the support they have given me throughout my studies. In a special way, I thank my fiancée, Linnet Akomire, for her company and encouragement during the lonely and hard times that I had during my studies.

Dedications

To my mother, Nabirye Irene, and my father, Bafunte Kalema Stephen (1961–1999).
Mummy, thank you for nurturing me into the person I am today; you are my hero, you are a superwoman!
Daddy, the pain of your untimely demise will stay with me forever, RIP.

Contents

List of Figures

List of Tables

Nomenclature

Notation

t	Index of discrete time intervals, $t = 1, 2, \ldots$
$N[\mu, \sigma^2]$	Univariate normal distribution with mean μ and variance σ^2
$N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$	Multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$E[X]$	Expected value of a random variable X
$V[X]$	Variance of a random variable X
$\text{Cov}[X, Y]$	Covariance of the random variables X and Y
$\text{Ga}[n, d]$	Gamma distribution with shape n and rate d, mean n/d and variance n/d^2
$T_\nu(\mu, \sigma^2)$	Univariate Student's t distribution with degrees of freedom ν, mode μ, and scale σ^2
$T_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}, h)$	Multivariate Student's t distribution with degrees of freedom ν, mode $\boldsymbol{\mu} \in \mathbb{R}^h$, and scale matrix $\boldsymbol{\Sigma}$
$\text{Be}(\alpha, \beta)$	Beta distribution with parameters α and β, mean $= \frac{\alpha}{\alpha+\beta}$
$\text{tr}(\boldsymbol{A})$	Trace of an $m \times m$ matrix $\boldsymbol{A}$

Abbreviations

JSE	Johannesburg Stock Exchange
WDLM	Wishart Dynamic Linear Model
DDNM	Dynamic Dependence Network Model
DLM	Dynamic Linear Model
CPU	Central Processing Unit
SGDLM	Simultaneous Graphical Dynamic Linear Model
ESS	Effective Sample Size
SP	Simultaneous Parent
MFVB	Mean-field Variational Bayes
KL	Kullback-Leibler
RMSE	Root Mean Square Error
MAD	Mean Absolute Deviation

Chapter 1

Introduction

This chapter introduces our study by first giving the background, followed by the problem statement, the aim and the objectives, the significance of the study, the justification of the study, the methodology overview and, lastly, the thesis outline.

1.1 Background

This section gives a brief history about models used in univariate and multivariate time series forecasting, daily stock data being the time series of interest.

Forecasting returns on investment in stocks is of paramount importance to stock market investors. There are many models for forecasting these returns if the focus is on a single stock. These univariate models are easy to construct and have been used by [25, 20, 33] to forecast stock prices for individual stocks. These models focus on a single stock without considering the impact of other stocks in the market on the stock under study. For example, if interest is in stock **A**, forecasting the prices of this stock is done without paying attention to changes in prices of the other stocks that co-exist with **A** in the market.

In practice – in the stock market – the price of any stock partly depends on the changes in the prices of the other stocks. It thus becomes unrealistic to say that, if interest is in the performance of stock **A** over time, the prices of stock **A** should be forecast in isolation without taking into consideration the changes in prices of the stocks that co-exist with stock **A**. A change in the price of stock **B** can impact the price of stock **A**, stock **C**, or any other stock in the market. So, for an investor interested in stock **A**, the construction of a model that predicts the prices of stock **A** need not be limited to the prices of this stock alone, but also the prices of all the other stocks in the market since these prices impact that of stock **A**. In other words, issues that cut across all the stocks should be considered while modelling individual stocks. Therefore, modelling should be done at the individual stock level and across all the stocks concurrently. The current research addresses the need for capturing the inter-stock dependencies in multivariate forecasting of stock data, Johannesburg Stock Exchange (JSE) stock returns being the data to which our approach is applied.

The challenge of including cross-series dependencies in the analysis calls for construction of a multivariate model. Construction of multivariate financial time series models, Bayesian or non-Bayesian, is not new. Many Bayesian multivariate time series models have been applied to financial data in the past, for example, Cholesky-style and factor models (e.g., [2]), dynamic graphical models of precision matrices (e.g., [4]), the standard Wishart dynamic linear model (WDLM) (e.g., [7] and [33, Section 16.4]), and dynamic dependence network models (DDNMs) (e.g., [36]). However, these models have shortcomings when applied to high-dimensional time series, for example,

- DDNMs require that the modeller arranges the time series in a particular order. How do you choose the most suitable ordering with 40 stocks? This would call for looking at all possible orderings, which are 40!.
- Also, for Cholesky-style and factor models, ordering of the time series matters.
- The WDLM does not allow for different predictors for each univariate time series.

In a recent study, [7] introduced more robust models, *Simultaneous Graphical Dynamic Linear Models*, that eradicate the shortcomings of the aforementioned models. In their study, they used the SGDLM to forecast returns of a multivariate time series consisting of 400 stocks of the S&P 500 index. The building blocks of SGDLMs are DLMs. In SGDLMs, the focus is on two things: *customising and scaling up* of models. Attention is first put on building a distinct predictive model for each of the univariate time series that exists in the multivariate system; this is what we call customising. Then, the analysis is extended to deal with many univariate time series concurrently; we call this scaling up. At each time, the univariate models are brought together to capture issues that cut across all series.

1.2 Problem statement

In a multivariate time series, the focus should not only be on constructing distinct predictive models for individual univariate time series, but also on the contemporaneous cross-series dependencies. Cross-series relationships in a multivariate time series have a significant impact on the accuracy of predictive values of individual series. In the stock market, dependency is when a change in the price of one stock affects the price of another. These inter-stock dependencies which occur in the stock market need inclusion in the forecasting model(s) to improve forecast accuracy. SGDLMs are a Bayesian class of models introduced by [7] to, in addition to predicting future values for individual univariate series, capture the multivariate dependencies with the aim of improving forecast accuracy.

1.3 Aim and objectives of the study

1.3.1 Aim

The aim of this thesis is to forecast the returns of a multivariate financial time series made up of 40 stocks, using the SGDLM.

1.3.2 Objectives

The objectives are:

1. To summarise the standard theory of DLMs and link it to SGDLMs.
2. To describe the structure of SGDLMs.
3. To outline the SGDLM algorithm.
4. To implement the SGDLM algorithm in Python, on a CPU-based local machine.
5. To assess whether the SGDLM forecasts the stock data accurately.

1.4 Justification of the study

Many Bayesian multivariate models, as mentioned in Section 1.1 with the old models, encounter difficulties when applied to high-dimensional time series. Models becoming over-parameterised and the inability to scale up computationally, with increasingly high dimensions, are some of the difficulties. SGDLMs are multivariate time series models that are more parsimonious. However, not many studies have been conducted on these recently introduced models. Therefore, more research is needed, especially in elucidating further the structure of SGDLMs and the SGDLM algorithm. Capturing cross-series dependencies in high-dimensional time series usually presents challenges in computing the big calculations involved, for both the old models and the novel SGDLMs; this is the basis upon which [7] used GPUs while introducing SGDLMs. Using the unconventional GPU-accelerated parallelisation greatly speeds up the computations in SGDLMs. Unlike in [7] where the authors worked with 400 stocks, using the common CPU-based computers should suffice when one has a relatively smaller number of time series, for example, the 40 stocks of the current study. There is therefore a need to investigate the feasibility of using CPU-based computers in situations when the dimension of the multivariate time series is smaller compared to that in [7].

1.5 Significance of the study

The details we have given about the structure of SGDLMs and those given in the algorithm are additions to the standard theory about SGDLMs for researchers and academics to follow up. Computing the filtering solution of the DLM analytically is relevant theory that aids the understanding of DLMs. Our use of CPU-based computers, though low-level compared to using GPU-based computers, provides some insight into the complexity of the SGDLM computations.

1.6 Methodology overview

The SGDLM analysis is a canonical Bayesian analysis. The SGDLM strategy starts with decoupled independent priors, for the state parameters and precisions, for each of the 40 series on day t. Using these priors, forecasts of the returns of all stocks on day t are obtained jointly. The independent priors are updated to decoupled naive posteriors, series by series, and the product of the decoupled naive posteriors gives the naive joint posterior. Importance sampling is used to obtain the exact joint posterior from the naive joint posterior in a technique known as recoupling. Mean-field variational Bayes (MFVB) is applied to the importance sample-based posterior to give a product of the independent conjugate forms in their posterior state; this is called decoupling and is done while minimising the Kullback-Leibler divergence. The variational Bayes posteriors are evolved independently to give independent priors for day $t+1$. The procedure keeps repeating each day for the entire forecasting period. The forecasting and recoupling steps are computationally intensive – they both involve forming dozens of $m \times m$ matrices, and respectively involve computing inverses and determinants of $m \times m$ matrices dozens of times (m is the number of stocks).

1.7 Thesis outline

This thesis consists of six chapters. In Chapter 2, we review the literature on SGDLMs and identify the gaps upon which we build our research. Chapter 3 summarises the standard theory of dynamic linear models (DLMs) as presented, mainly, by [33], while tailoring the summary in the direction of SGDLMs. In Chapter 4, building on the work of [7], a detailed explanation of the structure of SGDLMs is given together with the SGDLM algorithm. Chapter 5 implements the SGDLM algorithm. In Chapter 5, we give details of how we implement the algorithm to enable interested researchers to follow up easily. We give the results of our SGDLM analysis at the end of the chapter. Chapter 6 concludes the thesis and presents recommendations for further research. Appendices are given at the end of the thesis to give extra information that is useful for our study.

Chapter 2

Literature Review

The purpose of this chapter is to give an overview of the research that has been conducted on SGDLMs. We start with a brief history of DLMs (the progenitors of SGDLMs) followed by previous research that is specific to SGLDMs. At the end of the chapter, we wrap up with a summary of the focus of this thesis. At the end of every subsection in Section 2.2, we mention the gap that the current study addresses depending on what has been presented in that subsection.

2.1 Dynamic linear models

Dynamic linear models have a long history; pioneer knowledge about them dates way back to the 1880s [20, Chapter 1]. In the 1960s, dynamic linear models were used in the engineering field to control and monitor dynamic systems ([20, Chapter 1] and [32]). The use of dynamic linear models in time series modelling came into the limelight in the 1970s [20, Chapter 1]. Since then, dynamic linear models have gained traction for application in many areas, for example, finance, economics, engineering, genetics, et cetera [20, Chapter 1]. The prominence of DLMs has been largely due to their ability to handle computational difficulties associated with time series by using Monte Carlo methods in a Bayesian approach.

Literature on DLMs has been published profusely over the years. The main reference for dynamic linear models is [33], but extensive theory and applications exist in other references like [25, 20, 24, 28]. Computing software for DLMs has also been widely published (e.g., [20, 22, 23]). Dynamic linear models are used in modelling financial time series data (e.g., [15, 18, 34, 11]). Typical financial applications of dynamic linear models (and extensions of the models) include forecasting returns on investment in stocks and foreign currency (e.g., [2, 26, 7, 36]) and portfolio analysis (e.g., [2, 37, 10]). Time series applications of dynamic linear models in other areas can be found in [34, 11, 19, 16, 12]. To date, DLMs remain an active area of research, SGDLMs being one of the recent fascinating advancements.

2.2 Simultaneous graphical dynamic linear models

In this section, we summarise and synthesise some of the studies that have been done on SGDLMs. In the literature, the top studies on these models have been conducted by [7, 10, 31, 6, 35]. Therefore, our review will refer mainly to work done in the aforementioned references. In this section, we review the literature by reporting how the authors (i) described the structure of SGDLMs, (ii) applied SGDLMs, and (iii) stated and implemented the SGDLM algorithm.

2.2.1 Structure of SGDLMs

While introducing SGDLMs, [7] gave the notation and the structural forms of SGDLMs, starting from the definition of the stochastic observational variance DLM. In addition to this, [7] explained the structure of the SGDLM, and stated formulae for the joint likelihood and the joint posterior. In another study, [35] derived the equations for the joint likelihood and the joint posterior. While reviewing the work in [7] and [10], [31] dug deep into the challenges and opportunities presented by the structure of SGDLMs. Among other things, [31] discussed dynamic dependence network models (e.g., [36]), a special case of SGDLMs, where the order in which you arrange the time series matters. It is difficult to scale up DDNMs to high dimensions, because, to work with m time series, one has to choose the most suitable ordering of the series from the total number of orderings, which is $m!$. Whereas the ordering of the series does not matter in SGDLMs, SGDLMs are far more computationally intensive than DDNMs.

Series-specific contemporaneous predictors (simultaneous parents) constitute part of the structure of SGDLMs. The approach in which simultaneous parents are chosen is thus critical. In [7], each stock is assigned 10 simultaneous parents which are selected from the remaining 399 stocks by choosing those with the highest effect sizes. The method of choosing simultaneous parents in [7] is such that, for a particular stock, simultaneous parents remain the same for the entire period of analysis. The issue of using the same simultaneous parents for the entire period of the analysis is not realistic since the performance of stocks in the market is dynamic. So, while working with empirical data, it is ideal to keep changing simultaneous parents over time depending on the current data – this is something that needed to be included in the first version of SGDLMs. In a later study, [10] introduced a more practical method of selecting simultaneous parents which was adopted by [31, 6]. In this method, they divided potential predictors of a particular stock, say, stock j, into three categories: the core set, the up set, and the down set. The core set contains predictors of stock j at time t, the up set contains core set candidates – stocks that qualify to be promoted to the core set, and the down set contains stocks that were previously in either the core set or the up set. The size of each of the three sets is specified by the modeller. Stocks do not stay in these sets permanently, different stocks keep moving from one set to another. A multivariate Wishart dynamic linear model, which runs in parallel to SGDLMs, is used to keep refreshing the stocks in the sets.

Whereas a combination of the studies above elucidates the general structure of SGDLMs, none of those studies gives a detailed link from DLMs to SGDLMs. The studies discussed above are largely SGDLMs-focused. Before understanding SGDLMs, one needs to have a good general understanding of DLMs. Concepts like model specification, discount factors, and the Kalman filter are central in understanding DLMs as well as SGDLMs. However, such concepts appear in the literature on DLMs, which was written before the SGDLMs advancement. This thesis addresses the need of linking DLMs to SGDLMs, by summarising the standard theory on DLMs while directing the summary towards SGDLMs.

2.2.2 Application of SGDLMs to data sets

We start with the study of [7] which analysed daily log-returns of 400 stocks of the S&P 500 index and the index itself, from October 2000 to October 2013. The data for the first 845 days was used as a training set to choose simultaneous parents for each stock; the data for the next 522 days was also used as a training set to select discount factors; and the data for the last 2,044 days was used to test the models by forecasting returns on a daily basis. In both the training and test data sets, the authors used 10,000 Monte Carlo samples for forecasting and joint posterior approximation. Analyses were done using three different methods: the *full SGDLM*, the *no recouple-decouple SGDLM*, and the standard *Wishart dynamic linear model*. The authors constructed prediction intervals to evaluate the resultant forecasts across all the stocks and for a few selected stocks, for all the three methods. The full SGDLM approach emerged as the most accurate, followed by the no recouple-decouple SGDLM, and the WDLM came last. They also made plots of 60-day moving averages of empirical returns for six selected stocks. They compared the trend of the moving averages with trends obtained under the full SGDLM, the no recouple-decouple SGDLM, and the WDLM. All the three models performed in a similar way and their trends were similar to trends of empirical returns. In another plot, they plotted 60-day tracking moving averages of volatility (standard deviation) of returns for the same (six) companies and compared them with volatilities estimated by each of the three models. Volatilities of the full SGDLM tracked well the empirical volatilities. The volatilities measured by the no recouple-decouple SGDLM tracked well the empirical volatilities up to the start of the 2008 market crash; thereafter, the model volatilities moved away from the empirical volatility trend. The WDLM overestimated the volatilities before and after the market crash, and it underestimated them during the market crash. In this study, the authors evaluated the effectiveness of importance sampling using effective sample size and that of mean-field variational Bayes using the entropy of the importance sample, an approximation of Kullback-Leibler divergence.

In [10], the authors focused on portfolio investment decisions. This paper advanced the SGDLM methodology of [7] by (i) implementing the more robust method of selecting simultaneous parents that was mentioned in Section 2.2.1, and (ii) comparing an entropy measure with the St. Louis Federal Reserve Bank Financial Stress Index in the context of measuring market stress. The study used the same 400 stocks together with the S&P 500 index like [7]. Six investment strategies and two models (the full SGDLM and the standard WDLM) were used. For each pair of a model and a strategy, the investment analysis proceeds sequentially as follows: (i) the observations $\boldsymbol{y}_{t-1}$ are used to update the model distributions at the market close of day $t-1$; (ii) the one-step ahead forecast distribution for $\boldsymbol{y}_t$ is simulated or computed, the appropriate optimisation rule is solved, and the investment weights are adjusted to the optimised ones; (iii) observations at time t are taken and the realised returns are calculated, and the analysis proceeds to day $t+1$. The best investment strategy under the SGDLM performed far better than the best investment strategy under the standard WDLM. For the best investment strategy under the SGDLM, an investment of $1,000 grew to $3,862 over 11 years, whereas an investment of $1,000 grew to $1,168 over the same period with the best investment strategy under the standard WDLM – the WDLM is constrained by the high dimensions. A passive investment in the S&P 500 grew from $1,000 to $1,996 over the same period. Unlike [7], [10] compared the scaled entropy of the importance sample with the standard St. Louis Fed and Kansas City Fed Indices. The entropy of an importance sample is a measure of financial market stress. In comparison with traditional measures of financial market stress, the scaled entropy led both the standard market stress index measure (the St. Louis Fed index) and the Kansas City Fed index.

In a later study, [31] overviewed the SGDLM methodology with a focus on the recouple-decouple strategy. He elucidated the strategy using diagrammatic illustrations. Using the same number of stocks like [7, 10], he measured the effectiveness of the importance sample – the recoupling technique – using the entropy of the importance sample. Like [10], [31] compared this entropy (in its capacity as a measure of financial market stress) with the St. Louis Fed Index. Using cumulative density function plots of the realised one-step ahead forecast errors, he demonstrated the danger associated with ignoring recoupling – ignoring the determinant term and you handle SGDLMs as DDNMs where the determinant term is equal to 1.

Lastly, unlike the previous authors, [35] applied SGDLMs to economic time series. They forecast the United States' microeconomic time series. However, we will not detail this application given that the focus of our study is financial time series, stock data in particular.

From the summaries above, we notice that the authors of [7, 10, 31] applied SGDLMs to the same data set, the 400 S&P 500 stocks, to draw all the conclusions. Much as [35] applies the models to a different data set, this is not a financial time series data set. It is therefore necessary to apply SGDLMs to another stock data set.

2.2.3 Algorithm and its implementation

Algorithm

In the introductory study to SGDLMs, [7] gave a quite detailed (SGDLM) algorithm in six steps, together with a summary of the six steps. In later studies, [10, 31, 6] just summarise the algorithm. However, none of these authors derived the formulae on which the algorithm is premised. In [8, Appendix 5.B], the thesis where [7] was extracted, the formulae for the mean-field variational approximation were derived. In [35], an alternative derivation of the mean-field variational approximation formulae is given as well as the derivation of the joint posterior formula. A proof of the positive definiteness of the covariance matrix of the joint predictive distribution is also given in [35, Appendix A].

Whereas the algorithm is outlined by some authors as mentioned above, none of those authors gives sufficient detail for someone (especially a new person in the field) to understand the algorithm easily. In this study, following [7], we give a more detailed algorithm to make it suitable for almost direct implementation. Additionally, we give more detailed explanations of the derivation of the joint posterior formula and that of the mean-field variational Bayes formulae.

Implementation

SGDLMs, like any other multivariate models, are computationally intensive. But, for SGDLMs, the sequential forecasting, filtering, and evolution are parallelisable. Since the computation of the predictive distributions, updating to the naive posteriors, and simulation of the naive posteriors are done at the level of the decoupled univariate series, the entire computation strategy can exploit GPU computing where there is access to several cores. For this reason, [7, 10, 31, 35, 6] use the GPU-accelerated approach to do computations.

In [7], C++/CUDA programming (on a local machine) was used to speed up computations via GPU parallelisation. In a later study, [6] used the TensorFlow library in combination with Google GPUs to do the GPU-accelerated computations. The authors of [7] developed the R package "rSGDLM: An R Package for Simultaneous Graphical DLMs [9]" that works in conjunction with GPUs. The use of GPUs by these authors is realistic given that [7, 10, 31] worked with 400 stocks and [6] worked with 487 stocks.

Much as GPUs significantly speed up computations, none of the authors above used a standard laptop or desktop computer that uses CPU hardware. In the current study, we use a standard desktop computer to explore the relative slowness that one might encounter. This approach to computation is useful for two reasons: (i) GPU-based computers are not as common as CUP-based ones (so, not all researchers can be in position to use GPU computing), and (ii) in situations where one has to apply SGDLMs to a relatively lower number of time series, CPU-based computers should suffice. Instead of using the R package or the TensorFlow library (e.g., [6]), in this study, we implement the algorithm from scratch in Python. We highly optimise the code to ensure that the runtime is reduced significantly.

2.3 Our research

To wrap up this chapter, we restate, in short form, the gaps that our research addresses, as highlighted in the previous sections of this chapter. In the current study, we summarise DLMs while directing the summary in the direction of SGDLMs. We present the algorithm of the first version of the SGDLM in detail and explain further the derivations of the formula for the joint posterior and the formulae used in the mean-field variational approximation. We apply the SGDLM to forecast daily log-returns of a multivariate financial time series consisting of 40 JSE stocks and assess the truth associated with the forecasts. We do extra analyses, for example, a comparison between the forecasts of the DLM and those of the SGDLM for a particular stock, and the effect of the number of simultaneous on forecast accuracy. We implement the code in Python and run it on a local machine with CPU hardware.

Chapter 3

Dynamic Linear Models

This chapter starts with an overview of Bayesian analysis and the Bayes' rule. Univariate Gaussian state space models are then presented, followed by dynamic linear models – a special case of the former. Some ideas on model specification are discussed. An analytic solution to the filtering/updating problem in dynamic linear models is given and related to the Kalman filter. The Kalman filter is given in three different scenarios. Standard notation from [33] dominates throughout the chapter.

3.1 A primer on Bayesian analysis

A *Bayesian analysis* is a statistical analysis that combines prior information about a population parameter with evidence from the observed data to guide the statistical inference process. In Bayesian analysis, the parameters, constant or stochastic, are treated as random variables, and full probability distributions are postulated for them. Using probability distributions, summaries of means, variances, and confidence intervals can be given. Bayesian analysis follows the *Bayes' rule*, the rule combines prior information with the observed data to get posterior information in the form of a distribution called the posterior distribution.

The Bayesian approach is different from the more conventional frequentist approach. In the latter, population parameters are considered unknown but fixed and estimated from samples. Frequentist statistics provides point estimates of the population parameters together with their standard errors and confidence intervals based on sampling but the sample probability distributions are rarely known, although, in most cases, they are assumed to be normal.

3.1.1 Bayes' rule/theorem

Bayes' rule forms the foundation of Bayesian analysis. The rule has four parts as given below[1].

- Posterior density function. This is the density of the parameter(s) after observing the data. It is simply referred to as the *posterior*[2].
- Prior density function. This is the density of the parameter(s) before observing the data, simply called the *prior*.
- Likelihood function. This is the conditional (conditioned on the parameter(s)) density of the observed data. It is simply called the *likelihood*.
- Predictive density function of the observed data under any circumstances. This is also referred to as the *evidence*.

The four are related by

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}, \tag{3.1.1}$$

which is usually written as

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

The predictive density is a normalising constant; it normalises the numerator of Equation (3.1.1) to make the posterior a true density function. Unfortunately, with most models, the explicit formulae for these densities are difficult to deal with analytically. So, Monte Carlo methods have to be used to approximate the densities through simulation.

Let y_t be the scalar value of a time series at time t and $\boldsymbol{\theta}_t$ be the vector of parameters at that time. The joint density of y_t and $\boldsymbol{\theta}_t$, $p(y_t, \boldsymbol{\theta}_t)$, can be expressed in two ways:

$$p(y_t, \boldsymbol{\theta}_t) = p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t) \tag{3.1.2}$$

and

$$p(y_t, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_t|y_t)p(y_t) \tag{3.1.3}$$

[1] In the Bayesian approach, the rule is expressed in terms of density functions; in the frequentist approach, the rule is expressed in terms of probabilities.

[2] In the literature, the word posterior may be used to mean posterior density or posterior distribution. So, the reader needs to figure out whether by writing posterior, the writer is referring to density or distribution. A similar usage applies to prior. We also follow this convention.

From Equations (3.1.2) and (3.1.3), $p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_t|y_t)p(y_t)$. Then

$$p(\boldsymbol{\theta}_t|y_t) = \frac{p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t)}{p(y_t)}, \tag{3.1.4}$$

where $p(y_t) = \int p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t)d\boldsymbol{\theta}$, is the Bayes' rule.

3.1.2 Steps of a Bayesian analysis

Typically, a Bayesian analysis goes through the following steps.

1. Today's prior is obtained from yesterday's posterior.
2. A prediction of the value of the time series is made based on today's prior.
3. Today's data is observed.
4. Today's prior is combined with today's data to update the prior into the posterior.
5. Loop into step 1 (i.e. obtain tomorrow's prior) by evolving the posterior one day forward.

3.2 State space models

State space models are dynamic models that deal with dynamic time series problems which involve unobservable parameters that describe the evolution of the state of the underlying time series. State space models allow time-varying parameters and hence account for temporal nature of data. State space models use the *Bayesian approach*.

Definition 3.2.1. *For time $t = 1, 2, \ldots T$, the Gaussian state space model for a univariate time series y_t is defined by two equations,*

$$\begin{array}{lll} \textit{Observation equation:} & y_t = h(\boldsymbol{\theta}_t) + \nu_t, & \nu_t \sim N[0, v_t], \qquad (3.2.1) \\ \textit{State equation:} & \boldsymbol{\theta}_t = g(\boldsymbol{\theta}_{t-1}) + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim N[\mathbf{0}, \boldsymbol{W}_t], \qquad (3.2.2) \\ \textit{Initial information:} & (\boldsymbol{\theta}_0|\mathcal{D}_0) \sim N[\boldsymbol{m}_0, \boldsymbol{C}_0], & \end{array}$$

where

- y_t is a scalar value called the *observation* of the time series at time t and is a Gaussian process;
- $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{pt})^T$ is a $p \times 1$ Gaussian vector of parameters at time t, known as the *state vector*;

- ν_t, a zero mean scalar with variance v_t, is called the *observational error*; and
- $\boldsymbol{\omega}_t$, a $p \times 1$ zero mean vector with covariance matrix $\boldsymbol{W}_t$, is called the *evolution error*.

Equations (3.2.1) and (3.2.2) are respectively referred to as the *observation equation* and the *state equation/evolution equation/system equation*. In the most general univariate state space model, the distributions involved do not necessarily have to be Gaussian, but Definition 3.2.1 has been tailored with the condition of Gaussianity to suit the models being studied in this thesis. The state vector $\boldsymbol{\theta}_t$ is an unobservable variable which represents the inherent properties of the time series. The state is a representation of quantitative information that summarises the history of y_t and is enough to predict the future of y_t. At time t, the information set available, $\mathcal{D}_t$, is defined by

$$\mathcal{D}_t = \{y_t, \mathcal{D}_{t-1}\}.$$

In other words, $\mathcal{D}_t$ is the set containing all past values of the time series up to and including y_t.

For $t = 1, 2, \ldots$, a state space model satisfies the following assumptions [20, Section 2.3].

1. $\boldsymbol{\theta}_t$ is a first-order Markov chain, that is, $\boldsymbol{\theta}_t$ depends on $\boldsymbol{\theta}_{t-1}$ only but independent of all state vectors before time $t - 1$.
2. Given $\boldsymbol{\theta}_t$, y_t depends on $\boldsymbol{\theta}_t$ only, and all the y_t terms are independent.

Given the two assumptions above, a state space model is completely specified if we have the distribution of $\boldsymbol{\theta}_0$ and the densities $p(y_t|\boldsymbol{\theta}_t)$ and $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ for $t = 1, 2, \ldots$.

3.2.1 Bayesian analysis scheme in state space models

$$\begin{tikzcd}
& \begin{array}{c} p(y_t|\mathcal{D}_{t-1}) \\ \text{Predictive density} \end{array} & \\
\begin{array}{c} p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \\ \text{Yesterday's posterior} \end{array} \arrow[r, "\text{Evolution}"] & \begin{array}{c} p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \\ \text{Today's prior} \end{array} \arrow[u, "\text{Prediction}"'] \arrow[r, "\text{Updating}"] & \begin{array}{c} p(\boldsymbol{\theta}_t|\mathcal{D}_t) \\ \text{Today's posterior} \end{array}
\end{tikzcd}$$

Figure 3.1: Schematic diagram showing Bayesian analysis steps in state space models.

Bayesian inference in state space models proceeds via three main steps: evolution, prediction, and updating. A typical Bayesian inference starts with the posterior distribution at time $t-1$. Using Equation (3.2.2), the posterior at time $t-1$ is *evolved* into the prior for the next day. The prior is then used to make the *prediction* of the time series at time t. Finally, the prior $p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$, the predictive density $p(y_t|\mathcal{D}_{t-1})$, and the likelihood $p(y_t|\boldsymbol{\theta}_t)$ are substituted in Bayes' theorem to *update* to the posterior $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$. Explicit equations that move through the three steps exist and are given in Proposition 3.2.2.

State updating/filtering

State vector updating, also referred to as *filtering*, is the computing of the posterior density $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$. This density helps in obtaining the value of the state vector at time t based on observations up to time t. The posterior at time t is used to evolve to time $t+1$ before a forecast of the time series at time $t+1$, y_{t+1}, is obtained.

As stated by [20, Section 2.7.1], given $p(\boldsymbol{\theta}_0|\mathcal{D}_0)$, we can recursively compute, for $t = 1, 2, \ldots$

1. The prior density $p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$ using the posterior density $p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})$ and the conditional density $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ as given by the evolution equation of the model.

2. The one-step ahead predictive density for the value of the time series at time t. This is the density $p(y_t|\mathcal{D}_{t-1})$.

3. The posterior density $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$ using Bayes' rule.

These recursions are stated more formally in the following proposition.

Proposition 3.2.2 (Filtering/updating recursions). *For the state space model defined in Definition 3.2.1, the following statements hold.*

(i) The prior is computed from the posterior $p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})$ according to

$$p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})d\boldsymbol{\theta}_{t-1}. \tag{3.2.3}$$

(ii) The one-step ahead predictive density for the observation is computed from the prior as

$$p(y_t|\mathcal{D}_{t-1}) = \int p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})d\boldsymbol{\theta}_t. \tag{3.2.4}$$

(iii) The posterior (filtered density) is computed using

$$p(\boldsymbol{\theta}_t|\mathcal{D}_t) = \frac{p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})p(y_t|\boldsymbol{\theta}_t)}{p(y_t|\mathcal{D}_{t-1})}. \tag{3.2.5}$$

Proof. First notice that, for any two random variables X and Y with joint density $p(x,y)$,

(a) the marginal densities of X and Y are respectively given by

$$p(x) = \int p(x,y)dy \quad \text{and} \quad p(y) = \int p(x,y)dx,$$

(b) the conditional densities $p(x|y)$ and $p(y|x)$ are given by

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad \text{and} \quad p(y|x) = \frac{p(x,y)}{p(x)}.$$

Therefore, (i) is proved by writing

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) &= \int p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})d\boldsymbol{\theta}_{t-1} \\ &= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathcal{D}_{t-1})p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})d\boldsymbol{\theta}_{t-1} \end{aligned}$$

Due to the Markovian nature of $\boldsymbol{\theta}_t$, given $\boldsymbol{\theta}_{t-1}$, $\boldsymbol{\theta}_t$ is independent of $\mathcal{D}_{t-1}$. Consequently,

$$p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})p(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})d\boldsymbol{\theta}_{t-1}.$$

To prove (ii), notice that

$$\begin{aligned} p(y_t|\mathcal{D}_{t-1}) &= \int p(y_t, \boldsymbol{\theta}_t|\mathcal{D}_{t-1})d\boldsymbol{\theta}_t \\ &= \int p(y_t|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})d\boldsymbol{\theta}_t \end{aligned}$$

As part of the definition of a state space model, given $\boldsymbol{\theta}_t$, y_t is independent of $\mathcal{D}_{t-1}$. Therefore,

$$p(y_t|\mathcal{D}_{t-1}) = \int p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})d\boldsymbol{\theta}_t.$$

Part (iii) follows trivially from Bayes' rule, that is, from

$$\begin{aligned} \text{posterior } &= \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}, \\ p(\boldsymbol{\theta}_t|\mathcal{D}_t) &= \frac{p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})p(y_t|\boldsymbol{\theta}_t)}{p(y_t|\mathcal{D}_{t-1})}. \end{aligned}$$

□

In the next section, a special case – and the most important form – of state space models called *dynamic linear models* is introduced.

3.3 Dynamic linear models

3.3.1 Introduction

Dynamic linear models, also called *Gaussian linear state space models*, are *state space models* that are conditionally linear and conditionally Gaussian – they are a special form of state space models, being special in the sense of linearity and Gaussianity. For DLMs, the functions $h(\boldsymbol{\theta}_t)$ and $g(\boldsymbol{\theta}_{t-1})$ in Definition 3.2.1 are linear in $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t-1}$ respectively, and all distributions involved are Gaussian. They are mainly for short-term forecasting, monitoring, and intervention analysis.

Definition 3.3.1. *The general DLM for a univariate time series y_t is defined by*

$$\begin{array}{llll} \textit{Observation equation:} & y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, & \nu_t \sim N[0, v_t] & (3.3.1) \\ \textit{State equation:} & \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim N[\mathbf{0}, \mathbf{W}_t], & (3.3.2) \\ \textit{Initial information:} & (\boldsymbol{\theta}_0 | \mathcal{D}_0) \sim N[\boldsymbol{m}_0, \boldsymbol{C}_0], & & \end{array}$$

where $t = 1, 2, \ldots, T$.

- y_t is the *observation* at time t;
- $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{pt})^T$ is the *state vector* at time t;
- $\mathbf{F}_t^T$ (transpose of $\mathbf{F}_t$) is a $1 \times p$ vector of known constants at time t;
- $\mathbf{G}_t$ is a $p \times p$ matrix of known coefficients, known as the *evolution matrix*, or the *state matrix*, or the *system matrix*, or the *transition matrix*;
- v_t, the variance of observation error ν_t, is known as the *observational variance*; and
- $\boldsymbol{W}_t$, the $p \times p$ covariance matrix of evolution error $\boldsymbol{\omega}_t$, is called the *evolution variance*.

In the simplest case, the moments $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$ as well as the sequences of the moments v_t and $\boldsymbol{W}_t$, for all t, are known. However, as discussed in Sections 3.3.4, 3.3.5, and 3.3.6, v_t and $\boldsymbol{W}_t$ may be unknown.

The DLM above is also occasionally represented as

$$\begin{aligned} (y_t | \boldsymbol{\theta}_t) &\sim N[\mathbf{F}_t^T \boldsymbol{\theta}_t, v_t], \\ (\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) &\sim N[\mathbf{G}_t \boldsymbol{\theta}_{t-1}, \mathbf{W}_t], \\ (\boldsymbol{\theta}_0 | \mathcal{D}_0) &\sim N[\boldsymbol{m}_0, \boldsymbol{C}_0]. \end{aligned}$$

The sequences $\{\nu_t\}$ and $\{\omega_t\}$ are assumed to be independent internally and mutually. That is $\text{Cov}[\nu_t, \nu_s] = \text{Cov}[\boldsymbol{\omega}_t, \boldsymbol{\omega}_s] = 0 \ \forall \ t \neq s$ and $\text{Cov}[\nu_t, \boldsymbol{\omega}_s] = 0 \ \forall \ t$ and s. They are also independent of the initial information $(\boldsymbol{\theta}_0|\mathcal{D}_0)$. In most cases, $\mathbf{F}_t$ and $\mathbf{G}_t$ are time-invariant, that is $\mathbf{F}_t = \mathbf{F}$ and $\mathbf{G}_t = \mathbf{G}$; this is the case in the current study. Specification of $\mathbf{F}_t$ and $\mathbf{G}_t$ in dynamic linear models depends on the type of the DLM being constructed. Details of their choice, together with those for the choice of v_t and $\mathbf{W}_t$, will be discussed later in this chapter. The state vector $\boldsymbol{\theta}_t$ evolves via a linear, Gaussian, first-order Markov evolution equation. Definition 3.3.1 follows directly from Definition 3.2.1 by writing the functions $h(\boldsymbol{\theta}_t)$ and $g(\boldsymbol{\theta}_{t-1})$ explicitly.

3.3.2 Local-level DLM

In this section, we give the mathematical structure of the simplest and most important DLM. This is the most widely used dynamic linear model in financial studies. Other examples of dynamic linear models are discussed in [24, Section 3.5].

Definition 3.3.2. *The local-level model or the random walk plus noise model is a univariate, uniparametric DLM defined, for each time $t \geq 1$, by*

$$\begin{array}{lll} \textit{Observation equation:} & y_t = \theta_t + \nu_t, & \nu_t \sim N[0, v_t], \\ \textit{State equation:} & \theta_t = \theta_{t-1} + \omega_t, & \omega_t \sim N[0, W_t], \\ \textit{Initial information:} & (\theta_0|\mathcal{D}_0) \sim N[m_0, C_0], & \end{array}$$

where: y_t is the *observation* at time t; θ_t is the only model parameter, referred to as, *level* of the series at time t; ν_t is the *observational error*; ω_t is the *evolution error*; v_t is the *observational variance*; and W_t is the *evolution variance*. The values of m_0, C_0, v_t, and W_t are known.

The terms y_t, θ_t, ν_t, ω_t, v_t, and W_t are all scalars. Again, the sequences $\{\nu_t\}$ and $\{\omega_t\}$ are mutually and internally independent normal random variables, and are independent of the initial information $(\theta_0|\mathcal{D}_0)$. The regression vector $\mathbf{F}_t$ and the system matrix $\mathbf{G}_t$ are given by

$$\boldsymbol{F}_t = 1 \quad \text{and} \quad \boldsymbol{G}_t = 1.$$

In other words, in this model, $\mathbf{F}_t$ and $\mathbf{G}_t$ are also scalars and both equal to unit.

3.3.3 Kalman filter for known v_t and $\boldsymbol{W}_t$

We start this section by evaluating the integrals in Proposition 3.2.2 for the local-level DLM – this solves the filtering problem for the local-level DLM. This evaluation is an insight unique to this thesis; it is not found in the literature the way we have presented it. Afterwards, we state the Kalman filter – an equivalent way of solving the filtering problem for DLMs. Throughout this section, the observational variance v_t and the evolution variance $\boldsymbol{W}_t$ are assumed to be known.

Analytic solution to the filtering problem for the local-level model

We evaluate the integrates in Equations (3.2.3), (3.2.4), and (3.2.5), for the local-level model. The limits move from $-\infty$ to $+\infty$ for all integrals since $\mathbb{R}$ is the support for the normal distribution. All quantities involved are scalars.

The integral for the prior. We have the distribution $(\theta_t|\theta_{t-1}) \sim N[\theta_{t-1}, W_t]$ from the state equation and the distribution $(\theta_{t-1}|\mathcal{D}_{t-1}) \sim N[m_{t-1}, C_{t-1}]$ from the time $t-1$ posterior. The moments W_t, m_{t-1}, and C_{t-1}, as well as the observational variance v_t, are assumed to be known. For the local-level model, Equation (3.2.3) is written as

$$\begin{aligned}
p(\theta_t|\mathcal{D}_{t-1}) &= \int_{\mathbb{R}} p(\theta_t|\theta_{t-1})p(\theta_{t-1}|\mathcal{D}_{t-1})d\theta_{t-1} \\
&= \int \frac{1}{\sqrt{2\pi W_t}}\exp\left\{\frac{-\frac{1}{2}(\theta_t-\theta_{t-1})^2}{W_t}\right\}\frac{1}{\sqrt{2\pi C_{t-1}}}\exp\left\{\frac{-\frac{1}{2}(\theta_{t-1}-m_{t-1})^2}{C_{t-1}}\right\}d\theta_{t-1} \\
&= \frac{1}{\sqrt{(2\pi)^2 W_t C_{t-1}}}\int \exp\left\{\frac{-\frac{1}{2}(\theta_t^2-2\theta_t\theta_{t-1}+\theta_{t-1}^2)}{W_t}\right\}\times \\
&\qquad \exp\left\{\frac{-\frac{1}{2}(\theta_{t-1}^2-2\theta_{t-1}m_{t-1}+m_{t-1}^2)}{C_{t-1}}\right\}d\theta_{t-1}
\end{aligned}$$

$$\begin{aligned}
p(\theta_t|\mathcal{D}_{t-1}) = \frac{1}{\sqrt{(2\pi)^2 W_t C_{t-1}}}\int \exp\Bigg\{-\frac{1}{2}\Big(\big(\frac{1}{W_t}+\frac{1}{C_{t-1}}\big)\theta_{t-1}^2 - 2\big(\frac{\theta_t}{W_t}+\frac{m_{t-1}}{C_{t-1}}\big)\theta_{t-1}+ \\
\frac{\theta_t^2}{W_t}+\frac{m_{t-1}^2}{C_{t-1}}\Big)\Bigg\}d\theta_{t-1} \qquad (3.3.3)
\end{aligned}$$

Equation (3.3.3) simplifies to (see Appendix B for details)

$$p(\theta_t|\mathcal{D}_{t-1}) = \frac{1}{\sqrt{2\pi R_t}}\exp\left\{\frac{-\frac{1}{2}(\theta_t-m_{t-1})^2}{R_t}\right\}, \qquad (3.3.4)$$

where $R_t = C_{t-1} + W_t$. Equation (3.3.4) is the density of the normal distribution with mean m_{t-1} and variance R_t. Therefore, we can write

$$(\theta_t|\mathcal{D}_{t-1}) \sim N[a_t, R_t],$$

where $a_t = m_{t-1}$ and $R_t = C_{t-1} + W_t$. Therefore, the prior distribution is Gaussian with $a_t = m_{t-1}$ as the first moment and $R_t = C_{t-1} + W_t$ as the second moment.

The integral for the predictive density. The observation equation gives the distribution $(y_t|\theta_t) \sim N[\theta_t, v_t]$. We have just proved that $(\theta_t|\mathcal{D}_{t-1}) \sim N[a_t, R_t]$. So, for the local-level model, we can write Equation (3.2.4) as

$$\begin{aligned}
p(y_t|\mathcal{D}_{t-1}) &= \int_{\mathbb{R}} p(y_t|\theta_t)p(\theta_t|\mathcal{D}_{t-1})d\theta_t \\
&= \int \frac{1}{\sqrt{2\pi v_t}}\exp\left\{\frac{-\frac{1}{2}(y_t-\theta_t)^2}{v_t}\right\}\frac{1}{\sqrt{2\pi R_t}}\exp\left\{\frac{-\frac{1}{2}(\theta_t-m_{t-1})^2}{R_t}\right\}d\theta_t \\
&= \frac{1}{\sqrt{(2\pi)^2 v_t R_t}}\int \exp\left\{\frac{-\frac{1}{2}(y_t^2-2y_t\theta_t+\theta_t^2)}{v_t}\right\}\times \\
&\qquad \exp\left\{\frac{-\frac{1}{2}(\theta_t^2-2\theta_t m_{t-1}+m_{t-1}^2)}{R_{t-1}}\right\}d\theta_t \\
&= \frac{1}{\sqrt{(2\pi)^2 v_t R_t}}\int \exp\left\{-\frac{1}{2}\Big((\frac{1}{v_t}+\frac{1}{R_t})\theta_t^2 - 2(\frac{y_t}{v_t}+\frac{m_{t-1}}{R_t})\theta_t + \right. \\
&\qquad \left.\frac{y_t^2}{v_t}+\frac{m_{t-1}^2}{R_t}\Big)\right\}d\theta_t
\end{aligned} \tag{3.3.5}$$

In a way similar to that of the integral of the prior (see Appendix B for details), Equation (3.3.5) simplifies to

$$p(y_t|\mathcal{D}_{t-1}) = \frac{1}{\sqrt{2\pi q_t}}\exp\left\{\frac{-\frac{1}{2}(y_t-m_{t-1})^2}{q_t}\right\},$$

where $q_t = R_t + v_t$. Thus, we get the density of the normal distribution with mean m_{t-1} and variance q_t. This means that the one-step ahead predictive distribution is a normal distribution with mean m_{t-1} and variance $q_t = R_t + v_t$. So, we can write

$$(y_t|\mathcal{D}_{t-1}) \sim N[f_t, q_t],$$

where $f_t = m_{t-1} = a_t$ and $q_t = R_t + v_t$.

The posterior via Bayes' rule. Equation (3.2.5) is the Bayes' rule. For the local-level model, it is written as

$$p(\theta_t|\mathcal{D}_t) = \frac{p(y_t|\theta_t)p(\theta_t|\mathcal{D}_{t-1})}{p(y_t|\mathcal{D}_{t-1})}.$$

Note that, en route to the answer we make use of $q_t = R_t + v_t$, which implies that $1 - R_t q_t^{-1} = v_t q_t^{-1}$ and $1 - v_t q_t^{-1} = R_t q_t^{-1}$, and $e_t = y_t - f_t = y_t - m_{t-1}$. The scalar e_t is called the *forecast error*, the difference between the observation and the forecast made. Using the results from the two integrals above, we write

$$\begin{aligned} p(\theta_t|\mathcal{D}_t) &= \frac{\frac{1}{\sqrt{2\pi v_t}}\exp\left\{\frac{-\frac{1}{2}\left(y_t-\theta_t\right)^2}{v_t}\right\}\frac{1}{\sqrt{2\pi R_t}}\exp\left\{\frac{-\frac{1}{2}\left(\theta_t-m_{t-1}\right)^2}{R_t}\right\}}{\frac{1}{\sqrt{2\pi q_t}}\exp\left\{\frac{-\frac{1}{2}\left(y_t-m_{t-1}\right)^2}{q_t}\right\}} \\ &= \frac{1}{\sqrt{2\pi R_t v_t q_t^{-1}}}\exp\left\{-\frac{1}{2}\left(\frac{y_t^2-2\theta_t y_t+\theta_t^2}{v_t}+\frac{\theta_t^2-2\theta_t m_{t-1}+m_{t-1}^2}{R_t}-\right.\right. \\ &\quad \left.\left.\frac{(y_t^2-2y_t m_{t-1}+m_{t-1}^2)}{q_t}\right)\right\} \end{aligned} \tag{3.3.6}$$

Equation (3.3.6) can be simplified to give (Appendix B gives details)

$$p(\theta_t|\mathcal{D}_t) = \frac{1}{\sqrt{2\pi C_t}}\exp\left\{\frac{-\frac{1}{2}\left(\theta_t - m_t\right)^2}{C_t}\right\},$$

where $m_t = m_{t-1} + R_t q_t^{-1} e_t$ and $C_t = R_t v_t q_t^{-1}$. Thus, the posterior is a normal distribution with mean m_t and variance C_t. We can therefore write

$$(\theta_t|\mathcal{D}_t) \sim N[m_t, C_t],$$

with

$$m_t = m_{t-1} + A_t e_t \quad \text{and} \quad C_t = A_t v_t = R_t - A_t^2 q_t,$$

where

$$e_t = y_t - f_t \quad \text{and} \quad A_t = R_t q_t^{-1}.$$

In DLMs, the filtering problem is solved using the *Kalman filter*. The Kalman filter is the algorithm used to evolve, forecast, and update while using dynamic linear models. It gives a summary of all the distributions encountered in the integrals above. It also gives formulae for moving from the posterior $p(\theta_{t-1}|\mathcal{D}_{t-1})$ to the posterior $p(\theta_t|\mathcal{D}_t)$ just the way we have done by integration. In DLMs, computing the integrals above explicitly is equivalent to simply applying the Kalman filter.

Theorem 3.3.3 (Kalman filter for the local-level model). *In the local-level DLM of Definition 3.3.2, the initial information, the prior distribution, the one-step ahead predictive distribution, and the posterior distribution are, for each $t \geq 1$, given by*

(a) Initial information (at time $t-1$):

$$(\theta_0|\mathcal{D}_0) \sim N[m_0, C_0],$$

for known m_0 and C_0.

(b) Prior distribution at time t:

$$(\theta_t|\mathcal{D}_{t-1}) \sim N[a_t, R_t],$$

where $a_t = m_{t-1}$ and $R_t = C_{t-1} + W_t$.

(c) One-step ahead predictive distribution at time t:

$$(y_t|\mathcal{D}_{t-1}) \sim N[f_t, q_t],$$

where $f_t = a_t$ and $q_t = R_t + v_t$.

(d) Posterior distribution at time t:

$$(\theta_t|\mathcal{D}_t) \sim N[m_t, C_t],$$

where

$$m_t = m_{t-1} + A_t e_t, \quad C_t = R_t - A_t^2 q_t$$

and

$$A_t = \frac{R_t}{q_t}, \quad e_t = y_t - f_t.$$

In the next section, we give the statement of the Kalman filter for the general univariate dynamic linear model for known v_t and $\boldsymbol{W_t}$ and give its proof.

Theorem 3.3.4 (Kalman filter for the general univariate DLM). *In the univariate dynamic linear model of Definition 3.3.1, the initial information, the prior distribution, the one-step ahead predictive distribution, and the posterior distribution are, for each $t \geq 1$, given by*

(a) Initial information (at time $t-1$):

$$(\boldsymbol{\theta}_0|\mathcal{D}_0) \sim N[\boldsymbol{m}_0, \boldsymbol{C}_0],$$

for known mean $\boldsymbol{m}_0$ and covariance matrix $\boldsymbol{C}_0$.

(b) *Prior distribution at time t:*

$$(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \sim N[\boldsymbol{a}_t, \boldsymbol{R}_t],$$

where

$$\boldsymbol{a}_t = \boldsymbol{G}_t\boldsymbol{m}_{t-1} \quad \textit{and} \quad \boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T + \boldsymbol{W}_t.$$

(c) *One-step ahead predictive distribution at time t:*

$$(y_t|\mathcal{D}_{t-1}) \sim N[f_t, q_t],$$

where

$$f_t = \boldsymbol{F}_t^T\boldsymbol{a}_t \quad \textit{and} \quad q_t = \boldsymbol{F}_t^T\boldsymbol{R}_t\boldsymbol{F}_t + v_t.$$

(d) *Posterior distribution at time t:*

$$(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim N[\boldsymbol{m}_t, \boldsymbol{C}_t],$$

where

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t \quad \textit{and} \quad \boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{A}_t q_t \boldsymbol{A}_t^T,$$

with

$$\boldsymbol{A}_t = \frac{\boldsymbol{R}_t\boldsymbol{F}_t}{q_t} \quad \textit{and} \quad e_t = y_t - f_t.$$

Proof. The proofs of (b) and (c) proceed by induction, whereas that of (d) is obtained from Bayes' rule. To prove (b), start by assuming the validity of the distribution of (a), that is,

$$(\boldsymbol{\theta}_0|\mathcal{D}_0) \sim N[\boldsymbol{m}_0, \boldsymbol{C}_0]. \tag{3.3.7}$$

We initialise time at $t-1$ and write Equation (3.3.7) as

$$(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \sim N[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}].$$

Note that, $\boldsymbol{\theta}_t$ is the sum of $\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\omega}_t$. By the first property of the multivariate normal distribution in Appendix A, $\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}$ follows the normal distribution. Thus, $\boldsymbol{\theta}_t$, being a sum of two independent normal quantities $\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\omega}_t$, is itself normal. It thus suffices to compute the mean and the variance of $(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$.
The mean $\boldsymbol{a}_t$ is given by

$$\begin{aligned}
\boldsymbol{a}_t &= E[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}] \\
&= E[(\boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t)|\mathcal{D}_{t-1}] \\
&= E[\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}] + E[\boldsymbol{\omega}_t] \; (\boldsymbol{\omega}_t \text{ does not depend on } \mathcal{D}_{t-1}) \\
&= \boldsymbol{G}_t E[\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}] \\
&= \boldsymbol{G}_t\boldsymbol{m}_{t-1}
\end{aligned}$$

and the variance $\boldsymbol{R}_t$ by

$$\begin{aligned}\boldsymbol{R}_t &= V[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}]\\ &= V[(\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}+\boldsymbol{\omega}_t)|\mathcal{D}_{t-1}]\\ &= V[\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}]+V[\boldsymbol{\omega}_t] \text{ (since } \boldsymbol{\theta}_{t-1} \text{ and } \boldsymbol{\omega}_t \text{ are independent)}\\ &= \boldsymbol{G}_t V[\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}]\boldsymbol{G}_t^T+\boldsymbol{W}_t\\ &= \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T+\boldsymbol{W}_t.\end{aligned}$$

To prove (c), notice that, y_t is a sum of two independent normal quantities $\boldsymbol{F}_t^T\boldsymbol{\theta}_t$ and ν_t, so is itself normal. It then suffices to compute the moments.
The mean f_t is given by

$$\begin{aligned}f_t &= E[y_t|\mathcal{D}_{t-1}]\\ &= E[(\boldsymbol{F}_t^T\boldsymbol{\theta}_t+\nu_t)|\mathcal{D}_{t-1}]\\ &= E[\boldsymbol{F}_t^T\boldsymbol{\theta}_t|\mathcal{D}_{t-1}]+E[\nu_t] \ (\nu_t \text{ does not depent on } \mathcal{D}_{t-1})\\ &= \boldsymbol{F}_t^T E[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}]\\ &= \boldsymbol{F}_t^T\boldsymbol{a}_t\end{aligned}$$

and the variance q_t by

$$\begin{aligned}q_t &= V[y_t|\mathcal{D}_{t-1}]\\ &= V[(\boldsymbol{F}_t^T\boldsymbol{\theta}_t+\nu_t)|\mathcal{D}_{t-1}]\\ &= V[\boldsymbol{F}_t^T\boldsymbol{\theta}_t|\mathcal{D}_{t-1}]+V[\nu_t] \text{ (since } \boldsymbol{\theta}_t \text{ and } \nu_t \text{ are independent)}\\ &= \boldsymbol{F}_t^T V[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}]\boldsymbol{F}_t+v_t\\ &= \boldsymbol{F}_t^T\boldsymbol{R}_t\boldsymbol{F}_t+v_t.\end{aligned}$$

To prove (d), first notice that, the observation equation provides the probability density function

$$p(y_t|\boldsymbol{\theta}_t) \propto \exp\Big\{-(y_t-\boldsymbol{F}_t^T\boldsymbol{\theta}_t)^T v_t^{-1}(y_t-\boldsymbol{F}_t^T\boldsymbol{\theta}_t)/2\Big\}.$$

Also, the conditional distribution $(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})$ has density

$$p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \propto \exp\Big\{-(\boldsymbol{\theta}_t-\boldsymbol{a}_t)^T\boldsymbol{R}_t^{-1}(\boldsymbol{\theta}_t-\boldsymbol{a}_t)/2\Big\}.$$

From Bayes' rule, the posterior is then given by

$$\begin{aligned}p(\boldsymbol{\theta}_t|\mathcal{D}_t) &\propto p(\boldsymbol{\theta}_t|\mathcal{D}_{t-1})p(y_t|\boldsymbol{\theta}_t)\\ &\propto \exp\Big\{-(\boldsymbol{\theta}_t-\boldsymbol{a}_t)^T\boldsymbol{R}_t^{-1}(\boldsymbol{\theta}_t-\boldsymbol{a}_t)/2-(y_t-\boldsymbol{F}_t^T\boldsymbol{\theta}_t)^T v_t^{-1}(y_t-\boldsymbol{F}_t^T\boldsymbol{\theta}_t)/2\Big\}.\end{aligned}$$

We concentrate on $p(\boldsymbol{\theta}_t|\mathcal{D}_t)$ as a function of $\boldsymbol{\theta}_t$ only and take all multiplicative factors as constants. We take natural logarithm on both sides and multiply by -2. For constants K_1 and K_2, we get

$$\begin{aligned}
-2\ln p(\boldsymbol{\theta}_t|\mathcal{D}_t) &= (\boldsymbol{\theta}_t - \boldsymbol{a}_t)^T \boldsymbol{R}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{a}_t) + (y_t - \boldsymbol{F}_t^T\boldsymbol{\theta}_t)^T v_t^{-1}(y_t - \boldsymbol{F}_t^T\boldsymbol{\theta}_t) + K_1 \\
&= (\boldsymbol{\theta}_t^T \boldsymbol{R}_t^{-1} - \boldsymbol{a}_t^T \boldsymbol{R}_t^{-1})(\boldsymbol{\theta}_t - \boldsymbol{a}_t) + (y_t^2 - y_t\boldsymbol{F}_t^T\boldsymbol{\theta}_t - y_t\boldsymbol{\theta}_t^T\boldsymbol{F}_t + \boldsymbol{\theta}_t^T\boldsymbol{F}_t\boldsymbol{F}_t^T\boldsymbol{\theta}_t)v_t^{-1} + K_1 \\
&= \boldsymbol{\theta}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{a}_t - \boldsymbol{a}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{\theta}_t + \boldsymbol{a}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + y_t^2 v_t^{-1} - y_t\boldsymbol{F}_t^T\boldsymbol{\theta}_t v_t^{-1} - \\
&\quad y_t\boldsymbol{\theta}_t^T\boldsymbol{F}_t v_t^{-1} + \boldsymbol{\theta}_t^T\boldsymbol{F}_t\boldsymbol{F}_t^T\boldsymbol{\theta}_t v_t^{-1} + K_1 \\
&= \boldsymbol{\theta}_t^T(\boldsymbol{R}_t^{-1} + \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1})\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{a}_t - \boldsymbol{a}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{\theta}_t - y_t\boldsymbol{F}_t^T\boldsymbol{\theta}_t v_t^{-1} - \\
&\quad y_t\boldsymbol{\theta}_t^T\boldsymbol{F}_t v_t^{-1} + K_2.
\end{aligned}$$

Notice that, because the products $\boldsymbol{a}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{\theta}_t$ and $\boldsymbol{F}_t^T\boldsymbol{\theta}_t$ yield scalars,

(i) $\boldsymbol{F}_t^T\boldsymbol{\theta}_t = (\boldsymbol{F}_t^T\boldsymbol{\theta}_t)^T = \boldsymbol{\theta}^T\boldsymbol{F}_t$,

(ii) $\boldsymbol{a}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{\theta}_t = (\boldsymbol{a}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{\theta}_t)^T = \Big((\boldsymbol{a}_t^T\boldsymbol{R}_t^{-1})\boldsymbol{\theta}_t\Big)^T = \boldsymbol{\theta}_t^T(\boldsymbol{a}_t^T\boldsymbol{R}_t^{-1})^T = \boldsymbol{\theta}_t^T(\boldsymbol{R}_t^{-1})^T\boldsymbol{a}_t = \boldsymbol{\theta}_t^T(\boldsymbol{R}_t^T)^{-1}\boldsymbol{a}_t = \boldsymbol{\theta}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{a}_t$.
$\boldsymbol{R}_t^T = \boldsymbol{R}_t$ because $\boldsymbol{R}_t$ is a covariance matrix so it is symmetric.

Therefore,

$$\begin{aligned}
-2\ln p(\boldsymbol{\theta}_t|\mathcal{D}_t) &= \boldsymbol{\theta}_t^T(\boldsymbol{R}_t^{-1} + \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1})\boldsymbol{\theta}_t - 2\boldsymbol{\theta}_t^T\boldsymbol{R}_t^{-1}\boldsymbol{a}_t - 2\boldsymbol{\theta}_t^T\boldsymbol{F}_t y_t v_t^{-1} + K_2 \\
&= \boldsymbol{\theta}_t^T(\boldsymbol{R}_t^{-1} + \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1})\boldsymbol{\theta}_t - 2\boldsymbol{\theta}_t^T(\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1}) + K_2. \qquad (3.3.8)
\end{aligned}$$

Let us relate the expressions $\boldsymbol{R}_t^{-1} + \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1}$ and $\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1}$ to the statement of the theorem. Let us first show that, given $(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim N[\boldsymbol{m}_t, \boldsymbol{C}_t]$ (as in the theorem statement), $\boldsymbol{C}_t^{-1} = \boldsymbol{R}_t^{-1} + \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1}$ and $\boldsymbol{C}_t^{-1}\boldsymbol{m}_t = \boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1}$. To show the former, we do the following. From $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}_t q_t^{-1}$ and $\boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{A}_t q_t\boldsymbol{A}_t^T$ (which implies that $\boldsymbol{R}_t = \boldsymbol{C}_t + \boldsymbol{A}_t q_t\boldsymbol{A}_t^T$), we get

$$\begin{aligned}
\boldsymbol{A}_t &= (\boldsymbol{C}_t + \boldsymbol{A}_t q_t\boldsymbol{A}_t^T)\boldsymbol{F}_t^T q_t^{-1} \\
&= \boldsymbol{C}_t\boldsymbol{F}_t q_t^{-1} + \boldsymbol{A}_t\boldsymbol{A}_t^T\boldsymbol{F}_t.
\end{aligned}$$

This leads to

$$\boldsymbol{A}_t(1 - \boldsymbol{A}_t^T\boldsymbol{F}_t) = \boldsymbol{C}_t\boldsymbol{F}_t q_t^{-1},$$

which implies

$$\boldsymbol{A}_t = (1 - \boldsymbol{A}_t^T\boldsymbol{F}_t)^{-1}\boldsymbol{C}_t\boldsymbol{F}_t q_t^{-1}. \qquad (3.3.9)$$

From $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}_t q_t^{-1}$, we have $\boldsymbol{R}_t\boldsymbol{F}_t = \boldsymbol{A}_t q_t$. Using this result, $q_t = \boldsymbol{F}_t^T\boldsymbol{R}_t\boldsymbol{F}_t + v_t$ gives $q_t = \boldsymbol{F}_t^T\boldsymbol{A}_t q_t + v_t$, which implies $q_t(1 - \boldsymbol{F}_t^T\boldsymbol{A}_t) = v_t$, which implies

$$q_t v_t^{-1} = (\boldsymbol{I} - \boldsymbol{F}_t^T\boldsymbol{A}_t)^{-1}. \tag{3.3.10}$$

Substitute Equation (3.3.10) in Equation (3.3.9) to get

$$\begin{aligned}\boldsymbol{A}_t &= q_t v_t^{-1}\boldsymbol{C}_t\boldsymbol{F}_t q_t^{-1}\\ &= \boldsymbol{C}_t\boldsymbol{F}_t v_t^{-1}. \qquad (3.3.11)\end{aligned}$$

From $\boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{A}_t q_t \boldsymbol{A}_t^T$ and $\boldsymbol{A}_t q_t = \boldsymbol{R}_t\boldsymbol{F}_t$, we get

$$\begin{aligned}\boldsymbol{C}_t &= \boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{A}_t^T\\ &= \boldsymbol{R}_t(\boldsymbol{I} - \boldsymbol{F}_t\boldsymbol{A}_t^T). \qquad (3.3.12)\end{aligned}$$

From Equation (3.3.11),

$$\begin{aligned}\boldsymbol{A}_t^T &= (\boldsymbol{C}_t\boldsymbol{F}_t v_t^{-1})^T\\ &= \boldsymbol{F}_t^T\boldsymbol{C}_t^T v_t^{-1}. \qquad (3.3.13)\end{aligned}$$

Substituting Equation (3.3.13) in Equation (3.3.12) followed by subsequent manipulation gives

$$\begin{aligned}\boldsymbol{C}_t &= \boldsymbol{R}_t(\boldsymbol{I} - \boldsymbol{F}_t\boldsymbol{F}_t^T\boldsymbol{C}_t^T v_t^{-1})\\ \boldsymbol{C}_t &= \boldsymbol{R}_t - v_t^{-1}\boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{F}_t^T\boldsymbol{C}_t \quad (\boldsymbol{C}_t \text{ is symmetric})\\ \boldsymbol{I} &= \boldsymbol{R}_t\boldsymbol{C}_t^{-1} - \boldsymbol{R}_t\boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1}\\ \boldsymbol{I} &= \boldsymbol{R}_t(\boldsymbol{C}_t^T - \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1})\\ \boldsymbol{C}_t^{-1} &= \boldsymbol{R}_t^{-1} + \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1}.\end{aligned}$$

To show that $\boldsymbol{C}_t^{-1}\boldsymbol{m}_t = \boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1}$, first notice that from $\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t$, $e_t = y_t - f_t$, and $\boldsymbol{A}_t = \boldsymbol{C}_t\boldsymbol{F}_t v_t^{-1}$. We then obtain

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{C}_t\boldsymbol{F}_t v_t^{-1}(y_t - f_t),$$

which implies

$$\begin{aligned}\boldsymbol{C}_t^{-1}\boldsymbol{m}_t &= \boldsymbol{C}_t^{-1}\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1} - \boldsymbol{F}_t v_t^{-1} f_t\\ &= \boldsymbol{C}_t^{-1}\boldsymbol{a}_t - \boldsymbol{F}_t\boldsymbol{F}_t^T\boldsymbol{a}_t v_t^{-1} + \boldsymbol{F}_t y_t v_t^{-1}\\ &= (\boldsymbol{C}_t^{-1} - \boldsymbol{F}_t\boldsymbol{F}_t^T v_t^{-1})\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1}\\ &= \boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \boldsymbol{F}_t y_t v_t^{-1}.\end{aligned}$$

Thus, Equation (3.3.8) can be written as

$$\begin{aligned}
-2\ln p(\boldsymbol{\theta}_t|\mathcal{D}_t) &= \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{\theta}_t - 2\boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t + K_2 \\
&= \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t - \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t + K_2 \\
&= \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t - \boldsymbol{m}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{\theta}_t + K_2 \\
&= \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t - \boldsymbol{m}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{\theta}_t + \boldsymbol{m}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t - \boldsymbol{m}_t^T \boldsymbol{C}_t^{-1}\boldsymbol{m}_t + K_2 \\
&= (\boldsymbol{\theta}_t^T - \boldsymbol{m}_t^T)\boldsymbol{C}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{m}_t) + K_3 \\
&= (\boldsymbol{\theta}_t - \boldsymbol{m}_t)^T\boldsymbol{C}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{m}_t) + K_3 \qquad (3.3.14)
\end{aligned}$$

Exponentiation of Equation (3.3.14) gives

$$p(\boldsymbol{\theta}_t|\mathcal{D}_t) \propto \exp\Big\{ - (\boldsymbol{\theta}_t - \boldsymbol{m}_t)^T\boldsymbol{C}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{m}_t)/2\Big\},$$

such that

$$(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim N[\boldsymbol{m}_t, \boldsymbol{C}_t],$$

where $\boldsymbol{m}_t$ and $\boldsymbol{C}_t$ are as defined in the theorem statement. □

3.3.4 Specification of the evolution variance $\mathbf{W}_t$ using discount factors

Unlike Section 3.3.3 where $\boldsymbol{W}_t$ has been assumed known, in a situation where $\boldsymbol{W}_t$ is not known – which is the case with most practical applications – $\boldsymbol{W}_t$ is specified by the modeller using discount factors. Evolution variance $\boldsymbol{W}_t$ is a measure of information lost about the state vector in moving from $t-1$ to t due to the presence of the stochastic error term $\boldsymbol{\omega}_t$. The information lost depends on the magnitude of $\boldsymbol{W}_t$. If $\boldsymbol{W}_t = \boldsymbol{0}$, then we have got a static model where there is no loss of information about the state in moving from $t-1$ to t. In such a model, $\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1}$. If $\boldsymbol{W}_t$ is large, then there is high uncertainty in the state evolution and a lot of information is lost in moving from $t-1$ to t.

At time $t-1$, we have the posterior distribution $(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \sim N[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}]$. Evolution variance $\boldsymbol{W}_t$ is used in obtaining

$$\boldsymbol{R}_t = V[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}] = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T + \boldsymbol{W}_t.$$

Let $\boldsymbol{P}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T$ so that

$$\boldsymbol{R}_t = \boldsymbol{P}_t + \boldsymbol{W}_t. \qquad (3.3.15)$$

The term $\boldsymbol{P}_t$ can then be defined as the prior variance of the state, $V[\boldsymbol{\theta}_t|\mathcal{D}_{t-1}]$, in a static model. Intuitively, the uncertainty reflected in $\boldsymbol{R}_t$ is expected to be bigger than or equal to the uncertainty reflected in $\boldsymbol{P}_t$ – they are only equal when $\boldsymbol{W}_t = \boldsymbol{0}$. So, for some $0 < \delta \leq 1$, we postulate that

$$\boldsymbol{R}_t = \frac{1}{\delta}\boldsymbol{P}_t. \tag{3.3.16}$$

From Equations (3.3.15) and (3.3.16),

$$\boldsymbol{P}_t + \boldsymbol{W}_t = \frac{1}{\delta}\boldsymbol{P}_t,$$

and hence

$$\boldsymbol{W}_t = \frac{1-\delta}{\delta}\boldsymbol{P}_t. \tag{3.3.17}$$

Equation (3.3.17) gives the entire sequence $\{\boldsymbol{W}_t\}$ once δ and $\boldsymbol{C}_0$ have been given.

The scalar $\delta \in (0,1]$ is called a *discount factor*. It is a correction factor which inflates $\boldsymbol{P}_t$ to make it equal to $\boldsymbol{R}_t$. The case for which $\delta = 1$ means that all information about the state is retained during the evolution for $t-1$ to t. This is the situation when $\boldsymbol{W}_t = \boldsymbol{0}$. The term $1-\delta$ is the information lost, for example, if the information lost from $t-1$ to t is 5%, then $\delta = 0.95$. Low values of δ imply high volatility in the trajectory of the time series, and the model will be sensitive to outliers. However, high discount factors lead to stable trajectories of the time series but the model will not be sensitive to outliers; such a model may not adjust to actual changes. The optimal value of a discount factor involves a balance between flexibility and stability. Typically, practical discount factors are in the range $[0.9, 0.99]$[33, Section 6.3].

3.3.5 Kalman filter when observational variance is unknown and constant

This section gives a learning mechanism for a constant unknown variance v_t and the corresponding Kalman filter formulation. The evolution variance $\boldsymbol{W}_t$ is specified via discount factors, although it can be assumed to be known.

Focus is restricted on the special case when $v_t = v$ for all t, where v is unknown. Other formulations for a constant unknown variance exist (e.g., [33, Section 10.7]). Working with the precision $\lambda := v^{-1}$ simplifies the proceedings – we work interchangeably between v and λ. A full conjugate Bayesian analysis for learning v is developed here under certain conditions. The idea is that: v is unknown and constant, but we can design a mechanism that leads to its value. The approach involves starting with an initial estimate of the value of v, s_0. As time elapses, the value of

s_t converges to the unknown value v. A conjugate analysis is when the prior and the posterior belong to the same family of distributions so that the analysis remains tractable throughout time.

To achieve conjugacy, two conditions are imposed [33, Section 4.5].

1. All variances and covariances of the DLM are scaled by $v = \lambda^{-1}$, that is, instead of working with $\boldsymbol{C}_t$, $\boldsymbol{R}_t$, $\boldsymbol{W}_t$, and q_t, we respectively work with $\lambda^{-1}\boldsymbol{C}_t^*$, $\lambda^{-1}\boldsymbol{R}_t^*$, $\lambda^{-1}\boldsymbol{W}_t^*$, and $\lambda^{-1}q_t^*$. Generality of the DLM is not lost in doing this because we can set $\boldsymbol{C}_t = \lambda^{-1}\boldsymbol{C}_t^*$, $\boldsymbol{R}_t = \lambda^{-1}\boldsymbol{R}_t^*$, $\boldsymbol{W}_t = \lambda^{-1}\boldsymbol{W}_t^*$, and $q_t = \lambda^{-1}q_t^*$ to retain the general DLM structure given in Definition 3.3.1.

2. A gamma distribution for λ, or equivalently, an inverse gamma distribution for v, for all t.

Definition 3.3.5. *For each $t \geq 1$, the constant unknown variance DLM is defined by*

$$\begin{array}{lll} \textit{Observation equation:} & y_t = \mathbf{F}_t^T\boldsymbol{\theta}_t + \nu_t, & \nu_t \sim N[0, \lambda^{-1}], \\ \textit{State equation:} & \boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim N[\mathbf{0}, \lambda^{-1}\mathbf{W}_t^*], \\ \textit{Initial information:} & (\boldsymbol{\theta}_0|\mathcal{D}_0, v) \sim N[\boldsymbol{m}_0, \lambda^{-1}\boldsymbol{C}_0^*], & (\lambda|\mathcal{D}_0) \sim Ga\left[n_0/2, d_0/2\right], \end{array}$$

where

- n_0 stands for the *degrees of freedom* of the initial information;
- $\boldsymbol{m}_0$, $\boldsymbol{C}_0^*$, n_0, and d_0 are initial quantities; and
- $E[\lambda|D_0] = \frac{n_0}{2}/\frac{d_0}{2} = n_0/d_0 = 1/s_0$, where s_0 is obtained by computing the harmonic mean[3] of v.

Kalman filter

All equations involved are analogous to those given in Theorem 3.3.4. A complete proof of the results is shown in [33, Section 4.5].

Initial information. This is information given at time $t-1$ for $t \geq 1$. It is what we feed into the model in order to work with equations at time t.

[3] Harmonic mean refers to the reciprocal of the mean of reciprocal(s). Let harmonic mean of v at time t be s_t. Then, $s_t = 1/E[1/v] = 1/E[\lambda] = d_t/n_t$, which implies $n_t/d_t = 1/s_t$. For a single scalar like v, arithmetic mean is equal to harmonic mean.

Initial information for both the state vector $\boldsymbol{\theta}_t$ and the scale parameter λ should be provided. For the former, there are two situations: conditional on v and unconditional on v. Conditional on v, as introduced in Definition 3.3.5, the initial distribution of $\boldsymbol{\theta}_t$ is of the form

$$(\boldsymbol{\theta}_0|\mathcal{D}_0, v) \sim N[\boldsymbol{m}_0, \lambda^{-1}\boldsymbol{C}_0^*]. \tag{3.3.18}$$

Unconditional on v, initial information for the state is the multivariate Student's t distribution on n_0 degrees of freedom, with mode $\boldsymbol{m}_0$ and scale matrix $\boldsymbol{C}_0$, that is,

$$(\boldsymbol{\theta}_0|\mathcal{D}_0) \sim T_{n_0}[\boldsymbol{m}_0, \boldsymbol{C}_0]. \tag{3.3.19}$$

The Student's t distribution in Equation (3.3.19) is an outcome of standard normal-gamma theory, given the distribution in Equation (3.3.18) and the distribution $(\lambda|\mathcal{D}_0) \sim \mathrm{Ga}\left[n_0/2, d_0/2\right]$. A brief discussion of this theory is given in Appendix A. Details can be found in [33, Section 17.3] and [20, Appendix A].

Initial information for the scale parameter is given by

$$(\lambda|\mathcal{D}_0) \sim \mathrm{Ga}\left[n_0/2, d_0/2\right].$$

Prior information at time t. This is also discussed for both the state $\boldsymbol{\theta}_t$ and the scale parameter λ. Conditional on v, the state's prior is given by

$$(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}, v) \sim N[\boldsymbol{a}_t, \lambda^{-1}\boldsymbol{R}_t^*]$$

and unconditional on v, by

$$(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \sim T_{n_{t-1}}[\boldsymbol{a}_t, \boldsymbol{R}_t],$$

where $\boldsymbol{a}_t = \boldsymbol{G}_t\boldsymbol{m}_{t-1}$ and $\boldsymbol{R}_t^* = \boldsymbol{G}_t\boldsymbol{C}_{t-1}^*\boldsymbol{G}_t^T + \boldsymbol{W}_t^*$. The precision $\lambda = v^{-1}$ takes on the usual gamma form, that is,

$$(\lambda|\mathcal{D}_{t-1}) \sim \mathrm{Ga}\left[n_{t-1}/2, d_{t-1}/2\right].$$

Note that the posterior at time $t-1$ is the same as prior at time t, for λ.

Forecasts at time t. Forecasts are of two catogories: one conditional on v and the other unconditional on v. Analogous to $(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}, v) \sim N[\boldsymbol{a}_t, \lambda^{-1}\boldsymbol{R}_t^*]$, the conditional forecast distribution is of the form

$$(y_t|\mathcal{D}_{t-1}, v) \sim N[f_t, \lambda^{-1}q_t^*],$$

where $f_t = \boldsymbol{F}_t^T\boldsymbol{a}_t$ and $q_t^* = 1 + \boldsymbol{F}_t^T\boldsymbol{R}_t^*\boldsymbol{F}_t$.

By the standard theory of the normal-gamma distribution and linear regression models (see Appendix A), the unconditional forecast has a univariate Student's t distribution on n_{t-1} degrees of freedom. That is,

$$(y_t|\mathcal{D}_{t-1}) \sim T_{n_{t-1}}[f_t, q_t],$$

where $q_t = s_{t-1} + \boldsymbol{F}_t^T \boldsymbol{R}_t^* s_{t-1} \boldsymbol{F}_t$. The equation $q_t = s_{t-1} + \boldsymbol{F}_t^T \boldsymbol{R}_t^* s_{t-1} \boldsymbol{F}_t$ is derived from $q_t = \boldsymbol{F}_t^T \boldsymbol{R}_t \boldsymbol{F}_t + v_t$ where the unknown $v_t = v$ is at this stage equal to its prior expected value s_{t-1} and $\boldsymbol{R}_t = v\boldsymbol{R}_t^* = s_{t-1}\boldsymbol{R}_t^*$ by definition. Note that $E(y_t|\mathcal{D}_{t-1}) = E(y_t|\mathcal{D}_{t-1}, v)$ but the variances differ.

Posterior information at time t**.** The posterior, just like the prior, is summarised for both the state parameter $\boldsymbol{\theta}_t$ and the scale parameter λ. The distributions are given by

$$(\boldsymbol{\theta}_t|\mathcal{D}_t, v) \sim N[\boldsymbol{m}_t, \lambda^{-1}\boldsymbol{C}_t^*]$$

and

$$(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim T_{n_t}[\boldsymbol{m}_t, \boldsymbol{C}_t],$$

with

$$e_t = y_t - f_t, \qquad \boldsymbol{A}_t = \boldsymbol{R}_t^* \boldsymbol{F}_t / q_t^*,$$

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t, \qquad \boldsymbol{C}_t^* = \boldsymbol{R}_t^* - \boldsymbol{A}_t \boldsymbol{A}_t^T q_t^*.$$

For the precision λ,

$$(\lambda|\mathcal{D}_t) \sim \text{Ga}\left[n_t/2, d_t/2\right],$$

where $n_t = n_{t-1} + 1$ and $d_t = d_{t-1} + e_t^2/q_t^*$. Notice that $d_t = n_t s_t$ turns $d_t = d_{t-1} + e_t^2/q_t^*$ into the common form for calculations, $s_t = s_{t-1} + \frac{s_{t-1}}{n_t}\left(\frac{e_t^2}{q_t} - 1\right)$.

To get the Kalman filter operational equations, we make the substitutions $\boldsymbol{R}_t^* = \boldsymbol{R}_t/s_{t-1}$, $q_t^* = q_t/s_{t-1}$, and $\boldsymbol{C}_t^* = \boldsymbol{C}_t/s_t$ in the equations above to get the equations below. We use the unconditional distributions for the state and forecast (e.g., [33]).

Initial information

This is made up of the known values of $\boldsymbol{m}_0, \boldsymbol{C}_0, n_0$, and s_0.

Then, for $t \geq 1$, *the following are computed*

Evolution equations

Prior mean vector:	$\boldsymbol{a}_t = \boldsymbol{G}_t \boldsymbol{m}_{t-1}$
Evolution covariance matrix:	$\boldsymbol{W}_t = \frac{1-\delta}{\delta} \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t^T$
Prior covariance matrix factor:	$\boldsymbol{R}_t = \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t^T + \boldsymbol{W}_t$

Forecasting equations

One-step ahead forecast:	$f_t = \boldsymbol{F}_t^T \boldsymbol{a}_t$
One-step ahead forecast variance factor:	$q_t = \boldsymbol{F}_t^T \boldsymbol{R}_t \boldsymbol{F}_t + s_{t-1}$

Updating/filtering equations

One-step ahead forecast error:	$e_t = y_t - f_t$
Adaptive coefficient vector:	$\boldsymbol{A}_t = \boldsymbol{R}_t \boldsymbol{F}_t / q_t$
Posterior degrees of freedom:	$n_t = n_{t-1} + 1$
Posterior observational variance estimate:	$s_t = s_{t-1} + \frac{s_{t-1}}{n_t}\left(\frac{e_t^2}{q_t} - 1\right)$
Posterior mean vector:	$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t$
Posterior covariance matrix factor:	$\boldsymbol{C}_t = \frac{s_t}{s_{t-1}}(\boldsymbol{R}_t - \boldsymbol{A}_t \boldsymbol{A}_t^T q_t)$

3.3.6 Kalman filter when observational variance is stochastic

The model described in Section 3.3.5 assumes that the unknown variance v_t is constant throughout time – this is not always true. In many practical applications, v_t changes stochastically and unpredictably over time. A discounted variance learning model for time-changing v_t is described here. The description follows the approach of [25, Section 4.3.7] and [33, Section 10.8].

We start by adjusting Definition 3.3.5 to accommodate the fact that observational variance is now time-variant. We therefore use λ_t instead of λ. Also, to simplify the notation, we get rid of all starred variances since we have seen that they are not used in the Kalman filter equations, but are only relevant in defining the structure that was imposed on the DLM of Definition 3.3.5. Specifically, we make the substitutions $\boldsymbol{W}_t^* = \boldsymbol{W}_t E[\lambda_t]$ and $\boldsymbol{C}_0^* = \boldsymbol{C}_0 E[\lambda_0]$ in the definition. We then define the resultant DLM as follows.

Definition 3.3.6. *For each $t \geq 1$, the stochastic observational variance DLM is defined by*

$$\begin{array}{llll}
\textit{Observation equation:} & y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t, & & \nu_t \sim N[0, \lambda_t^{-1}], \\
\textit{State equation:} & \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & & \boldsymbol{\omega}_t \sim N\left[\mathbf{0}, \dfrac{\boldsymbol{W}_t}{\lambda_t / E[\lambda_t]}\right], \\
\textit{Precision equation:} & \lambda_t = \dfrac{\lambda_{t-1}\eta_t}{\beta}, & & \eta_t \sim Be\left[\dfrac{\beta n_{t-1}}{2}, \dfrac{(1-\beta)n_{t-1}}{2}\right], \\
\textit{Initial information:} & (\boldsymbol{\theta}_0 | \mathcal{D}_0, v_0) \sim N\left[\boldsymbol{m}_0, \dfrac{\boldsymbol{C}_0}{\lambda_0 / E[\lambda_0]}\right], & & (\lambda_0 | \mathcal{D}_0) \sim Ga\left[n_0/2, d_0/2\right],
\end{array}$$

where

- $\beta \in (0, 1]$ is a *discount factor*;

- $\eta_t \in (0, 1]$ is a random variable named *random shock* with the beta distribution and is independent of λ_{t-1}; and
- $\boldsymbol{m}_0$, $\boldsymbol{C}_0$, n_0, and d_0 are *initial quantities*.

The posterior for the scale parameter/precision at time $t-1$ is the gamma distribution

$$(\lambda_{t-1}|\mathcal{D}_{t-1}) \sim \text{Ga}\big[n_{t-1}/2, d_{t-1}/2\big].$$

For the variance analysis to remain tractable, the prior distribution for the scale must also be gamma. Starting with the precision's posterior above, the resulting prior distribution of λ_t, following the precision equation, is the gamma distribution

$$(\lambda_t|\mathcal{D}_{t-1}) \sim \text{Ga}\Big[\frac{\beta n_{t-1}}{2}, \frac{\beta d_{t-1}}{2}\Big] \text{ [25, Section 4.3.7].}$$

The mean estimate of the precision remains unchanged, that is,

$$E[\lambda_{t-1}|\mathcal{D}_{t-1}] = E[\lambda_t|\mathcal{D}_{t-1}] = 1/s_{t-1},$$

but the prior has a bigger variance since

$$V[\lambda_t|\mathcal{D}_{t-1}] = \frac{1}{\beta}V[\lambda_{t-1}|\mathcal{D}_{t-1}].$$

When $\beta = 1$, the constant variance model where $v_t = v$ is obtained.

After observing y_t, the precision's posterior is still a gamma distribution, $(\lambda_t|\mathcal{D}_t) \sim \text{Ga}\big[n_t/2, d_t/2\big]$, but now the scale updating equations have discount factors, that is,

$$n_t = \beta n_{t-1} + 1 \quad \text{and} \quad d_t = \beta d_{t-1} + s_{t-1}\frac{e_t^2}{q_t}.$$

Kalman filter

We now give the equations for evolution, prediction, and updating for a DLM with a stochastic observational variance. Note that these equations are the same as those of the unknown constant variance model; the only change appears in the equations of the scale parameter.

Before giving the Kalman filter equations, let us make a few modifications in

$$d_t = \beta d_{t-1} + s_{t-1}\frac{e_t^2}{q_t} \tag{3.3.20}$$

and

$$\boldsymbol{C}_t = \frac{s_t}{s_{t-1}}(\boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{A}_t^T q_t)$$

to create a quantity called *volatility update factor*, like in [7]. We just make these two equations become three but we do not remove or add anything else to the equations. Notice that, by using $n_t = \beta n_{t-1} + 1$ and defining

$$r_t = \beta n_{t-1} \quad \text{and} \quad c_t = s_{t-1},$$

Equation (3.3.20) can be modified as follows.

$$\begin{aligned} d_t &= \beta d_{t-1} + s_{t-1} e_t^2 / q_t \\ n_t s_t &= \beta n_{t-1} s_{t-1} + s_{t-1} e_t^2 / q_t \\ n_t s_t &= r_t s_{t-1} + s_{t-1} e_t^2 / q_t \\ s_t &= (\frac{r_t}{n_t} + \frac{e_t^2}{n_t q_t}) s_{t-1} \\ s_t &= c_t \big(r_t + \frac{e_t^2}{q_t} \big) \frac{1}{r_t + 1} \\ s_t &= c_t z_t, \end{aligned}$$

where $z_t = (r_t + e_t^2/q_t)/(r_t + 1)$ is called the *volatility update factor*. Also, observe that $s_t = c_t z_t$ gives $z_t = s_t / s_{t-1}$ such that $\boldsymbol{C}_t = (\boldsymbol{R}_t - \boldsymbol{A}_t \boldsymbol{A}_t^T q_t) z_t$. The term c_t is the expected value of the of the variance v_t at the prior stage.

Also, before giving the analytic equations of the Kalman filter, let us summarise the distributions involved. With the DLM in Definition 3.3.6, for $t \geq 1$, we obtain the following distributions.

- Conditional on v_t,

$$(\boldsymbol{\theta}_{t-1} | \mathcal{D}_{t-1}, v_{t-1}) \sim N\Big[\boldsymbol{m}_{t-1}, \frac{\boldsymbol{C}_{t-1}}{\lambda_{t-1} s_{t-1}}\Big],$$

$$(\boldsymbol{\theta}_t | \mathcal{D}_{t-1}, v_t) \sim N\Big[\boldsymbol{a}_t, \frac{\boldsymbol{R}_t}{\lambda_t c_t}\Big],$$

$$(y_t | \mathcal{D}_{t-1}, v_t) \sim N\Big[f_t, \frac{q_t}{\lambda_t c_t}\Big],$$

$$(\boldsymbol{\theta}_t | \mathcal{D}_t, v_t) \sim N\Big[\boldsymbol{m}_t, \frac{\boldsymbol{C}_t}{\lambda_t s_t}\Big],$$

  where

$$s_{t-1} = \frac{1}{E[\lambda_{t-1} | \mathcal{D}_{t-1}]} \quad \text{and} \quad c_t = \frac{1}{E[\lambda_t | \mathcal{D}_{t-1}]} = s_{t-1}.$$

- Unconditional on v_t,

$$(\boldsymbol{\theta}_{t-1} | \mathcal{D}_{t-1}) \sim T_{n_{t-1}}[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}],$$

$$(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \sim T_{r_t}[\boldsymbol{a}_t, \boldsymbol{R}_t],$$
$$(y_t|\mathcal{D}_{t-1}) \sim T_{r_t}[f_t, q_t],$$
$$(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim T_{n_t}[\boldsymbol{m}_t, \boldsymbol{C}_t].$$

With the background above, it remains to give the analytic equations.

Initial information

Initial mean vector $\boldsymbol{m}_0$, initial covariance matrix factor $\boldsymbol{C}_0$, initial degrees of freedom n_0, and initial observational variance estimate s_0.

Then, for $t \geq 1$, we compute the following

Evolution equations

Prior mean vector:	$\boldsymbol{a}_t = \boldsymbol{G}_t\boldsymbol{m}_{t-1}$
Evolution covariance matrix:	$\boldsymbol{W}_t = \frac{1-\delta}{\delta}\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T$
Prior covariance matrix factor:	$\boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T + \boldsymbol{W}_t$
Prior observational variance estimate:	$c_t = s_{t-1}$
Prior degrees of freedom:	$r_t = \beta n_{t-1}$

Forecasting equations

One-step ahead forecast:	$f_t = \boldsymbol{F}_t^T\boldsymbol{a}_t$
One-step ahead forecast variance factor:	$q_t = \boldsymbol{F}_t^T\boldsymbol{R}_t\boldsymbol{F}_t + c_t$

Updating/filtering equations

One-step ahead forecast error:	$e_t = y_t - f_t$
Adaptive coefficient vector:	$\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}_t/q_t$
Volatility update factor:	$z_t = (r_t + e_t^2/q_t)/(r_t + 1)$
Posterior mean vector:	$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t$
Posterior covariance matrix factor:	$\boldsymbol{C}_t = (\boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{A}_t^T q_t)z_t$
Posterior degrees of freedom:	$n_t = r_t + 1$
Posterior observational variance estimate:	$s_t = c_t z_t$

Chapter 4

Simultaneous Graphical Dynamic Linear Models

In this chapter, we start by presenting the SGDLM strategy in plain language. We then introduce SGDLMs from a mathematical point of view and link them to the DLMs of Chapter 3. We follow this with the joint form of SGDLMs. Unlike [7, 8, 31], we discuss importance sampling and mean-field variational Bayes, first, in their generality and then in a way that tailors the two techniques to the current context. Finally, we present the SGDLM algorithm like in [7] but in a more elaborate form.

4.1 Introduction

In Chapter 3, we discussed models that can forecast a time series in isolation (univariate models). For example, in the stock market, such models can predict returns on investing in, say, stock **A**, without considering the effect of the changes in the prices of the other stocks that co-exist with stock **A** on the performance of stock **A**. In a multivariate system, for example, the stock market, trajectories of variables are partly directed by changes in the values of some of the other variables. So, to be realistic enough while modelling a multivariate system, models that capture the multivariate dependencies should be used. *Simultaneous Graphical Dynamic Linear Models (SGDLMs)* were introduced recently by [7] to address the need for capturing dependencies among time series while maintaining the flexibility of customising models at the level of individual time series.

In a nutshell, SGDLMs use the following way of thinking. Suppose that there are five time series: **A**, **B**, **C**, **D**, and **E**. Suppose further that these time series co-exist in a system and that there are inter-series dependencies. In particular, suppose that there is a causal relationship between **A** and **B**. The task is to forecast the value of series **A** tomorrow. Now, suppose we knew the value of series **B** tomorrow, would not this be useful information in predicting the value of series **A** tomorrow? Of course yes, it would be. So, while building a model that will forecast the value of **A** tomorrow, tomorrow's value of **B** is included in the model for **A**. We call **B** the simultaneous parent (contemporaneous predictor) of **A**. This kind of thinking can be applied while predicting tomorrow's value of any of the other stocks. So, every time series will have its customised model with customised simultaneous parent(s). This is one of the ways in which SGDLMs capture dependencies. The number of simultaneous parents is normally kept low to have parsimonious models and to avoid overfitting. For instance, we cannot have all the four series: **B**, **C**, **D**, and **E** as predictors of **A**. It is however possible to have **A** as a simultaneous predictor of **B**, and **B** as a predictor of **A**. SGDLMs involve constructing a distinct univariate DLM for each time series, where the series' predictor(s) is/are included. In addition to using simultaneous predictors, dependencies are also captured by bringing together all the DLMs at the posterior stage on a daily basis and through joint forecasting. Importance sampling is used to bring together all the DLMs – something that is referred to as *recouple*, and mean-field variational Bayes is used to separate the DLMs – to *decouple*.

4.2 Structure of simultaneous graphical dynamic linear models

In this section, we define SGDLMs and give the notation that we work with. Thereafter, we give the structure of the joint model – the model for the entire multivariate system.

4.2.1 Definition of SGDLMs and notation

An SGDLM is a joint model that consists of stochastic observational variance DLMs that were introduced in Definition 3.3.6. The observation equation in Definition 3.3.6 can be written as

$$y_t = \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t = \boldsymbol{x}_t^T \boldsymbol{\phi}_t + \boldsymbol{y}_{sp,t}^T \boldsymbol{\gamma}_t + \nu_t,$$

where $\boldsymbol{x}_t, \boldsymbol{\phi}_t, \boldsymbol{y}_{sp,t}$, and $\boldsymbol{\gamma}_t$ are column vectors. This implies that the column vectors $\boldsymbol{F}_t$ and $\boldsymbol{\theta}_t$ have been catenated as $\boldsymbol{F}_t = (\boldsymbol{x}_t^T, \boldsymbol{y}_{sp,t}^T)^T$ and $\boldsymbol{\theta}_t = (\boldsymbol{\phi}_t^T, \boldsymbol{\gamma}_t^T)^T$. Consider an m-variate time series represented by the column vector $\boldsymbol{y}_t = (y_{1t}, \dots, y_{mt})^T$, where $t = 1, 2, \dots$. Each univariate time series y_{jt}, $j = 1 : m$, is represented via a customised, stochastic variance DLM. Therefore, the full SGDLM form is defined as follows.

Definition 4.2.1. *For each $t \geq 1$, each univariate time series y_{jt} is defined by:*

$$y_{jt} = \boldsymbol{x}_{jt}^T \phi_{jt} + \boldsymbol{y}_{sp(j),t}^T \gamma_{jt} + \nu_{jt}, \qquad \nu_{jt} \sim N[0, \lambda_{jt}^{-1}], \tag{4.2.1}$$

$$\boldsymbol{\theta}_{jt} = \mathbf{G}_{jt} \boldsymbol{\theta}_{j,t-1} + \boldsymbol{\omega}_{jt}, \qquad \boldsymbol{\omega}_{jt} \sim N\Big[\mathbf{0}, \frac{\boldsymbol{W}_{jt}}{\lambda_{jt}/E[\lambda_{jt}]}\Big], \tag{4.2.2}$$

$$\lambda_{jt} = \frac{\lambda_{j,t-1}\eta_{jt}}{\beta_j}, \qquad \eta_{jt} \sim Be\Big[\frac{\beta_j n_{j,t-1}}{2}, \frac{(1-\beta_j)n_{j,t-1}}{2}\Big], \tag{4.2.3}$$

$$(\boldsymbol{\theta}_{j0}|\lambda_{j0}, \mathcal{D}_{j0}) \sim N\Big[\boldsymbol{a}_{j0}, \frac{\boldsymbol{R}_{j0}}{\lambda_{j0}c_{j0}}\Big], \qquad (\lambda_{j0}|\mathcal{D}_{j0}) \sim Ga\big[r_{j0}/2, r_{j0}c_{j0}/2\big], \tag{4.2.4}$$

where

- y_{jt} is the *scalar value* of the j^{th} time series at time t;
- $\boldsymbol{x}_{jt}$ is a vector of dimensions $p_{j\phi} \times 1$ consisting of the *lag predictor values* for series y_{jt};
- ϕ_{jt} is a vector of dimensions $p_{j\phi} \times 1$ consisting of the *regression coefficients* for the lag predictors;
- $\boldsymbol{y}_{sp(j),t}$ is a $p_{j\gamma} \times 1$ vector consisting of the time t contemporaneous values of some of the other time series, called the *simultaneous parents* (SP);
- γ_{jt} is a $p_{j\gamma} \times 1$ vector of *regression coefficients* for the simultaneous parents;
- λ_{jt}^{-1} is the *stochastic variance* of the observational error ν_{jt}, λ_{jt} is therefore the time-varying *precision*;
- $\boldsymbol{W}_{jt}$ is the *covariance matrix* of the evolution error $\boldsymbol{\omega}_{jt}$;
- $\boldsymbol{G}_{jt}$ is the *evolution matrix* of the j^{th} time series at time t;
- $\beta_j \in (0, 1]$ is a *discount factor* for stock j and η_j is a *random shock*; and
- Equation (4.2.4) gives the *initial prior* information.

In the current study, we work with the local-level model, which is very common in financial studies (e.g., [7, 36, 6]). Therefore, the lag predictors vector $\boldsymbol{x}_{jt} = 1$ such that $\boldsymbol{F}_{jt} = (1, \boldsymbol{y}_{sp(j),t}^T)^T$; $\boldsymbol{x}_{jt} = 1$ implies that $p_{j\phi} = 1$, which makes ϕ_{jt} a scalar. The state vector $\boldsymbol{\theta}_{jt}$ now has dimensions $p_j = 1 + p_{j\gamma}$, where $p_{j\gamma}$ is the number of simultaneous parents for series y_{jt}. And lastly, $\boldsymbol{G}_{jt} = \boldsymbol{I}_{p_j \times p_j}$. Thus, Equations (4.2.1) and (4.2.2) can be written as

$$y_{jt} = \phi_{jt} + \boldsymbol{y}_{sp(j),t}^T \gamma_{jt} + \nu_{jt},$$
$$\boldsymbol{\theta}_{jt} = \boldsymbol{\theta}_{j,t-1} + \boldsymbol{\omega}_{jt}.$$

The simultaneous parental set for time series y_{jt}, $sp(j)$, is given by $sp(j) \subseteq \{1,\ldots,m\} \setminus \{j\}$. In other words, the j^{th} time series has all the remaining $m-1$ time series as potential simultaneous parents. A time series is a simultaneous parent of time series y_{jt} only if it affects the behaviour of y_{jt}. As part of the model specification under the current study, simultaneous parents for a particular series remain the same for the entire period of analysis.

4.2.2 Structure of the joint model

In this section, we describe the joint model for the entire multivariate system. We start by writing the observation equation of each of the m univariate series $y_{1t}, y_{2t}, \ldots, y_{mt}$. That is,

$$\left.\begin{aligned} y_{1t} &= \phi_{1t} + \boldsymbol{y}_{sp(1),t}^T \gamma_{1t} + \nu_{1t} \\ y_{2t} &= \phi_{2t} + \boldsymbol{y}_{sp(2),t}^T \gamma_{2t} + \nu_{2t} \\ \vdots \qquad & \qquad\quad \vdots \\ y_{mt} &= \phi_{mt} + \boldsymbol{y}_{sp(m),t}^T \gamma_{mt} + \nu_{mt} \end{aligned}\right\} \tag{4.2.5}$$

Let $\boldsymbol{\mu}_t = (\phi_{1t}, \ldots, \phi_{mt})^T$, $\boldsymbol{\nu}_t = (\nu_{1t}, \ldots, \nu_{mt})^T$ and

$$\boldsymbol{\Gamma_t} = \begin{pmatrix} 0 & \gamma_{1,2,t} & \gamma_{1,3,t} & \cdots & \gamma_{1,m,t} \\ \gamma_{2,1,t} & 0 & \gamma_{2,3,t} & \cdots & \gamma_{2,m,t} \\ \gamma_{3,1,t} & \gamma_{3,2,t} & 0 & \cdots & \gamma_{3,m,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{m,1,t} & \gamma_{m,2,t} & \gamma_{m,3,t} & \cdots & 0 \end{pmatrix},$$

where the matrix $\boldsymbol{\Gamma_t}$ has zeros on its diagonal and $\gamma_{j,k,t} = 0$ if k is not in the simultaneous parental set $sp(j)$, for $j = 1 : m$. Typically, parsimonious models require $|sp(j)| << m$; this makes the matrix $\boldsymbol{\Gamma}_t$ to be sparse.

Then, Equations (4.2.5) can be represented using one equation as

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Gamma}_t \boldsymbol{y}_t + \boldsymbol{\nu}_t, \tag{4.2.6}$$

where $\boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_t^{-1})$ with $\boldsymbol{\Lambda}_t = \text{diag}(\lambda_{1t}, \ldots, \lambda_{mt})$ since the idiosyncratic terms $\nu_{1t}, \ldots, \nu_{mt}$ are mutually independent and each $\nu_{jt} \sim N[0, \lambda_{jt}^{-1}]$ for $j = 1, \ldots, m$. Equation (4.2.6) is called the *structural form* [31] of the model and can be modified to give the *reduced form* [31] (Equation (4.2.7)) as follows

$$\begin{aligned} \boldsymbol{y}_t - \boldsymbol{\Gamma}_t \boldsymbol{y}_t &= \boldsymbol{\mu}_t + \boldsymbol{\nu}_t \\ (\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{y}_t &= \boldsymbol{\mu}_t + \boldsymbol{\nu}_t \\ \boldsymbol{y}_t &= (\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}(\boldsymbol{\mu}_t + \boldsymbol{\nu}_t) \end{aligned} \tag{4.2.7}$$

Note that

$$(\boldsymbol{\mu}_t + \boldsymbol{\nu}_t) \sim N[\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t^{-1}]$$

and

$$(\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}(\boldsymbol{\mu}_t + \boldsymbol{\nu}_t) \sim N\Big[(\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}\boldsymbol{\mu}_t, (\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}\boldsymbol{\Lambda}_t^{-1}\big((\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}\big)^T\Big]$$

(by the first property of the multivariate normal distribution in Appendix A).

Therefore, conditional on $\boldsymbol{\Theta}_t = \{\boldsymbol{\theta}_{1t}, \ldots, \boldsymbol{\theta}_{mt}\}$ and $\boldsymbol{\Lambda}_t = \{\lambda_{1t}, \ldots, \lambda_{mt}\}$, where $\boldsymbol{\Theta}_t$ is the set of all the state vectors and $\boldsymbol{\Lambda}_t$ is in this situation defined as the set of all precisions, we say that $\boldsymbol{y}_t$ is conditionally distributed as follows

$$\boldsymbol{y}_t | \boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t \sim N[\boldsymbol{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t], \tag{4.2.8}$$

where

$$\boldsymbol{A}_t = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1} \text{ and } \boldsymbol{\Sigma}_t = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}\boldsymbol{\Lambda}_t^{-1}\big((\boldsymbol{I} - \boldsymbol{\Gamma}_t)^{-1}\big)^T = \Big((\boldsymbol{I} - \boldsymbol{\Gamma}_t)^T\boldsymbol{\Lambda}_t(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\Big)^{-1}.$$

4.3 Techniques used in the recouple/decouple strategy

In this section, short notes on importance sampling (the recoupling technique) and mean-field variational Bayes (the decoupling technique) are given. We introduce each of the two techniques in a general sense and then apply them to the context of SGDLMs.

4.3.1 Importance sampling

Importance sampling is a Monte Carlo method where the expectation of a function with respect to a particular distribution (target distribution) is approximated by using weighted random samples from another distribution (proposed distribution) [29]. Importance sampling is used when: (i) it is difficult or not possible to sample from the target distribution directly, and (ii) the Monte Carlo estimate is required with a smaller variance relative to a value obtained through naive/direct Monte Carlo estimation – importance sampling is, therefore, a variance reduction technique.

Consider a situation where we wish to find the expectation of a function of a random variable $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ may be a vector or scalar), with respect to a target distribution whose probability density function is $p(\boldsymbol{\theta})$. We denote this expectation as $E_p[h(\boldsymbol{\theta})]$. Naive Monte Carlo approximates this expectation using the formula

$$E_p[h(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}_i), \tag{4.3.1}$$

where $\boldsymbol{\theta}_i$, $i = 1 : N$, is sampled from $p(\boldsymbol{\theta})$.

In importance sampling, another distribution from which $\boldsymbol{\theta}_i$ is sampled is proposed. Let $g(\boldsymbol{\theta})$, hereafter referred to as the *importance density*, be the density of this distribution. The importance density $g(\boldsymbol{\theta})$ should be easy to sample from and have the same support with the target density $p(\boldsymbol{\theta})$. By definition,

$$E_p[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{4.3.2}$$

Equation (4.3.2) can be written as

$$E_p[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int h(\boldsymbol{\theta}) \cdot \frac{p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \cdot g(\boldsymbol{\theta}) d\boldsymbol{\theta} = E_g\Big[h(\boldsymbol{\theta})\frac{p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\Big] = E_g[w^*(\boldsymbol{\theta})h(\boldsymbol{\theta})],$$

where $w^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta})/g(\boldsymbol{\theta})$ is called the *importance sampling weight*. By Equation (4.3.1),

$$E_p[h(\boldsymbol{\theta})] = E_g[w^*(\boldsymbol{\theta})h(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} w^*(\boldsymbol{\theta}_i)h(\boldsymbol{\theta}_i), \tag{4.3.3}$$

where the random sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ is drawn from $g(\boldsymbol{\theta})$.

In many Bayesian applications, the target density can be obtained only up to a certain constant, that is, $p(\boldsymbol{\theta}) = C\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ can be evaluated but the constant C is unknown. In this situation, the self-normalising form of Equation (4.3.3) is used, like in [20, Section 5.1]. This is given by

$$\begin{aligned} E_p[h(\boldsymbol{\theta})] &\approx \frac{1}{\sum_{i=1}^{N} w^*(\boldsymbol{\theta}_i)} \sum_{i=1}^{N} w^*(\boldsymbol{\theta}_i)h(\boldsymbol{\theta}_i) \\ &= \sum_{i=1}^{N} w(\boldsymbol{\theta}_i)h(\boldsymbol{\theta}_i), \end{aligned}$$

where $w(\boldsymbol{\theta}_i) = \frac{w^*(\boldsymbol{\theta}_i)}{\sum_{i=1}^{N} w^*(\boldsymbol{\theta}_i)}$. The weights $w(\boldsymbol{\theta}_i)$, $i = 1 : N$, sum to 1, and can be looked at as probabilities corresponding to the samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. Therefore, the weights $w(\boldsymbol{\theta}_1), \ldots, w(\boldsymbol{\theta}_N)$ and the random sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ represent a *discrete distribution* that approximates the target distribution.

In Bayesian applications, importance sampling can be used to approximate analytically intractable posterior distributions, as well as moments like posterior expectation with respect to such distributions. We now demonstrate how posterior expectation can be obtained with respect to the posterior distribution by importance sampling without simulating the posterior distribution. We use the approach of [3, Section 13.4]. The task is to evaluate the expectation of a function of a parameter, say, $\boldsymbol{\theta}$, or any other parameter associated with the distribution, with respect to the posterior density $p(\boldsymbol{\theta}|y)$. By definition,

$$E_p[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}.$$

Let the proposed distribution have density $g(\boldsymbol{\theta})$. Then,

$$E_p[h(\boldsymbol{\theta})] = \int h(\boldsymbol{\theta}) \cdot \frac{p(\boldsymbol{\theta}|y)}{g(\boldsymbol{\theta})} \cdot g(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.3.4}$$

By Bayes' theorem, the posterior $p(\boldsymbol{\theta}|y)$ is given by

$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Therefore, Equation (4.3.4) can be written as

$$\begin{aligned}
E_p[h(\boldsymbol{\theta})] &= \int h(\boldsymbol{\theta}) \cdot \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})\int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \cdot g(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \frac{1}{\int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \int h(\boldsymbol{\theta}) \cdot \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \cdot g(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \frac{\int h(\boldsymbol{\theta}) \cdot \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \cdot g(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \cdot g(\boldsymbol{\theta})d\boldsymbol{\theta}}
\end{aligned}$$

Then, by the definition of direct Monte Carlo sampling,

$$\begin{aligned}
E_p[h(\boldsymbol{\theta})] &\approx \frac{\frac{1}{N}\sum_{i=1}^{N}\left(h(\boldsymbol{\theta}_i) \cdot \frac{p(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}\right)}{\frac{1}{N}\sum_{i=1}^{N}\frac{p(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}} \\
&= \frac{\sum_{i=1}^{N}\left(h(\boldsymbol{\theta}_i) \cdot \frac{p(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}\right)}{\sum_{i=1}^{N}\frac{p(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}},
\end{aligned} \tag{4.3.5}$$

where $\boldsymbol{\theta}_i$, $i = 1 : N$, is drawn from $g(\boldsymbol{\theta})$. Thus, given the likelihood $p(y|\boldsymbol{\theta})$, the prior $p(\boldsymbol{\theta})$, and the importance density $g(\boldsymbol{\theta})$, the value of $E_p[h(\boldsymbol{\theta})]$ can be approximated using Equation (4.3.5) without simulating the posterior $p(\boldsymbol{\theta}|y)$. Note that the product of the likelihood and the prior divided by the importance density gives the weights, that is, $\frac{p(y|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)} = w^*(\boldsymbol{\theta}_i)$.

Effective sample size

Effective sample size (ESS) is one measure of the effectiveness of importance sampling [17]. It refers to the corresponding number of independent samples that must be drawn from the exact distribution to produce the same level of efficiency like in the estimation obtained when a Monte Carlo approximation like importance sampling is used [17]. The most common formula for calculating ESS in importance sampling is

$$\text{ESS} = \frac{1}{\sum_{i=1}^{N} w_i^2},$$

where w_i are the normalised importance weights (e.g., [17, 31]). If 1000 samples are used in the importance sample approximation, and ESS is computed to be 900, then the approximation is 90% effective.

4.3.2 Mean-field variational Bayes

Variational Bayes methods turn the analytic approximation of an intractable problem into an optimisation problem [8, Section 4.6]. A typical problem is the analytic approximation of an intractable posterior distribution by a more tractable distribution. The idea is to posit a family of distributions, then select a distribution from the family whose parameters minimise the difference between the exact distribution and the approximating distribution [8, Section 4.6]. The difference between the exact distribution and the approximating distribution is measured by a loss function, *Kullback-Leibler (KL) divergence* (e.g., [33, Section 12.3.4]) being the most common such measure.

Mean-field variational Bayes, also called *mean-field variational inference* or *mean-field approximation*, is a variational Bayes method which assumes that the variational family approximating the intractable problem factorises into a product of independent forms [30]. For example, the mean-field approximation of a joint posterior is the product of the marginal forms which are assumed to be independent. Let p be the target distribution and q be the approximating distribution, with $p(\cdot)$ and $q(\cdot)$ as the respective probability density functions. Let $\mathbf{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ be a collection of random vectors. Then, the mean-field approximation of the joint density $p(\mathbf{\Theta})$ is given by

$$p(\mathbf{\Theta}) \approx q(\mathbf{\Theta}) = \prod_{j=1}^{m} q(\boldsymbol{\theta}_j),$$

where the vectors $\boldsymbol{\theta}_j$, $j = 1 : m$, are independent random variables.

Kullback-Leibler divergence

This is a standard measure of the efficacy of mean-field approximation. It is a non-symmetric measure that quantifies numerically the difference between the two probability distributions p and q. It can be interpreted as a measure of information lost when we approximate p with q. It is valid for both discrete and continuous distributions. We denote Kullback-Leibler divergence of q from p as $KL(p||q)$ and give its definition like in [30] as

$$KL(p||q) = E_p\Big[\log\Big(\frac{p(\mathbf{\Theta})}{q(\mathbf{\Theta})}\Big)\Big], \tag{4.3.6}$$

where $E_p[\cdot]$ stands for expectation with respect to distribution p and $\mathbf{\Theta}$ is a random variable. An assumption on the densities p and q is that they should have the same support. $KL(p||q) \geq 0$, and $KL(p||q) = 0$ iff $p = q$. It should be noted that the term divergence is a misnomer; in the context of MFVB, it does not take on its meaning in vector calculus. KL is also not a distance measure, therefore $KL(p||q) \neq KL(q||p)$. The strategy for mapping p to q is to find parameters of q that minimise $KL(p||q)$ or $KL(q||p)$.

When the distribution being approximated is a posterior, minimising KL divergence using Equation (4.3.6) becomes difficult because the expansion of the right-hand side of the equation gives rise to *evidence*, which is, in most situations, intractable due to high-dimensional integration. Instead, a term referred to as *evidence lower bound* is maximised (e.g., [30]). Maximising evidence lower bound is the same as minimising KL divergence [30]. Fortunately, if the distribution that is being approximated is represented in form of a Monte Carlo sample, like an importance sample, minimisation of KL divergence can be done trivially using an entropy measure (e.g., [7]), where entropy in this situation means uncertainty associated with a random variable. Entropy is another measure of the efficacy of importance sampling; in other words, it is a measure of the uncertainty associated with the importance sample. When the distribution being approximated is represented in form of an importance sample, KL minimisation can be done using entropy of the importance sample. A good importance sample should have a low value of entropy. In [7], the entropy H_N of importance sampling weights relative to uniformity is defined as $H_N = \sum_{i=1}^{N} w_i \log_e(Nw_i)$, where N is the size of the importance sample. As $N \to \infty$, $H_N \to KL(p||q)$ [7]. It can be proved that $H_N \leq N/ESS - 1$ [7].

4.4 The SGDLM algorithm: sequential forecasting, filtering, and evolution

In this section, following [7], we give the equations for forecasting, filtering, and evolution in SGDLMs. Unlike [7, 10, 31, 6], we present the algorithm with sufficient detail to make it easier for the reader to understand it and implement if it deems necessary. This is the algorithm we implement in Chapter 5 to do the stock data analysis. Unless otherwise stated, parameters like $\boldsymbol{a}_t, \boldsymbol{R}_t, r_t$, et cetera, and similar terms remain as defined in Chapter 3, with the only extension that they are now considered for each series separately.

The SGDLM algorithm can be given in six coherent steps.

1. *Initial prior at time t.* Start with decoupled, conjugate, normal-gamma priors $p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1})$, for each $j = 1 : m$, at time t. A modelling assumption is that these priors are independent. Each of these independent normal-gamma forms is jointly represented as

$$(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1}) \sim NG[\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}, r_{jt}, c_{jt}].$$

The precision λ_{jt} follows the gamma distribution with the representation

$$(\lambda_{jt}|\mathcal{D}_{j,t-1}) \sim \text{Ga}[r_{jt}/2, r_{jt}c_{jt}/2]. \tag{4.4.1}$$

Conditional on the precision λ_{jt}, the state vector $\boldsymbol{\theta}_{jt}$ follows the multivariate normal distribution, that is,

$$(\boldsymbol{\theta}_{jt}|\lambda_{jt}, \mathcal{D}_{j,t-1}) \sim N[\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}/(\lambda_{jt}c_{jt})]. \tag{4.4.2}$$

Unconditional on λ_{jt}, the state $\boldsymbol{\theta}_{jt}$ follows the multivariate Student's t distribution, represented as

$$(\boldsymbol{\theta}_{jt}|\mathcal{D}_{j,t-1}) \sim T_{r_{jt}}[\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}]. \tag{4.4.3}$$

Therefore, to start the analysis at time t, values of the parameters $\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}, r_{jt}$, and c_{jt} must be known. Typically, the analysis starts by setting $t = 0$, so $\boldsymbol{a}_{j0}, \boldsymbol{R}_{j0}, r_{j0}$, and c_{j0}, referred to as initial values, must be known. Since the individual priors are independent, the implied joint prior is a product of the independent forms. The joint prior is therefore given by

$$p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_{t-1}) = \prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1}), \tag{4.4.4}$$

where $\boldsymbol{\Theta}_t = \{\boldsymbol{\theta}_{1t}, \ldots, \boldsymbol{\theta}_{mt}\}$, $\boldsymbol{\Lambda}_t = \{\lambda_{1t}, \ldots, \lambda_{mt}\}$, and $\mathcal{D}_{t-1} = \{\mathcal{D}_{1,t-1}, \ldots, \mathcal{D}_{m,t-1}\}$.

Please note that the theory discussed in Chapter 3 starts analyses at the posterior stage. However, for the current algorithm, the analysis is started at the prior stage. This difference does not matter. What matters is using the evolution and updating equations at the right place. Standard theory on DLMs like in [33, 24, 25, 20] starts analyses at the posterior stage. However, SGDLMs were introduced by [7] with analysis starting at the prior stage. In the current study, we adopt the conventional approach of starting at the posterior stage in Chapter 3 (for DLMs) but switch to the approach of starting at the prior stage in Chapters 4 and 5 for SGDLMs (like [7, 10, 6, 31]). One could as well maintain the conventional DLMs approach under SGDLMs (e.g., [35]).

2. *Time t joint forecasting.* At this step, time t forecasts for all the time series are obtained in a *recoupling* approach that exploits Equation (4.2.8). The precisions and states are simulated independently across $j = 1 : m$ using the distributions in (4.4.1) and (4.4.2) respectively. That is, for some large K, K values of λ_{jt} are sampled from its distribution in Equation (4.4.1). Then, each sampled value of λ_{jt} is plugged in Equation (4.4.2) to sample one value of $\boldsymbol{\theta}_{jt}$. In the process, K samples of λ_{jt} and K samples of $\boldsymbol{\theta_{jt}}$ are generated. Samples for all the time series are combined to form the multivariate Monte Carlo samples $\{\boldsymbol{\Theta}_t^k, \boldsymbol{\Lambda}_t^k\}$, $k = 1 : K$. The samples $\{\boldsymbol{\Theta}_t^k, \boldsymbol{\Lambda}_t^k\}$ are transformed to form K first moments and K second moments for the reduced form of Equation (4.2.8), that is, the moments $(\boldsymbol{A}_t\boldsymbol{\mu}_t)^k$ and $\boldsymbol{\Sigma}_t^k$. The moments $(\boldsymbol{A}_t\boldsymbol{\mu}_t)^k$ are averaged to obtain the mean $\boldsymbol{A}_t\boldsymbol{\mu}_t$ of the vector $\boldsymbol{y}_t$ and a similar thing is done to $\boldsymbol{\Sigma}_t^k$ to obtain the corresponding covariance matrix $\boldsymbol{\Sigma}_t$. Using this mean and covariance matrix, K values of the multivariate normal distribution $\boldsymbol{y}_t \sim N[\boldsymbol{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t]$ are drawn and averaged to obtain the forecasts for all time series at time t.

The following should be noted:

(i) Simulation of the states, precisions, and forecasts is necessary to ensure that uncertainties associated with expected values are captured in the analysis. Obtaining the moments $\boldsymbol{A}_t\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ directly from the prior of step 1 without simulation would give the coherent forecasts and the covariance matrix, but this does not suffice as it ignores the uncertainties.

(ii) Instead of simulating the states from the normal form of Equation (4.4.2) using values of λ_{jt} obtained from Equation (4.4.1), one could just sample these states from the multivariate Student's t form of Equation (4.4.3) – this does not require knowing λ_{jt} values.

(iii) Recoupling is necessary in step 2 to enable capturing of relationships among time series for forecasting.

(iv) The conditioning in $\boldsymbol{y}_t|\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t \sim N[\boldsymbol{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t]$ is dropped for notational purposes. In the current study, like in [33], known quantities are made implicit in conditioning distributions. Values of $(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t)|\mathcal{D}_{t-1}$ are obtained in step 1, so step 2 is done when they are already known, this leads to writing $\boldsymbol{y}_t \sim N[\boldsymbol{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t]$ instead of $\boldsymbol{y}_t|\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t \sim N[\boldsymbol{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t]$. Otherwise one could still carry on with making the conditioning explicit. This kind of notation is mentioned in [33, Section 4.2].

3. *Updating to time t naive posterior.* The priors $p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1})$ of step 1 are updated to the *naive* posteriors $\tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})$ using the updating equations given in Section 3.3.6; the tilde notation stands for naive. At this stage, these posteriors are naive because they do not account for cross-series dependencies. Exact posteriors will be discussed in steps 4 and 5. Note that, jointly

$$(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt}) \sim NG[\tilde{\boldsymbol{m}}_{jt}, \tilde{\boldsymbol{C}}_{jt}, \tilde{n}_{jt}, \tilde{s}_{jt}],$$

where $\tilde{\boldsymbol{m}}_{jt}, \tilde{\boldsymbol{C}}_{jt}, \tilde{n}_{jt},$ and $\tilde{s}_{jt}$ are the parameters of the naive normal-gamma posteriors. These parameters are analogous to the prior parameters: $\boldsymbol{a}_{jt}, \boldsymbol{R}_{jt}, r_{jt},$ and c_{jt}. The normal-gamma form of the naive posteriors is a consequence of

$$(\lambda_{jt}|\mathcal{D}_{jt}) \sim \text{Ga}[\tilde{n}_{jt}/2, \tilde{n}_{jt}\tilde{s}_{jt}/2] \tag{4.4.5}$$

and

$$(\boldsymbol{\theta}_{jt}|\lambda_{jt}, \mathcal{D}_{jt}) \sim N[\tilde{\boldsymbol{m}}_{jt}, \tilde{\boldsymbol{C}}_{jt}/(\lambda_{jt}\tilde{s}_{jt})]. \tag{4.4.6}$$

Equations for doing this update are the same like those in Section 3.3.6, but they should now have the index j to emphasise that they are applied series by series. They are given by:

First compute the quantities: $f_{jt} = \boldsymbol{F}_{jt}^T\boldsymbol{a}_{jt}$, $q_{jt} = \boldsymbol{F}_{jt}^T\boldsymbol{R}_{jt}\boldsymbol{F}_{jt} + c_{jt}$, $e_{jt} = y_{jt} - f_{jt}$, $\boldsymbol{A}_{jt} = \boldsymbol{R}_{jt}\boldsymbol{F}_{jt}/q_{jt}$, and $z_{jt} = (r_{jt} + e_{jt}^2/q_{jt})/(r_{jt}+1)$.

Then compute the naive posterior parameters: $\tilde{\boldsymbol{m}}_{jt} = \boldsymbol{a}_{jt} + \boldsymbol{A}_{jt}e_{jt}$, $\tilde{\boldsymbol{C}}_{jt} = (\boldsymbol{R}_{jt} - \boldsymbol{A}_{jt}\boldsymbol{A}_{jt}^Tq_{jt})z_{jt}$, $\tilde{n}_{jt} = r_{jt}+1$, and $\tilde{s}_{jt} = c_{jt}z_{jt}$.

The product of the independent naive posteriors $\tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})$ gives the naive joint posterior $\tilde{p}(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_t)$ which is, as we demonstrate in step 4, very close to the exact joint posterior $p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_t)$.

$$\tilde{p}(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_t) = \prod_{j=1}^{m}\tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt}) \tag{4.4.7}$$

4. *Recouple by importance sampling to obtain the time t exact joint posterior.* The exact joint posterior is given by the formula

$$p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_t) \propto |\boldsymbol{I} - \boldsymbol{\Gamma}_t|\prod_{j=1}^{m}\tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt}) \tag{4.4.8}$$

(see Appendix C for the proof). The form of Equation (4.4.8) presents a challenge in the analysis; it is not possible to simulate the exact posterior series by series using this equation. Whereas the product $\prod_{j=1}^{m} \tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})$ factorises compositionally into the independent conjugate normal-gamma forms, the term $|\boldsymbol{I} - \boldsymbol{\Gamma}_t|$ does not – this determinant is coupled across all series. As a result, the product $|\boldsymbol{I} - \boldsymbol{\Gamma}_t| \prod_{j=1}^{m} \tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})$ makes the independent conjugate forms inaccessible for simulation. Recall, SGDLMs leverage simulation of individual series.

This is therefore a typical situation when importance sampling becomes handy. An alternative distribution for simulation of the posterior has to be proposed. As mentioned in Section 4.2.2, practically workable models require that the matrix $\boldsymbol{\Gamma}_t$ is sparse. This make the absolute value of the determinant $|\boldsymbol{I} - \boldsymbol{\Gamma}_t|$ to be close to 1 [31]. This implies that the naive joint posterior in Equation (4.4.7) is almost the exact posterior, so it can be used as an alternative distribution from which precisions and states can be sampled for importance sampling to take effect. The unknown proportionality constant will be taken care of in the importance sampling strategy as discussed in Section 4.3.1. The product form of Equation (4.4.7) is exploited for independent sampling. Using the naive posteriors, precisions and states are simulated independently for all series, in away that is similar to the simulation approach used in step 2. Samples from all the series are combined to form joint Monte Carlo samples $\{\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i\}$, $i = 1 : N$, for some large N.

From *posterior* $\propto$ *likelihood* $\times$ *prior*, we conclude that the term

$$|\boldsymbol{I} - \boldsymbol{\Gamma}_t| \prod_{j=1}^{m} \tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})$$

is the product of the joint likelihood and the joint prior. Using Equation (4.3.5), it was concluded that the product of the likelihood and the prior divided by importance density gives the importance weight. Therefore, using the forms in Equations (4.4.7) and (4.4.8), we can write

$$w^*(\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i) = \frac{|\boldsymbol{I} - \boldsymbol{\Gamma}_t^i| \prod_{j=1}^{m} \tilde{p}(\boldsymbol{\theta}_{jt}^i, \lambda_{jt}^i|\mathcal{D}_{jt})}{\prod_{j=1}^{m} \tilde{p}(\boldsymbol{\theta}_{jt}^i, \lambda_{jt}^i|\mathcal{D}_{jt})} = |\boldsymbol{I} - \boldsymbol{\Gamma}_t^i|.$$

The normalised weights $w(\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i)$ are given by

$$w(\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i) = \frac{1}{\sum_{i=1}^{N} |\boldsymbol{I} - \boldsymbol{\Gamma}_t^i|} |\boldsymbol{I} - \boldsymbol{\Gamma}_t^i|.$$

The exact joint posterior is then approximated by the importance sample

$$\{\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i, w(\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i), i = 1 : N\}. \tag{4.4.9}$$

Note that $w(\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i) = w(\boldsymbol{\theta}_{jt}^i) = w(\lambda_{jt}^i)$ for all $j = 1 : m$. For simplicity of notation, hereafter, we write $w(\boldsymbol{\Theta}_t^i, \boldsymbol{\Lambda}_t^i)$ as w_{it}.

Effective sample size is calculated at every value of t using the formula

$$\mathrm{ESS}_t = \frac{1}{\sum_{i=1}^N w_{it}^2}$$

to monitor the efficiency of the importance sample.

5. *Decouple by mean-field variational Bayes to get back the independent, decoupled normal-gamma posteriors.* To achieve conjugacy, the importance sample-based joint posterior (p_{MC}) is mapped to a product of independent normal-gamma forms (q). The MFVB-based joint posterior is now written as

$$q(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t | \mathcal{D}_t) = \prod_{j=1}^m q(\boldsymbol{\theta}_{jt}, \lambda_{jt} | \mathcal{D}_{jt}), \tag{4.4.10}$$

with each $(\boldsymbol{\theta}_{jt}, \lambda_{jt} | \mathcal{D}_{jt}) \sim NG[\boldsymbol{m}_{jt}, \boldsymbol{C}_{jt}, n_{jt}, s_{jt}]$, where $\boldsymbol{m}_{jt}$, $\boldsymbol{C}_{jt}$, n_{jt}, and s_{jt} are the parameters of the resultant MFVB-based decoupled posteriors. By denoting $E_{p_{MC}}[\cdot]$ as expectation with respect to the importance sample, the formulae (whose derivations are given in Appendix C) for obtaining the parameters are:

- Obtain $\boldsymbol{m}_{jt}$ from

$$\boldsymbol{m}_{jt} = E_{p_{MC}}[\lambda_{jt}\boldsymbol{\theta}_{jt}] / E_{p_{MC}}[\lambda_{jt}] = \sum_{i=1}^N (w_{it}\lambda_{jt}^i \boldsymbol{\theta}_{jt}^i) \Big/ \sum_{i=1}^N (w_{it}\lambda_{jt}^i).$$

- Calculation of n_{jt}: First obtain the intermediate quantities

$$\boldsymbol{V}_{jt} = E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T] = \sum_{i=1}^N \left(w_{it}\lambda_{jt}^i (\boldsymbol{\theta}_{jt}^i - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt}^i - \boldsymbol{m}_{jt})^T \right)$$

  and

$$d_{jt} = E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T \boldsymbol{V}_{jt}^{-1} (\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})] = \sum_{i=1}^N w_{it}\lambda_{jt}^i (\boldsymbol{\theta}_{jt}^i - \boldsymbol{m}_{jt})^T \boldsymbol{V}_{jt}^{-1} (\boldsymbol{\theta}_{jt}^i - \boldsymbol{m}_{jt}),$$

  then calculate n_{jt} by solving the numerical equation

$$\log_e(n_{jt} + p_j - d_{jt}) - \psi(\frac{n_{jt}}{2}) - \frac{(p_j - d_{jt})}{n_{jt}} - \log_e(2E_{p_{MC}}[\lambda_{jt}]) + E_{p_{MC}}[\log_e \lambda_{jt}] = 0,$$

which implies

$$\log_e(n_{jt}+p_j-d_{jt})-\psi(\frac{n_{jt}}{2})-\frac{(p_j-d_{jt})}{n_{jt}}-\log_e(2\sum_{i=1}^{N}w_{it}\lambda_{jt}^{i})+\sum_{i=1}^{N}w_{it}\log_e\lambda_{jt}^{i}=0,$$

where $\psi(\cdot)$ is the digamma function.

- Obtain s_{jt} from

$$s_{jt}=(n_{jt}+p_j-d_{jt})/n_{jt}E_{p_{MC}}[\lambda_{jt}]=(n_{jt}+p_j-d_{jt})\Big/n_{jt}\sum_{i=1}^{N}w_{it}\lambda_{jt}^{i}.$$

- Finally, obtain $\boldsymbol{C}_{jt}$ from $\boldsymbol{C}_{jt}=s_{jt}\boldsymbol{V}_{jt}$.

Calculate $KL(p_{MC}||q)_t$ from $KL(p_{MC}||q)_t \approx \sum_{i=1}^{N}w_{it}\log_e(Nw_{it})$ and check whether $KL(p_{MC}||q)_t \leq N/\text{ESS}_t-1$ to monitor the efficacy of the mean-field approximation at every value of t.

6. *Independent state and precision evolution to time $t+1$.* The time t MFVB posteriors are evolved independently to time $t+1$ priors by means of evolution equations given in Section 3.3.6; this takes us back to the point where we started. This is the evolution from $(\boldsymbol{\theta}_{jt},\lambda_{jt}|\mathcal{D}_{jt})$ to $(\boldsymbol{\theta}_{j,t+1},\lambda_{j,t+1}|\mathcal{D}_{jt})$, with

$$(\boldsymbol{\theta}_{j,t+1},\lambda_{j,t+1}|\mathcal{D}_{jt})\sim NG[\boldsymbol{a}_{j,t+1},\boldsymbol{R}_{j,t+1},r_{j,t+1},c_{j,t+1}].$$

Evolution equations are: $\boldsymbol{a}_{j,t+1}=\boldsymbol{m}_{jt}$, $\boldsymbol{R}_{j,t+1}=\boldsymbol{C}_{jt}+\boldsymbol{W}_{j,t+1}$, $c_{j,t+1}=s_{jt}$, and $r_{j,t+1}=\beta n_{jt}$. Evolution variance $\boldsymbol{W}_{j,t+1}$ is specified using two discount factors: δ_ϕ for the local-level component and δ_γ for the simultaneous parents component, following the standard block discounting approach of [33, Sections 6.3.2 and 10.2.2].

4.5 Overview of the analytic solution to the SGDLM analysis

In this section, we lay out the integrals that, if computed, give the analytic solution that we approximated using the algorithm in Section 4.4. This is our original insight into the problem, as we found no study that looks at the problem in this direction.

We wish to state how the SGDLM analysis can be handled analytically as we did with the DLM analysis in Section 3.3.3. By Proposition 3.2.2 and assuming independence of the time series at the prior, posterior, and evolution stages,

(a) The prior distribution for all the states and precisions is given by

$$\begin{aligned}p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_{t-1}) &= \int p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Lambda}_{t-1})p(\boldsymbol{\Theta}_{t-1}, \boldsymbol{\Lambda}_{t-1}|\mathcal{D}_{t-1})d\boldsymbol{\Theta}_{t-1}d\boldsymbol{\Lambda}_{t-1}\\ &= \underbrace{\int \ldots \int}_{m(p_j+1)} \Bigg(\prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\boldsymbol{\theta}_{j,t-1}, \lambda_{j,t-1}) \prod_{j=1}^{m} p(\boldsymbol{\theta}_{j,t-1}, \lambda_{j,t-1}|\mathcal{D}_{j,t-1})\\ &\quad d\boldsymbol{\theta}_{1,t-1} \cdots d\boldsymbol{\theta}_{m,t-1} d\lambda_{1,t-1} \cdots d\lambda_{m,t-1}\Bigg),\end{aligned}$$

(b) The one-step ahead predictive distribution is given by

$$\begin{aligned}p(\boldsymbol{y}_t|\mathcal{D}_{t-1}) &= \int p(\boldsymbol{y}_t|\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t)p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_{t-1})d\boldsymbol{\Theta}_t d\boldsymbol{\Lambda}_t\\ &= \underbrace{\int \cdots \int}_{m(p_j+1)} \Bigg(\bigg(|\boldsymbol{I} - \boldsymbol{\Gamma}_t| \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{\theta}_{jt}, \lambda_{jt})\bigg) \times \prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1})\\ &\quad d\boldsymbol{\theta}_{1,t-1} \cdots d\boldsymbol{\theta}_{m,t-1} d\lambda_{1,t-1} \cdots d\lambda_{m,t-1}\Bigg),\end{aligned}$$

(c) The posterior distribution is given by

$$\begin{aligned}p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_t) &= \frac{p(\boldsymbol{y}_t|\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t)p(\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t|\mathcal{D}_{t-1})}{p(\boldsymbol{y}_t|\mathcal{D}_{t-1})}\\ &= \frac{\left(|\boldsymbol{I} - \boldsymbol{\Gamma}_t| \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{\theta}_{jt}, \lambda_{jt})\right) \times \prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1})}{\underbrace{\int \cdots \int}_{m(p_j+1)} \Bigg(\left(|\boldsymbol{I} - \boldsymbol{\Gamma}_t| \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{\theta}_{jt}, \lambda_{jt})\right) \times \prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1})}\\ &\quad d\boldsymbol{\theta}_{1,t-1} \cdots d\boldsymbol{\theta}_{m,t-1} d\lambda_{1,t-1} \cdots d\lambda_{m,t-1}\Bigg).\end{aligned}$$

Chapter 5

Data Analysis and Results

In this chapter, we apply the SGDLM to stock data. We start by describing the data set. We follow this by stating how we implement the SGDLM algorithm, clearly stating how we divide the implementation into three phases. We then give the results of the test data analysis. The results include: the coverage of prediction intervals by the SGDLM; a comparison between the performances of the SGDLM and the DLM; a comparison between the observed trend of returns and the SGDLM trend; an assessment of the success of the recouple/decouple strategy; a look into the effect of the number of simultaneous parents on model accuracy; and finally, a look into the computation time.

5.1 Data set

Our data set is the *daily log-returns* of 40 JSE stocks that were selected from the Top 100 JSE index. Using the Industry Classification Benchmark method of categorising companies, each stock can be categorised into one of the sectors of financials, basic materials, consumer goods, consumer services, technology, telecommunications, industrials, and health care. The stocks, and their sector categorisations, are given in Table 5.1. We download the daily closing prices of all the stocks from Yahoo! Finance [1] for the period 01/01/2014 to 30/06/2022. Apart from Anglo American Platinum Limited, which has a missing closing price for 28/02/2020, and Sasol Limited, which has a missing closing price for 30/12/2020, all the other companies have all the closing prices for the entire period. All closing prices are in units of South African Rand (ZAR). The missing closing prices for the two companies are in each case filled by taking the mean of the two closing prices flanking the missing value. We replace the closing prices of all the stocks on 14/06/2022 by the respective averages of 13/06/2022 and 15/06/2022 after identifying that nine companies have closing prices that are much smaller on 14/06/2022, for example, Shoprite

Holdings Limited has ZAR 20998 on 13/06/2022, ZAR 209 on 14/06/2022, and ZAR 21744 on 15/06/2022. By using the formula

$$\text{Log-return of stock } j \text{ on day } t, y_{jt} = \log_e \left(\frac{P_{jt}}{P_{j,t-1}} \right),$$

where P_{jt} and $P_{j,t-1}$ are the respective closing prices for stock j on day t and day $t-1$, we calculate the log-returns for all stocks for the entire period. For the entire period, the total number of observations (the daily-log returns) is 2161 for each stock.

Table 5.1: The selected 40 JSE companies and their sector categorisations.

Financials	**Basic materials**
FirstRand Limited	Glencore plc
Standard Bank Group	Anglo American plc
Capitec Bank Holdings	Anglo American Platinum Limited
ABSA Group Limited	Sasol Limited
Nedbank Group Limited	Kumba Iron Ore
Discovery Limited	Impala Platinum Holdings Limited
Remgro Limited	AngloGold Ashanti
PSG Group Limited	Exxaro Resources Limited
Nepi Rockcastle plc	African Rainbow Minerals
Santam Limited	Sappi Limited
Transaction Capital Limited	
Investec Limited	

Consumer services	**Consumer goods**
Shoprite Holdings Limited	British American Tobacco
Clicks Group Limited	Compagnie Fin Richemont
Woolworths Holdings Limited	Tiger Brands Limited
Mr Price Group	AVI Limited
Pick n Pay Stores Limited	
Spar Group Limited	
Truworths International Limited	

Telecommunications	**Industrials**
MTN Group Limited	Bidvest Group
Vodacom Group Limited	Barloworld Limited
Telkom SA Limited	

Technology	**Health care**
Naspers	Aspen Pharmacare Holdings Limited

5.2 Implementation of the SGDLM

We implement the SGLDM algorithm which we outlined in Section 4.4 by writing down the code in Python from scratch. The data set is divided into the training set and the test set. The training set is further divided into two subsets, one for selecting simultaneous parents and the other for selecting discount factors. The data from 01/01/2014 to 31/12/2016 (782 observations) is used to select simultaneous parents; the data from 01/01/2017 to 31/12/2018 (506 observations) is used to select discount factors and obtain starting values for the test data analysis; and the data from 01/01/2019 to 30/06/2022 (873 observations) is the test set. Therefore, we divided the implementation into three phases:

- Phase 1: Selection of simultaneous parents.
- Phase 2: Selection of discount factors and initial priors for phase 3.
- Phase 3: Stock return forecasting.

5.2.1 Selection of simultaneous parents

Here, we ran the Kalman filter equations for each of the 40 stocks using the equations given in Section 4.4. This phase entails implementing the steps 1, 3, and 6 of the algorithm. Note that the recouple/decouple steps are not included in this phase, rather the analysis involves simply running the Kalman filter for each of the decoupled series. We adopt the initial priors of [7] for this phase; these are $\boldsymbol{a}_{j0} = (0,\ldots,0)^T$, $\boldsymbol{R}_{j0} = \text{diag}(0.0001, 0.01, \ldots, 0.01)$, $r_{j0} = 5$, and $c_{j0} = 0.001$, where $\boldsymbol{a}_{j0}$ is a 40×1 vector and $\boldsymbol{R}_{j0}$ is a 40×40 diagonal matrix whose first diagonal entry is 0.0001 but the rest are 0.01. All the stocks use the same initial prior. Evolution to the next day (step 6) uses the naive posteriors of step 3. In this phase, every stock has all the remaining 39 stocks as simultaneous parents.

We specify the evolution variance $\boldsymbol{W}_{jt}$ using two discount factors via block discounting. First notice that the matrix $\tilde{\boldsymbol{C}}_{jt}$ is of the form

$$\tilde{\boldsymbol{C}}_{jt} = \left(\begin{array}{c|cccc} \tilde{c}_{1,1,j,t} & \tilde{c}_{1,2,j,t} & \cdots & \tilde{c}_{1,40,j,t} \\ \hline \tilde{c}_{2,1,j,t} & \tilde{c}_{2,2,j,t} & \cdots & \tilde{c}_{2,40,j,t} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{c}_{40,1,j,t} & \tilde{c}_{40,2,j,t} & \cdots & \tilde{c}_{40,40,j,t} \end{array}\right)$$

Let us denote the upper-left block of $\tilde{\boldsymbol{C}}_{jt}$ by $\tilde{\boldsymbol{C}}_{jt}[1,1]$ and the lower-right block by $\tilde{\boldsymbol{C}}_{jt}[2:,2:]$. The upper-left block $\tilde{\boldsymbol{C}}_{jt}[1,1]$ is the local-level component whereas the lower-right block $\tilde{\boldsymbol{C}}_{jt}[2:,2:]$ is the simultaneous parents component. Then, by block discounting, we define $\boldsymbol{W}_{jt}$ as

$$\boldsymbol{W}_{jt} = \begin{pmatrix} \frac{1-\delta_\phi}{\delta_\phi}\tilde{\boldsymbol{C}}_{jt}[1,1] & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1-\delta_\gamma}{\delta_\gamma}\tilde{\boldsymbol{C}}_{jt}[2:,2:] \end{pmatrix}.$$

With the current dimension of 40 stocks, we found out that the SGDLM analysis is most accurate if every stock has just one simultaneous parent (see Section 5.3.5). However, for studies that involve higher dimensions, e.g., [7, 10], several simultaneous parents are required to achieve the highest level of accuracy. On the last day of the period 01/01/2014 to 31/12/2016, for each stock j, we chose the stock's simultaneous parent from the other 39 stocks, depending on the absolute values of the posterior means of the vector γ_{jt}. As it can be seen from $y_{jt} = \phi_{jt} + \boldsymbol{y}^T_{sp(j),t}\gamma_{jt} + \nu_{jt}$, the entries of the vector γ_{jt} are a measure of the effect of each of the other 39 stocks on stock j (effect size). The simultaneous parent to stock j is the stock that corresponds to the biggest effect size.

In Table 5.2, we give some selected stocks together with their simultaneous parents as generated by our analysis. We have underlined the simultaneous parent if it falls in the same sector with the stock it predicts.

Table 5.2: Simultaneous parents for some selected stocks.

Stock	**Simultaneous parent**
FirstRand Limited	Standard Bank Group
Standard Bank Group	Nedbank Group Limited
MTN Group Limited	ABSA Group Limited
British American Tobacco	Investec Limited
Compagnie Fin Richemont	Mr Price Group
Naspers	Aspen Pharmacare Holdings Limited
Truworths International Limited	Mr Price Group
Shoprite Holdings Limited	Nedbank Group Limited
Glencore plc	Anglo American plc
Anglo American plc	Clicks Group Limited

We notice that some of the simultaneous parents fall in the same category with the stock being predicted – this causal relationship is expected. However, in some situations, the predictor and the stock being predicted fall in different sectors. This is still fine because dependencies in an economy can cut across sectors.

5.2.2 Selection of discount factors and obtaining initial priors for phase 3

In this section, we outline how we selected the discount factors and obtained initial priors for the test phase. The discount factors to be selected are (i) β (for learning the stochastic variance) (Section 3.3.6), and (ii) δ_ϕ and δ_γ (for specifying the evolution variance). Phase 2 involves running all the steps of the SGDLM algorithm save for step 2. The discount factors are selected using the decoupled DLMs by maximising the log-likelihood function series by series (e.g., [25, Section 4.3.6]). The intial priors are similar to those in phase 1, but since the analysis of the current phase uses only one simultaneous parent, $\boldsymbol{a}_{j0} = (0,0)^T$, $\boldsymbol{R}_{j0} = \mathrm{diag}(0.0001, 0.01)$, $r_{j0} = 5$, and $c_{j0} = 0.001$. The evolution variance is now of the form

$$\boldsymbol{W}_{jt} = \begin{pmatrix} \frac{1-\delta_\phi}{\delta_\phi} c_{1,1,j,t} & 0 \\ 0 & \frac{1-\delta_\gamma}{\delta_\gamma} c_{2,2,j,t} \end{pmatrix},$$

where the scalars $c_{1,1,j,t}$ and $c_{2,2,j,t}$ are the diagonal entries of the covariance matrix $\boldsymbol{C}_{jt}$ of the exact posterior obtained in step 5 of the algorithm.

Let us explain how we determined δ_γ. The other two discount factors were determined in a similar way. Using the standard theory of Section 3.3.6, we can write the predictive distribution of each of the decoupled time series as $(y_{jt}|\mathcal{D}_{j,t-1}) \sim T_{r_{jt}}[f_{jt}, q_{jt}]$. The log-likelihood for stock j is then defined as

$$\begin{aligned} \log_e p(y_{j,783:1288}|\mathcal{D}_{j,782}, \delta_{\gamma j}) &= \log_e \prod_{t=783}^{1288} p(y_{jt}|\mathcal{D}_{j,t-1}, \delta_{\gamma j}) \\ &= \sum_{t=783}^{1288} \log_e p(y_{jt}|\mathcal{D}_{j,t-1}, \delta_{\gamma j}) \\ &= \sum_{t=783}^{1288} \log_e \left\{ \frac{\Gamma(\frac{r_{jt}+1}{2})}{\Gamma(\frac{r_{jt}}{2})\sqrt{\pi r_{jt} q_{jt}}} \left(1 + \frac{(y_{jt}-f_{jt})^2}{r_{jt}q_{jt}}\right)^{-(\frac{r_{jt}+1}{2})} \right\} \end{aligned} \tag{5.2.1}$$

We keep β_j and $\delta_{\phi j}$ constant and vary $\delta_{\gamma j}$. (The values of β_j and $\delta_{\phi j}$ are uniform across all stocks.) For different values of $\delta_{\gamma j}$, we obtain the sum in Equation (5.2.1) at the level of individual stocks. For the running example, after inspection, we observed that most of the values of $\delta_{\gamma j}$ were on the interval $[0.859, 0.999]$. So, we varied $\delta_{\gamma j}$ on this interval for each stock. In Table 5.3, we show the log-likelihood values that correspond to the different values of $\delta_{\gamma j}$ for two companies, Standard Bank and MTN Group.

Table 5.3: Log-likelihood values at different values of $\delta_{\gamma j}$.

$\delta_{\gamma j}$	Log-likelihood	
	Standard Bank	MTN Group
0.859	1480	1280
0.894	**1495**	1285
0.929	1480	1288
0.964	1471	**1292**
0.999	1404	1287

The optimal value of the discount factor is the one that corresponds to the maximum log-likelihood. Therefore, for Standard Bank, $\delta_\gamma = 0.894$ and for MTN Group, $\delta_\gamma = 0.964$. We obtained the value of δ_γ for the remaining stocks in a similar way and computed the average across all stocks. This average now serves as the discount factor for each stock. With the current example, this average is 0.953. Thus, $\delta_\gamma = 0.953$, which is taken uniform across all stocks.

In a similar way, by keeping δ_γ and β constant, we obtained δ_ϕ as 0.993. And by keeping δ_γ and δ_ϕ constant, we obtained β as 0.922. Then, using these optimal values of the discount factors and the same initial priors, steps 1, 3, 4, 5, and 6 were re-run to obtain starting values for phase 3. The size of the importance sample in this phase was kept at $N = 2{,}000$.

5.2.3 Stock return forecasting

In the test phase, we ran the all the six steps of the SGDLM algorithm for the last three and half years of our study. This phase used the discount factors and initial priors obtained in phase 2. The analysis used $K = N = 2{,}000$.

5.3 Results from the test data analysis

5.3.1 Coverage of prediction intervals

In the context of time series forecasting, a *prediction interval*, aka *forecast interval*, is the interval which is constructed around the forecast, within which the observation is expected to lie with a specified probability [13, Section 3.5]. For example, the 95% prediction interval $[a, b]$ (constructed around the forecast $\hat{y}_{jt}$) means that, according to the predicting model, there is a 95% probability that the observation y_{jt} will lie within the interval $[a, b]$. We calculated prediction intervals at the level of individual stocks using the large sample formula (e.g., [27, Section 14.5])

$$\hat{y}_{jt} \pm z_{\alpha/2}\sqrt{\Sigma_{j,j,t}}\sqrt{1+\frac{1}{K}}, \tag{5.3.1}$$

where

- $\hat{y}_{jt}$ is the *forecast* that corresponds to the observation y_{jt};
- $z_{\alpha/2}$ is the *critical value* of the standard normal distribution and $1 - \alpha$ is the *degree of confidence*;
- $\Sigma_{j,j,t}$ is the j^{th} diagonal element of the covariance matrix $\mathbf{\Sigma}_t$, which is the *variance* of $\hat{y}_{jt}$; and
- K is the forecasting *simulation sample size*.

For example, using the formula in Equation (5.3.1), we constructed ten 50% prediction intervals for Standard Bank over the period indicated in Figure 5.1. Theoretically, the 50% prediction intervals imply that we expect 5 of the 10 observations to fall within the prediction intervals and the other five to fall outside the prediction intervals. For the chosen period, the resultant coverage of prediction intervals agrees with what the theory says.

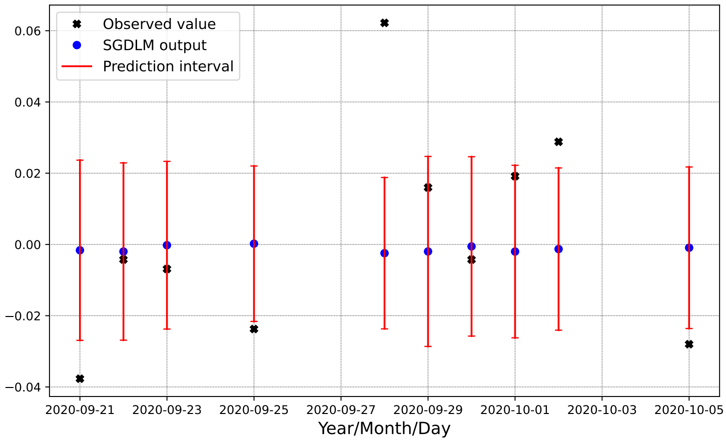


Figure 5.1: Coverage of the 50% prediction intervals for Standard Bank over a section of the test period.

For a perfect model, empirical coverage is equal to theoretical coverage. Because of the noise in the data and sometimes errors in the model, empirical coverage is not always equal to theoretical coverage. In practice, outputs of models portray under-coverage or over-coverage of intervals. The closer the output of the model to the theoretical coverage, the more accurate is the model. Over-coverage is preferred to under-coverage of the same magnitude because the former is coverage that is more than what is enough.

We calculated the interval coverages at different levels of confidence for all the stocks throughout the entire test period. From $\boldsymbol{y}_t \sim N[\boldsymbol{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t]$, each $y_{jt} \sim N[\hat{y}_{jt}, \Sigma_{j,j,t}]$. Using this result, we calculated the prediction intervals for all the stocks at the following levels of confidence: 99% ($z_{\alpha/2} = 2.58$), 95% ($z_{\alpha/2} = 1.96$), 90% ($z_{\alpha/2} = 1.64$), 80% ($z_{\alpha/2} = 1.28$), 50% ($z_{\alpha/2} = 0.67$), 20% ($z_{\alpha/2} = 0.25$), and 10% ($z_{\alpha/2} = 0.13$). In Table 5.4, we give the average interval coverages across all stocks and the interval coverages for eight of the forty stocks, for the entire test period. We also include the aggregate interval coverages of [7] as benchmark values.

Table 5.4: Average interval coverage across all stocks/aggregate interval coverage and interval coverage for some selected stocks, for the entire test period.

Prediction interval (%)	**99**	**95**	**90**	**80**	**50**	**20**	**10**
	Aggregate interval coverage						
Coverage (%)	98.4	95.7	92.6	86.0	60.4	26.5	14.4
	Benchmark aggregate interval coverage						
Coverage (%)	98.4	95.6	92.4	85.5	59.7	27.2	14.4
	Standard Bank Group						
Coverage (%)	98.9	95.5	92.3	85.9	60.5	27.1	14.2
	FirstRand Limited						
Coverage (%)	99.2	96.3	93.4	86.3	59.8	25.0	14.0
	Glencore plc						
Coverage (%)	99.2	96.0	92.1	84.4	58.2	22.6	11.5
	Anglo American plc						
Coverage (%)	98.7	95.0	92.6	85.8	58.8	24.5	12.5
	British American Tobacco						
Coverage (%)	98.5	95.0	92.1	84.3	59.6	27.6	14.1
	MTN Group Limited						
Coverage (%)	97.7	96.0	94.3	86.8	63.0	28.3	16.0
	Naspers						
Coverage (%)	97.9	95.0	90.8	84.8	61.1	26.6	14.5
	Shoprite Holdings Limited						
Coverage (%)	97.7	95.5	93.1	87.5	62.0	27.8	15.9

According to Table 5.4, the aggregate interval coverages from 10% to 95% are bigger than the theoretical values. Nevertheless, these interval coverages are more precise compared to outputs of other multivariate models (e.g., see [7]). The 99% prediction interval is under-estimated in both the aggregate analysis and for most of the individual stocks, but empirical coverage remains close to the nominal one. We also observe that the interval coverages for each of the eight stocks are similar to those of the aggregate analysis. These realised SGDLM interval coverages are therefore literally tolerable. Finally, our aggregate interval coverage estimates compare nicely with those of the benchmark study.

5.3.2 Comparison between the SGDLM and the DLM

In addition to predicting the daily log-returns using the SGDLM, we independently predicted the returns of each of the eight stocks in Table 5.4 using the stochastic volatility local-level DLM (Section 3.3.6). In the DLM analysis, we partitioned the data in a way that is similar to that of the SGDLM analysis. We used the data from 01/01/2017 to 31/12/2018 (506 observations) to select the discount factors β_j and δ_j, and to get the initial values for the testing phase. The initial values of this training period were taken as $a = 0$, $R = 0.0001$, $c = 0.001$, and $r = 5$. The test data is from 01/01/2019 to 30/06/2022 (873 observations). We never used the data from 01/01/2014 to 31/12/2016 (782 observations) as this was purposely for selecting simultaneous parents in the SGDLM case.

If, for example, we are interested in forecasting the price of Standard Bank on a daily basis, we can use either the SGDLM where Standard Bank will be modelled together with other stocks or the DLM that will focus on Standard Bank alone. So, in the SGDLM we track Standard Bank only and then compare results with those from the DLM of Standard Bank. We computed two measures of forecast accuracy, root mean square error (RMSE) and mean absolute deviation (MAD), for each of the stocks, in the SGDLM case and the DLM case, and made comparisons. Notice that, for each stock,

$$RMSE = \sqrt{\frac{1}{873}\sum_{t=1{,}289}^{2{,}161} e_{jt}^2} \quad \text{and} \quad MAD = \frac{1}{873}\sum_{t=1{,}289}^{2{,}161} |e_{jt}|.$$

Table 5.5 summarises the results. For each stock, we bold the smaller value of the error to indicate the better model.

Table 5.5: Comparison of measures of forecast accuracy (RMSE and MAD) between the SGDLM and the DLM.

Standard Bank Group			**FirstRand Limited**		
	SGDLM	DLM		SGDLM	DLM
RMSE	**0.023407**	0.023459	RMSE	**0.023118**	0.023208
MAD	**0.016465**	0.016469	MAD	0.016520	**0.016510**
Glencore plc			**Anglo American plc**		
	SGDLM	DLM		SGDLM	DLM
RMSE	0.024470	**0.024337**	RMSE	0.025138	**0.024968**
MAD	0.018218	**0.018092**	MAD	0.017994	**0.017911**
British American Tobacco			**MTN Group Limited**		
	SGDLM	DLM		SGDLM	DLM
RMSE	**0.017620**	0.017646	RMSE	**0.031298**	0.031244
MAD	**0.012833**	0.012860	MAD	0.020034	**0.019962**
Naspers			**Shoprite Holdings Limited**		
	SGDLM	DLM		SGDLM	DLM
RMSE	**0.222441**	0.223146	RMSE	**0.222292**	0.222671
MAD	0.028722	**0.028694**	MAD	0.026352	**0.026126**

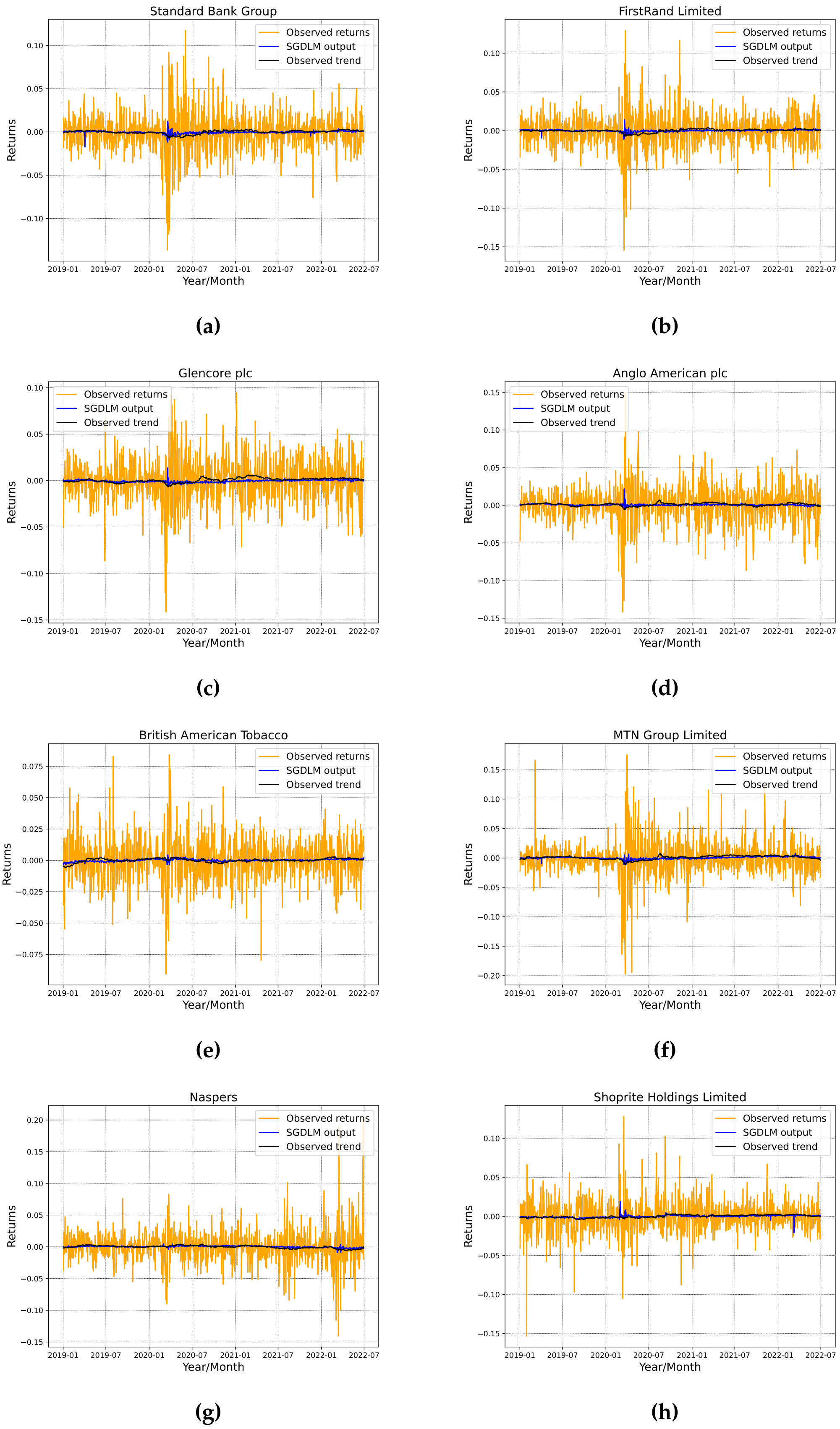
According to the results in Table 5.5, none of the two models outperforms the other in all cases. The SGDLM performs better than the DLM in the case of Standard Bank and British American Tobacco given that it gives smaller values of both the RMSE and the MAD, but the exact opposite occurs for Glencore plc and Anglo American plc. For the remaining stocks, the SGDLM produces smaller values of RMSE whereas the DLM produces smaller values of MAD. For these selected stocks, we observe a tie between the two models. It should be seen that the differences between the errors of the SGDLM and the DLM are very small. The figures in Table 5.5 are run-dependent; they keep changing slightly each time you run the analysis, but the comparison between the two models generally remains as in the table.

In principle, we expect the SGDLM to outperform the DLM because it is a model framework that captures dependencies among stocks. So, for all the stocks, we expect the SGDLM to give more accurate forecasts. This is not the case in the current example. We propose that to make the SGDLM perform better universally than the DLM, there is a need to improve the formulation of the SGDLM. One aspect here is the selection of simultaneous parents. In the current study, we picked the simultaneous parent of each of the stocks at $t = 782$ and maintained it to the end of the analysis. This is unrealistic because the market is dynamic; a good simultaneous parent to Standard Bank today may not remain good to Standard Bank after, say, one year. So, there is a need to use a more robust method of selecting simultaneous parents which involves refreshing the parents as the analysis proceeds (e.g., [10]).

5.3.3 Comparison of the empirical returns trend with the SGDLM trend

We calculated the empirical/observed 100-day simple moving averages for the observed returns and compared the resultant trend with that of the SGDLM, for each of the eight stocks. For each stock, we calculated the first moving average using the formula $(y_{j1} + y_{j2} + \cdots + y_{j100})/100$ (positioned at $t = 100$), the second using $(y_{j2} + y_{j3} + \cdots + y_{j101})/100$ (positioned at $t = 101$), the third using $(y_{j3} + y_{j4} + \cdots + y_{j102})/100$ (positioned at $t = 102$), and so on. Figure 5.2 summarises the outcomes.

For any model that fits data well, the observed trend of the data and the trend of the model forecasts should follow each other closely, if there is no stock market stress. Up to around March 2020 the observed trend and the SGDLM trend follow each other closely for almost all the eight stocks; only British American Tobacco has a clear discrepancy between the two trends during this period. This discrepancy is a reflection of the up and down movements of the price of British American Tobacco in 2019 (see Figure 5.3e). The SGDLM overestimates the returns during the market crash that started in March 2020 for all the stocks that were hit hard by the crash. This overestimation is literally visible in the case of Standard Bank, FirstRand, Glencore plc, Anglo American plc and MTN Group Limited, and is a reflection of the profuse drop in the prices in March 2020 (see Figure 5.3). The SGDLM trend generally trails below the observed trend just after the period of the intense market stress – this is expected because the formula for calculating the 100-moving averages carries along the radically below zero values of the returns for a couple of months after the market crash. Generally, the two trends track each other closely after the impact of the intense market crash.



(a)

(b)

(c)

(d)

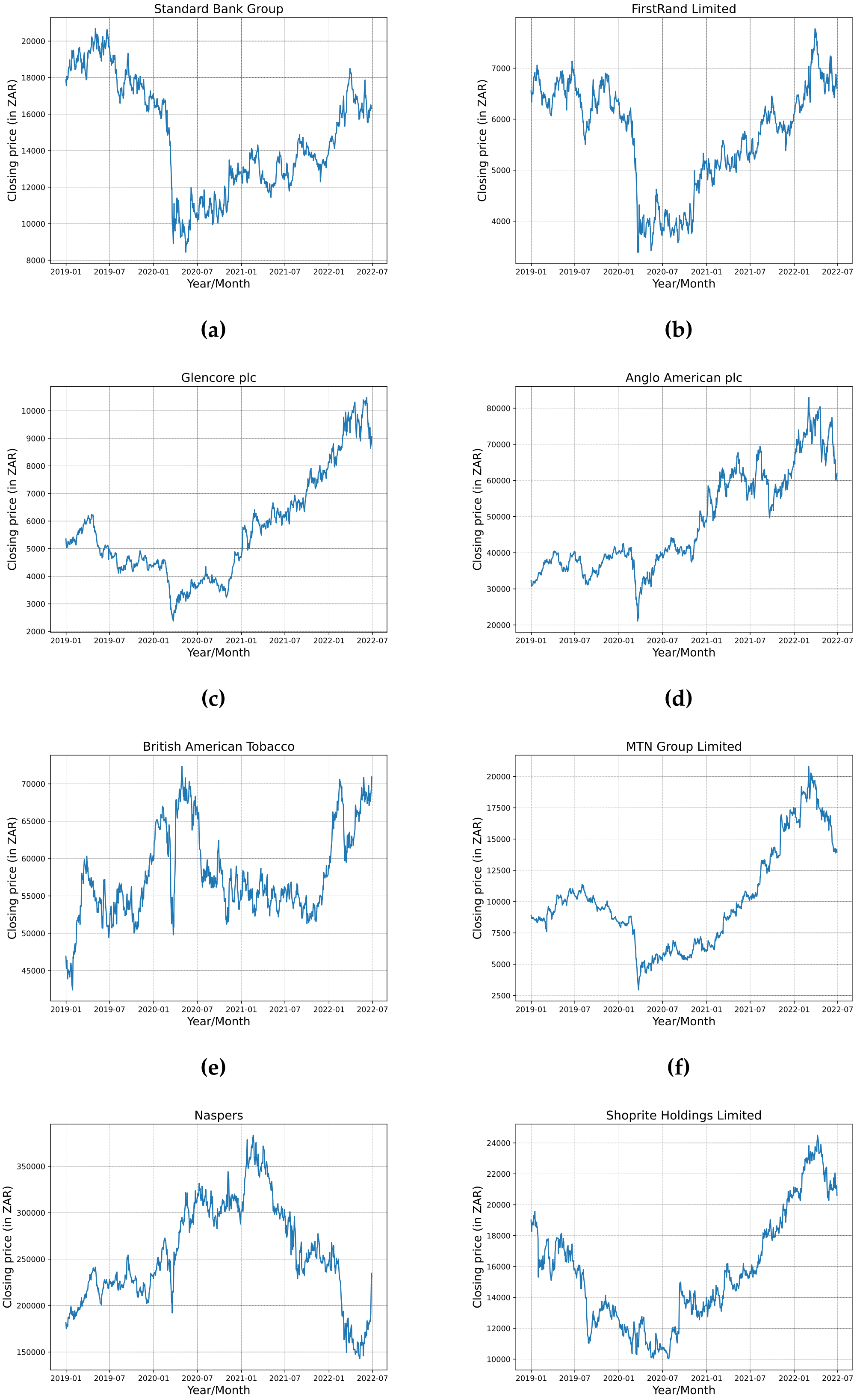
(e)

(f)

(g)

(h)

Figure 5.2: Comparison of the observed trend of the returns with the SGDLM trend.



(a) **(b)**

(c) **(d)**

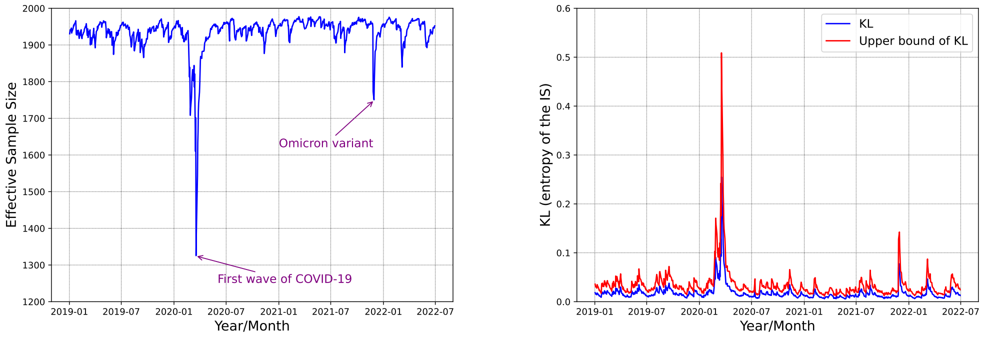
(e) **(f)**

(g) **(h)**

Figure 5.3: Closing prices of some stocks over the test data period.

5.3.4 Efficiency of importance sampling and MFVB

In Figure 5.4 below, we illustrate the evaluation of the efficiency of the importance sample-based approximation of the exact posterior. The efficiency of the MFVB approach to obtaining the decoupled conjugate forms from the approximated posterior is also evaluated. It can be seen in Figure 5.4a that for the bigger part of the test period, the effective sample size is above 1,900, which means that the importance sample-based approximation of the posterior is more than 95% effective. The most worrying period starts towards the end of February 2020 up to around mid-April 2020, during which the effective sample size nosedives to 1325 or so (about 66% effective). It should be noted that the first case of COVID-19 was announced in South Africa in early March 2020, and as Figure 5.3 shows, the result of this announcement was a plunge in the prices of most of the stocks, which subsequently caused a temporary breakdown of the SGDLM and hence the drastic fall in ESS. The other quite radical unexpected fall of the effective sample size is seen in late November 2021 due to the outbreak of the Omicron variant. The SGDLM however recovers from both short-term breakdowns and the importance sample-based posterior approximation is generally good throughout the test period. Correspondingly, Figure 5.4b shows the KL divergence as a measure of the effectiveness of MFVB. Since the KL divergence is approximated by the entropy of the importance sample, it is expected that whenever ESS is high, KL divergence is low and vice versa. So, the periods when the ESS drops are the very periods when KL divergence goes up. Generally, KL divergence remains small throughout the test period. It is interesting to see that the realised KL divergence does not exceed its theoretical upper bound at any point of the test period.



(a) **(b)**

Figure 5.4: **(a)** Measurement of the efficacy of the importance sample (IS) and **(b)** measurement of the efficacy of the MFVB approximation.

5.3.5 Effect of the number of simultaneous parents on forecast accuracy

We re-ran the full SGDLM analysis for a bigger number of simultaneous parents. We ran the analysis with two and five simultaneous parents. With two simultaneous parents, we found out that the optimal values of discount factors are $\beta = 0.919$, $\delta_\phi = 0.990$, and $\delta_\gamma = 0.970$. With five simultaneous parents, we obtained $\beta = 0.909$, $\delta_\phi = 0.983$, and $\delta_\gamma = 0.984$. In both situations, $K = N = 2{,}000$. In Table 5.6, we present the aggregate coverage of prediction intervals for the different numbers of simultaneous parents. Also, in Table 5.7, we compare the RMSE and MAD values for three stocks, for the different parental sizes.

Table 5.6: Aggregate interval coverage for different parental sizes.

Prediction interval (%)	**99**	**95**	**90**	**80**	**50**	**20**	**10**
	One simultaneous parent						
Coverage (%)	98.4	95.7	92.6	86.0	60.4	26.5	14.4
	Two simultaneous parents						
Coverage (%)	98.5	96.0	93.0	86.5	61.3	27.5	15.1
	Five simultaneous parents						
Coverage (%)	98.6	96.2	93.3	87.0	62.0	28.1	15.7

Table 5.7: Comparison of RMSE and MAD across different parental sizes for the SGDLM.

Standard Bank Group			
	1 SP	2 SP	5 SP
RMSE	0.0234	0.0236	0.0308
MAD	0.0165	0.0166	0.0179
British American Tobacco			
	1 SP	2 SP	5 SP
RMSE	0.0176	0.0192	0.0220
MAD	0.0128	0.0132	0.0139
Naspers			
	1 SP	2 SP	5 SP
RMSE	0.2224	0.2243	0.3684
MAD	0.0287	0.0290	0.0532

In Table 5.6, we observe that the SGDLM of one simultaneous parent produces the most concise prediction intervals, followed by the one with two, and the one of five comes last. The percentages in the table are the averages across all stocks; the percentages for individual stocks across different parental sizes compare similarly.

Also, in Table 5.7, the SGDLM of one simultaneous parent gives the smallest errors. The errors produced by the SGDLM of two simultaneous parents are bigger than those of the SGDLM of one simultaneous parent but smaller than those of the SGDLM of five simultaneous parents. Results from both tables suggest that the SGDLM with one simultaneous parent is the most accurate, followed by the one with two, and the one of five comes last. This supports our use of one simultaneous parent in the analyses of the preceding sections. Thus, with the current dimension of 40 stocks, using one simultaneous parent produces the most accurate results. The results of the table also suggest that accuracy reduces as the number of simultaneous parents increases. However, it should be noted that, in higher dimensions, the most accurate results may be obtained when using more than one simultaneous parent (e.g., [7]).

5.3.6 Computation time

In this section, we present the runtimes for the SGDLM analysis, for the different parental sizes. We did all analyses on a 2017 desktop computer with a CPU of 3.20 GHz, four cores, and 8GB RAM. In all analyses, $K = N = 2{,}000$. Phase 1 of the SGDLM implementation took about 9 seconds. However, phases 2 and 3 had much longer runtimes. Table 5.8 shows the approximate number of hours taken for the analysis to execute. It should be noted that we took less time than what is shown in the table because we could run three Jupyter Notebooks at once to select the discount factors; for example, for the analysis that involves using one simultaneous parent, the total runtime for phases 2 and 3 was $19 + 4 + 14 = 37$ hours. This computation time is much higher than that realised when using GPU-accelerated computing, e.g., [7].

Table 5.8: Runtime (in hours) of the SGDLM implementation for the different parental sizes.

		1 SP	2 SP	5 SP
Phase 2	Selection of discount factors	19×3	23×3	29×3
	Obtaining initial priors	4	5	6
Phase 3		14	15	19
Total		**75**	**89**	**112**

Chapter 6

Conclusion and Future Work

This chapter concludes the thesis by first giving the main takeaway points from our study. The chapter then lists the areas that are original to our study and proposes directions for further research.

6.1 Conclusion

The aim of this study was to forecast the returns of 40 stocks using the SGDLM. We found out that the SGDLM forecasts the returns accurately. In addition, with a dimension of 40 stocks or less, our results suggest that the most accurate forecasts are obtained with one simultaneous parent. Furthermore, our insights into the efficiency of the recoupling/decoupling techniques indicate that the techniques perform well generally and that SGDLMs respond well to the changes in the market. Lastly, we found out that, the use of a computer with CPU hardware for computations is however much more time-consuming compared to the use of GPU-based computers.

6.2 Contribution

Our major contributions are:

1. Firstly, we obtained the analytic solution to the filtering problem of the local-level DLM by integration. Although standard theory on DLMs is well-documented by many authors, we never found any author(s) that present the analytic solution the way we did.

2. Secondly, we presented the integrals which, when evaluated, give the analytic solution to the SGDLM analysis. The attempt to obtain this analytic solution is something that was introduced in this thesis.

3. Besides the above, we used a CPU desktop computer for computations, which none of the references noted herein did.

4. Furthermore, we did extra analyses, most importantly, the analysis that makes a comparison between the forecasts obtained from the DLM with those obtained from the SGDLM, for a particular stock.

5. Lastly, but ultimately, we are drafting a paper that we wish to submit to a journal for publication.

6.3 Future work

Firstly, by adopting the approach of refreshing simultaneous parents depending on the prevailing market conditions (e.g., [10]), we recommend re-doing the comparison of RMSE and MAD between the DLM and the SGDLM. This will perhaps make the SGDLM to outperform the DLM universally.

The current study and the references herein apply the SGDLM to stock data. We recommend applying the SGDLM to a presumably more challenging financial time series, cryptocurrencies. Cryptocurrencies are much more volatile and deviate from the normal distribution than stock data. An exploration of how the SGDLM will forecast cryptocurrencies looks to be a fascinating direction.

Lastly, further research needs to be done on evaluating the integrals that lead to the exact solution of the SGDLM, as we did with the DLM. This involves evaluating the integrals that we stated in Section 4.5. This analytic evaluation however appears to be difficult because the dimension of the integrals to be evaluated grows very fast as the number of stocks increases. A case involving 2 stocks and 1 simultaneous parent should be a good starting point as this gives rise to a 6-dimensional integral. The current example of 40 stocks and one simultaneous parent gives rise to an integral in 120 dimensions!

Appendix A

Distributions

In this appendix, standard theory about some distributions is presented.

A.1 Normal distribution

Univariate normal distribution

A random variable θ has a normal/Gausssian distribution with mean μ and variance σ^2 if it has probability density function

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Bigg\{-\frac{1}{2}\Big(\frac{\theta-\mu}{\sigma}\Big)^2\Bigg\}, \qquad -\infty < \theta < \infty.$$

Multivariate normal distribution

A random p-vector $\boldsymbol{\theta}$ follows the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ if its density function is given by

$$p(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{p/2}}\exp\Big\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\Big\}.$$

Properties of the multivariate normal distribution

1. Suppose that a random column p-vector $\boldsymbol{\theta} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. If $\boldsymbol{A}$ is a matrix of constants with many columns as the components of $\boldsymbol{\theta}$, then $\boldsymbol{A\theta} \sim N[\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T]$. If $\boldsymbol{c}$ is a $p \times 1$ vector of constants, then $\boldsymbol{c} + \boldsymbol{\theta} \sim N[\boldsymbol{c} + \boldsymbol{\mu}, \Sigma]$.

2. If the univariate random variables $\theta_1, \ldots, \theta_p$ are jointly normal, then each θ_i is normally distributed but not conversely. The converse is only true if the random variables $\theta_1, \ldots, \theta_p$ are independent (see property 3 below).

3. If $\theta_1, \dots, \theta_p$ are independent random variables with each $\theta_i \sim N[\mu_i, \sigma_i^2]$, then $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ follows the distribution $N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^T$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$.

A.2 Gamma distribution

A random variable λ has a gamma distribution with parameters $n > 0$ and $d > 0$, denoted by $\lambda \sim \text{Ga}[n, d]$, if and only if it has probability density function

$$p(\lambda) = \frac{d^n}{\Gamma(n)} \lambda^{n-1} \exp\{-d\lambda\}, \quad \lambda > 0.$$

Note that $E[\lambda] = \frac{n}{d}$ and $V[\lambda] = \frac{n}{d^2}$.

A.3 Student's t distribution

Univariate Student's t distribution

A random variable y follows a (generalised) univariate Student's t distribution with degrees of freedom ν, mode μ, and scale σ^2, written as

$$y \sim T_\nu(\mu, \sigma^2),$$

if its probability density function is

$$p(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-(\frac{\nu+1}{2})}.$$

Note that $E[y] = \mu$ for $\nu > 1$ and $V[y] = \frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$.

Multivariate Student's t distribution

A random p-vector $\boldsymbol{\theta}$ is said to have a multivariate Student's t distribution with degrees of freedom ν, mode $\boldsymbol{\mu} \in \mathbb{R}^p$, and scale matrix $\Sigma \in \mathbb{R}^{p \times p}$, denoted by

$$T_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}, p),$$

if it has density

$$p(\boldsymbol{x}) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{p/2}|\Sigma|^{1/2}} \left(1 + \frac{(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}{\nu}\right)^{-(\frac{\nu+p}{2})}.$$

The mean is given by $E[\boldsymbol{\theta}] = \boldsymbol{\mu}, \nu > 1$ and the covariance by $V[\boldsymbol{\theta}] = \frac{\nu}{\nu-2}\Sigma, \nu > 2$.

$T_\nu(0, 1, 1) = t_\nu$ is the standard Student's t distribution with ν degrees of freedom.

A.4 Normal-gamma distribution

Univariate normal-gamma distribution

For the random variables λ and θ, let $\lambda \sim \text{Ga}[n/2, d/2]$ and $(\theta|\lambda) \sim N[m, C\lambda^{-1}]$ for some $m \in \mathbb{R}$ and $n, d, C \in \mathbb{R}^+$. The joint distribution of θ and λ is called univariate normal-gamma and is denoted by

$$(\theta, \lambda) \sim NG[m, C, n, d],$$

with density given by

$$p(\theta, \lambda) = p(\theta|\lambda)p(\lambda).$$

The marginal density of θ is Student's t with n degrees of freedom, mode m, and scale $R = C(\frac{d}{n}) = C/E(\lambda)$. The marginal is denoted by

$$\theta \sim T_n[m, R].$$

Multivariate normal-gamma distribution

Let $\lambda \sim \text{Ga}[n/2, d/2]$ and suppose that, for a random p-vector $\boldsymbol{\theta}$, $(\boldsymbol{\theta}|\lambda) \sim N[\boldsymbol{m}, \boldsymbol{C}\lambda^{-1}]$, for some $n, d \in \mathbb{R}^+, \boldsymbol{m} \in \mathbb{R}^p$, and the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}$. The joint distribution of $\boldsymbol{\theta}$ and λ is the multivariate normal-gamma

$$(\boldsymbol{\theta}, \lambda) \sim NG[\boldsymbol{m}, \boldsymbol{C}, n, d].$$

The random vector $\boldsymbol{\theta}$ has a marginal multivariate Student's t distribution in p dimensions with degrees of freedom n, mode $\boldsymbol{m}$, and scale matrix $\boldsymbol{R} = \boldsymbol{C}(d/n) = \boldsymbol{C}/E[\lambda]$, denoted by

$$\boldsymbol{\theta} \sim T_n[\boldsymbol{m}, \boldsymbol{R}].$$

If θ_j is the j^{th} element of $\boldsymbol{\theta}$ with mean m_j and C_{jj} the corresponding diagonal element of $\boldsymbol{C}$, then

$$\theta_j \sim T_n[m_j, R_{jj}],$$

where $R_{jj} = C_{jj}(\frac{d}{n})$.

By using $\boldsymbol{C} = \boldsymbol{R}E[\lambda]$ and letting $d = ns$ $(s > 0)$, which implies that $s = 1/E[\lambda]$, we give an equivalent definition of the multivariate normal-gamma distribution. Let $\lambda \sim \text{Ga}[\frac{n}{2}, \frac{ns}{2}]$ and suppose that the conditional distribution of a random p-vector $\boldsymbol{\theta}$ is given by $(\boldsymbol{\theta}|\lambda) \sim N[\boldsymbol{m}, \boldsymbol{R}/(s\lambda)]$, for some $n, s > 0$, $\boldsymbol{m} \in \mathbb{R}^p$ and the $p \times p$ symmetric positive definite matrix $\boldsymbol{R}$. The joint distribution of $\boldsymbol{\theta}$ and λ is the multivariate normal-gamma denoted by

$$(\boldsymbol{\theta}, \lambda) \sim NG[\boldsymbol{m}, \boldsymbol{R}, n, s].$$

Again, the marginal distribution of $\boldsymbol{\theta}$ is multivariate Student's t with mode $\boldsymbol{m}$, scale matrix $\boldsymbol{R}$, and degrees of freedom n, denoted by $\boldsymbol{\theta} \sim T_n[\boldsymbol{m}, \boldsymbol{R}]$.

Normal-gamma distribution and linear regression models

The normal-gamma distribution is vital in ensuring a conjugate Bayesian analysis for linear regression models with unknown scale parameters. With λ defined as a scale parameter, suppose that the dependent variable y is related to the independent variable $\boldsymbol{\theta}$ (a p-vector) by

$$(y|\boldsymbol{\theta}, \lambda) \sim N[\boldsymbol{F}^T\boldsymbol{\theta}, \lambda^{-1}],$$

where $\boldsymbol{F}$ is a p-vector. Suppose further the existence of the distributions $(\boldsymbol{\theta}|\lambda) \sim N[\boldsymbol{a}, \boldsymbol{R}\lambda^{-1}]$ and $\lambda \sim \text{Ga}[n/2, d/2]$, for scalars $n, d \in \mathbb{R}^+$, $\boldsymbol{a} \in \mathbb{R}^p$, and a covariance matrix $\boldsymbol{R}$. Let $s = d/n = 1/E[\lambda]$. Then,

- conditional on λ, $(y|\lambda) \sim N[f, q\lambda^{-1}]$, where $f = \boldsymbol{F}^T\boldsymbol{a}$ and $q = \boldsymbol{F}^T\boldsymbol{R}\boldsymbol{F} + 1$, and
- unconditional on $\boldsymbol{\theta}$ or λ, $y \sim T_n[f, qs]$.

A.5 Beta distribution

A random variable η follows a beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$, denoted by $\eta \sim \text{Be}[\alpha, \beta]$, if its density function is given by

$$p(\eta) = \frac{1}{B(\alpha, \beta)}\eta^{\alpha-1}(1-\eta)^{\beta-1},$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Note that $E[\eta] = \frac{\alpha}{\alpha+\beta}$ and $V[\eta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Appendix B

Analytic solution to the filtering/updating problem

We give a step-by-step simplification of the integrals in Section 3.3.3.

The integral for the prior

Let us start from

$$p(\theta_t|\mathcal{D}_{t-1}) = \frac{1}{\sqrt{(2\pi)^2 W_t C_{t-1}}} \int \exp\Bigg\{ -\frac{1}{2}\Big((\frac{1}{W_t} + \frac{1}{C_{t-1}})\theta_{t-1}^2 - 2(\frac{\theta_t}{W_t} + \frac{m_{t-1}}{C_{t-1}})\theta_{t-1} + \frac{\theta_t^2}{W_t} + \frac{m_{t-1}^2}{C_{t-1}}\Big) \Bigg\} d\theta_{t-1}. \tag{B.0.1}$$

We first simplify the integrand. Let

$$M = \Big((\frac{1}{W_t} + \frac{1}{C_{t-1}})\theta_{t-1}^2 - 2(\frac{\theta_t}{W_t} + \frac{m_{t-1}}{C_{t-1}})\theta_{t-1} + \frac{\theta_t^2}{W_t} + \frac{m_{t-1}^2}{C_{t-1}}\Big).$$

Then,

$$M = \frac{1}{C_{t-1}W_t}\Bigg\{ (C_{t-1} + W_t)\Big(\theta_{t-1}^2 - \frac{2(\theta_t C_{t-1} + m_{t-1}W_t)}{C_{t-1} + W_t}\theta_{t-1}\Big) + \theta_t^2 C_{t-1} + m_{t-1}^2 W_t \Bigg\}$$

We complete squares and factorise to get

$$\begin{aligned} M &= \frac{C_{t-1} + W_t}{C_{t-1}W_t}\Bigg\{ \Big(\theta_{t-1} - \frac{(\theta_t C_{t-1} + m_{t-1}W_t)}{C_{t-1} + W_t}\Big)^2 - \frac{(\theta_t C_{t-1} + m_{t-1}W_t)^2}{(C_{t-1} + W_t)^2} + \frac{\theta_t^2 C_{t-1} + m_{t-1}^2 W_t}{C_{t-1} + W_t} \Bigg\} \\ &= \frac{C_{t-1} + W_t}{C_{t-1}W_t}\Big(\theta_{t-1} - \frac{(\theta_t C_{t-1} + m_{t-1}W_t)}{C_{t-1} + W_t}\Big)^2 + \frac{\theta_t^2 C_{t-1} + m_{t-1}^2 W_t}{C_{t-1}W_t} - \frac{(\theta_t C_{t-1} + m_{t-1}W_t)^2}{C_{t-1}W_t(C_{t-1} + W_t)} \end{aligned}$$

Let $N = \frac{\theta_t^2 C_{t-1} + m_{t-1}^2 W_t}{C_{t-1}W_t} - \frac{(\theta_t C_{t-1} + m_{t-1}W_t)^2}{C_{t-1}W_t(C_{t-1} + W_t)}$. Then,

$$
\begin{aligned}
N &= \frac{\theta_t^2 C_{t-1}^2 + m_{t-1}^2 W_t C_{t-1} + \theta_t^2 C_{t-1} W_t + m_{t-1}^2 W_t^2 - \theta_t^2 C_{t-1}^2 - 2\theta_t C_{t-1} m_{t-1} W_t - m_{t-1}^2 W_t^2}{C_{t-1} W_t (C_{t-1} + W_t)} \\
&= \frac{\theta_t^2 - 2\theta_t m_{t-1} + m_{t-1}^2}{C_{t-1} + W_t} \\
&= \frac{(\theta_t - m_{t-1})^2}{C_{t-1} + W_t}
\end{aligned}
$$

Therefore, $M = \frac{C_{t-1} + W_t}{C_{t-1} W_t}\Big(\theta_{t-1} - \frac{(\theta_t C_{t-1} + m_{t-1} W_t)}{C_{t-1} + W_t}\Big)^2 + \frac{(\theta_t - m_{t-1})^2}{C_{t-1} + W_t}$. Equation (B.0.1) can then be written as

$$
\begin{aligned}
p(\theta_t|\mathcal{D}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^2 W_t C_{t-1}}} \int \exp\Bigg\{ -\frac{1}{2}\Big[\frac{C_{t-1} + W_t}{C_{t-1} W_t}\Big(\theta_{t-1} - \frac{(\theta_t C_{t-1} + m_{t-1} W_t)}{C_{t-1} + W_t}\Big)^2 + \\
&\qquad \frac{(\theta_t - m_{t-1})^2}{C_{t-1} + W_t}\Big]\Bigg\} d\theta_{t-1} \\
&= \frac{1}{\sqrt{(2\pi)^2 W_t C_{t-1}}} \exp\Bigg\{ \frac{-\frac{1}{2}(\theta_t - m_{t-1})^2}{C_{t-1} + W_t}\Bigg\} \times \\
&\qquad \int_{-\infty}^{+\infty} \exp\Bigg\{ -\frac{1}{2}\Big[\frac{C_{t-1} + W_t}{C_{t-1} W_t}\Big(\theta_{t-1} - \frac{(\theta_t C_{t-1} + m_{t-1} W_t)}{C_{t-1} + W_t}\Big)^2\Big]\Bigg\} d\theta_{t-1}
\end{aligned}
\tag{B.0.2}
$$

The standard integral

$$
\int_{-\infty}^{+\infty} \exp\{ -\frac{1}{2} a (x - K)^2 \} dx = \sqrt{\frac{2\pi}{a}}, \quad a > 0, \tag{B.0.3}
$$

for some constant K, enables us to write Equation (B.0.2) as

$$
\begin{aligned}
p(\theta_t|\mathcal{D}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^2 W_t C_{t-1}}} \exp\Bigg\{ \frac{-\frac{1}{2}(\theta_t - m_{t-1})^2}{C_{t-1} + W_t}\Bigg\} \sqrt{\frac{2\pi(C_{t-1} W_t)}{C_{t-1} + W_t}} \\
&= \frac{1}{\sqrt{2\pi R_t}} \exp\Bigg\{ \frac{-\frac{1}{2}(\theta_t - m_{t-1})^2}{R_t}\Bigg\},
\end{aligned}
\tag{B.0.4}
$$

where $R_t = C_{t-1} + W_t$.

The integral for the predictive density

Starting with

$$
\begin{aligned}
p(y_t|\mathcal{D}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^2 v_t R_t}} \int \exp\Bigg\{ -\frac{1}{2}\Big((\frac{1}{v_t} + \frac{1}{R_t})\theta_t^2 - 2(\frac{y_t}{v_t} + \frac{m_{t-1}}{R_t})\theta_t + \\
&\qquad \frac{y_t^2}{v_t} + \frac{m_{t-1}^2}{R_t}\Big)\Bigg\} d\theta_t,
\end{aligned}
\tag{B.0.5}
$$

we let

$$P = \Big((\frac{1}{v_t} + \frac{1}{R_t})\theta_t^2 - 2(\frac{y_t}{v_t} + \frac{m_{t-1}}{R_t})\theta_t + \frac{y_t^2}{v_t} + \frac{m_{t-1}^2}{R_t} \Big),$$

which simplifies to

$$P = \frac{1}{R_t v_t}\Bigg\{ (R_t + v_t)\Big(\theta_t^2 - \frac{2(y_t R_t + m_{t-1} v_t)}{R_t + v_t}\theta_t\Big) + y_t^2 R_t + m_{t-1}^2 v_t \Bigg\}.$$

We complete squares, factorise, and rearrange to get

$$P = \frac{R_t + v_t}{R_t v_t}\Big(\theta_t - \frac{(\boldsymbol{y}_t R_t + m_{t-1} v_t)}{R_t + v_t}\Big)^2 + \frac{\boldsymbol{y}_t^2 R_t + m_{t-1}^2 v_t}{R_t v_t} - \frac{(\boldsymbol{y}_t R_t + m_{t-1} v_t)^2}{R_t v_t (R_t + v_t)}.$$

We let $Q = \dfrac{\boldsymbol{y}_t^2 R_t + m_{t-1}^2 v_t}{R_t v_t} - \dfrac{(\boldsymbol{y}_t R_t + m_{t-1} v_t)^2}{R_t v_t (R_t + v_t)}$, which simplifies to

$$Q = \frac{(y_t - m_{t-1})^2}{R_t + v_t}.$$

We can then write P as

$$P = \frac{R_t + v_t}{R_t v_t}\Big(\theta_t - \frac{(y_t R_t + m_{t-1} v_t)}{R_t + v_t}\Big)^2 + \frac{(y_t - m_{t-1})^2}{R_t + v_t},$$

which enables us to write Equation (B.0.5) as

$$\begin{aligned}
p(y_t|\mathcal{D}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^2 v_t R_t}} \int \exp\Bigg\{ -\frac{1}{2}\Big[\frac{R_t + v_t}{R_t v_t}\Big(\theta_t - \frac{(y_t R_t + m_{t-1} v_t)}{R_t + v_t}\Big)^2 + \\
&\qquad \frac{(y_t - m_{t-1})^2}{R_t + v_t}\Big]\Bigg\} d\theta_t \\
&= \frac{1}{\sqrt{(2\pi)^2 v_t R_t}} \exp\Bigg\{\frac{-\frac{1}{2}(y_t - m_{t-1})^2}{R_t + v_t}\Bigg\} \times \\
&\qquad \int_{-\infty}^{+\infty} \exp\Bigg\{ -\frac{1}{2}\Big[\frac{R_t + v_t}{R_t v_t}\Big(\theta_t - \frac{(y_t R_t + m_{t-1} v_t)}{R_t + v_t}\Big)^2\Big]\Bigg\} d\theta_t \\
&= \frac{1}{\sqrt{(2\pi)^2 v_t R_t}} \exp\Bigg\{\frac{-\frac{1}{2}(y_t - m_{t-1})^2}{R_t + v_t}\Bigg\} \sqrt{\frac{2\pi(R_t v_t)}{R_t + v_t}} \\
&= \frac{1}{\sqrt{2\pi q_t}} \exp\Bigg\{\frac{-\frac{1}{2}(y_t - m_{t-1})^2}{q_t}\Bigg\},
\end{aligned}$$

where $q_t = R_t + v_t$.

The posterior via Bayes' rule

We start with the integral

$$p(\theta_t|\mathcal{D}_t) = \frac{1}{\sqrt{2\pi R_t v_t q_t^{-1}}}\exp\Bigg\{ -\frac{1}{2}\Big(\frac{y_t^2 - 2\theta_t y_t + \theta_t^2}{v_t} + \frac{\theta_t^2 - 2\theta_t m_{t-1} + m_{t-1}^2}{R_t} - \frac{(\boldsymbol{y}_t^2 - 2y_t m_{t-1} + m_{t-1}^2)}{q_t}\Big)\Bigg\} \tag{B.0.6}$$

We let $L = \Big(\frac{y_t^2 - 2\theta_t y_t + \theta_t^2}{v_t} + \frac{\theta_t^2 - 2\theta_t m_{t-1} + m_{t-1}^2}{R_t} - \frac{(\boldsymbol{y}_t^2 - 2y_t m_{t-1} + m_{t-1}^2)}{q_t}\Big)$. This simplifies as follows

$$\begin{aligned}
L &= \frac{\begin{array}{r} R_t q_t y_t^2 - 2R_t q_t \theta_t y_t + R_t q_t \theta_t^2 + v_t q_t \theta_t^2 - 2v_t q_t \theta_t m_{t-1} + v_t q_t m_{t-1}^2 - v_t R_t y_t^2 \\ + 2v_t R_t y_t m_{t-1} - v_t R_t m_{t-1}^2 \end{array}}{R_t v_t q_t} \\
&= \frac{\begin{array}{r} (R_t q_t + v_t q_t)\theta_t^2 - 2(R_t q_t y_t + v_t q_t m_{t-1})\theta_t + R_t q_t y_t^2 + v_t q_t m_{t-1}^2 - v_t R_t y_t^2 \\ + 2v_t R_t y_t m_{t-1} - v_t R_t m_{t-1}^2 \end{array}}{R_t v_t q_t} \\
&= \frac{\begin{array}{r} q_t^2 \theta_t^2 - 2q_t(R_t y_t + v_t m_{t-1})\theta_t + q_t(R_t y_t^2 + v_t m_{t-1}^2) - v_t R_t y_t^2 + 2v_t R_t y_t m_{t-1} \\ - v_t R_t m_{t-1}^2 \end{array}}{R_t v_t q_t} \\
&= \frac{\theta_t^2 - 2q_t^{-1}(R_t y_t + v_t m_{t-1})\theta_t + q_t^{-1}(R_t y_t^2 + v_t m_{t-1}^2) + q_t^{-2}(2v_t R_t y_t m_{t-1} - v_t R_t y_t^2 - v_t R_t m_{t-1}^2)}{R_t v_t q_t^{-1}} \\
&= \frac{\begin{array}{r} \theta_t^2 - 2q_t^{-1}(R_t y_t + v_t m_{t-1})\theta_t + q_t^{-1} R_t y_t^2 - q_t^{-2} v_t R_t y_t^2 + 2v_t R_t y_t m_{t-1} q_t^{-2} + v_t m_{t-1}^2 q_t^{-1} \\ - q_t^{-2} v_t R_t m_{t-1}^2 \end{array}}{R_t v_t q_t^{-1}} \\
&= \frac{\begin{array}{r} \theta_t^2 - 2q_t^{-1}(R_t y_t + v_t m_{t-1})\theta_t + (1 - q_t^{-1} v_t) q_t^{-1} R_t y_t^2 + (1 - R_t q_t^{-1}) v_t m_{t-1}^2 q_t^{-1} + \\ 2v_t R_t y_t m_{t-1} q_t^{-2} \end{array}}{R_t v_t q_t^{-1}} \\
&= \frac{\theta_t^2 - 2q_t^{-1}(R_t y_t + v_t m_{t-1})\theta_t + q_t^{-1} R_t R_t y_t^2 q_t^{-1} + q_t^{-1} v_t v_t m_{t-1}^2 q_t^{-1} + 2v_t R_t y_t m_{t-1} q_t^{-2}}{R_t v_t q_t^{-1}} \\
&= \frac{\theta_t^2 - 2q_t^{-1}(R_t y_t + v_t m_{t-1})\theta_t + q_t^{-2}(R_t^2 y_t^2 + 2v_t R_t y_t m_{t-1} + v_t^2 m_{t-1}^2)}{R_t v_t q_t^{-1}} \\
&= \frac{\theta_t^2 - 2q_t^{-1}(R_t y_t + v_t m_{t-1})\theta_t + q_t^{-2}(R_t y_t + v_t m_{t-1})^2}{R_t v_t q_t^{-1}} \\
&= \frac{\Big(\theta_t^2 - q_t^{-1}(R_t y_t + v_t m_{t-1})\Big)^2}{R_t v_t q_t^{-1}} \\
&= \frac{\Big(\theta_t^2 - \big(R_t y_t q_t^{-1} + (1 - R_t q_t^{-1}) m_{t-1}\big)\Big)^2}{R_t v_t q_t^{-1}}
\end{aligned}$$

$$\begin{aligned}
L &= \frac{\left(\theta_t - (m_{t-1} + R_t q_t^{-1} y_t - R_t q_t^{-1} m_{t-1})\right)^2}{R_t v_t q_t^{-1}} \\
&= \frac{\left(\theta_t - \left(m_{t-1} + R_t q_t^{-1} (y_t - m_{t-1})\right)\right)^2}{R_t v_t q_t^{-1}} \\
&= \frac{\left(\theta_t - (m_{t-1} + R_t q_t^{-1} e_t)\right)^2}{R_t v_t q_t^{-1}} \\
&= \frac{(\theta_t - m_t)^2}{R_t v_t q_t^{-1}},
\end{aligned}$$

where $m_t = m_{t-1} + R_t q_t^{-1} e_t$. Thus, Equation (B.0.6) can be written as

$$\begin{aligned}
p(\theta_t|\mathcal{D}_t) &= \frac{1}{\sqrt{2\pi R_t v_t q_t^{-1}}} \exp\left\{ \frac{-\frac{1}{2}(\theta_t - m_t)^2}{R_t v_t q_t^{-1}} \right\} \\
&= \frac{1}{\sqrt{2\pi C_t}} \exp\left\{ \frac{-\frac{1}{2}(\theta_t - m_t)^2}{C_t} \right\},
\end{aligned}$$

where $C_t = R_t v_t q_t^{-1}$. Thus, the posterior is a normal distribution with mean m_t and variance C_t. We can therefore write

$$(\theta_t|\mathcal{D}_t) \sim N[m_t, C_t],$$

with

$$m_t = m_{t-1} + A_t e_t \quad \text{and} \quad C_t = A_t v_t,$$

where

$$e_t = y_t - f_t \quad \text{and} \quad A_t = R_t q_t^{-1}.$$

Appendix C

Derivation of formulae used in the recouple/decouple strategy

C.1 Derivation of the joint posterior density formula

We use the approach of [35, Appendix A], but with more detailed explanations of the modifications. En route to the answer, we use, among others, the following properties of matrices: (i) if $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are non-singular matrices, then $(\boldsymbol{ABC})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$; (ii) let $\boldsymbol{A}$ be a square matrix, then $|\boldsymbol{A}^m| = |\boldsymbol{A}|^m$; (iii) let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $m \times m$ matrices, then $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$; (iv) $|\boldsymbol{A}| = |\boldsymbol{A}^T|$, where $\boldsymbol{A}$ is a square matrix; and (v) the determinant of $\mathrm{diag}(c_1, \ldots, c_m)$ is the product $\prod_{i=1}^{m} c_i$.

Proof. We obtain the desired formula from

$$\text{Posterior } \propto \text{Likelihood} \times \text{Prior}. \tag{C.1.1}$$

The likelihood is obtained from the form

$$\boldsymbol{y}_t|\boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t \sim N\Big[(\boldsymbol{I}-\boldsymbol{\Gamma}_t)^{-1}\boldsymbol{\mu}_t, \Big((\boldsymbol{I}-\boldsymbol{\Gamma}_t)^T\boldsymbol{\Lambda}_t(\boldsymbol{I}-\boldsymbol{\Gamma}_t)\Big)^{-1}\Big], \tag{C.1.2}$$

which was introduced in Equation (4.2.8). We remove the index t to simplify notation. By the general formula of the density of the multivariate normal distribution given in Appendix A,

$$p(\boldsymbol{y}|\boldsymbol{\Theta},\boldsymbol{\Lambda}) = (1/2\pi)^{m/2}\Big|\Big((\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\Big)^{-1}\Big|^{-1/2} \times \\ \exp\Big\{-\frac{1}{2}\big(\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\big)^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\big(\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\big)\Big\}, \tag{C.1.3}$$

where $\boldsymbol{y}$ is an m-vector. We first simplify $\Big|\Big((\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\Big)^{-1}\Big|^{-1/2}$. Note that $\Big|\Big((\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\Big)^{-1}\Big|^{-1/2} = \Big|(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\Big|^{1/2} = \Big(|(\boldsymbol{I}-\boldsymbol{\Gamma})^T||\boldsymbol{\Lambda}||\boldsymbol{I}-\boldsymbol{\Gamma}|\Big)^{1/2} =$

$|\boldsymbol{I}-\boldsymbol{\Gamma}||\boldsymbol{\Lambda}|^{1/2} = |\boldsymbol{I}-\boldsymbol{\Gamma}|\prod_{j=1}^{m}\lambda_{jt}^{1/2}$. We now turn to the exponential term. Let $Q = \big(\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\big)^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\big(\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\big)$. Then,

$$\begin{aligned}
Q &= \Big(\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma}) - \boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^{-T}(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\Big)\Big(\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\Big)\\
&= \Big(\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma}) - \boldsymbol{\mu}^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\Big)\Big(\boldsymbol{y}-(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\Big)\\
&= \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}+\\
&\quad \boldsymbol{\mu}^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})(\boldsymbol{I}-\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}\\
&= \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y} + \boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\\
&= \Big(\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{\Gamma})^T - \boldsymbol{\mu}^T\Big)\Big(\boldsymbol{\Lambda}(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y} - \boldsymbol{\Lambda}\boldsymbol{\mu}\Big)\\
&= \Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)^T\boldsymbol{\Lambda}\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)
\end{aligned}$$

Note that $\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)^T$ is a $1 \times m$ matrix; $\boldsymbol{\Lambda}$ is an $m \times m$ diagonal matrix; and $\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)$ is an $m \times 1$ matrix. So, the product

$$\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)^T\boldsymbol{\Lambda}\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)$$

is a scalar. The j^{th} ($j = 1 : m$) element of $\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)^T$ is a scalar, which multiplies the j^{th} diagonal element of $\boldsymbol{\Lambda}$ to give a $1 \times m$ matrix. Then, the j^{th} element of $\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)^T\boldsymbol{\Lambda}$, which is a scalar, multiplies the j^{th} element of $\Big((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}-\boldsymbol{\mu}\Big)$, which is also a scalar, to give a scalar. Notice that the last matrix multiplication operation involves a sum over $j = 1 : m$. Notice further that the j^{th} element of $\boldsymbol{\mu}$ is $\boldsymbol{x}_j^T\boldsymbol{\phi}_j$; the j^{th} diagonal element of $\boldsymbol{\Lambda}$ is λ_j; and from, $\boldsymbol{y}-\boldsymbol{\Gamma}\boldsymbol{y}$, the j^{th} element of $(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{y}$ is $y_j - \boldsymbol{y}_{sp(j)}^T\gamma_j$. This leads to

$$Q = \sum_{j=1}^{m}(y_j - \boldsymbol{y}_{sp(j)}^T\gamma_j - \boldsymbol{x}_j^T\boldsymbol{\phi}_j)^T\lambda_j(y_j - \boldsymbol{y}_{sp(j)}^T\gamma_j - \boldsymbol{x}_j^T\boldsymbol{\phi}_j) = \sum_{j=1}^{m}(y_j - \boldsymbol{F}_j^T\boldsymbol{\theta}_j)^T\lambda_j(y_j - \boldsymbol{F}_j^T\boldsymbol{\theta}_j).$$

Consequently,

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\Theta},\boldsymbol{\Lambda}) &= (1/2\pi)^{m/2}|\boldsymbol{I}-\boldsymbol{\Gamma}|\Big(\prod_{j=1}^{m}\lambda_{jt}^{1/2}\Big)\exp\Big\{-\frac{1}{2}\sum_{j=1}^{m}(y_j - \boldsymbol{F}_j^T\boldsymbol{\theta}_j)^T\lambda_j(y_j - \boldsymbol{F}_j^T\boldsymbol{\theta}_j)\Big\}\\
&= |\boldsymbol{I}-\boldsymbol{\Gamma}|\prod_{j=1}^{m}(1/2\pi)^{1/2}\lambda_{jt}^{1/2}\exp\Big\{-\frac{1}{2}(y_j - \boldsymbol{F}_j^T\boldsymbol{\theta}_j)^T\lambda_j(y_j - \boldsymbol{F}_j^T\boldsymbol{\theta}_j)\Big\}\\
&= |\boldsymbol{I}-\boldsymbol{\Gamma}|\prod_{j=1}^{m}p(y_j|\boldsymbol{\theta}_j,\lambda_j)
\end{aligned}$$

Thus, from Equation (C.1.1),

$$\begin{aligned}
p(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t) &\propto \left(|\boldsymbol{I} - \mathbf{\Gamma}_t| \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{\theta}_{jt}, \lambda_{jt})\right) \times p(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_{t-1}) \\
&= \left(|\boldsymbol{I} - \mathbf{\Gamma}_t| \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{\theta}_{jt}, \lambda_{jt})\right) \times \prod_{j=1}^{m} p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1}) \\
&= |\boldsymbol{I} - \mathbf{\Gamma}_t| \prod_{j=1}^{m} p(y_{jt}|\boldsymbol{\theta}_{jt}, \lambda_{jt}) \times p(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{j,t-1}) \\
&\propto |\boldsymbol{I} - \mathbf{\Gamma}_t| \prod_{j=1}^{m} \tilde{p}(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})
\end{aligned}$$

□

C.2 Derivation of formulae for obtaining parameters under mean-field variational Bayes

Here, following the approach of [33, Section 12.3.4] and [35, Appendix A], we derive formulae for obtaining the parameters $\boldsymbol{m}_{jt}$, $\boldsymbol{C}_{jt}$, n_{jt}, and s_{jt} as used in mean-field variational Bayes in Section 4.4. Recall that p denotes the exact joint posterior distribution, p_{MC} the importance sample-based approximation of p, and q the MFVB-based posterior distribution. We wish to find parameters $\boldsymbol{m}_{jt}$, $\boldsymbol{C}_{jt}$, n_{jt}, and s_{jt}, of q that minimise the Kullback-Leibler divergence between q and p_{MC}. We work with the divergence of q from p_{MC}, $KL(p_{MC}||q)$. Notice that

$$\begin{aligned}
KL(p_{MC}||q)_t &= E_{p_{MC}}\left[\log_e \left(\frac{p_{MC}(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)}{q(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)}\right)\right] \\
&= E_{p_{MC}}[\log_e p_{MC}(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)] - E_{p_{MC}}[\log_e q(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)].
\end{aligned}$$

According to [33, Section 12.3.4], minimising $KL(p_{MC}||q)_t$ is equivalent to maximising $E_{p_{MC}}[\log_e q(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)]$. This is true because, since the distribution p_{MC} is known, the expectation $E_{p_{MC}}[\log_e p_{MC}(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)]$ is known and constant, unlike $E_{p_{MC}}[\log_e q(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)]$ whose value depends on the choice of q. So, it suffices to maximise $E_{p_{MC}}[\log_e q(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)]$. Since $q(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)]$ is required as a product of independent forms, we can minimise KL divergence series by series. Thus, we seek to find the parameters

$$\boldsymbol{m}_{jt}, \boldsymbol{C}_{jt}, n_{jt}, s_{jt} = \underset{\boldsymbol{m}_{jt},\ \boldsymbol{C}_{jt},\ n_{jt},\ s_{jt} \in q}{\operatorname{argmax}} E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})].$$

The joint density of the normal-gamma form of $(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})$ is given by

$$\begin{aligned} q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt}) &= (2\pi)^{-p_j/2}|(s_{jt}\lambda_{jt})^{-1}\boldsymbol{C}_{jt}|^{-1/2}\exp\Big\{-\frac{1}{2}(s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})\Big\}\times \\ &\quad \frac{(\frac{n_{jt}s_{jt}}{2})^{n_{jt}/2}}{\Gamma(\frac{n_{jt}}{2})}\lambda_{jt}^{n_{jt}/2-1}\exp\Big\{-\lambda_{jt}(\frac{n_{jt}s_{jt}}{2})\Big\} \\ &= (2\pi)^{-p_j/2}(s_{jt}\lambda_{jt})^{p_j/2}|\boldsymbol{C}_{jt}|^{-1/2}\exp\Big\{-\frac{1}{2}(s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})\Big\}\times \\ &\quad \frac{(\frac{n_{jt}s_{jt}}{2})^{n_{jt}/2}}{\Gamma(\frac{n_{jt}}{2})}\lambda_{jt}^{n_{jt}/2-1}\exp\Big\{-\lambda_{jt}(\frac{n_{jt}s_{jt}}{2})\Big\}. \qquad \text{(C.2.1)} \end{aligned}$$

(We have used the property: $|aA|^m = a^m|A|$, where a is a scalar and A is an $m \times m$ matrix.) From Equation (C.2.1), for some constant C,

$$\begin{aligned} \log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt}) &= \frac{1}{2}\Big\{C + p_j\log_e(s_{jt}\lambda_{jt}) - \log_e|\boldsymbol{C}_{jt}| - (s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt}) + \\ &\quad n_{jt}\log_e(\frac{n_{jt}s_{jt}}{2}) + (n_{jt}-2)\log_e\lambda_{jt} - \lambda_{jt}n_{jt}s_{jt} - 2\log_e\Gamma(\frac{n_{jt}}{2})\Big\} \\ &= \frac{1}{2}\Big\{C - \log_e|\boldsymbol{C}_{jt}| - (s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt}) + (p_j+n_{jt})\log_e s_{jt} \\ &\quad + n_{jt}\log_e n_{jt} + (p_j+n_{jt}-2)\log_e\lambda_{jt} - n_{jt}\log_e 2 - 2\log\Gamma(\frac{n_{jt}}{2}) - n_{jt}s_{jt}\lambda_{jt}\Big\} \\ &\qquad \text{(C.2.2)} \end{aligned}$$

Derivation of the equation for $\boldsymbol{m}_{jt}$

We differentiate $E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})]$ with respect to $\boldsymbol{m}_{jt}$ and equate the result to zero. For simplicity, from Equation (C.2.2), we pick the term

$$-\frac{1}{2}(s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})$$

only, since the other terms become zero when we differentiate with respect to $\boldsymbol{m}_{jt}$. We make use of: (i) If $\boldsymbol{C}$ is a symmetric matrix, then $\boldsymbol{C}^{-1}$ is also symmetric; (ii) Let $\boldsymbol{m} \in \mathbb{R}^m$ and $\boldsymbol{C}$ be an $m \times m$ matrix. If $\boldsymbol{C}$ is symmetric and not a function of $\boldsymbol{m}$, then $\frac{\partial}{\partial \boldsymbol{m}}(\boldsymbol{m}^T\boldsymbol{C}\boldsymbol{m}) = 2\boldsymbol{C}\boldsymbol{m}$ [14]; and (iii) $\frac{\partial}{\partial \boldsymbol{m}_{jt}}E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})] = E_{p_{MC}}\Big[\frac{\partial}{\partial \boldsymbol{m}_{jt}}\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})\Big]$ (Leibniz rule for differentiation under the integral sign (see [5] and [35, Appendix A])). Thus,

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{m}_{jt}}E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})] \\ &= E_{p_{MC}}\Big[\frac{\partial}{\partial \boldsymbol{m}_{jt}}\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})\Big] \\ &= E_{p_{MC}}\Big[\frac{\partial}{\partial \boldsymbol{m}_{jt}}\Big(-\frac{1}{2}(s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})\Big)\Big] \\ &= E_{p_{MC}}[\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})(s_{jt}\lambda_{jt})] \\ &= s_{jt}\boldsymbol{C}_{jt}^{-1}E_{p_{MC}}[\boldsymbol{\theta}_{jt}\lambda_{jt} - \boldsymbol{m}_{jt}\lambda_{jt}] \end{aligned}$$

This implies that $\mathbf{0} = E_{p_{MC}}[\boldsymbol{\theta}_{jt}\lambda_{jt}] - E_{p_{MC}}[\boldsymbol{m}_{jt}\lambda_{jt}]$, which leads to $\boldsymbol{m}_{jt} = {E_{p_{MC}}[\boldsymbol{\theta}_{jt}\lambda_{jt}]}/{E_{p_{MC}}[\lambda_{jt}]}$.

Derivation of the equation for $\boldsymbol{C}_{jt}$

We use the following properties of matrices.

(i) If $\boldsymbol{C}$ is a symmetric matrix, then $\frac{\partial}{\partial \boldsymbol{C}} \log_e |\boldsymbol{C}| = C^{-1}$ [14].

(ii) $\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA}) = \text{tr}(\boldsymbol{CAB})$ [21, Section 1.1].

(iii) For an $m \times m$ matrix $\boldsymbol{C}$ and an $m \times m$ covariance matrix $\boldsymbol{A}$, $\frac{\partial}{\partial \boldsymbol{C}}\text{tr}(\boldsymbol{C}^{-1}\boldsymbol{A}) = -\boldsymbol{C}^{-2}\boldsymbol{A}$ [35, Appendix A].

In an approach similar to that of deriving the equation for $\boldsymbol{m}_{jt}$, we differentiate $E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})]$ with respect to $\boldsymbol{C}_{jt}$ by picking the term

$$\frac{1}{2}\Big(- \log_e |\boldsymbol{C}_{jt}| - (s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T \boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})\Big)$$

only (from Equation (C.2.2)) and equating the result to $\mathbf{0}$. First notice that

$$\begin{aligned}\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{C}_{jt}} E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})] \\ &= E_{p_{MC}}\Big[\frac{\partial}{\partial \boldsymbol{C}_{jt}}\Big(- \frac{1}{2}\log_e |\boldsymbol{C}_{jt}| - \frac{1}{2}(s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T \boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})\Big)\Big]\end{aligned}$$

Notice further that, since the result of the product $(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T \boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})$ is a scalar, we can write

$$\begin{aligned}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T \boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt}) &= \text{tr}\Big((\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T \boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})\Big) \\ &= \text{tr}\Big(\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T\Big).\end{aligned}$$

Because $(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T$ is a covariance matrix,

$$\frac{\partial}{\partial \boldsymbol{C}_{jt}}\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T = -\boldsymbol{C}_{jt}^{-2}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T.$$

Consequently,

$$\begin{aligned}\mathbf{0} &= E_{p_{MC}}[-\boldsymbol{C}_{jt}^{-1} + (s_{jt}\lambda_{jt})\boldsymbol{C}_{jt}^{-2}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T] \\ \boldsymbol{C}_{jt}^{-1} &= s_{jt}\boldsymbol{C}_{jt}^{-2}E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T] \\ \boldsymbol{C}_{jt} &= s_{jt}E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T]\end{aligned}$$

We let $\boldsymbol{V}_{jt} = E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})(\boldsymbol{\theta}_{jt} - \boldsymbol{m}_{jt})^T]$ to have $\boldsymbol{C}_{jt} = s_{jt}\boldsymbol{V}_{jt}$.

Derivation of the equation for s_{jt}

$$\begin{aligned}
0 &= \frac{\partial}{\partial s_{jt}} E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})] \\
&= E_{p_{MC}}\Big[\frac{\partial}{\partial s_{jt}}\frac{1}{2}\Big(-(s_{jt}\lambda_{jt})(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt}) + (p_j+n_{jt})\log_e s_{jt} - n_{jt}s_{jt}\lambda_{jt}\Big)\Big] \\
&= E_{p_{MC}}\Big[-\frac{1}{2}\lambda_{jt}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{C}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt}) + \frac{p_j+n_{jt}}{2s_{jt}} - \frac{1}{2}n_{jt}\lambda_{jt}\Big] \\
&= E_{p_{MC}}\Big[-\frac{1}{2}\lambda_{jt}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T(s_{jt}\boldsymbol{V}_{jt})^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt}) + \frac{p_j+n_{jt}}{2s_{jt}} - \frac{1}{2}n_{jt}\lambda_{jt}\Big]
\end{aligned}$$

This leads to $s_{jt}n_{jt}E_{p_{MC}}[\lambda_{jt}] = p_j + n_{jt} - E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{V}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})]$. We let $d_{jt} = E_{p_{MC}}[\lambda_{jt}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})^T\boldsymbol{V}_{jt}^{-1}(\boldsymbol{\theta}_{jt}-\boldsymbol{m}_{jt})]$ to obtain

$$s_{jt} = \frac{n_{jt}+p_j-d_{jt}}{n_{jt}E_{p_{MC}}[\lambda_{jt}]}.$$

Derivation of the equation for n_{jt}

$$\begin{aligned}
0 &= \frac{\partial}{\partial n_{jt}} E_{p_{MC}}[\log_e q(\boldsymbol{\theta}_{jt}, \lambda_{jt}|\mathcal{D}_{jt})] \\
0 &= E_{p_{MC}}\Big[\frac{\partial}{\partial n_{jt}}\frac{1}{2}\Big((p_j+n_{jt})\log_e s_{jt} + n_{jt}\log_e n_{jt} - (p_j+n_{jt}-2)\log_e\lambda_{jt} - n_{jt}\log_e 2 + \\
&\quad 2\log_e\Gamma(\frac{n_{jt}}{2}) - n_{jt}s_{jt}\lambda_{jt}\Big)\Big]
\end{aligned}$$

This leads to

$$\log_e s_{jt} + 1 + \log_e n_{jt} + E_{p_{MC}}[\log_e\lambda_{jt}] - \log_e 2 - \psi(\frac{n_{jt}}{2}) - s_{jt}E_{p_{MC}}[\lambda_{jt}] = 0 \quad \text{(C.2.3)}$$

We substitute $s_{jt} = (n_{jt}+p_j-d_{jt})/(n_{jt}E_{p_{MC}}[\lambda_{jt}])$ in Equation (C.2.3) and simplify to obtain

$$\log_e(n_{jt}+p_j-d_{jt}) - \psi(\frac{n_{jt}}{2}) - \frac{(p_j-d_{jt})}{n_{jt}} - \log_e(2E_{p_{MC}}[\lambda_{jt}]) + E_{p_{MC}}[\log_e\lambda_{jt}] = 0.$$

Appendix D

Our Python code for the SGDLM implementation

All Python codes for the analyses in this thesis are available on my github page: github.com/nelsonkyakutwika/SGDLM

List of References

[1] Yahoo! Finance. https://finance.yahoo.com/, [Online; accessed July 2022].

[2] O. Aguilar and M. West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

[3] A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[4] C. M. Carvalho and M. West. Dynamic matrix-variate graphical models. *Bayesian analysis*, 2(1):69–97, 2007.

[5] K. Conrad. Differentiating under the integral sign, 2019.

[6] T. Griveau-Billion and B. Calderhead. A dynamic bayesian model for interpretable decompositions of market behaviour. *arXiv preprint arXiv:1904.08153*, 2019.

[7] L. Gruber and M. West. GPU-accelerated bayesian learning and forecasting in simultaneous graphical dynamic linear models. *Bayesian Analysis*, 11(1):125–149, 2016.

[8] L. F. Gruber. *Bayesian Modeling of General Multivariate Problems and High-Dimensional Time Series*. PhD thesis, Technische Universität München, 2015.

[9] L. F. Gruber. rSGDLM: An R Package for Simultaneous Graphical DLMs. https://github.com/lutzgruber/gpuSGDLM, [Online; accessed July 2022].

[10] L. F. Gruber and M. West. Bayesian forecasting and scalable multivariate volatility analysis using simultaneous graphical dynamic models. *arXiv preprint arXiv:1606.08291*, 2016.

[11] J. Harrison and M. West. Practical bayesian forecasting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 36(2-3):115–125, 1987.

[12] J. Harrison and M. West. Dynamic linear model diagnostics. *Biometrika*, 78(4):797–808, 1991.

[13] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[14] H. Kamper. Vector and matrix calculus, 2013.

[15] K. R. Keane and J. J. Corso. Dynamically mixing dynamic linear models with applications in finance. In *ICPRAM (2)*, pages 295–302, 2012.

[16] M. Laine. Introduction to dynamic linear models for time series analysis. In *Geodetic time series analysis in Earth sciences*, pages 139–156. Springer, 2020.

[17] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[18] H. S. Migon, D. Gamerman, H. F. Lopes, and M. A. Ferreira. Dynamic models. *Handbook of statistics*, 25:553–588, 2005.

[19] J. Oakley et al. *Bayesian forecasting and dynamic linear models*. PhD thesis, Durham University, 2019.

[20] G. P. a. Patrizia Campagnoli, Sonia Petrone. *Dynamic Linear Models with R*. Use R! Springer, 1 edition, 2009.

[21] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[22] G. Petris. dlm: an r package for bayesian analysis of dynamic linear models. *University of Arkansas*, 2009.

[23] G. Petris. An r package for dynamic linear models. *Journal of statistical software*, 36:1–16, 2010.

[24] A. Pole, M. West, and P. J. Harrison. *Applied Bayesian Forecasting and Time Series Analysis*. Chapman-Hall, 1994.

[25] R. Prado and M. West. *Time Series: Modeling, Computation & Inference*. Chapman & Hall/CRC Press, 1st edition, 2010.

[26] J. M. Quintana and M. West. An analysis of international exchange rates using multivariate dlm's. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 36(2-3):275–281, 1987.

[27] K. Ramachandran and C. P. Tsokos. *Mathematical Statistics with Applications in R*. Elsevier, 2014.

[28] A. M. Schmidt and H. F. Lopes. Dynamic models. In *Handbook of environmental and ecological statistics*, pages 57–80. Chapman and Hall/CRC, 2019.

[29] S. T. Tokdar and R. E. Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

[30] M.-N. Tran, T.-N. Nguyen, and V.-H. Dao. A practical tutorial on variational bayes. *arXiv preprint arXiv:2103.01327*, 2021.

[31] M. West. Bayesian forecasting of multivariate time series: scalability, structure uncertainty and decisions. *Annals of the Institute of Statistical Mathematics*, 72(1):1–31, 2020.

[32] M. West and J. Harrison. Bayesian forecasting. *Encyclopedia of Statistical Sciences*, 1996.

[33] M. West and P. J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 2nd edition, 1997.

[34] M. West, P. J. Harrison, and H. S. Migon. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985.

[35] M. Xie. *Multivariate Dynamic Modeling and Bayesian Decision Analysis for Macroeconomic Policy*. PhD thesis, Duke University, 2021.

[36] Z. Y. Zhao, M. Xie, and M. West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32(3):311–332, 2016.

[37] X. Zhou, J. Nakajima, and M. West. Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting*, 30(4):963–980, 2014.